Univerza v Ljubljani
FILOZOFSKA FAKULTETA

Zbornik konference

Slovenščina na spletu in v novih medijih

Ljubljana, 25. — 27. november 2015

Uredila Darja Fišer

JANES

Slovenščina na spletu in v novih medijih
Zbornik konference

Uredila: Darja Fišer

Založila: Znanstvena založba Filozofske fakultete Univerze v Ljubljani
Izdal: Oddelek za prevajalstvo
Za založbo: Branka Kalenić Ramšak, dekanja Filozofske fakultete

Ljubljana, 2015
Prva izdaja
Elektronska izdaja

Publikacija je brezplačno dostopna na spletni strani:
http://nl.ijs.si/janes/wp-content/uploads/2015/11/Konferenca2015.pdf

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

811.163.6(082)(0.034.2)

KONFERENCA Slovenščina na spletu in v novih medijih (2015 ; Ljubljana)
Zbornik konference Slovenščina na spletu in v novih medijih, Ljubljana, 25.-27. november 2015 [Elektronski vir] / uredila Darja Fišer. - 1. izd., elektronska izd. - El. knjiga. - Ljubljana : Znanstvena založba Filozofske fakultete, 2015

Način dostopa (URL):
http://nl.ijs.si/janes/wp-content/uploads/2015/11/Konferenca2015.pdf

ISBN 978-961-237-798-4 (pdf)
1. Dodat. nasl. 2. Fišer, Darja, 1978-
282290176

Predgovor k zborniku konference Slovenščina na spletu in v novih medijih

V pričujočem zborniku so objavljeni prispevki z znanstvene konference Slovenščina na spletu in v novih medijih, ki je potekala od 25. do 27. novembra 2015 v Ljubljani. Konferenco smo priredili v okviru temeljnega raziskovalnega projekta JANES (http://nl.ijs.si/janes/), ki ga med letoma 2014 in 2017 financira Javna agencija za raziskovalno dejavnost Republike Slovenije.

Zbornik vsebuje 15 izvirnih znanstvenih prispevkov 23 avtorjev s področja korpusnega in računalniškega jezikoslovja, ki pokrivajo 4 tematske sklope. Prvi se posveča gradnji virov in razvoju orodij za analizo računalniško posredovane komunikacije. V drugi sklop uvrščamo prispevke, ki karakteristike slovenščine v računalniško posredovani komunikaciji primerjajo s pisno normo in govorjenim jezikom. Tretji sklop zajema prispevke s področja leksikografije, frazeologije in terminologije, četrti pa prinaša sociolingvistične raziskave o regionalni zaznamovanosti jezika uporabniških spletnih vsebin, ekspresivnih prvinah v jeziku moških in žensk ter o žaljivem govoru.

Organizatorji se zahvaljujemo vsem, ki so prispevali k uspehu konference: vabljeni predavateljici Maji Miličević z Univerze v Beogradu za poučen predkonferenčni tutorial statistike za jezikoslovce, vabljenemu predavatelju Michaelu Beißwengerju s Tehniške univerze v Dortmundu za inspirativno predstavitev rezultatov sorodnega projekta Empirikom v Nemčiji, panelistom Špeli Arhar Holdt, Marku Stabeju, Heleni Dobrovoljc, Simonu Kreku, Poloni Gantar in Damjanu Popiču za živahno in dragoceno razpravo o podobi, vlogi in pomenu spletne slovenščine, avtorjem prispevkov za zanimive, drzne in temeljite raziskave sodobne slovenščine, Heleni Dobrovoljc, Jerneji Fridl in Primožu Gašperiču z Znanstvenoraziskovalnega centra Slovenske akademije znanosti in umetnosti, da so nam prijazno odstopili konferenčno dvorano, še posebej pa programskemu odboru za izjemno predano recenzentsko delo, ki ni le pomembno izboljšalo kvalitete prispevkov na tej konferenci, temveč je omogočilo tudi zrelejše raziskovalno delo nove generacije obetavnih jezikoslovcev v prihodnje.

Darja Fišer
Ljubljana, november 2015

Recenzenti in programski odbor

Predsednica

Darja Fišer, Filozofska fakulteta Univerze v Ljubljani

Člani

Helena Dobrovoljc, Znanstvenoraziskovalni center Slovenske akademije znanosti in umetnosti in Fakulteta za humanistiko, Univerza v Novi Gorici

Polona Gantar, Filozofska fakulteta Univerze v Ljubljani

Vojko Gorjanc, Filozofska fakulteta Univerze v Ljubljani

Matjaž Juršič, Institut »Jožef Stefan«

Simon Krek, Institut »Jožef Stefan« in Univerza v Ljubljani

Nataša Logar, Fakulteta za družbene vede Univerze v Ljubljani

Lanko Marušič, Fakulteta za humanistiko Univerze v Novi Gorici

Dunja Mladenić, Institut »Jožef Stefan«

Marko Stabej, Filozofska fakulteta Univerze v Ljubljani

Marko Robnik Šikonja, Fakulteta za računalništvo in informatiko Univerze v Ljubljani

Darinka Verdonik, Fakulteta za elektrotehniko, računalništvo in informatiko Univerze v Mariboru

Organizacijski odbor

Predsednica

Darja Fišer, Filozofska fakulteta Univerze v Ljubljani

Člana

Jaka Čibej, Filozofska fakulteta Univerze v Ljubljani

Katja Zupan, Institut »Jožef Stefan«

Organizatorji

Filozofska fakulteta Univerze v Ljubljani

Slovensko društvo za jezikovne tehnologije

Slovenska raziskovalna infrastruktura
za jezikovne vire in tehnologije Clarin.si

Regional Linguistic Data Initiative

Vabljena predavatelja

Michael Beißwenger, Tehniška univerza v Dortmundu, Nemčija

Maja Miličević, Filološka fakulteta Univerze v Beogradu, Srbija

Panelisti

Vodja

Špela Arhar Holdt, Zavod za uporabno slovenistiko Trojina in Filozofska fakulteta Univerze v Ljubljani, Slovenija

Sodelujoči

Marko Stabej, Filozofska fakulteta Univerze v Ljubljani

Helena Dobrovoljc, Znanstvenoraziskovalni center Slovenske akademije znanosti in umetnosti in Fakulteta za humanistiko, Univerza v Novi Gorici

Simon Krek, Institut »Jožef Stefan« in Univerza v Ljubljani

Polona Gantar, Filozofska fakulteta Univerze v Ljubljani

Damjan Popič, Filozofska fakulteta Univerze v Ljubljani

Program konference

Sreda, 25. 11. 2015 *Predkonferenčni tutorial*

9.00–15.30 Maja Miličević: **»Beyond example extraction: Quantitative analysis of the Janes corpus«**

Četrtek, 26. 11. 2015 *1. sekcija: »Viri, orodja in metode za analizo računalniško posredovane komunikacije«*
vodi Marko Stabej

9.00–10.00 Michael Beißwenger: **»Linguistic annotation of social media corpora: To what extent do we have to adapt existing encoding standards and tag sets?«** (vabljeno predavanje)

10.00–10.30 Tomaž Erjavec, Darja Fišer, Nikola Ljubešić: **»Razvoj korpusa slovenskih spletnih uporabniških vsebin Janes«**

2. sekcija: »Karakteristike slovenščine v računalniško posredovani komunikaciji«
vodi Polona Gantar

11.00–11.30 Špela Arhar Holdt, Kaja Dobrovoljc: **»Zveze samostalnika z nesklonljivim levim prilastkom v korpusih Janes in Kres«**

11.30–12.00 Jaka Čibej, Nikola Ljubešić: **»S kje pa si? Metapodatki o regionalni pripadnosti uporabnikov družbenega omrežja Twitter«**

12.00–12.30 Špela Vintar: **»Terminologija v spletnih forumih«**

3. sekcija: »Izrazna moč slovenščine v uporabniških spletnih vsebinah«
vodi Nataša Logar

14.00–14.30 Senja Pollak: **»Identifikacija spletno specifičnih kolokacij pogostega besedišča«**

14.30–15.00 Eneja Osrajnik, Darja Fišer, Damjan Popič: **»Ekspresivna raba ločil v tvitih slovenskih uporabnic in uporabnikov«**

15.00–15.30 Martin Justin, Nejc Hirci, Polona Gantar: **»Rana ura, slovenskih fantov grob: analiza frazeoloških prenovitev v spletni slovenščini«**

15.30–16.00 Urška Vranjek Ošlak, Ajda Centa: **»Prava frekvenca – analiza žaljivega govora v spletnih komentarjih«**

Program konference

Petek, 27. 11. 2015

4. sekcija: »Računalniško posredovana komunikacija in slovenistika«
vodi Špela Arhar Holdt

09.00–10.00 Panel: **»Slovenščina Janes: pogovorna, nestandardna, spletna ali spretna?«**
Panelisti: Marko Stabej, Helena Dobrovoljc, Simon Krek, Polona Gantar, Damjan Popič

10.00–10.30 Teja Rebernik: **»Slovenščina pod palcem interneta: vezajne in dvozačetniške tvorjenke«**
(nagrada za najboljši študentski prispevek)

5. sekcija: »Novi mediji in norma«
vodi Helena Dobrovoljc

11.00–11.30 Iza Škrjanec, Darja Fišer, Damjan Popič: **»Arheologija začetnice pri stvarnih lastnih imenih«**

11.30–12.00 Teja Goli, Darja Fišer, Damjan Popič: **»Velika in mala dilema pri imenih industrijskih izdelkov na družbenem omrežju Twitter«**

12.00–12.30 Anja Krajnc, Marko Robnik-Šikonja: **»Postavljanje vejic v slovenščini s pomočjo strojnega učenja in izboljšanega korpusa Šolar«**

6. sekcija: »Splet in leksikografija«
vodi Iztok Kosem

14.00–14.30 Ana Zwitter Vitez, Darja Fišer: **»Elementi interakcije v govorjenih in spletnih besedilih«**

14.30–15.00 Kaja Dolar: **»Kaj nam o slovenščini lahko pove kolaborativni spletni slovar?«**

15.30–16.00 Mija Michelizza: **»Leksika na spletu (in kje jo iskati)«**

Kazalo vsebine

Linguistic annotation of social media corpora: To what extent do we have to adapt existing encoding standards and tag sets?

Michael Beißwenger*

* Technical University Dortmund, Germany

Abstract

The talk gives an overview of challenges and open issues in annotating linguistic corpora of social media and computer-mediated communiation (CMC). On the example of an ongoing corpus project in the context of the German CLARIN-D initiative ('ChatCorpus2CLARIN', http://de.clarin.eu/en/curation-project-1-3-german-philology) it presents intermediate results from work dedicated to the modeling and linguistic annotation of CMC. It discusses the question to what extent a modification of existing encoding standards and NLP resources is needed and practical in order to meet two requirements: (1) The resulting schemas and tag sets should allow for an adequate representation of the structural and lingusitic peculiarities of social meda and CMC genres, while at the same time (2) they should not complicate comparative analyses of the language of social media/CMC with the language given in corpora of genres of edited text and of spoken interaction.

In the project ChatCorpus2CLARIN, an existing corpus of German chat communication, the 'Dortmund Chat Corpus' (Beißwenger 2013), and samples of other social media/CMC re-sources will be restructured to conform to current standards for the representation of corpora in the Digital Humanities context. The main goal of this work is to pave the way for the inclu-sion of linguistically annotated CMC resources into CLARIN-D corpus infrastructures and to create the prerequisites for investigating linguistic peculiarities of CMC with state-of-the art corpus technology.

The focus of the talk is on the following aspects:
(a) on adapting the encoding guidelines of the Text Encoding Initiative (TEI, http://tei-c.org) for the modeling of structural and linguistic peculiarities of CMC and social media genres,
(b) on adapting a part-of-speech (PoS) tag set for written German (the 'Stuttgat-Tübingen Tagset', Schiller et al. 1999) and using it for adding a layer with part-of-speech annota-tions to the corpus.

The talk will present and discuss a customized TEI schema that has been developed for representing the corpus data and highlight the ideas behind the main modeling decisions, especially with respect to models which are different from TEI-P5 and which have been added or modified in order to capture the peculiarities of CMC.[1].

As a second step, the talk will present and discuss a PoS tag set (Beißwenger et al. 2015) which has been extended with tags for CMC-specific phenomena as well as for phenomena which are particular of spontaneous interactional language (and, thus, for different types of "non-standardness" on the token level, cf. Ljubešić et al. 2015). A basic PoS annotation of the corpus could be achieved by using tagging models developed in the BMBF project "Schreibgebrauch" (http://www.schreibgebrauch.de/) at the University of Saarbrücken (Horbach et al. 2014). For manual post-processing the project uses the editor 'OrthoNormal' in FOLKER (Schmidt 2012) which has originally been developed and applied for the manual normalisation and correction of PoS-tagged spoken language transcripts in the FOLK corpus at the Institute for the German Language (IDS) Mannheim (http://agd.ids-mannheim.de/folk.shtml) and which, for use in the ChatCorpus2CLARIN, has been adapted for editing PoS-tagged chat data.

In an outlook, the talk will give an overview of current initiatives and activities in Germany and in the TEI for creating annotation standards for genres of social media / CMC.

[1] Besides the annotation of the corpus resources in the CLARIN-D project, the TEI schema serves as a contribution to the work of the TEI special interest group (SIG) „Computer-mediated communication" which is preparing a proposal for a TEI standard for the representation of CMC. It is based on previous schema versions created and discussed in Beißwenger et al. (2012), Chanier et al. (2014) and Margaretha/Lüngen (2014). The schema and its documentation will be made available in the form of an ODD document on the SIG pages in the TEI wiki as of October, 23): http://wiki.tei-c.org/index.php/SIG:Computer-Mediated_Communication

References

Beißwenger, Michael (2013): Das Dortmunder Chat-Korpus. In: Zeitschrift für germanistische Linguistik 41 (1), 161-164. Extended version: http://tinyurl.com/chatkorpus

Beißwenger, Michael; Ermakova, Maria; Geyken, Alexander; Lemnitzer, Lothar; Storrer, Angelika (2012): A TEI Schema for the Representation of Computer-mediated Communication. In: Journal of the Text Encoding Initiative (jTEI) 3. http://jtei.revues.org/476 (DOI: 10.4000/jtei.476).

Beißwenger, Michael; Bartz, Thomas; Storrer, Angelika; Westpfahl, Swantje (2015): Tagset und Richtlinie für das PoS-Tagging von Sprachdaten aus Genres internetbasierter Kommunikation. Guideline Document, Dortmund 2015. https://sites.google.com/site/empirist2015/home/annotation-guidelines

Chanier, Thierry; Poudat, Celine; Sagot, Benoit; Antoniadis, Georges; Wigham, Ciara; Hriba, Linda; Longhi, Julien; Seddah, Djamé (2014): The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. In: Journal of Language Technology and Computational Linguistics JLCL 29 (2), 1-30. http://www.jlcl.org/2014_Heft2/1Chanier-et-al.pdf

Horbach, Andrea; Steffen, Diana; Thater, Stefan; Pinkal, Manfred (2014): Improving the Performance of Standard Part-of-Speech Taggers for Computer-Mediated Communication. Proceedings of KONVENS 2014, 171-177.

Ljubešić, Nikola; Fišer, Darja; Erjavec, Tomaž; Čibej, Jaka; Marko, Dafne; Pollak, Senja; Škrjanec, Iza (2015): Predicting the Level of Text Standardness in User-generated Content. In: Proceedings of Recent Advances in Natural Language Processing, Hissar, Bulgaria, Sep 7–9 2015, 371–378, http://lml.bas.bg/ranlp2015/docs/RANLP_main.pdf

Margaretha, Eliza; Lüngen, Harald (2014): Building Linguistic Corpora from Wikipedia Articles and Discussions. In: Journal of Language Technology and Computational Linguistics (JLCL) 29 (2), 59-82. http://www.jlcl.org/2014_Heft2/3MargarethaLuengen.pdf

TEI Consortium (2015): TEI P5: Guidelines for Electronic Text Encoding and Interchange. Available online at: http://www.tei-c.org/Guidelines/P5/

Schiller, Anne; Teufel, Simone; Stöckert, Christine (1999): Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset). University of Stuttgart: Institut für maschinelle Sprachverarbeitung.

Schmidt, Thomas (2012): EXMARaLDA and the FOLK tools – two toolsets for transcribing and annotating spoken language. In: Proceedings of the Eighth conference on International Language Resources and Evaluation (LREC'12), Istanbul, Turkey: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2012/pdf/529_Paper.pdf.

Beyond example extraction: Quantitative analysis of the JANES corpus

Maja Miličevič*

* Faculty of Philology, University of Belgrade, Serbia

Short description

The goal of the workshop is to provide an introduction to quantitative analysis of corpus data using the R environment. The rationale is that (1) quantitative analysis is needed to properly describe corpus data, and in particular to generalise from one language sample to other similar samples and language in general; (2) R is one of the most powerful tools for quantitative analysis out there, and is freely available. The workshop will be divided in three sessions, dedicated in turn to basic considerations of corpus data and R, sample description and statistical inference. All sessions will use (meta)data from JANES.

Prerequisites: Experience in work with corpora will be assumed. No previous knowledge of statistics or R is required; an introductory handout will be provided about a week before the workshop to help participants brush up some basic math concepts and form expectations about R.

Session 1: "Obtaining data from corpora: How and why?"

- introduction to quantitative corpus studies
- formulating linguistic hypotheses testable on corpus data
- the R environment: installing R, setting working directory, installing packages
- importing data into R: defining and coding variables, file formats

Session 2: "Describing and visualising corpus data"

- descriptive statistics: counts, frequency distributions; mean, median; standard deviation, interquartile range
- graphs: scatter plots, line charts, bar charts, histograms, box plots

Session 3: "Generalising from corpus data"

- basics of statistical hypothesis testing: intro to probability; null hypothesis, significance levels and their meaning; parametric vs. non-parametric statistics
- some specific tests: chi-square, correlation, (intro to) regression

Zveze samostalnika z nesklonljivim levim prilastkom v korpusih Janes in Kres

Špela Arhar Holdt,*♦ Kaja Dobrovoljc*

* Zavod za uporabno slovenistiko Trojina, Dunajska 116, 1000 Ljubljana
♦ Filozofska fakulteta Univerze v Ljubljani, Aškerčeva 2, 1000 Ljubljana
spela.arhar@trojina.si, kaja.dobrovoljc@trojina.si

Povzetek

Korpus Janes prinaša uporabniško generirane spletne vsebine (tvite, forume, bloge, komentarje), ki so za razliko od gradiva v ostalih slovenskih korpusih večinoma nekorigirane s strani druge osebe, npr. lektorja ali urednika. Prispevek preverja vrednost novega korpusnega gradiva za normativistične raziskave, in sicer z analizo pogostosti in zapisovanja zvez samostalnika z nesklonljivim levim prilastkom (*solo petje, RTV prispevek*) v korpusih Janes in Kres. Gradivo korpusa Janes razkrije jasnejše trende zapisovanja tovrstnih zvez narazen, v primerjavi s korpusom Kres, kjer je število zvez znatno nižje, trendi v zapisu pa so bolj heterogeni. Rezultati podatkovnega luščenja so v analizi vsebinsko kategorizirani in natančneje preučeni. Na drugi strani izbrana metodologija razgali težave, ki jih pri luščenju podatkov povzroči neenotnost avtomatskega označevanja, in s tem pokaže na zadrege s kategorizacijo nesklonljivih levih prilastkov, ki se ob samostalniku pojavljajo.

Noun phrases with uninflected premodifiers in the Janes and Kres corpora

Unlike other existing Slovenian corpora, the Janes corpus of user-generated content (tweets, forums, blogs, comments) mostly consists of texts that have not been modified by any third party, such as a proofreading expert or an editor. The aim of this paper is to explore the potential of this newly available corpus data for normative language research in a case study of usage frequency and orthography of nominal phrases with uninflected premodifiers, such as *solo petje* and *RTV prispevek*, in the Janes and Kres corpus. In comparison with the Kres reference corpus, which contains significantly less phrases of this type and a more heterogeneous orthography, language data in Janes reveal clearer tendencies towards writing such nominal phrases as two separate words. A subset of the extracted data is further categorized and analysed in more detail, while the methodology itself reveals inconsistencies in automatic POS tagging due to the challenging task of linguistic categorisation of uninflected premodifiers in general.

1 Uvod

Besedilni korpusi kot vzorčene zbirke napredno označenega jezikovnega gradiva predstavljajo izhodišče za raziskovanje avtentične jezikovne rabe, s tem pa nepogrešljivo orodje za izvedbo vseh vrst jezikoslovnih raziskav. Za raziskovanje jezikovnih prvin sodobne pisne slovenščine se (poleg specializiranih virov) uporabljajo referenčni korpusi, danes predvsem Gigafida in Kres, v preteklosti so bili v podobni vlogi korpusi FIDA, FidaPLUS in Nova beseda.

Za naštete vire je značilno, da vsebujejo velik (in težko natančno določljiv) delež lektoriranih besedil. Lektorski posegi v besedila, namenjena javni objavi, so del slovenske jezikovne prakse in njihov obstoj v korpusih ustrezna odslikava realnega stanja. Uporaba lektoriranega gradiva pa je lahko problematična oz. nezadostna v primerih, ko raziskovalca zanimajo primarne, nekorigirane tendence jezikovne rabe, na področju normativistike denimo pri ocenjevanju intuitivnosti določenega jezikovnega pravila za jezikovno skupnost.

Novonastali korpus Janes, ki prinaša uporabniško generirane spletne vsebine (tvite, uporabniške komentarje, bloge in zapise z uporabniških forumov), ponuja možnost za vpogled v jezikovno produkcijo brez lektorskih oz. uredniških posegov, seveda ob upoštevanju specifik v korpus zajetih besedilnih vrst.[1] Kot primer raziskovalnega vprašanja, ki mu tovrstni podatki lahko koristijo, smo v prispevku izbrali rabo zvez samostalnika z nesklonljivim (samostalniškim) levim prilastkom.[2]

2 Predstavitev problema

Vprašanje zapisovanja, sočasno pa tudi jezikovnosistemskega uvrščanja zvez, kot so *alfa samec, servo volan, RTV prispevek* (oz. kot medponskoobrazilne zloženke zapisano skupaj: *alfasamec, servovolan* oz. z vezajem *RTV-prispevek*), je v slovenskem jezikoslovnem prostoru prisotno že desetletja in v tem času so bile temi posvečene – običajno v povezavi s pripravo oz. izidom jezikovnih priročnikov – številne razprave.[3] Ker na tem mestu ni prostora za izčrpen povzetek argumentov, napotujemo bralca k obstoječim pregledom diskusije, npr. v Logar (2005).

Pojav besedilnih korpusov je raziskovalcem ponudil možnost obsežnejših in hitrejših podatkovnih analiz, ne pa tudi enoznačnega odgovora na zgoraj opredeljeno vprašanje. Dobrovoljc in Jakop (2011: 113–114) tako ugotavljata, da se dvojnice glede zapisa v normi ne prekrivajo z dvojnicami v jezikovni rabi, da so obstoječa pravila mestoma nejasna, jezikovna raba pa izrazito neustaljena. Neustaljenost v rabi in neskladje s predpisom so izkazale tudi raziskave N. Logar, ki jih povzema Logar

[1] V prispevku puščamo ob strani sicer zanimivo vprašanje vpliva prisotnosti oz. odsotnosti jezikovnega pregleda s strani druge osebe na količino in naravo avtorjevih samokorekcij.

[2] Za opazovano skupino zvez, za katero so v sorodni literaturi glede na kategorizacijska izhodišča predlagana različna poimenovanja, v prispevku uporabljamo krovni izraz *zveze samostalnika z nesklonljivim levim prilastkom*. Pri tem ne želimo sugerirati absolutne nesklonljivosti prilastkov v tovrstnih zvezah, kakršna bi zahtevala dodatne korpusne analize, temveč predvsem njihovo statistično izstopajočo paradigmatsko fiksiranost v danih zvezah. Nadaljnja zamejitev na *samostalniške* nesklonljive leve prilastke pa opozarja na metodološka izhodišča raziskave, ki so podrobneje razložena v razdelku 3.

[3] Referenčni prispevki k tematiki so mdr. (Rigler, 1971; Toporišič 1971; Vidovič Muha, 1988) v povezavi s slovarjem SSKJ, (Gložančev, 2012) v povezavi s SP 2001, rešitve v novejših slovarskih virih pa predstavljata npr. (Gantar, 2015; Kern, 2012).

(2012). Dosedanje korpusne analize so sicer okrepile argumente za zapis narazen, vendar se ob podatkih izpostavlja vprašanje relevantnosti uporabljenih korpusnih virov, saj na osnovi lektoriranih besedil ni mogoče realno oceniti obsežnosti in narave obravnavanega problema.[4]

V tem prispevku raziskujemo, kako se na ravni obravnavane vrste besed oz. zvez razlikujeta dva za raziskave prosto dostopna korpusna vira: uravnoteženi referenčni korpus Kres (Logar et al., 2012) in korpus uporabniških vsebin Janes (Fišer et al., 2015). S tem preverjamo vrednost korpusa Janes za normativistične raziskave in trenutne možnosti za izvedbo širših, sintetičnih korpusnih raziskav izbrane tematike.

3 Luščenje korpusnih podatkov

Kot potencialne zveze samostalnikov z nesklonljivim levim samostalniškim prilastkom smo iz obeh korpusov izluščili tiste nize dveh zaporednih samostalnikov, pri katerih se dana oblika prvega samostalnika[5] ne glede na velikost črk v celotnem korpusu pojavi pred vsaj tremi različnimi oblikami leme jedrnega samostalnika (npr. *RTV prispevek, RTV prispevka, RTV prispevkom*). Če je bil ta pogoj izpolnjen, je bil niz oblike prilastka in leme jedra prepoznan kot potencialna zveza samostalnika z nesklonljivim levim samostalniškim prilastkom (npr. *RTV prispevek*).

Ker se izbrana metoda luščenja deloma opira na besednovrstne oznake, kakršne so bile pojavnicam pripisane v postopku strojnega oblikoskladenjskega označevanja, je pri načrtovanju in interpretaciji izluščenih podatkov tako potrebno upoštevati dve metodološki omejitvi.

Prva izhaja iz dejstva, da sta korpusa označena z različnima (statističnima) označevalnikoma: korpus Kres z označevalnikom Obeliks (Grčar et al., 2012) in korpus Janes z označevalnikom ToTaLe (Erjavec et al., 2005). Čeprav oba označevalnika svoj model znanja gradita na istih jezikovnih virih, leksikonu besednih oblik Sloleks (Dobrovoljc et al., 2015) in učnem korpusu ssj500k (Krek et al., 2013), med njima lahko prihaja do razlik pri tokenizaciji[6] besedila ali obravnavi nekaterih specifičnih jezikovnih sredstev.[7]

Druga, vsebinska, omejitev strojnega označevanja je posledica nedosledne obravnave nesklonljivih prilastkov v obeh omenjenih jezikovnih virih (z izjemo kratic in lastnih imen), zlasti pri vprašanjih besednovrstne kategorizacije (kako ločujemo med pridevniki in samostalniki) in njihove obravnave v besedilnem kontekstu (kako slovnične lastnosti jedra vplivajo na označevanje spola, sklona in števila nesklonljivih pridevnikov oz. sklona nesklonljivih samostalnikov v vlogi pridevnika).[8]

Izpostavljeni omejitvi z vidika kvantitativnih primerjav v pričujočem prispevku sicer nista problematični, saj predpostavljamo, da glede na prekrivnost izhodiščnih jezikovnih virov označevalnika obravnavane skladenjske strukture označujeta s podobno natančnostjo, zaradi česar sta delež in nabor nerelevantnih oz. manjkajočih zadetkov v obeh korpusih primerljiva. Kot podrobneje izpostavimo pri opisu kvalitativne kategorizacije izluščenih zvez (razdelek 5), pa bi veljalo ob nadaljnjih analizah posameznih podskupin nesklonljivih levih prilastkov označevanje poenotiti in iskanje razširiti tudi na pojavnice z nesamostalniškimi oznakami.

4 Kvantitativna primerjava rezultatov

4.1 Pogostost zvez z nesklonljivim levim prilastkom

Kot prikazujejo podatki v Tabeli 1, smo iz korpusa Kres z opisano metodo izluščili 3.054, iz korpusa Janes pa 7.840 različnih potencialnih zvez z nesklonljivim levim prilastkom. Primerjava njihove relativne pogostosti v obeh korpusih razkriva, da se v korpusu Janes pojavlja skoraj enkrat več tovrstnih zvez kot v korpusu Kres, kar kaže na izrazito pogostejšo rabo tega skladenjskega mehanizma v nelektoriranih uporabniških spletnih vsebinah.

	Kres različnice		Kres pojavnice		Janes različnice		Janes pojavnice		Prekrivne različnice
Pogostost	Abs.	Rel.	Abs.	Rel.	Abs.	Rel.	Abs.	Rel.	Abs.
zveze (npr. *alfa samec*)	3.054	**31**	95.897	987	7.840	**61**	212.808	1.662	**888**
prilastki (npr. *alfa*)	1.432	**15**	95.897	987	2.851	**22**	212.808	1.662	**719**

Tabela 1: Pogostost zvez in prilastkov v obeh korpusih.

888 je zvez, ki se pojavljajo v obeh korpusih, kar predstavlja približno tretjino izluščenih zvez v korpusu Kres, a le desetino izluščenih zvez v korpusu Janes. Nadaljnja analiza zvez, ki se pojavljajo zgolj v korpusu Janes, kaže, da lahko to razliko deloma pripišemo dejstvu, da je v besedilih korpusa Janes tudi sicer pogosteje rabljeno prevzeto besedišče, ki običajno nastopa v vlogi levega nesklonljivega prilastka, npr. *stand-up* (*komedija, scena*), *kickstarter* (*projekt, kampanja*), *live* (*stream, prenos*). Drugo dejstvo je, da se v korpusu Janes kot nesklonljivi prilastki pojavljajo samostalniki, ki so razmeroma pogosti tudi v korpusu Kres, a v njem redko nastopajo v tej

[4] Med običajnimi lektorskimi posegi so popravki na ravni zapisa narazen/skupaj, kot tudi preoblikovanja tovrstnih zvez v obliko, ki je v slovenski skladnji pogostejša.

[5] V korpusu Janes so bili kot nerelevantni že v postopku luščenja izločeni samostalniki, ki se začnejo z znakoma @ ali #.

[6] Zloženke z vezajem (npr. *C-vitamin*) označevalnik Obeliks denimo tokenizira kot niz treh pojavnic (C, -, vitamin), označevalnik ToTaLe pa kot eno samo pojavnico (C-vitamin).

[7] Krajšava *html* je denimo v korpusu Kres v vseh pojavitvah označena kot samostalnik, v korpusu Janes pa tudi kot napaka tokenizacije (oznaka Np).

[8] O označevalnih težavah besed, kot so *latino, afro, mini,* piše (Gantar, 2015: 118–120) in izpostavlja, da v jezikovni rabi pri tovrstnih lemah običajno izstopa bodisi samostalniška bodisi pridevniška vloga, kar je mogoče oz. je treba v slovarskem opisu (in posledično pri označevanju) ustrezno upoštevati.

skladenjski vlogi. Med njimi izstopajo zveze s stvarnimi in osebnimi imeni, npr. *Fiat Panda*, *Harry Potter*, v katerih se torej sklanjajo samo priimki, pa tudi zveze z nekaterimi časovno in funkcijsko manj zaznamovanimi prevzetimi občnimi imeni, npr. *privat* (*firma, sporočilo*), *kasko* (*zavarovanje, kritje*), *placebo* (*efekt, tabletka*), *rally* (*voznik, avto*).

4.2 Zapisovanje zvez z nesklonljivim levim prilastkom

V drugem koraku kvantitativne primerjave rabe zvez z nesklonljivim levim prilastkom nas je zanimalo, v kolikšni meri pri prepoznanih zvezah z levim prilastkom v korpusih prihaja do variantnosti pri njihovem zapisovanju. Rezultati kažejo, da je delež zvez z zapisovalnimi dvojnicami oz. trojnicami (zvez, ki se poleg zapisa narazen v korpusu vsaj enkrat pojavijo tudi v zapisu skupaj in/ali z vezajem) v obeh korpusih približno enak, a presenetljivo nekoliko pogostejši v besedilih korpusa Kres (29 % v korpusu Kres in 25 % v korpusu Janes).

Medtem ko se v korpusu Kres kaže predvsem preklapljanje med zapisoma narazen in z vezajem oz. narazen in skupaj, je variantnost zapisovanja v korpusu Janes enakomerneje porazdeljena med vse tri tipe variantnosti, vključno z variantnostjo vseh treh načinov zapisa. Ob rezultatih, ki jih predstavlja Tabela 2, pa je treba upoštevati specifike luščenja podatkov, ki trenutno ne zajema zvez, ki se v korpusih pojavljajo zgolj v zapisu skupaj in/ali z vezajem.[9]

Zapis	Kres	Janes
samo zapis narazen (npr. *loto številka*)	71 %	75 %
zapis narazen in z vezajem (npr. *tv film, tv-film*)	13 %	8 %
zapis narazen in skupaj (npr. *špas teater, špasteater*)	11 %	9 %
zapis narazen, z vezajem in skupaj (npr. *new york, newyork, new-york)*	5 %	7 %

Tabela 2: Primerjava variantnosti zapisovanja zvez z nesklonljivim levim prilastkom.

5 Kategorizacija prekrivnih zvez

Da bi lahko natančneje določili vsebino izluščenih podatkov, smo 888 zvez, ki se pojavljajo v obeh korpusih, razvrstili v pet robustnih kategorij.[10]

[1] **Nerelevantni rezultati**: raznovrstne kombinacije, ki so ustrezale pogojem luščenja, vendar niso relevantne za raziskavo (*york city, pearl jama, družba človek*).

[2] **Lastna imena** (zemljepisna, stvarna), tako domača (*butan plin, ford fiesta*) kot tuja (*financial times*), v podatkih pa se pojavljajo tudi osebna imena (npr. *indiana jones, chuck norris*).

[3] **Citatna oz. polcitatna poimenovanja,** npr. *after party, bad boy, fair play, press center, team building*.

[4] **Kratične zveze,** npr. *rtv prispevek, usb ključek, c vitamin, lcd zaslon,* tudi *zf film, fb stran.*

[5] **Občna imena z nekratičnim prilastkom,** tako z nesklonljivim samostalniškim prilastkom, ki je bodisi lastno (*android telefon*) bodisi občno ime (*joga studio*), kot tudi zveze z okrajšano prvo sestavino (*info točka*) ali nesklonljivim pridevniškim prilastkom (*mikro podjetje*).

Rezultati luščenja v točki [5] so v jezikoslovnem smislu precej heterogeni. Medtem ko so zveze tipa *android telefon* in *joga studio* glede na metodologijo pričakovane (in med seboj tudi jasno ločljive), so se zveze tipa *info točka* in *mikro podjetje* med podatki znašli zaradi specifik obravnave v označevalnih virih (gl. pogl. 3). Slednja odslikava težave pri enoznačnem ločevanju med zvezami z okrajšano prvo sestavino (*eko šola*), nesklonljivimi pridevniki (*mini krilo*) in samostalniki v pridevniški rabi (*golf igrišče*).[11] Težave z razmejevanjem, kot tudi želja ugotoviti morebitne tendence v rabi, ki bi razmejevanje lahko utemeljile, so razlog, da v nadaljevanju raziskave raznovrstne zveze obravnavamo skupaj. Ker pa za različne od naštetih skupin veljajo različne normativne smernice glede zapisa, je pri razumevanju in posploševanju podatkov potrebna dodatna previdnost. Rezultate kategorizacije prikazuje Tabela 3.

Kategorija	Število zvez
Nerelevantni rezultati	150
Lastna imena	205
Citatna oz. polcitatna imena	37
Kratične zveze	187
Zveze z nesklonljivim prilastkom	309

Tabela 3: Kategorije prekrivnih zvez.

V nadaljevanju prispevka se od predstavitve rezultatov pogostosti rabe različnih vrst nesklonljivih prilastkov v obeh korpusih premikamo k natančnejšemu pregledu zapisovanja zvez dveh izbranih podskupin: kratičnih zvez [4] in občnih imen z nekratičnim prilastkom [5].

6 Zapisovanje zvez: Janes vs. Kres

V zadnjem koraku raziskave nas je zanimalo, v kolikšni meri se obravnavana korpusa razlikujeta glede trendov v zapisu besednih zvez tipa *USB ključek/USB-ključek* in *joga studio/jogastudio.* Za vse ustrezajoče podatke so bila izračunana razmerja, v kolikšnem deležu se posamezna zveza pojavlja zapisana narazen, skupaj ali z vezajem. Nato smo deleže primerjali med obema korpusoma in zveze razvrstili v štiri skupine:

[9] Pilotni poskus luščenja zvez z vezajem, ki se v korpusu Janes nikoli ne pojavijo v zapisu narazen, sicer kaže, da med tistimi s pogostostjo nad 100 pojavitev kot zveze z nesklonljivim prilastkom pojavljajo samo zveze s krajšavo *e-* (npr. *e-volitve*).

[10] Luščenje podatkov je potekalo neobčutljivo na velike začetnice (skupaj obravnavamo *fb stran, FB stran* in *Fb stran*). Pri navajanju zgledov v poglavjih 5 in 6 ne zapisujemo vseh evidentiranih oblik, ampak navajamo vse zapise z malimi črkami. Prav tako v zgledih ne navajamo vseh variant zapisa skupaj / narazen / z vezajem: privzeta oblika zapisa pri zgledih je narazen, izjeme od tega načela pa so v besedilu posebej napovedane.

[11] Kategorije in primeri po (Dobrovoljc in Jakop, 2011: 114).

[A] Zveze, pri katerih **ne prihaja do razlik,** npr. *loto številka, tempera barva, pat pozicija*, ki se v obeh korpusih pišejo izključno narazen.
[B] Zveze, pri katerih se posamezni deleži **razlikujejo do 25 odstotnih točk,** npr. *pop pevka* se v Janesu zapisuje narazen v 99,3 %, v Kresu pa v 90,2 % primerov.
[C] Zveze, pri katerih je **razhajanje med 25 in 50 odstotnimi točkami,** npr. *solo petje* se v korpusu Janes zapisuje narazen v 71,7 %, v Kresu v 45,1 % primerov.
[D] Zveze, pri katerih so **razhajanja večja od 50 odstotnih točk,** npr. *lcd zaslon* je v korpusu Janes zapisan narazen v 97,7 %, v Kresu pa v 47 % primerov.

Čeprav pri redko rabljenih zvezah nekoliko manj zanesljive, so se na tovrsten način opredeljene razlike izkazale za ustrezno izhodišče ugotavljanja smiselnosti uporabe korpusa Janes kot komplementarni vir ob korpusu Kres, omogočile pa so tudi osnovno identifikacijo trendov jezikovne rabe, ki se v korpusu Janes kažejo drugače kot v korpusu Kres. V nadaljevanju razlike med korpusoma predstavljamo ločeno glede na tip zveze.

6.1 Kratične zveze

Zvez, pri katerih je na prvem mestu kratica, je med podatki 187. Glede na Pravopis (§ 496) naj bi se tovrstne zapisovale z vezajem. Tabela 4 prikazuje, kolikšne so razlike v deležu narazen zapisanih zvez v obeh korpusih.

Rang	Delež	Primeri iz korpusa Janes
[A] ni razlik	14 % zvez	*jv evropa, pdf datoteka, html koda, uefa liga, sv vojna*
[B] majhne razlike	34 % zvez	*rtv slovenija, eu poslanec, dsg menjalnik, nba liga, sms donacija*
[C] srednje razlike	36 % zvez	*tv program, rtv prispevek, led dioda, usb ključek, c vitamin*
[D] velike razlike	16 % zvez	*tv oddaja, mp3 predvajalnik, iq test, 3d model, g točka*

Tabela 4: Razlike zapisa kratičnih zvez.

Pri zvezah, ki jih najdemo v skupinah [C] in [D], v Janesu po večini prevladuje zapis narazen, v korpusu Kres pa se zapis narazen giblje med 30 in 75 % v skupini [C] oz. med 12 in 49 % v skupini [D] – razlike so, po pričakovanjih, na račun zapisa z vezajem. Korpus Janes kaže nekoliko velikodušnejšo rabo vezaja pri zvezah, kjer je na prvem mestu posamezna črka, vendar tudi pri slednjih ne dosledno: več kot 50-odstotno pojavitev z vezajem v korpusu Janes izkazujeta samo primera *e-naslov* (v 93,1 %) in *b-vitamin* (v 61,5 % primerov).[12]

Različne zveze torej prinašajo različna razmerja v zapisu, pri čemer so nedoslednosti v rabi precej višje v korpusu Kres, kar prikazuje Tabela 5, v kateri so prikazana razmerja za zveze s kratico *USB*.

Zveza	Zapis narazen Kres	Zapis narazen Janes
usb disk	70,0 %	100,0 %
usb kabel	68,4 %	100,0 %
usb ključ	52,6 %	95,4 %
usb ključek	62,2 %	96,0 %
usb modem	83,3 %	91,2 %
usb vhod	63,6 %	100,0 %
usb vmesnik	41,7 %	100,0 %

Tabela 5: Narazen zapisane zveze s kratico *USB*.[13]

6.2 Občna imena z nekratičnim prilastkom

Raznovrstnih zvez z nesklonljivim levim prilastkom je med podatki 309. Trenutna jezikovna pravila za te zveze predvidevajo zapis skupaj ali narazen, kar predstavljata (Dobrovoljc in Jakop, 2011: 113–122). Tabela 6 prikazuje razlike v deležu narazen zapisanih zvez v obeh korpusih.

Rang	Delež	Primeri iz korpusa Janes
[A] ni razlik	34 % zvez	*mainstream medij, android telefon, diesel motor, jazz klub, beta verzija*
[B] majhne razlike	54 % zvez	*stereo zvočnik, rock legenda, pleksi steklo, kino spored, spin doktor*
[C] srednje razlike	11 % zvez	*solo akcija, video predvajalnik, alfa samec, tapas bar, elektro omarica*
[D] velike razlike	1 % zvez	*porno film, avdio sistem, video zaslon, video film*

Tabela 6: Razlike zapisa občnih imen z nekratičnim prilastkom.

Če pri kratičnih zvezah v kategorijah [C] in [D] najdemo 52 % zvez, je v Tabeli 6 ta delež le 12 %. Splošno gledano sta torej v zapisovanju občnih imen z nekratičnim prilastkom korpusa skladnejša in po večini gre za skladnost v zapisu narazen, ki povprečno gledano v podatkih močno prevladuje. Vendar pa zapis narazen ni dominanten pri prav vseh posameznih primerih: v Janesu več kot 50-odstotno pojavitev zapisa skupaj izkazuje 18 primerov: *avtocesta, videoposnetek, fotogalerija, videospot, avtošola, avtohiša, kinodvorana, elektromotor, motošport, fotozgodba, videokaseta, turbomotor, betablokator, avtosalon, fotodelavnica, elektroinženir, narkokartel* in *videokonferenca*. V korpusu Kres je takih primerov 40.

Značilno za podatke je, da se raba posameznega prilastka v različnih zvezah razlikuje. Če si ogledamo skupine zvez, ki vsebujejo (vsaj tri različne) primere z enakim prilastkom, dobimo naslednje rezultate:

[1] V obeh korpusih se dokaj dosledno zapisuje narazen skupina zvez, kjer je prilastek lastno ime (*android, erasmus, linux*). Podobno velja za zveze s prilastki *fitnes, golf, reli, welness, vikend, house, jazz, latino* in *metal*.

[12] V primerjavi s 43 tovrstnimi primeri v korpusu Kres (razlog, da jih ni več, gre iskati tudi v specifikah izbrane metodologije).

[13] Število pojavitev v korpusu Kres (prva številka v oklepaju) in Janes (druga številka v oklepaju): *usb disk* (7; 33), *usb kabel* (13; 120), *usb ključ* (30; 187), *usb ključek* (28; 243), *usb modem* (25; 83), *usb vhod* (14; 74), *usb vmesnik* (5; 10).

[2] V nekaterih skupinah zvez se najdejo pri posameznih primerih glede zapisovanja odstopanja od splošnega trenda, vendar so te razlike relativno skladne v obeh korpusih, npr. pri zvezah s prilastki *avto, bas, beta, foto, kino, seks, pop* in *rock.*

[3] Nekatere skupine pa prinašajo heterogene zapise, ki se tudi med korpusoma razlikujejo (skupini [C] in [D] v Tabeli 6), npr. zveze s prilastki *avdio, elektro, evro, makro, moto, solo* in *video*. Razlika je praviloma na račun zvez, ki se v Kresu pišejo skupaj, v Janesu pa narazen. Zveze s prilastkom *solo* prikazuje Tabela 7.[14]

Zveza	Zapis narazen Kres	Zapis narazen Janes
solo akcija	75,0 %	100,0 %
solo album	87,5 %	100,0 %
solo kariera	88,0 %	100,0 %
solo kitara	95,2 %	100,0 %
solo nastop	100,0 %	100,0 %
solo petje	45,1 %	71,7 %
solo projekt	81,8 %	88,9 %

Tabela 7: Narazen zapisane zveze s prilastkom *solo*.[15]

7 Sklep

V prispevku predstavljena analiza je potrdila tezo, da se referenčni korpus Kres in korpus uporabniško generiranih spletnih besedil Janes glede rabe zvez samostalnika z nesklonljivim levim prilastkom pomembno razlikujeta. Kvantitativni del analize je potrdil hipotezo, da je raba tovrstnih zvez v korpusu Janes bistveno pogostejša kot v korpusu Kres in da se v obeh korpusih pojavlja visok delež zvez, ki v rabi izkazujejo variantnost v zapisovanju (narazen ali skupaj oz. z vezajem).

Natančnejši pregled zvez, ki se pojavljajo v obeh korpusih, je pokazal, da so izluščeni podatki različnih vrst: lastna imena, občna citatna oz. polcitatna poimenovanja, kratične zveze in občna imena z nekratičnim prilastkom. Zadnja skupina je v jezikoslovnem smislu heterogena, saj vsebuje tako primere z nesklonljivim samostalniškim prilastkom, kot tudi primere z nesklonljivim pridevnikom oz. okrajšano prvo sestavino. Raznorodnost rezultatov opozarja na potrebo po nadaljnjih gradivno utemeljenih raziskavah tematike, s korenitim premislekom jezikovnosistemskega razvrščanja nesklonljivih prilastkov, ki trenutno vpliva tudi na priklic podatkov v oblikoskladenjsko označenem gradivu.

Drugi del raziskave je natančneje osvetlil razlike v zapisovanju kratičnih imen in občnih imen z nekratičnim prilastkom v korpusih Janes in Kres. Korpusa se razlikujeta predvsem v zapisovanju kratičnih zvez, kjer so bistvene razlike prisotne kar pri 52 % analiziranega gradiva in pretežno enoznačne: v korpusu Janes prevladuje zapis brez vezaja, v lektorsko reguliranem korpusu Kres pa je rabe vezaja več, vendar slednja ne prevladuje dosledno. Pri zapisovanju občnih imen z nekratičnim prilastkom sta korpusa skladnejša, bistveno se razlikujeta v 12 % analiziranih podatkov. Kljub temu je mogoče tudi pri teh podatkih zaključiti, da korpus Janes izkazuje višji delež zapisovanja narazen kot korpus Kres, Kres pa sorazmerno višji, vendar v splošnem še vedno ne prevladujoč delež zapisovanja skupaj. Poseben izziv za normativistiko, kakor tudi za jezikovni opis, je dejstvo, da se v rabi variantnost pogosto izkazuje že na ravni posameznega prilastka, ki v različnih zvezah kaže različne zapisovalne tendence. Te so, kot smo pokazali v prispevku, v določenih primerih med korpusoma skladne, v določenih primerih različne, vsaj na osnovi obravnavanih podatkov pa se ne zdijo povezane z obstoječo jezikoslovno tipologizacijo prilastkov.

Za dokončno opredelitev stanja bi bilo metodo luščenja treba razširiti, da bi zajela tudi podatke, ki se vedno zapisujejo z vezajem oz. skupaj, po možnosti pa tudi del tipičnih lektorskih besednozveznih parafraz (*tenis igrišče* > *teniško igrišče, igrišče za tenis*), ob tem pa seveda ustrezno zaobiti v prispevku izpostavljene označevalne zadrege. V kontekstu predhodnih raziskav bi bilo v nadaljevanju zanimivo primerjati tudi razmerje med skupaj in narazen pisanimi zvezami/zloženkami pri tistih primerih, kjer je v prvem delu že zloženka (*stan-up komedija, kickstarter projekt*), na drugi strani pa gradivo osvetliti tudi s časovnega vidika in primerjati trende pri zapisu novih besed oz. zvez s tistimi, ki so v jeziku prisotne že dlje časa.

Nekoliko posplošena ugotovitev pričujočega prispevka, da je nelektorirana jezikovna produkcija v rabi doslednejša (četudi gre pri tem mestoma za odstop od obstoječega jezikovnega predpisa), da torej lektorska jezikovna regulacija pravzaprav krepi variantnost v jezikovni rabi, vsekakor predstavlja pomemben argument v razpravi o bodočih normativističnih odločitvah glede obravnavane tematike.

8 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.

9 Literatura

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

Kaja Dobrovoljc, Simon Krek, Peter Holozan, Tomaž Erjavec in 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 Pouliquen, Ralf Steinberger. 2005. Massive multi-lingual corpus compilation: Acquis Communautaire and totale. V *Proceedings of the 2nd Language & Technology Conference*, str. 32–36. Poznan, Poland.

Darja Fišer, Tomaž Erjavec, Jaka Čibej in Nikola Ljubešić, 2015. Gradnja in analiza korpusa spletne slovenščine JANES. V: *Slovnica in slovar - aktualni jezikovni opis (Obdobja 34)*.

[14] Dodati je mogoče, da se v SSKJ in slovarju Pravopisa (www.fran.si, dostop 30. 8. 2015) od navedenih primerov pojavita v zapisu skupaj dve iztočnici, *soloakcija* in *solopetje* (slednja z omembo narazen pisane dvojnice), kar sovpada z nižjim deležem zapisov narazen v Kresu, vendar bi bilo za ugotavljanje neposrednih povezav med referenčnimi priročniki in rabo treba pregledati več gradiva.

[15] Število pojavitev v korpusu Kres (prva številka v oklepaju) in Janes (druga številka v oklepaju): *solo akcija* (6; 73), *solo album* (21; 50), *solo kariera* (22; 46), *solo kitara* (20; 20), *solo nastop* (10; 17), *solo petje* (46; 43), *solo projekt* (9; 16).

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

Alenka Gložančev. 2012. Novejša slovenska leksika v luči obravnave samostalniških zloženk v Slovenskem pravopisu 2001. V: N. Jakop in H. Dobrovoljc, ur., *Pravopisna stikanja: Razprave o pravopisnih vprašanjih*, str. 125–39. Ljubljana: Založba ZRC.

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. Ljubljana: Institut Jožef Stefan.

Boris Kern. 2012. Pisanje skupaj in narazen v Slovarju novejšega besedja slovenska jezika. V: N. Jakop in H. Dobrovoljc, ur., *Pravopisna stikanja: Razprave o pravopisnih vprašanjih*, str. 141–49. Ljubljana: Založba ZRC.

Simon Krek, Tomaž Erjavec, Kaja Dobrovoljc, Sara Može, Nina Ledinek in Nanika Holz. 2013. Training corpus ssj500k 1.3, *Slovenian language resource repository CLARIN.SI*, http://hdl.handle.net/11356/1029.

Nataša Logar. 2005. Filter vrečka ali filtervrečka, foto posnetek ali fotoposnetek, ISDN paket ali ISDN-paket? V: M. Jesenšek, ur., *Knjižno in narečno besedoslovje slovenskega jezika*, str. 222–49. Maribor: Slavistično društvo.

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

Nataša Logar. 2012. Razmejitev med besednimi zvezami in zloženkami v sodobnem jezikovnem gradivu. V: N. Jakop in H. Dobrovoljc, ur., *Pravopisna stikanja: Razprave o pravopisnih vprašanjih*, str. 113–23. Ljubljana: Založba ZRC.

Jakob Rigler. 1971. H kritikam pravopisa, pravorečja in oblikoslovja v SSKJ. *Slavistična revija*, 19(4): 433–62.

Jože Toporišič. 1971. Pravopis, pravorečje in oblikoslovje v SSKJ I. *Slavistična revija*, 19(1): 55–75.

Ada Vidovič Muha. 1988. *Slovensko skladenjsko besedotvorje ob primerih zloženk*. Ljubljana: Partizanska knjiga.

»S kje pa si?« – Metapodatki o regionalni pripadnosti uporabnikov družbenega omrežja Twitter

Jaka Čibej,* Nikola Ljubešić†‡

* Oddelek za prevajalstvo, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
jaka.cibej@ff.uni-lj.si
† Odsek za tehnologije znanja, Institut »Jožef Stefan«
Jamova cesta 39, 1000 Ljubljana
‡ Odsek za informacijske in komunikacijske znanosti, Filozofska fakulteta, Univerza v Zagrebu
Ivana Lučića 3, 10000 Zagreb
nljubesi@ffzg.hr

Povzetek

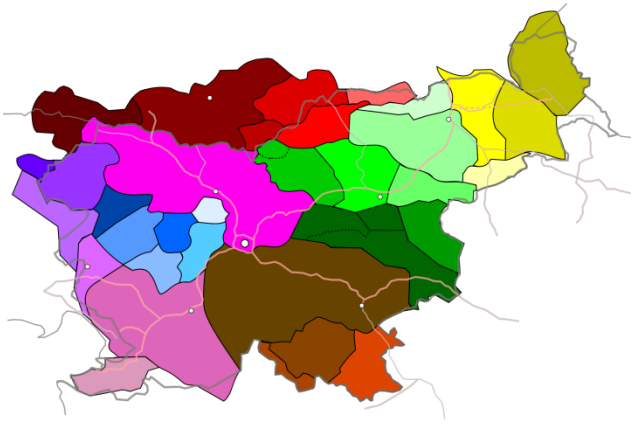
Slovenščina je kot narečno razčlenjen jezik z dialektološkega vidika že zelo dobro raziskana, a le v govoru. V dobi vsesplošne uporabe družbenih medijev se je slovenska regionalna jezikovna produkcija razširila tudi v računalniško posredovano komunikacijo prek številnih (pisnih) platform za sporazumevanje, kot sta npr. družbeni omrežji Facebook in Twitter. To nakazuje, da bo tudi spletno sporazumevanje odigralo vlogo v nadaljnjem razvoju slovenskih regionalnih jezikovnih različic, zato je nujno, da sodobne dialektološke raziskave upoštevajo tudi ta vidik jezikovne rabe. V prispevku zato predstavljamo prvo stopnjo v raziskavi slovenskih regionalnih jezikovnih različic na spletu, in sicer postopek dodajanja metapodatkov o regionalni pripadnosti uporabnikov Twitterja v korpus spletne slovenščine Janes. To bo omogočilo medregionalno kontrastivno primerjavo jezikovne produkcije in ugotavljanje specifik regionalnih jezikovnih različic na spletu.

"Where you from?" – Metadata on regional origin of Twitter users

From a dialectological perspective, Slovene as a dialectally diverse language has been researched to a considerable extent, but only in speech. Due to the pervasive presence of social media today, Slovene regional language production has expanded into computer-mediated communication through numerous (written) communication platforms, such as the social networks Facebook and Twitter. This indicates that internet communication will play a part in the future development of Slovene regional language variants, which is why modern dialectological research should also take this aspect of language use into account. In this paper, we present the first step in the research of Slovene regional language variants on the web: the addition of metadata on regional origin of tweeters in the Janes corpus of internet Slovene. The metadata will enable an inter-regional contrastive analysis of language production and a definition of specific characteristics of Slovene regional language variants on the web.

1 Uvod

Slovenščina je v jezikoslovnem smislu svojevrsten fenomen, saj je kljub relativno majhnemu številu govorcev in geografskemu ozemlju narečno zelo razčlenjena. Že ob začetku obširnejših dialektoloških raziskav v prvi polovici 20. stoletja je Fran Ramovš (1931) govorce slovenščine razdelil v 7 narečnih skupin s skupno več kot 30 narečji (Slika 1).



Slika 1: Slovenske narečne skupine.

Temu primerno je tudi dialektološki vidik slovenščine precej obširno raziskan, a kljub temu ni brez pomanjkljivosti. Tradicionalna dialektologija se je namreč pogosto opirala na tezo, da narečja izumirajo in se počasi zlivajo v standardno različico jezika (Kolarič, 1954) oziroma v najboljšem primeru v nestandardno različico, ki prevladuje v mestnih središčih (Ramovš, 1951). Raziskave so se zato omejevale na proučevanje t. i. »čistih« narečij, tj. na govorice tistih govorcev, ki so bili v svojem življenju čim bolj izolirani od drugih jezikovnih zvrsti in so kot taki predstavljali idealne govorce narečja (Kenda Jež, 2002: 26). Kriteriji, ki so predstavljali idealnega govorca, so bili v nekaterih primerih precej natančni, saj so poleg kraja bivanja določali tudi spol in stopnjo izobrazbe govorca, izpostavljenost drugim govoricam, govorico staršev ipd.[1]

Ob takšnem pristopu se porajajo številne kritike o reprezentativnosti vzorcev, obenem pa je iz nekaterih sodobnejših dialektoloških raziskav razvidno, da teza o izumiranju narečij ni povsem aksiomatična in da jezikovni razvoj narečja pelje v drugo smer, o kateri pa si dialektologi niso enotni: npr. da se bo raba narečij

[1] Logar (1958: 129) in Unuk (1997: 310) npr. zagovarjata, da so ženske boljši informatorji od moških, ker so večinoma bolj doma in tudi bolj konservativne v svoji govorici. Chambers in Trudgill (1994: 33) navajata, da je večina informatorjev v dialektoloških raziskavah starejših kmečkih moških, ki stalno živijo doma (ang. *NORM → Non-mobile Older Rural Male*).

omejevala sprva na ruralno okolje in nazadnje le še na kontekst folklornega kulturnega udejstvovanja ali da bodo zemljepisne jezikovne različice zamenjale družbene (Sgall et al., 1992), da se bodo lokalna narečja strnjevala v večja regionalna narečja (Niebaum in Macha, 1999; Kenda Jež, 2004) ali pa da bo v jezikovnem razvoju prišlo do t. i. nove dialektizacije (Labov, 1994), pri kateri bodo ključno vlogo odigrale jezikovne inovacije v govoricah jezikovnih skupnosti v urbanih središčih, ki se bodo postopoma razširile tudi v manjše jezikovne skupnosti na podeželju in tako ustvarile nov nabor narečij.[2] Tudi položaj narečnih jezikovnih različic ni povsem samoumeven. Jezikovne razmere na Norveškem so npr. zelo naklonjene rabi narečij v vseh funkcijskih zvrsteh (Jahr, 1997), govorci nasploh pa lahko jezikovne različice tudi zavestno kultivirajo ter jih uporabljajo v vedno večji meri (Reichan, 1999).

Pri proučevanju sprememb regionalnih jezikovnih različic v prihodnosti pa je treba v današnjih razmerah upoštevati še en vidik, ki prej ni bil prisoten. Z vzponom spleta in informacijske tehnologije v zadnjih 20 letih (še zlasti pa v zadnjem desetletju) so govorci pridobili številne nove platforme za (pisno) sporazumevanje, npr. spletne forume, novičarske portale in družbena omrežja, kot sta Facebook in Twitter. Jezik v računalniško posredovani komunikaciji (še posebej v klepetu in v drugih neformalnih kontekstih) pa se od standarda precej razlikuje (Crystal, 2001; Baron, 2010; Myslin in Gries, 2010; Erjavec in Fišer, 2013) in vsebuje tudi narečne prvine (Ueberwasser, 2013; Fišer et al., 2015).

Spletna komunikacija v jezikovni produkciji že dolgo več ne zajema zanemarljivega deleža, o čemer pričajo tudi podatki Statističnega urada Republike Slovenije:[3] v prvem četrtletju 2014 je v družbenih omrežjih sodelovalo skoraj 60 odstotkov oseb, 41 odstotkov pa je splet uporabljalo tudi za telefoniranje ali videotelefoniranje. Deleži iz leta v leto rastejo, kar nakazuje, da bo tudi spletna komunikacija odigrala vlogo pri nadaljnjem razvoju slovenskih regionalnih različic. Ključno je torej, da sodobne dialektološke raziskave upoštevajo tudi ta vidik jezikovne produkcije, proučevanje rabe regionalnih prvin v računalniško posredovani komunikaciji pa bo pripomoglo tudi k razvoju novih (ali izboljšanju že obstoječih) jezikovnih tehnologij za slovenščino, kot so označevalniki, lematizatorji, strojni prevajalniki ipd.

V prispevku zato predstavljamo prvi korak v raziskavi regionalne členjenosti v spletnem kontekstu. Najprej opravimo kratek pregled sorodnih raziskav, nato pa opišemo postopek, po katerem smo kategorizirali uporabnike Twitterja glede na regionalno pripadnost. Pridobljene metapodatke smo dodali v korpus spletne slovenščine Janes (Fišer et al., 2014) in tako ustvarili podkorpuse za proučevanje specifik rabe regionalnih jezikovnih različic[4] v pisnem sporazumevanju na spletu. Nazadnje še na kratko predstavimo preliminarne regionalne podkorpuse in navedemo predloge za prihodnje delo.

2 Pregled sorodnih raziskav

Za razliko od slovenščine so bile za številne tuje jezike že opravljene obsežne korpusne dialektološke raziskave, a najpogosteje na podlagi transkripcij govora v govornih korpusih, ki pa pogosto prvotno niso bili namenjeni dialektološkim raziskavam – British National Corpus npr. kljub razdelani taksonomiji vsebovanih narečij nudi samo standardizirano transkripcijo govora brez posnetkov. Za nizozemska narečja je bil zgrajen korpus DynaSAND (Kunst in Wesseling, 2010). Podoben projekt za nordijske jezike je Nordic Dialect Corpus (Johanessen et al., 2009), za angleščino pa Freiburg Corpus of English Dialects (Hernández, 2006).

Raziskovanje dialektalnih prvin v uporabniških spletnih vsebinah pa je tudi v tujini še precej sveže, deloma najbrž tudi zato, ker so bila jezikovnotehnološka orodja naučena na standardnih jezikovnih različicah in so se šele pred nedavnim začela prilagajati tudi šumnim besedilom, med katera lahko štejemo tudi narečno jezikovno produkcijo na spletu. Obenem je bilo to področje prej domena jezikovnih tehnologov kot jezikoslovcev. Predvsem z namenom gradnje novih jezikovnih tehnologij (za oblikoskladenjsko označevanje, strojno prevajanje, avtomatsko detekcijo regionalnih različic ipd.) je bilo opravljenih že veliko raziskav spletnih regionalnih različic arabščine (Harrat et al., 2013; Harrat et al., 2014; Cotterell in Callison-Burch, 2014) in ameriške angleščine (Eisenstein et al., 2010; Eisenstein et al., 2015), pa tudi manjših, jezikovnotehnološko manj podprtih jezikov, kot so npr. tatarsko narečje mišar (Khakimov et al., 2015), švicarska nemščina (Ruef in Ueberwasser, 2013) in alzaščina (Bernhard in Ligozat, 2013).

To kaže na svetovni trend, ki potrjuje, da bi bilo orodja za obdelavo regionalnih jezikovnih različic na spletu koristno razviti tudi za slovenščino, ki v tem smislu še ni jezikovnotehnološko podprta.

3 Metapodatki o regionalni pripadnosti uporabnikov Twitterja

V tem razdelku opisujemo postopek, po katerem smo klasificirali uporabnike Twitterja glede na njihovo regionalno pripadnost ter metodologijo in vire, ki so bili pri tem uporabljeni.

3.1 Korpus spletne slovenščine Janes

Trenutna različica[5] korpusa spletne slovenščine Janes, ki ga sestavljajo tviti, forumska sporočila, blogovski zapisi in komentarji na spletne novice, vsebuje približno 160 milijonov pojavnic. Od tega je skoraj 61 milijonov oz. 40 odstotkov tvitov, ki jih je spisalo približno 7.500 različnih avtorjev. Tviti v korpusu so že opremljeni z nekaterimi metapodatki (npr. sentiment, spol, ali gre za zasebnega uporabnika ali organizacijo, stopnja standardnosti besedila (Ljubešić et al., 2015)), kar omogoča natančnejše določanje ustreznega gradiva za številne jezikoslovne

[2] Labov (1994) predpostavlja, da so današnja narečja pravzaprav ostanki jezikovnega razvoja, ki se je začel v mestih in se nato postopoma razširil na podeželje.

[3] http://www.stat.si/StatWeb/glavnanavigacija/podatki/prikazistaronovico?IdNovice=6560

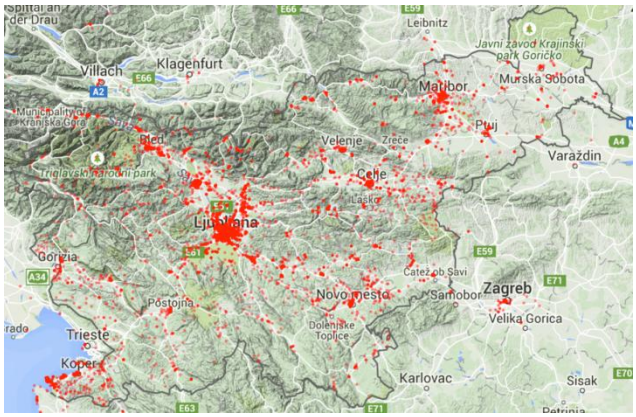
[4] Ker gre v primeru našega gradiva za pisni diskurz, ker se ne osredotočamo na konkretna narečja in ker pričakujemo, da se bo raba jezika na spletu precej razlikovala v primerjavi z rabo v govoru, se bomo izognili poimenovanju *narečje* in namesto tega za opazovane jezikovne zvrsti uporabljali manj specifični termin *regionalne jezikovne različice*, saj želimo jezik opisovati neodvisno od uveljavljene tipologije zvrsti brez vnaprejšnjega kategoriziranja.

[5] Različica 0.3 je bila zgrajena 5. marca 2015.

raziskave. Da bo korpus Janes uporaben tudi z geolingvističnega in dialektološkega vidika, smo uporabnikom Twitterja določili še regijo, iz katere najpogosteje pošiljajo tvite. To smo dosegli s pomočjo zbirke tvitov s podatki o geolokaciji, ki jo podrobneje predstavljamo v nadaljevanju.

3.2 Tviti s podatki o geolokaciji

Januarja 2015 smo začeli s pomočjo namenskega orodja TweetCat (Ljubešić et al., 2014) zajemati slovenske tvite s podatki o geolokaciji, tj. s koordinatami kraja, s katerega je bil tvit poslan. Do avgusta 2015 je bilo zajetih približno 130.000 tvitov, ki jih je napisalo 1.661 tviterašev po vsej Sloveniji (glej Sliko 2).



Slika 2: Razporeditev zajetih tvitov po Sloveniji. Vsaka rdeča pika predstavlja en tvit.

Izločili smo uporabnike, ki niso vključeni v korpus Janes, in tiste, ki niso zasebni uporabniki, saj organizacije na Twitterju v prevladujoči meri objavljajo v standardni slovenščini in bi v naših regionalnih podkorpusih predstavljale šum. Ostalo nam je 119.236 tvitov, ki jih je napisalo 1.461 uporabnikov. V korpusu Janes je uporabnikov, ki so označeni kot zasebni, 5.806. Z zbiranjem tvitov z geolokacijo smo do avgusta 2015 torej zajeli približno četrtino v korpus vključenih uporabnikov.

3.3 Razdelitev Slovenije na regije

V naslednjem koraku smo Slovenijo s pomočjo orodja Google Maps API v3 Tool[6] razdelili na 9 koordinatnih poligonov, ki predstavljajo 7 narečnih skupin[7] (gorenjsko, dolenjsko, štajersko, panonsko, koroško, rovtarsko in primorsko; glej Sliko 3) ter Ljubljano in Maribor.

Ljubljano in Maribor smo se odločili obravnavati posebej kot urbani središči, h katerima gravitira prebivalstvo iz številnih drugih krajev (tako okoliških kot bolj oddaljenih) in ki bi kot taki vnesli precejšnjo mero šuma v druge regije. Tak pristop zagovarja tudi Zemljarič Miklavčič (2008: 79).

[6] http://www.birdtheme.org/useful/v3tool.html

[7] Kategorizacija po Toporišiču (2000: 23–24) sicer v dolenjski skupini loči tudi posebno osmo, kočevsko skupino (na območju nekdanje nemške poselitve), a smo se v tem prispevku zaradi omejenega števila podatkov osredotočili le na 7 narečnih skupin po Ramovšu (1931), v katere smo vključili tudi slovenske manjšine v Italiji, v Avstriji in na Madžarskem.

Slika 3: Razdelitev Slovenije na koordinatne poligone.

3.4 Določitev regionalne pripadnosti uporabnikov

Za vsak tvit od preostalih 1.461 tviterašev iz baze tvitov z geolokacijo smo nato v programskem jeziku Python z metodo metanja žarka (ang. *ray-casting method*) preverili, iz katere regije je bil poslan. Metoda iz podane točke (v našem primeru so to koordinate tvita) pošlje žarek in preveri število presečišč med žarkom in robovi podanega poligona (regije) – če je število liho, točka leži v notranjosti poligona. Razporeditev po regijah je prikazana v Tabeli 1.

Regija	Število tvitov	Delež (%)
Gorenjska	22.070	18,51
Dolenjska	6.922	5,81
Štajerska	9.284	7,79
Panonska	2.512	2,11
Koroška	4.203	3,52
Primorska	5.748	4,82
Rovtarska	2.348	1,97
Ljubljana	43.018	36,08
Maribor	4.340	3,64
Tujina	18.791	15,76
Skupno	119.236	100,00

Tabela 1: Razporeditev tvitov po regijah.

Največ tvitov (36 %) je bilo poslanih iz Ljubljane, najmanj pa iz rovtarske (slaba 2 %) in panonske regije (dobra 2 %), ki sta tudi po površini med najmanjšimi.

Uporabnikom, ki so več kot 90 % tvitov z geolokacijo poslali iz ene same regije in so obenem poslali vsaj 3 tvite, smo pripisali metapodatek o regionalni pripadnosti. Takšnih uporabnikov je bilo 364, končni rezultati njihove kategorizacije pa so prikazani v Tabeli 2. Uporabniki, ki so tvite pošiljali večinoma iz tujine, za našo raziskavo niso relevantni, a jih kljub temu navajamo v tabeli, saj predstavljajo nezanemarljiv delež.

Regija	Število tviterašev	Delež (%)
Gorenjska	48	13,19
Dolenjska	22	6,04
Štajerska	42	11,54
Panonska	14	3,85
Koroška	5	1,37
Primorska	31	8,52
Rovtarska	7	1,92
Ljubljana	116	31,87
Maribor	14	3,85
Tujina	65	17,86
Skupno	364	100,00

Tabela 2: Število uporabnikov po regijah.

V povprečju je vsak uporabnik poslal približno 74 tvitov, mediana pa je 18 tvitov. Uporabnikov, ki so poslali zgolj 3 tvite, je bilo skupno 39 (2 dolenjska, 4 gorenjski, 8 ljubljanskih, 1 mariborski, 3 panonski, 4 primorski, 1 rovtarski, 6 štajerskih in 10 iz tujine). Več kot 74 tvitov je poslalo 89 uporabnikov, najproduktivnejši pa je poslal kar 1188 tvitov, in sicer iz Ljubljane.

Zanimivo je, da je koroških uporabnikov le 5, spisali pa so skupno 167 tvitov. Število tvitov iz te regije ni bilo majhno (približno 4.200), zato lahko sklepamo, da večina tamkajšnjih uporabnikov tvita tudi iz drugih regij (oziroma da so velik delež tamkajšnjih tvitov prispevali uporabniki iz drugih regij), zato zaradi strogih kriterijev (najmanj 90-odstotna pripadnost eni sami regiji) niso bili vključeni v končni nabor. Podobno je z rovtarsko skupino, pri kateri je uporabnikov le 7, poslali pa so skupno 956 tvitov. Število rovtarskih tvitov je bilo primerljivo s panonsko skupino, pri kateri pa je uporabnikov 14.

4 Regionalni podkorpusi

Uporabnikom Twitterja v korpusu Janes v smo pripisali metapodatke o regionalni pripadnosti in tako izdelali 9 regionalnih podkorpusov ter zabeležili število pojavnic v njih. Preverili smo tudi število pojavnic, če iščemo samo po tvitih, ki so bili v korpusu označeni kot nestandardni (L2 in L3), ter izračunali deleže nestandardnih tvitov. Rezultati so predstavljeni v Tabeli 3.

Regionalni podkorpus	Število pojavnic	Število pojavnic (L2, L3)	Delež L2 in L3 (%)
Gorenjska	37.683	16.679	44,26
Dolenjska	17.364	5.503	31,69
Štajerska	41.712	14.091	33,78
Panonska	5.020	1.345	26,79
Koroška	6.207	2.644	42,60
Primorska	13.917	3.579	25,72
Rovtarska	4.823	1.778	36,87
Ljubljana	92.104	27.036	29,35
Maribor	4.789	1.205	25,16

Tabela 3: Velikost regionalnih podkorpusov.

Kot je bilo pričakovano, je po številu pojavnic največji ljubljanski podkorpus, najmanjši pa so panonski, koroški, rovtarski in mariborski (kar je morda nekoliko presenetljivo, saj smo na začetku pričakovali, da bo kot drugo največje slovensko mesto v zbirko zajetih tvitov doprinesel mnogo več).

Po deležu nestandardnosti izstopata gorenjski in koroški podkorpus, oba z dobrimi 40 % nestandardnih tvitov. Najbolj standardni so mariborski, primorski in panonski podkorpus, pri katerih je kot nestandardnih označenih le dobrih 25 % tvitov. Ti podatki nam podajo grobo oceno, kateri podkorpusi vsebujejo največ nestandardnih (in potencialno tipično regionalnih) jezikovnih prvin, za podrobnejši vpogled v njihovo vsebino pa bo potrebna še temeljita jezikoslovna analiza.

5 Zaključek

V prispevku smo opisali postopek, po katerem smo uporabnikom s pomočjo zbirke tvitov s podatki o geolokaciji pripisali metapodatke o regionalni pripadnosti. V prihodnjem delu bomo poskušali obseg nastalih podkorpusov povečati z zajemanjem novih tvitov z geolokacijo (in novih uporabnikov), tiste podkorpuse, ki se bodo zaradi premajhnega števila podatkov izkazali za neuporabne, pa bomo po potrebi izključili iz nadaljnjih raziskav.

Poleg tega bomo temeljito preučili sestavo in vsebino regionalnih podkorpusov ter poskušali ugotoviti značilne razlike med regionalnimi jezikovnimi različicami spletne slovenščine, npr. z izdelavo seznamov ključnih besed za vsak regionalni podkorpus glede na celotni podkorpus tvitov korpusa Janes ter s kvalitativnim pregledom materiala z vidika regionalnih jezikovnih značilnosti (npr. regionalne razširjenosti variantnih različic zapisa besed, npr. *kaj*, *ka*, *kva*, *kwa*, *kuga*). Novi metapodatki bodo med drugim omogočili tudi primerjavo z govornim korpusom GOS, odkrite značilke pa bomo nato uporabili pri razvoju in učenju modela za avtomatsko prepoznavanje regionalnih jezikovnih različic slovenščine na spletu. Za primerjavo pa bomo avtorje poskušali razvrstiti tudi z gručenjem, ki ne bo odvisno od vnaprej določenih regij.

6 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.

Avtorja se za pomoč in nasvete pri pripravi prispevka iskreno zahvaljujeta Tomažu Erjavcu in Darji Fišer ter anonimnim recenzentom za konstruktivne opombe.

7 Literatura

Delphine Bernhard in Anne-Laure Ligozat. 2013. Hassle-free POS-Tagging for the Alsatian Dialects. V: Zampieri, M., S. Diwersy (ur.). Non-standard Data Sources in Corpus-based Research. Aachen: Shaker Verlag, str. 85–92.

Janne Bondi Johannessen, Joel Priestley, Kristin Hagen, Tor Anders Åfarli in Øystein Alexander Vangsnes. 2009. The Nordic Dialect Corpus – an Advanced Research Tool. V: K. Jokinen in E. Bick (ur.): Proceedings of the 17th Nordic Conference of

Computational Linguistics NODALIDA 2009. NEALT Proceedings Series Volume 4.

Ryan Cotterell in Chris Callison-Burch. 2014. A Multi-Dialect, Multi-Genre Corpus of Informal Written Arabic. V: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). Reykjavik: ELRA.

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

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith in Eric P. Xing. 2010. A latent variable model for geographic lexical variation. V: Proceedings of Empirical Methods for Natural Language Processing (EMNLP), str. 1277–1287. Stroudsburg, Pennsylvania: Association for Computational Linguistics.

Jacob Eisenstein. 2015. Written dialect variation in online social media. V: C. Boberg, J. Nerbonne in D. Watt (ur.): Handbook of Dialectology. Wiley.

Tomaž Erjavec in Darja Fišer. 2013. Jezik slovenskih tvitov: korpusna raziskava. V: Družbena funkcijskost jezika: (vidiki, merila, opredelitve), Obdobja 32. Ljubljana: Znanstvena založba Filozofske fakultete, str. 109–116.

Darja Fišer, Tomaž Erjavec, Ana Zwitter Vitez in Nikola Ljubešić. 2014. JANES se predstavi: metode, orodja in viri za nestandardno pisno spletno slovenščino. V: Zbornik Devete konference Jezikovne tehnologije. Ljubljana: Institut Jožef Stefan.

Darja Fišer, Tomaž Erjavec, Jaka Čibej in Nikola Ljubešić. 2015. Gradnja in analiza korpusa spletne slovenščine JANES. Obdobja 2015.

Ernst Håkon Jahr. 1997. On the Use of Dialects in Norway. V: Heinrich Ramisch in Kenneth Wyne (ur.): Language in Time and Space: Studies in Honour of Wolfgang Viereck on the Occasion of his 60th Birthday, str. 363–369. Stuttgart: Franz Steiner Verlag.

Salima Harrat, Karima Meftouh, Mourad Abbas in Kamel Smaili. 2014. Building Resources for Algerian Arabic Dialects.V: Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH 2014). Singapur.

Salima Harrat, Mourad Abbas, Karima Meftouh in Kamel Smaili. 2013. Diacritics restoration for Arabic dialect texts. V: Proceedings of the 14th Annual Conference of the International Speech Communication Association (INTERSPEECH 2013). Francija.

Nuria Hernández. 2006. User's Guide to FRED. Freiburg: University of Freiburg. http://www.freidok.uni-freiburg.de/volltexte/2489/

Jack K. Chambers in Peter Trudgill. 1994. Dialectology. Cambridge: University Press.

Karmen Kenda Jež. 2002. Cerkljansko narečje: teroetični model dialektološkega raziskovanja na zgledu besedišča in glasoslovja. Doktorska disertacija. Ljubljana: Filozofska fakulteta.

Karmen Kenda Jež. 2004. Narečje kot jezikovnozvrstna kategorija v sodobnem jezikoslovju. V: E. Kržišnik (ur.): Obdobja 22. Ljubljana: Filozofska fakulteta Univerze v Ljubljani, Center za slovenščino kot drugi/tuji jezik, Oddelek za slovenistiko, str. 263–276.

Bulat Khakimov, Farid Salimov in Dariya Ramazanova. 2015. Building dialectological corpora for Turkic languages: Mishar dialect of Tatar. V: Procedia – Social and Behavioral Sciences 198. str. 218–225.

Rudolf Kolarič. 1954. Die slowenische Mundartsforschung. V: Orbis: Bulletin International de Documentation Linguistique 3/1, str. 182–188. Louvain.

William Labov. 1994. Principles of Linguistic Change 1: Internal Factors. Oxford/Cambridge: Blackwell.

Nikola Ljubešić, Darja Fišer in Tomaž Erjavec. 2014. TweetCaT: a tool for building Twitter corpora of smaller languages. V: Zbornik konference Ninth International Conference on Language Resources and Evaluation Reykjavik, Iceland. LREC 2014: proceedings. 2279–2283. http://www.lrec-conf.org/proceedings/lrec2014/pdf/834_Paper.pdf

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. RANLP 2015.

Tine Logar. 1959. Iz priprav za lingvistični atlas. V: Jezik in slovstvo 4, str. 129–135.

Mark Myslín in Stefan T. Gries. 2010. k dixez? A corpus study of Spanish Internet orthography. V: Literacy and Linguistic Computing, 25 (1), str. 85–104.

Herman Niebaum in Jürgen Macha. 1999. Einführung in die Dialektologie des Deutschen. Tübingen: Max Niemeyer Verlag.

Jan Pieter Kunst in Franca Wesseling. 2010. Dialect Corpora Taken Further: The DynaSAND corpus and its application in newer tools. V: Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation, str. 759–767.

Fran Ramovš. 1931. Dialektološka karta slovenskega jezika. Ljubljana: Rektorat univerze kralja Aleksandra I. in J. Blaznika nasl. – Univerzitetna tiskarna.

Fran Ramovš. 1951. Osnovna črta v oblikovanju slovenskega vokalizma. V: Slavistična revija 4, str. 1–9.

Jerzy Reichan. 1999. Gwary polskie w końcu XX w. V: Polszczyzna 2000, str. 262–278.

Beni Ruef in Simone Ueberwasser. 2013. The Taming of a Dialect: Interlinear Glossing of Swiss German Text Messages. V: Zampieri, M., S. Diwersy (ur.). Non-standard Data Sources in Corpus-based Research. Aachen: Shaker Verlag, str. 61–68.

Naomi S. Baron. 2010. Always On: Language in an Online and Mobile World. Oxford University Press.

Petr Sgall, Jiři Hronek, Alexandr Stich in Ján Horecký. 1992. Variation in Language: Code Switching in Czech as a Challenge for Sociolinguistics. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Jože Toporišič. 2000. Slovenska slovnica: četrta, prenovljena in razširjena izdaja. Maribor: Obzorja 2000.

Simone Ueberwasser. 2013. Non-standard data in Swiss text messages with a special focus on dialectal forms. V: M. Zampieri in S. Diwersy (ur.): Non-standard Data Sources in Corpus-based Research. Aachen: Shaker Verlag, str. 7–24.

Drago Unuk. 1997. Dialektologija kot jezikoslovna disciplina. V: Jezik in slovstvo 43, str. 307–313.

Jana Zemljarič Miklavčič. 2008. Govorni korpusi. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani.

Kaj nam o slovenščini lahko pove kolaborativni spletni slovar?

Kaja Dolar

Université Paris Ouest Nanterre La Défense, Modyco
Avenue de la République, 92 000 Nanterre, Francija
dolar.kaja@gmail.com

Povzetek

V prispevku so predstavljeni kolaborativni slovarji kot zanimiv in pertinenten jezikovni vir. Posebej se posvečamo slovenskemu kolaborativnemu slovarju *Razvezani jezik*, kjer preučujemo težnje razvoja spletne slovenščine. V ospredju je problem neologizmov. Natančneje raziskujemo različne vrste formalnih neologizmov. Ti v slovenščini nastajajo na osnovi štirih osnovnih besedotvornih postopkov: izpeljave, sestavljanja, zlaganja in sklapljanja, ob čemer lahko opazimo tudi manj običajne vrste podstav (tujejezične podstave, lastna imena kot besedotvorne podstave ipd.). Poleg izpeljank, sestavljank, zloženk in sklopov v *Razvezanem jeziku* najdemo tudi neologizme, ki nastajajo s kratičenjem, krnjenjem in obračanjem zlogov. Ti različni postopki se pogosto tudi medsebojno kombinirajo v formalno kompleksnejših neologizmih, kjer je potrebno še posebej izpostaviti t. i. besedo-kovček (*mot-valise* oziroma *mot porte-manteau*; pojem je vpeljal Lewis Carroll), ki je kombinacija krnjenja in zlaganja. Zdi se, da je ta besedotvorni postopek v slovenščini izredno produktiven. Po drugi strani pa bi za nekatere primere iz *Razvezanega jezika* lahko rekli, da se nahajajo nekako med kategorijama formalnih in semantičnih neologizmov oziroma ju združujejo. Gre za nekakšne besedne igre, ki humorno izrabljajo formalne in semantične možnosti jezika za tvorjenje novih besed in besednih zvez.

What can be learned about Slovene language on basis of collaborative on-line dictionary?

In the present paper we argue that collaborative dictionaries are a pertinent and interesting language source. The paper focuses on Slovene collaborative dictionary *Razvezani jezik* as a source for the research of linguistic tendencies in online Slovene language with special emphasis on neology. We examine different types of formal neologisms. They are mostly based on derivation and compounding where one can notice also some less common lexical bases such as foreign bases or proper names. The corpus includes also some less standard cases of neologisms based on shortening, abbreviations, inversion of syllabi etc. All these lexical procedures can be combined in formally more complex neologisms. The combination of shortening (clipping) and compounding, the so called blending or *mot porte-manteau* (invented by Lewis Carroll), seems to be especially productive in Slovene language. On the other hand some cases are somewhere in between the categories of formal and semantic neologisms or combine both. They are kind of word-play that use formal and semantic linguistic features, treating them with humour to form new words and idioms.

1 Uvod

Jezik se spreminja pred našimi očmi, ne da bi se vedno zavedali; številne tradicionalne opisne kategorije ne ustrezajo več jezikovni realnosti, po drugi strani pa nekatere oblike niso splošno sprejete (Benveniste, 1974: 163). Lahko bi rekli, da to morda še posebej velja za spletni jezik, ki je deležen še hitrejšega razvoja in je stalno podvržen spremembam.

V prispevku bomo osvetlili kolaborativni slovar kot zanimiv in pertinenten jezikovni vir. Naša raziskava se osredotoča na neologizme, jezikovne novosti, saj so »inovacijski procesi v leksiki odsev dinamičnih teženj« (Stramljič Breznik, 2003: 105), in na problem neologije, ki jo lahko definiramo obenem kot nastajanje novih besed in kot vedo o njih (Veliki slovar tujk, 2002: 782).

Pruvost in Sablayrolles (2003: 10) ločita dve vrsti neologizmov. Formalni neologizmi nastanejo s pomočjo besedotvornih postopkov. Tako v slovenščini ločimo izpeljanke, zloženke, sklope in sestavljanke. Semantični neologizmi pa nastanejo, ko že obstoječa beseda dobi nov pomen preko metaforičnega ali metonimičnega prenosa.

Da bi osvetlili težnje jezikovnega razvoja spletne slovenščine, se bomo v prispevku posvetili formalnim neologizmom, različnim postopkom in njihovim kombinacijam, obravnavali pa bomo tudi nekatere manj pogoste primere neologizmov.

2 Kolaborativni slovar kot jezikovni vir

Kolaborativni spletni slovarji so vsem dostopni internetni slovarji, ki so poleg prebiranja namenjeni tudi aktivnemu sodelovanju – tako dodajanju novih gesel oziroma iztočnic in slovarskih člankov kot tudi spreminjanju že obstoječih člankov. Če tradicionalne slovarje ločujemo na enciklopedične in jezikovne (Polguère, 2003: 198), velja to tudi za kolaborativne: med prve sodi, denimo, *Wikipedija* (enciklopedični slovar), med druge pa *Wiktionary* (večjezični kolaborativni slovar) ali pa *Urban Dictionary* (kolaborativni slovar anglo-ameriškega slenga). Slovenski kolaborativni slovar *Razvezani jezik* je začel nastajati leta 2004 kot umetniški projekt in danes šteje 4800 gesel.[1]

Poudariti velja, da je takšen pristop v leksikografiji sorazmerno nov (zadnjih petnajst let, kar je povezano s široko dostopnostjo interneta), v ospredje pa postavlja kompetenco govorcev. Gre za nekakšen »bottom-up« pristop (Carr, 1997), ki bralce spodbuja k pisanju (Meyer in Gurevych, 2012: 259). Slovarja ne pišejo leksikografi, temveč ga lahko piše vsakdo, tako da imajo govorci kot končni uporabniki (*end users*, Cotter in Damaso, 2007: 8) neposreden vpliv na slovaropisje. Gre za fenomen t. i. modrosti množic (*wisdom of crowds*, Surowiecki, 2005).

Razvezani jezik se uvršča v sodobni trend t. i. participativne leksikografije, ki temelji na kolaborativnosti

[1] November 2015.

(*collaboration*, *collaborative*) oziroma participaciji, ki je del leksikografskega postopka in se razlikuje od t. i. množičenja (*crowdsourcing*), ki je leksikografski postopek, ki ga vsaj v določenih fazah nadzira leksikograf oziroma leksikografski kolektiv. Uporaba množičenja za leksikografske namene in kolaborativna leksikografija sta torej dva različna pristopa.

Kolaborativno leksikografijo lahko definiramo kot zbiranje podatkov na osnovi participativnosti oziroma prispevkov skupnosti (Granger, 2012: 5), kot slovarski žanr, ki združuje leksikografska načela in spletu lastne komunikacijske tehnologije. Tako nastane kontekst, v katerem uporabniki sodelujejo in se spodbujajo v ustvarjanju pomenov (Cotter in Damaso, 2007: 1).

Če se kolaborativni slovarji opirajo na kolaborativnost, takšna oblika participativne leksikografije lahko razkrije veliko o težnjah jezikovnega razvoja. Raziskujemo lahko, denimo, kateri postopki se pojavljajo pri nastajanju novih besed, na katera semantična polja se neologizmi najpogosteje nanašajo, od kod prihajajo tujejezične prvine itn. Ker gre za deskriptivne slovarje, to pomeni, da lahko raziskujemo tudi vprašanje obstoječe pravopisne norme – kako se nove besede zapisujejo, kako se spreminja zapis že obstoječih besed ipd.

Kolaborativni slovarji so lahko pertinenten in bogat jezikovni vir, ker gre za izredno ažurne leksikalne baze, saj jih lahko govorci sproti posodabljajo. Tako lahko v njih zasledimo posamezne neologizme mnogo prej, kot se ti pojavijo v tradicionalnih slovarjih. Do teh lahko namreč neologizmi potujejo več let. Ni pa nujno, da se beseda, zabeležena v kolaborativnemu slovarju, obdrži – lahko tudi zelo hitro izgine iz besednjaka. Toda določen neologizem se lahko razširi prav zato, ker je bil zabeležen v slovarju, čeprav se to dogaja redkeje. Prav tako gre za jezikovni vir, ki pokriva različne jezikovne zvrsti; v njih tako najdemo tudi slengovske in pogovorne izraze, kletvice, frazeme ipd. Pokriva torej vrsto nestandardnih rab.

Kljub temu, da gre za neustaljene besede, pa je potrebno razlikovati med večkrat uporabljenimi neustaljenimi izrazi oziroma pomeni, kot so denimo **lajkati**, **počutišče**, **pahorizem**, **kočerja** (slednja je navedena tudi v *Slovarju novejšega besedja slovenskega jezika*), ki jih lahko najdemo v splošnem jeziku, in popolnoma individualnimi rabami oziroma pojavnicami, kot denimo **imeti doma petro majdič**, **BTŠ** (Biotehniška šola Naklo; butci tut študirajo) ipd. V prispevku obravnavamo tako neologizme, ki se pojavljajo zgolj v *Razvezanem jeziku*, kot tiste, ki obstajajo zunaj njega in so se bodisi razširili s pomočjo *Razvezanega jezika* bodisi pa so tja lahko prišli naknadno.

Poudariti je potrebno, da ne gre za referenčni ali preskriptivni slovar, temveč za deskriptivni slovar. *Razvezani jezik* kaže dinamiko jezikovne ustvarjalnosti, ki ni vedno kolektivna; pravzaprav je predvsem individualna. Izraža torej jezikovni ustvarjalni potencial, vendar ne toliko splošnega jezika, pač pa jezika, ki je poseben in v slovarjih še ne opisan (če ima nek izraz opis v obstoječem slovarju, je za *Razvezani jezik* že v izhodišču manj zanimiv).

Ob tem ne moremo mimo nekaterih dilem, ki se porajajo pri uporabi kolaborativnega slovarja kot jezikovnega vira. Fišer in Čibej (2015) opozarjata, da se množičenje v leksikografiji še zmeraj sooča s številnimi predsodki, in lahko bi rekli, da to velja tudi za kolaborativne slovarje – morda celo v večji meri, saj leksikograf ni vključen v proces ustvarjanja slovarja. Pojavlja se tudi vprašanje legitimnosti in kakovosti vsebine; toda raziskave kažejo, da po kvaliteti kolaborativna leksikografija ne zaostaja nujno za tradicionalno (Meyer in Gurevych, 2012). Kvaliteta slovarskih člankov – ki ni vprašljiva zgolj zato, ker je ne ustvarjajo strokovnjaki – se regulira po eni strani z zunanjimi mehanizmi (ocenjevanje člankov »za« oziroma »proti«), po drugi pa obstaja tudi notranji proces regulacije, saj se pisci popravljajo med seboj.

Prav tako naletimo na vprašanja reprezentativnosti, velikosti in dejanske razširjenosti. Toda kolaborativni slovarji kot taki niso nujno zasnovani kot reprezentativni ali izčrpni. Z leksikografskega vidika sta po našem mnenju to vzporedna pristopa; tradicionalna in participativna leksikografija se ne izključujeta.

Kar zadeva očitek glede velikosti, omenimo, da je kolaborativne slovarje mogoče avtomatsko ali pol-avtomatsko povečati. Med tovrstne poskuse sodi tudi sistem Wisigoth, ki naj bi v večjezičnem kolaborativnem slovarju *Wiktionary* pripomogel k povečanju mreže sinonimov (Sajous et al., 2011: 24).

Ti možni ugovori sicer ne postavljajo pod vprašaj kolaborativnega slovarja kot jezikovnega vira, jih pa je potrebno pri analizi upoštevati.

V nadaljnji analizi nam bo slovenski kolaborativni slovar *Razvezani jezik* služil kot jezikovni vir za analizo neologizmov v spletni slovenščini.

3 Formalni neologizmi v slovenščini

S formalnega vidika ločimo v slovenščini štiri glavne besedotvorne načine: izpeljavo, sestavljanje, zlaganje in sklapljanje (Toporišič, 2000: 157–160; 2001: 156–160), pri čemer imamo eno-, dvo- ali večmorfemska obrazila (Vidovič-Muha, 2011: 24–25). Zdi se, da se formalni neologizmi v slovenščini porajajo na osnovi vseh naštetih postopkov, prav tako pa najdemo tudi primere kratičenja, krnjenja in obračanja vrstnega reda zlogov v besedah.

3.1 Izpeljanke

Pri izpeljavi »novo besedo dobimo tako, da podstavi (enodelni, in sicer nezloženi, predložni, zloženi) dodamo priponsko ali poponsko obrazilo: *sin-**ko**, proda-**ja**-ti*, *Obsotel-**je**, trdoglav-**ost**, ubiti **se***; pri tem prvotni podstavi pogosto spreminjamo slovnične značilnosti […]. Izpeljane besede so izpeljanke, način tvorjenja pa priponjanje (sufiksacija)« (Toporišič, 2000: 156–157). Izpeljanke so tako sestavljene iz podstave in pripone. Te so med neologizmi v *Razvezanem jeziku* sorazmerno pogoste. Omenimo nekaj primerov:[2] **kokičar** (oseba, ki običajno v

[2] Pomeni so iz spletne baze www.razvezanijezik.org zajeti septembra 2015. Slovarske razlage so po večini so močno okrnjene.

preddverjih kinematografov prodaja kokice), **lulčniti** (pomen je dovršen: iti se polulat), **očalnica** (z eno besedo povemo to, za kar potrebujemo ponavadi tri: etui za očala; besedotvorno izhaja iz podobnih izhodišč kot besede: denarnica, drvarnica), **vozičkati (se)** (sprehodi, na katere se odpravimo z dojenčkom v vozičku). Potrebno je poudariti, da imajo izpeljanke lahko tudi manj običajne podstave, saj lahko med drugim izhajajo iz tujih jezikov ali lasnih imen.

3.1.1 Tujejezične podstave in pripone

Nekatere izpeljanke vsebujejo tujejezične podstave. Primeri za to so **leftun** (volivec levice, levičar), **fejmič** (oseba, ki ima veliko lajkov na facebooku), **lajkati** (všeč biti), kjer so tuje podstave (*left*, *famous*, *like*) pravopisno prilagojene (fejm-, lajk-), pridane pa so jim slovenke pripone (-un, -ič, -ati). Nasprotno so tuje pripone na slovenskih podstavah redke. Takšen primer je, denimo, **gleding** (posebna vrsto šopinga, ki je izrazito značilna, kadar ga opravljamo brez denarja), kjer je slovenski podstavi gled- (gledati) pridana angleška pripona -ing, ki v tem primeru izraža dejanje.

Vendar je prav tako prisoten nasproten proces, ko tujezične podstave dobijo slovenske ekvivalente (ki niso nujno dobesedni prevodi, ampak so lahko tudi metaforične ali metonimične narave) in so nato deležne izpeljave. Takšni primeri so **brbotalnik** (*jacuzzi*), **ključnik** (*hashtag*), **kratkič** (SMS), **poskočnik** (*strat up*), **počutišče** (*wellness center*; lokal za dobro počutje), **čivkar** (*twiterer*), **čilišče** (*fitness*), **dotičnik** (*touchscreen*) itn. Številni neologizmi so narejeni po principu analogije: **sebič** in **sebček** kot dve varianti za angleški *selfie* vsebujeta seb-, ki je prevod podstave *self*, in slovenski priponi -ek oziroma -ič. Zdi se, da je pri tovrstnih neologizmih analogija ključnega pomena.

3.1.2 Lastno ime kot besedotvorna podstava

Nekatera lastna imena v *Razvezanem jeziku* delujejo metaforično (kot vrstna poimenovanja iz lastnih imen), kot denimo **bambi** (prikupno dekle manjše oziroma drobne rasti, zato pa zelo nežnega obraza in velikih, zvedavih oči) ali **imeti doma petro majdič** (občasno pasti v luknjo). V okviru dane raziskave se jim ne bomo podrobneje posvečali, saj so to v prvi vrsti semantičnimi neologizmi (nov pomen že obstoječega leksema), besedotvorno pa ostanejo nespremenjena (Dolar, 2014: 242–245).

Lahko pa lastno ime služi kot besedotvorna podstava pri izpeljankah. Takšni primeri so, denimo, **mazetniti** (suvereno, navdušujoče in gladko zmagati), **zakanglati** (potrošiti iz mestnega proračuna), **pahorizem** (smešna in neprimerna besedna zveza), **virantovanje** (cincanje, neodločnost, špekuliranje), **brecljevanje** (onemogočanje, oviranje ali nasprotovanje glasnim manifestacijam rimskokatoliške cerkve), **lidlast** (nizkocenovni, za enkratno uporabo), ipd.

Lastna imena, ki so osnova izpeljavi, so zlahka prepoznavna (Maze, Kangler, Pahor, Virant, Brecelj, Lidl). Pogosto gre za osebe iz sveta športa, politike ipd. Poleg formalne spremembe (izpeljave) je za te lekseme značilen tudi pomenski zdrs, saj lastno ime nima več izključno referenčne funkcije, temveč deluje metaforično (ibid.).

3.2 Sestavljanke

Sestavljanje je proces, v katerem »eno sestavino besednozvezne podstave zamenjamo s predponskim obrazilom, npr. ***nížji*** *kúhar* > ***pòd****kúhar*« (Slovenski pravopis, 2001: 109), »enodelni podstavi dodamo naglašeno predponsko obrazilo; slovnične značilnosti ostanejo neprizadete, zgubi se le njena začetnost« (Toporišič, 2000: 159). Sestavljanke tako vsebujejo predpono in podstavo. V nasprotju z izpeljankami pa so sestavljanke v *Razvezanem jeziku* sorazmerno redke. Omenimo lahko naslednje neologizme: **dekrasirati** (odstraniti ali odstranjevati okraske, npr. po božično-novoletnih praznikih), **pobožičnica** in **predbožičnica** (manična seansa nespečnih otrok, ki se odvija v otroški sobi na božično noč v velikem pričakovanju jutranjega odpiranja daril), **zavrjeti** (nasesti, pasti na finto), **zbudilka** (ura, ob njeni sprožitvi se dejansko zbudimo), **sosošolec** (učenec oziroma dijak drugega oddelka istega letnika tj. iz paralelke) ipd. Prav tako najdemo **stopnjevanje z nad-** (najboljši > naddober, najkrajši > nadkratek, kjer ima predpona nad- ekspresivno vrednost).

Ti neologizmi sestojijo iz sicer že obstoječih predpon (pred-, za-, z-) in podstav (božičnica, verjeti, budilka), nov pa je njihov spoj in prav tu lahko govorimo o inovativnosti. Omenjeni neologizmi namreč na neobičajen način združujejo že obstoječe jezikovne prvine. Nekoliko drugačen je primer **iBedarija** (obsedenost z izdelki določene znamke), kjer je predpona i- tujejezična prvina, ki aludira na Applove tehnološke izdelke.

3.3 Zloženke

Za zlaganje je značilno, da »dve sestavini (včasih tudi več) besednozvezne podstave povezujemo samo z medponskim obrazilom; če je besed v podstavi več, preostalo izrazimo še s priponskim obrazilom: *zdravnik živine* > *živin**o**zdravnik*« (Slovenski pravopis, 2001: 109); »večdelna govorna podstava se večinoma poveže z medpono -o-, -e-, -i-, z imenovalniško, tožilniško ali rodilniško končnico; končnemu delu podstave je bodisi dodana pripona ali pa v njem nastopa kar oblika prvotne govorne podstave« (Toporišič, 2000: 158). Zloženke so torej sestavljene iz podstave, medpone in podstave; tudi te se v *Razvezanem jeziku* pojavljajo veliko redkeje kot izpeljanke. Med njimi najdemo primere, kot so **muhotepec** (priprava za ubijanje muh z daljšim ročem in sploščenim, lopatičastim koncem), **picoklic** (ko pokličemo in naročimo dostavo pice), **tekmofehtar** (prosilec za tekmovanje) in **modrozob** (ki je po analogiji nastali prevod za *bluetooth*) ipd. V navedenih primerih medpona -o- združuje dva morfema, pomen pa je mogoče razbrati iz podstav. Medpono -i- najdemo v primeru **pešibus**, ki je nekoliko šaljiva sopomenka »iti peš« in namiguje na humorno razsežnost lingvistične inovativnosti.

3.4 Sklopi

Pri sklopih, ki nastajajo s sklapljanjem, »sestavine besednozvezne podstave združimo, npr. *dva in trideset* > *dvaintrideset*; *se ve da* > *seveda*« (Slovenski pravopis, 2001: 109) oziroma »enote večdelne podstave enostavno sklopimo v novo besedo: *sevé*, *bógvé*, *bojažêljen*, *tèmnordéč*, *zatém* < *se vé*, *bóg vé*, *bôja žêljen*, *tèmno rdèč*,

za tém. Pri tem se lahko izgubi kak naglas podstave ali pa se le-ta tudi kako drugače spremeni« (Toporišič, 2000: 160), pri čemer Logar opozarja, da podstav ni mogoče opredeliti po številu ali medsebojnem razmerju podstavnih besed (Logar, 2005: 190). Neologizmi v *Razvezanem jeziku* so le redko sklopi v strogem pomenu besede, lahko pa se sklapljanje kombinira z izpeljavo. Ker gre za kombinacijo dveh različnih postopkov, bomo takšne primere obravnavali pri kompleksnih neologizmih.

3.5 Kratice in krnjenje

Poleg štirih osnovnih postopkov Slovenski pravopis navaja tudi kratičenje. »Pri t. i. kraticah besede podstave pred sklapljanjem krnimo [...] nato pa jih združimo v navadno besedo«, kot denimo NOB (Slovenski pravopis, 2001: 121–122). Toporišič (2000: 158–159) sicer obravnava krnjenje kot posebno obliko izpeljave, kratičenje pa kot podvrsto složenk, toda v naši raziskavi bomo oboje obravnavali kot ločen besedotvorni postopek.

V *Razvezanem jeziku* najdemo kratice, ki pa so – poleg prvotnega pomena – ponovno motivirane:[3] **BTŠ** tako pomeni »Biotehniška šola Naklo« in »butci tut študirajo«. Podobno slovar navaja za NMS – poleg običajnega pomena »učenec ne dosega minimalnih standardov« – tudi »nije mama sretna«, »naj se mama sekira«, »najboljši med sošolci«, »ne mi srat«. Zdi se torej, da je proces dvosmeren, od leksemov h kratici in od kratice k leksemom.

Zapis črkovanih kratic je po Slovenskem pravopisu manj običajen (Slovenski pravopis, 2001: 121–122), vseeno pa najdemo v *Razvezanem jeziku* primere, kot so **gepes** (GPS), **ohape** (OHP – O, hudič, ponedeljek!), **pekape** (PKP – piši kući propalo).

Omeniti velja tudi druge oblike krnjenja, pri katerih izpustimo bodisi začetni bodisi končni del leksema (*troncation par aphérèse*, *troncation par apocope*, Apothéloz 2002: 117–123). Najdemo, denimo, geslo **ni blema**, kjer je leksem »problem« skrajšan v »blem«. Vendar se zdi, da je krnjenje samo nastopa sorazmerno redko, pogosto pa se kombinira z drugimi besedotvornimi postopki, kar bomo obravnavali pri kompleksnih neologizmih.

3.6 Obračanje vrstnega reda zlogov

Nove besede lahko nastajajo tudi z obračanjem vrstnega reda zlogov v besedi. V nekaterih jezikih, na primer v francoščini, je to izredno produktiven postopek, značilen za argo (Calvet 2007: 81–86); v slovenščini ga srečamo le redko. Omenimo primere, kot so **ganci** (cigan), **tikepa** (patike) ali pa **bazl**, ki izhaja iz besede »slaba«. Poleg obrnjenega vrstnega reda zlogov je tu prišlo tudi do spremembe nezvenečega glasu [s] v zvenečega [z].

4 Kompleksni neologizmi in besedne igre

V do sedaj omenjenih primerih je načeloma prisoten po en besedotvorni postopek. Besedotvorni postopki pa se lahko tudi medsebojno kombinirajo; v teh primerih govorimo o kompleksnih neologizmih. Tako najdemo v *Razvezanem jeziku* različne kombinacije, med katerimi se najpogosteje pojavlja kombinacija zlaganja in izpeljave:[4] **čezlužje** (ZDA), **mimovrstnik** (dijak, ki se je v 1. letnik gimnazije vpisal mimo uradnega izbora; ki so ga zaposleni na šoli stlačili mimo pravil), **našvašizem** (družbena ureditev, ki temelji na vladavini »naših« v nasprotju do »vaših«), **vhištvo** (vse, kar je v hiši), **ognjedražec** (grebljica za popravljanje živega ognja). Prav tako najdemo kompleksne neologizme, v katerih sta podstavi pridana tako predpona kot pripona (kombinacija izpeljave in sestave): **zastatusirati se** (zapisati v status na Facebooku nekaj, zaradi česar ti je potem žal), **uizica** (zabava, ki od nas zahteva manj priprav in naporov), kjer je tuja podstava (*easy*) pravopisno prilagojena. Lahko pa podstavo pred izpeljavo doleti še krnjenje, kot v primerih **hlado** (hladilnik), **zmrzo** (zmrzovalnik), **španljelizirati** (ob pravem trenutku pogledati tako milo, kot znajo samo koker španjeli, in na ta način izprositi kaj navidez nemogočega), **štrbunkič** (stranišče na štrbunk), **ležečko** (ležeči policaj), **digič** (digitalni fotoaparat), ali pa kratičenje, kot v primerih **nukati se** (izhaja iz kratice za Narodno in univerzitetno knjižnico, kamor študentje in študentke hodijo študirat, pogosto pa tudi pecat, flirtat in se pajsat v prepolnih čitalnicah) in **tozdirati** (netozdno družbo spreminjati v tozdno, TOZD – temeljna organizacija združenega dela).

4.1 Beseda-kovček

Posebej velja omeniti t. i. besedo-kovček. To je kompleksen besedotvoren postopek, kjer iz dveh leksemov najprej s krnjenjem nato pa s sklapljanjem dobimo enega, ki vsebuje dele enega in drugega. Vsebuje nekakšen skupni imenovalec prejšnjih leksemov. Med najbolj znane takšne primere sodijo, denimo, *brunch* (breakfast + lunch), *motel* (motor car + hotel) in *autobus* (automobil + omnibus), ki pa so leksikalizirani do te mere, da jih ne zaznavamo več kot besedo-kovček (Apothéloz, 2002: 20–21).

Izraz beseda-kovček izhaja iz francoskega *mot-valise*, sicer pa se uporablja tudi *mot porte-manteau*, ki ga je prvič uporabil pisatelj Lewis Carroll: »Well, 'SLITHY' means 'lithe and slimy'. 'Lithe' is the same as 'active'. You see it's like a portmanteau – there are two meanings packed up into one word«[5] (Carroll, 1978: 267), med njegove najbolj znane neologizme pa sodi snark (snake + shark).

Zanimivo je, da je ta morfo-leksikalni postopek v *Razvezanem jeziku* daleč najbolj prisoten. Najdemo lekseme, kot so **kočerja** (obed okrog 16.00 ali 17.00), **predboživčnost** (predbožična živčnost, mrzlica, fanatičnost), **strahotepec** (strahopetec, ki je poleg tega še neumen), **stakati**, **televinc**. V njih lahko poleg pomena razberemo prvotne lekseme (kosilo + večerja, predbožičen + živčnost, stati + čakati, strahopetec + tepec, televizija + dalinc oziroma daljinec).

Nasprotno je v primerih, kot so **najebnina** (oderuška najemnina) ali **izpirjen,** možna dvojna interpretacija (najemnina + najebati/zajebati, iztirjen/izprijen + pir oziroma pivo), kar pa ne vpliva na razumevanje leksema.

[3] Takšen proces opaža Calvet tudi v francoskem argoju (2007: 101).

[4] Takšne primere Toporišič obravnava kot zloženke (Toporišič, 2000: 158).

[5] »Hja, *tacne* je sestavljenka in pomeni 'brez tac'. Kot na primer 'garderoba' – v eno besedo sta stlačena dva pomena« (Carroll, 1990: 238, prevedla Gitica Jakopin).

Omenimo še neologizem **mariboring** (Maribor + boring; si v Mariboru in ti je dolgčas; dolgočasje na štajerski način), ki obenem združuje lastno ime in tujejezično prvino, poseben pa je tudi zaradi dejstva, da imata leksema neposredni skupni imenovalec (bor).

4.2 Besedne igre in aluzije

Večino neologizmov v *Razvezanem jeziku* lahko uvrstimo v kategoriji formalnih ali semantičnih neologizmov. Obstajajo pa tudi primeri, ki se nahajajo nekako med obema kategorijama oziroma ju združujejo. Tu največkrat najdemo neologizme, ki so osnovani na besednih igrah, aluzijah ipd. Povečini gre za stalne besedne zveze ali kolokacije, kjer manjša formalna sprememba vodi v večji pomenski premik. **Grimsove pravljice** (posebna oblika javnega nastopanja, v katerem politik v neskončnost ponavlja očitno prazne in nasprotujoče si obljube) tako aludirajo na Grimmove pravljice in ministra Branka Grimsa, njun spoj pa humorno namiguje na nerealnost predlaganih reform. **Poletje na školjki** (izlet z neprijetnim zapletom) se nanaša na film *Poletje v školjki*, kjer komičnost izvira iz polisemične narave leksema školjka (morska školjka, straniščna školjka). **Ambasada glavaboli** je nastala iz imena diskoteke (Ambasada Gavioli), ob čemer ni povsem jasno, kaj naj bi v diskoteki povzročalo glavobol (Techno glasba? Uživanje alkohola?).

Grudožer je prevod za besedo buldožer, ki posnema zvočne prvine izvirnika, obenem pa je neologizem posrečeno »motiviran«. Zloženka vsebuje leskema »gruda« in »žreti«, ki pa ga je potrebno razumeti metaforično. Tudi v nekaterih drugih »prevodih« je zaznati za neologizme značilno dvojnost, opredeljeno kot razmerje med analogijo in anomalijo (Stramljič Breznik, 2003: 105): **moja kolpa** (mea culpa), **klubovanje** (clubbing), **graber** (bager), pri čemer je slednji nepopoln anagram.

5 Zaključek

V prispevku je predstavljen *Razvezani jezik* kot jezikovni vir. V ospredju je problem neologizmov, nastajanje novih besed oziroma novih pomenov že obstoječih besed. Natančneje raziskujemo različne vrste formalnih neologizmov. V prispevku obravnavamo tako neologizme, ki (zaenkrat) obstajajo le v *Razvenaznem jeziku*, kot tiste, ki so razširjeni tudi v splošnem jeziku.

Formalni neologizmi nastajajo na podlagi štirih osnovnih besedotvornih postopkov: izpeljave, sestavljanja, zlaganja in sklapljanja. Poleg izpeljank, sestavljank, zloženk in sklopov v *Razvezanem jeziku* najdemo tudi neologizme, ki nastajajo s kratičenjem, krnjenjem in obračanjem zlogov.

Ti različni postopki se pogosto tudi medsebojno kombinirajo v formalno kompleksnejših neologizmih, kjer je potrebno še posebej izpostaviti t. i. besedo-kovček, kombinacijo krnjenja in zlaganja. Zdi se, da je ta besedotvorni postopek v slovenščini izredno produktiven.

Po drugi strani bi pa za nekatere primere iz *Razvezanega jezika* lahko rekli, da se nahajajo med obema tipoma neologizmov oziroma ju združujejo. Gre za neke vrste besedne igre, ki humorno izrabljajo možnosti jezika za tvorjenje novih besed in besednih zvez.

6 Literatura

Denis Apothéloz. 2002. La construction du lexique français. Editions Ophrys, Pariz.

Emile Benveniste. 1974. Problème de linguistique générale II. Gallimard, Pariz.

Michael Carr. 1997. Dictionary Use and Dictionary Needs of ESP Students: An Experimental Approach. International Journal of Lexicography, 15(3): 206–228.

Jean-Louis Calvet. 2007. L'agot, Que sais-je? PUF, Pariz.

Lewis Carroll. 1978. Alice's Adventres in Wonderland and Through the Looking Glass. Penguin, Harmondsworth.

Lewis Carroll. 1990. Aličine dogodivščine v Čudežni deželi in V ogledalu. Mladinska knjiga, Ljubljana.

Colleen Cotter in John Damaso. 2007. Online Dictionaries as Emergent Archives of Contemporary Usage and Collaborative Codification. Queen Mary's OPAL #9 (Occasional Papers Advancing Language). University of London, London.

Kaja Dolar. 2014. Kolaborativni slovar *Razvezani jezik*. Slavistična revija, 64 (2): 235–252.

Darja Fišer in Jaka Čibej. 2015. Potencial množičenja v sodobni leksikografiji. V V. Gorjanc, P. Gantar, I. Kosem in S. Krek, ur., Slovar sodobne slovenščine: problemi in rešitve, str. 542–564. Filozofska fakulteta, Univerza v Ljubljani, Ljubljana.

Sylviane Granger. 2012. Introduction: Electronic Lexicography – from Challange to Opportunity. V S. Granger in M. Paquot, ur., Electronic Lexicography, str. 1–14. Oxford University Press, Oxford.

Nataša Logar. 2005. Besedotvorni sklopi. Slavistična revija, 53 (2): 171–191.

Christian M. Meyer in Irina Gurevych. 2012. Wiktionary: A New Rival for Expert-built Lexicons? Exploring the Possibilities of Collaborative Lexicography. V S. Granger in M. Paquot, ur., Electronic Lexicography, str. 259–292. Oxford University Press, Oxford.

Alain Polguère. 2003. Lexicologie et sémantique lexicale. Notions fondamentales. Les Presses de l'Université de Montréal, Montréal.

Jean Pruvost in Jean-François Sablayrolles. 2003. Les néologismes, Que sais-je? PUF, Pariz.

Franck Sajous, Emmanuel Navarro in Bruno Gaume. 2011. Enrichissement de lexiques sémantiques approvisionnés par les foules: le système WISIGOTH appliqué à Wiktionary. TAL, 52 (1): 11–35.

Slovar novejšega besedja slovenskega jezika. 2012. Založba ZRC, ZRC SAZU, Ljubljana.

Slovenski pravopis. 2001. Založba ZRC, ZRC SAZU, Ljubljana.

Irena Stramljič Breznik. 2003. Besedotvorna tipologija novonastalega besedja s področja mobilne telefonije. Slavistična revija, 51 (posebna številka): 105–118.

James Surowiecki. 2005. The Wisdom of Crowds. Anchor Books, New York.

Jože Toporišič. 2000. Slovenska slovnica. Obzorja, Maribor.

Veliki slovar tujk. 2002. Cankarjeva založba, Ljubljana.

Ada Vidovič-Muha. 2011. Slovensko skladenjsko besedotvorje. Znanstvena založba Filozofske fakultete, Univerza v Ljubljani, Ljubljana.

Razvoj korpusa slovenskih spletnih uporabniških vsebin Janes

Tomaž Erjavec,† Darja Fišer,‡ Nikola Ljubešić*‡

† Odsek za tehnologije znanja, Institut »Jožef Stefan«, Jamova cesta 39, Ljubljana
tomaz.erjavec@ijs.si
‡ Oddelek za prevajalstvo, Univerza v Ljubljani, Aškerčeva 2, Ljubljana
darja.fiser@ff.uni-lj.si
∗ Odsek za informacijske znanosti, Fakulteta za humanistiko in družboslovje, Univerza v Zagrebu
nikola.ljubesic@ijs.si

Povzetek

V prispevku predstavimo trenutno različico korpusa spletne slovenščine Janes, ki vsebuje tvite, spletne forume, uporabniške komentarje na novice in blogovske zapise, postopek njihovega zajema ter jezikoslovnega označevanja. Podrobneje predstavimo trenutno različico korpusa tvitov, ki smo ga obogatili s številnimi metapodatki, kot so tip in spol avtorja ter sentiment posameznega besedila. Opišemo tudi postopek določanja stopnje tehnične in jezikovne standardnosti besedilom. Prispevek zaključimo z načrti za nadaljnje delo na korpusu.

The development of the Janes corpus of Slovene user-generated content

The paper presents the current version of the Slovene corpus of netspeak Janes which contains tweets, forum posts, news comments and blogs. We describe the harvesting procedure of the corpus and its linguistic annotation. We then focus on the latest version of the Tweet corpus that contains rich metadata, such as the type and sex of the authors and the sentiment of the tweets. We also describe the method of assigning a technical and linguistic standardness score to texts in the corpus.

1 Uvod

Kljub zgledni podprtosti slovenščine z referenčnimi in specializiranimi korpusi nobeden od njih ne vsebuje besedil, ki jih na spletu ustvarjajo uporabniki družbenih omrežij. Ker njihov pomen z razširjenostjo družbenih omrežij strmo narašča in ker številne tuje (Crystal, 2011; Baron, 2008; Beißwenger, 2013) pa tudi prve domače jezikoslovne raziskave kažejo, da se jezik v njih razlikuje od pisnega standarda (Michelizza, 2008; Dobrovoljc in Jakop, 2012; Erjavec in Fišer, 2013), smo se za celovito in podrobno proučevanje novomedijskega jezika odločili zgraditi korpus tvitov, forumskih sporočil, komentarjev na spletne novice in blogov. Poleg širokega nabora jezikoslovnih raziskav bo korpus namenjen tudi razvoju robustnejših jezikovnotehnoloških orodij, ki bi se s tem segmentom jezika uspešneje spopadala, kot to uspeva obstoječim, ki so bila naučena na standardni slovenščini (Ljubešić et al., 2014a).

V prispevku predstavimo korpus spletnih uporabniških vsebin, ki je še v delu in zato še ni uravnotežen in reprezentativen ter vsebuje še precej šuma, vendar je kljub temu kot edini tovrstni vir že dragocen in uporaben za jezikoslovne in jezikovnotehnološke raziskave spletne slovenščine.

V naslednjem razdelku opišemo zvrstnost korpusa, načela vključevanja virov, postopek zbiranja besedil in jezikoslovno označevanje korpusa Janes v0.3 ter podamo njegovo kvantitativno analizo. V tretjem razdelku predstavimo korpus Janes Tviti v0.3.4, ki smo ga osvežili z novo zbranimi tviti ter obogatili z bogatim naborom avtomatsko in ročno pripisanih metapodatkov. V četrtem razdelku predstavimo še postopek za avtomatsko pripisovanje stopnje tehnične in jezikovne nestandardnosti besedil, nato pa prispevek zaključimo s sklepnimi ugotovitvami in načrti za nadaljnji razvoj korpusa.

2 Korpus Janes

V razdelku opišemo vire in metode, ki smo jih uporabili za zajem posameznih zvrsti besedil, ki so zajeta v korpusu, jezikoslovno označevanje teh besedil ter podamo kvantitativno analizo korpusa Janes v0.3.

2.1 Izbor in zajem besedil

V trenutno različico korpusa Janes so vključene štiri zvrsti javno objavljenih uporabniških spletnih vsebin, in sicer tviti, forumska sporočila, komentarji na spletne novice in blogovski zapisi. Med slovenskimi uporabniki popularna družbena omrežja, kot so Facebook, Snapchat in WhatsApp, vsebujejo večinoma zasebno komunikacijo med uporabniki, zato jih v korpus nismo vključili.

Tviti so bili zajeti z namenskim orodjem TweetCat (Ljubešić et al., 2014b), ki je bilo izdelano prav za gradnjo korpusov tvitov manjših jezikov. Z orodjem smo s pomočjo začetnega seznama specifično slovenskih besed identificirali uporabnike, ki tvitajo pretežno v slovenščini, ter njihove prijatelje in sledilce. Orodje stalno širi nabor slovenskih uporabnikov tviterja, njihov zajem pa poteka že dve leti in še traja. Korpus tvitov poleg besedila posameznega tvita vsebuje tudi metapodatke, ki smo jih pridobili skupaj s tvitom, in sicer uporabniško ime avtorja, datum in čas pošiljanja ter število posredovanj (ang. *retweets*) in všečkov (ang. *favourites*) zajetega tvita. Podkorpus tvitov v Janes v0.3 smo nadgradili s korpusom Janes Tviti v0.3.4, ki vsebuje več tvitov, predvsem pa več metapodatkov; ta je podrobneje predstavljen v razdelku 2.3

Zaradi časovnih in finančnih omejitev popolne vključitve forumov in novičarskih portalov v korpus nismo mogli zagotoviti, zato smo za zajem forumskih sporočil in komentarjev z novičarskih portalov izbrali po tri vire, ki so v slovenskem spletnem prostoru najbolj priljubljeni, ponujajo največ jezikovne produkcije in/ali predstavljajo pomemben

del slovenskega spletnega prostora. Osrednji izbrani forum ima največje število registriranih uporabnikov in posledično pokriva tudi najširši nabor tem, poleg njega pa smo izbrali še dva sicer tudi množično uporabljana, a ožje specializirana, ki obravnavata zelo različni tematiki, imata precej različno ciljno publiko in izkazujeta tudi različne jezikovne specifike. Po podobnih načelih smo izbrali tudi novičarske portale, s katerih smo zajeli komentarje bralcev, in sicer osrednji nacionalni javni medij ter dva ožje usmerjena politična tednika, enega levičarskega, drugega pa desničarskega. Za vključitev vira v korpus je bila ključna tudi politika novičarskih portalov, saj številni portali dostop do novic zaračunavajo ali po določenem času komentarje avtomatsko izbrišejo, s čimer je zajem komentarjev tehnično onemogočen. Čeprav se zavedamo, da s tem nismo zajeli vseh tem, s katerimi se spletne uporabniške vsebine ukvarjajo, in besedišča, ki je na njih uporabljeno, smo prepričani, da smo zajeli zadovoljiv vzorec jezikovne rabe, ki je za ta način komunikacije med govorci slovenščine značilna.

V korpus smo vključili osrednji slovenski forum med.over.net ter specializirana foruma s področja avtomobilizma in znanosti avtomobilizem.com in kvarkadabra.net, s čimer smo želeli zajeti najaktivnejše forume, pokriti kar najširši nabor tem in zaobjeti čim bolj raznolike segmente jezikovne rabe. Komentarje na novice smo zajeli s spletnega portala nacionalne televizije RTV Slovenija, prav tako pa tudi levo orientiranega tednika mladina.si in njegovega ekvivalenta z desnega pola reporter.si. Ker se spletna mesta po sestavi med seboj razlikujejo, smo za vsak vir posebej napisali ekstraktor besedila, kar je bilo tudi ozko grlo pri nadaljnjem širjenju virov besedil. Iz zajetega materiala smo na ta način izluščili le tiste podatke, ki smo jih hoteli vključiti v korpus, in se tako izognili velikemu deležu šumnih prvin, kot so oglasna sporočila, nerelevantne povezave ipd. Pri zajetih komentarjih na novice smo izluščili tudi metapodatke, kot so naslov prispevka, naslov URL, identifikacijska številka pripadajočega članka, datum objave komentarja, uporabniško ime avtorja ter identifikacijska (zaporedna) številka komentarja. Vsi komentarji so z identifikacijskimi številkami razvrščeni glede na članke, ki jim pripadajo, zato jih je v korpusu mogoče opazovati v zaporedju. Pri forumskih sporočilih smo ohranili metapodatke o pripadajoči temi, naslovu URL posameznega vpisa, datumu objave, uporabniškem imenu avtorja in identifikacijski številki vpisa. Forumi so pogosto specifični in se osredotočajo na določeno temo (npr. zdravje, avtomobilizem, šport, vrtnarstvo), sestavljeni pa so iz več podforumov, ki obravnavajo različne vsebinske kategorije (npr. na forumu med.over.net najdemo podforume o vzgoji otrok, plastični kirurgiji ipd.). V korpusu zato lahko z iskanjem po identifikacijskih številkah tem ali podforumov opazujemo tudi značilnosti izbranih vsebinskih podsegmentov forumov.

Za gradnjo podkorpusa blogovskih zapisov smo zaenkrat uporabili kar deduplicirano različico splošnega korpusa slovenskega spleta slWaC 2.0 (Erjavec in Ljubešić, 2014), iz katerega smo zajeli vsa besedila, pri katerih se v domeni naslova URL pojavi niz “blog”, pri čemer se izkaže, da velika večina zajetih besedil prihaja s portala blog.siol.net. Rešitev je začasna, saj za razliko od ostalih podkorpusov za bloge zaenkrat še nismo izdelali ciljnega ekstraktorja, tako da nimamo ohranjene notranje strukture blogovskih zapisov, npr. razdelitve besedila na sam zapis in na komentarje pod njim, ravno tako pa ne zajamemo naslova posameznega bloga oz. njegovega avtorja in avtorjevega profila.

Vsi našteti podkorpusi so bili nato združeni v korpus Janes v0.3, ki poenoti in s tem tudi poenostavi metapodatke posameznih besedil. Podkorpusi in korpus Janes so zapisani v formatu XML, ki omogoča strukturiranje korpusa, zapis metapodatkov in konsistenten zapis znakov po standardu Unicode.

2.2 Jezikoslovno označevanje

Zajete vire smo avtomatsko jezikoslovno označili. Prvi korak označevanja sta bili tokenizacija in stavčna segmentacija, za kar smo uporabili rahlo prilagojeno standardno knjižnico mlToken za slovenski jezik, ki je del programa ToTaLe (Erjavec et al., 2005). Prilagoditve so zajemale pravilno obravnavo najpogostejših vrst posebnih pojavnic v besedilih, kot so emotikoni, klub temu pa se že na tem koraku pojavijo težave, npr. izpusti presledkov med ločilom in naslednjo besedo, ki v primeru “Virantova briljantna ideja.Zelo liberalno.” povzroči, da je besedilo razdeljeno na eno namesto na dve povedi in na pet namesto na sedem pojavnic.

V naslednjem koraku smo besedne pojavnice normalizirali z metodo, ki temelji na statističnem strojnem prevajanju črk, naučena pa je bila na 1.000 ključnih besedah iz korpusa tvitov glede na korpus KRES in na njihovih ročno normaliziranih oblikah (Ljubešić et al., 2014a). Čeprav s tem nismo zajeli vseh fenomenov nestandardnega zapisa besed, 1.000 najbolj ključnih pojavnic predstavlja tisto besedišče, ki se od standardne slovenščine najbolj razlikuje, in sicer 5,3 milijona oz. 3,3 % vseh pojavnic v korpusu. Ob predpostavki, da je treba normalizirati le manjšino pojavnic v korpusu, to niti ni tako zelo malo, bomo pa metodo v prihodnje še nadgradili s pomočjo ročno označenega učnega korpusa in strojnim učenjem normalizacije. Z orodji za standardno slovenščino programa ToTaLe smo nato normalizirane besede še oblikoskladenjsko označili in lematizirali.

V Ljubešić et al. (2014a) smo izvedli tudi evalvacijo točnosti lematizacije tvitov, pri čemer lematizacija potrebuje predhodno oblikoskladenjsko označevanje, zato smo implicitno evalvirali obe ravni označevanja. Na izvornih besedah v tvitih je bila točnost lematizacije 75 %, točnost na ročno normaliziranih tvitih 92 %, na avtomatsko normaliziranih pa 84 %; z drugimi besedami, avtomatska normalizacija zmanjša napako lematizacije za polovico.

Posamezne podkorpuse in celoten korpus smo uvozili tudi v spletni konkordančnik NoSketch Engine (Erjavec, 2013). Dostop do njih je trenutno omejen na sodelavce projekta, ob zaključku projekta pa načrtujemo tudi splošno (tako prosto kot odprto dostopno) različico korpusov, ki pa bo morala upoštevati avtorske pravice, pravico do zasebnosti in pogoje uporabe vključenih spletnih portalov.

(Pod) korpus	Št. besed	Št. besedil	Št. besed/ besedilo	Št. avtorjev	Št. besed/ avtorja	Št. besedil/ avtorja
Janes v0.3	134.543.613	4.819.558	27,9	85.428	1.574,9	56,4
tviti	50.148.724	3.684.909	13,6	7.590	6.607,2	485,5
forumi	39.576.432	772.953	51,2	63.543	622,8	12,2
avtomobilizem	21.776.486	569.594	38,2	12.793	1.702,2	44,5
medovernet	11.585.631	122.613	94,5	49.484	234,1	2,5
kvarkadabra	6.214.315	80.746	77,0	2.212	2.809,4	36,5
komentarji	12.542.551	299.420	41,9	14.295	877,4	20,9
rtvslo	10.350.937	267.932	38,6	12.921	801,1	20,7
mladina	1.898.780	26.084	72,8	1.276	1.488,1	20,4
reporter	292.834	5.404	54,2	237	1.235,6	22,8
blogi	32.275.906	62.276	518,3	-	-	-

Tabela 1: Velikost korpusa Janes v0.3.

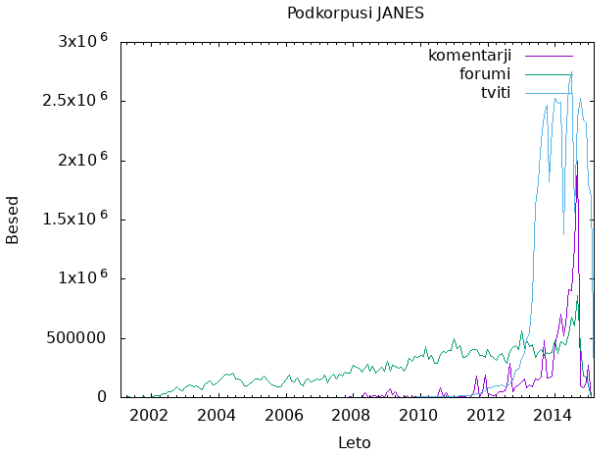
2.3 Sestava

Kot prikazuje Tabela 1, vsebuje korpus Janes v0.3 nekaj manj kot 135 milijonov besed (kar je okoli 161 milijonov pojavnic) in skoraj 5 milijonov besedil.

Največji je podkorpus tvitov s prek 50 milijoni besed in 3,6 milijona tvitov, sledita mu podkorpusa forumskih sporočil in blogov, najmanj pa je komentarjev na novice. Tabela poda tudi razdelitev po virih znotraj forumov in komentarjev, kjer lahko vidimo, da je med forumi največji Avtomobilizem, Medovernet je skoraj polovico manjši, Kvarkadabra pa še za polovico manjši. Pri komentarjih so razlike še večje, saj komentarji, zajeti s portala RTV Slovenija, vsebujejo prek 10 milijonov besed, kar je več kot petkrat več od števila zajetih komentarjev s portala Mladina, medtem ko nam je s portala Reporter uspelo zajeti zgolj 300 tisoč besed.

Besedila v korpusu so tipično zelo kratka, saj v povprečju vsebujejo samo 28 besed, kar seveda sledi iz narave zajetih besedil. Po pričakovanju so najdaljši blogovski zapisi z nekaj nad 500 besedami na besedilo, najkrajši pa, pričakovano, tviti, dolžina katerih je zaradi odločitve ponudnika omejena na največ 140 znakov. Zanimivo je, da med forumi s 50 besedami in komentarji z 42 ni bistvene razlike, saj bi pričakovali, da bodo forumska sporočila bistveno daljša. Podrobnejši pogled sicer razkrije, da so med posameznimi viri precejšnje razlike, tako so npr. v Medovernet besedila dolga skoraj 95 besed, kar je skoraj trikrat več kot pa pri Avtomobilizmu, vendar ta zaradi svoje velikosti bolj vpliva na povprečje celotnega podkorpusa forumov. Tudi znotraj komentarjev opazimo precejšnje razlike, saj so npr. komentarji na portalu RTV Slovenija skoraj dvakrat krajši kot pri Mladini.

Besedila v korpusu je napisalo več kot 85.000 avtorjev, kjer kot enega avtorja upoštevamo eno uporabniško ime znotraj enega podkorpusa. Število avtorjev je tako zgolj ocena, saj lahko ista oseba uporablja različna uporabniška imena. Poleg tega, kot že omenjeno, za podkorpus blogov trenutno nimamo podatkov o avtorjih. Posamezni avtor je v povprečju napisal nekaj čez 1.500 besed oz. 56 besedil, pri čemer se tudi tu številke zelo razlikujejo od vira do vira. Kar osemkrat več besedil, ki obenem vsebujejo štirikrat več besed od povprečja, objavljajo uporabniki omrežja



Slika 1: Število besed po podkorpusih Janes v0.3 glede na čas objave.

Twitter. Ne glede na spletni portal jih komentatorji sestavijo za slabo polovico glede na povprečje, pri čemer največ besed posamezni komentator prispeva na portalu Mladina, najmanj pa na portalu RTV Slovenija. Največ nihanja opazimo pri forumih, kjer posamezni uporabnik na forumu Avtomobilizem objavi kar 18-krat več besedil kot uporabnik foruma Medovernet, ki vanj prispeva tudi najmanj besed, in sicer več kot šestkrat manj od povprečja, po drugi strani pa posamezni avtor na forumu Kvarkadabra objavi skoraj dvakrat več besed od povprečja, s čimer po številu prispevanih besed na avtorja zaseda drugo mesto, tik za uporabniki omrežja Twitter.

Kot prikazuje Slika 1, so bila besedila, vključena v korpus, objavljena v obdobju 2001–2015, a jih je skoraj polovica (49 %) iz leta 2014. Najstarejši vir so forumi, ki so očitno dovolj stabilni, da je z njih možno pridobiti objave vse od 2001. Najstarejši komentarji na novice so iz leta 2008, vendar jih je velika večina iz 2014, kar je posledica tehničnih rešitev novičarskih portalov. Najmlajši vir besedil je družbeno omrežje Twitter, kjer se zajem starih tvitov začne z letom 2011, velika večina jih je iz let 2013 in 2014, ko je zajem besedil potekal. Nihanja v letu 2014 niso posledica začasne neuporabe Twitterja, temveč kažejo

na obdobja, ko zaradi težav s strežnikom zbiranje tvitov ni delovalo.

Ti podatki kažejo, da je zgrajeni korpus zelo heterogen tako glede na avtorstvo kot tudi glede na dolžino, količino in starost prispevanih besedil.

3 Korpus Janes Tviti

Korpus tvitov smo od izdelave korpusa Janes v0.3 že povečali in opremili z dodatnimi metapodatki. V Janes v0.3 se je namreč zajem tvitov zaključil z 2. marcem 2015, v trenutnem korpusu Janes Tviti v0.3.4 pa s 23. junijem 2015, s čimer smo pridobili še dodatnih 500.000 tvitov oz. 6 milijonov besed. Poleg obstoječih metapodatkov o uporabniškem imenu avtorja, datumu in času pošiljanja ter številu posredovanj in všečkov smo tvitom dodali še ročno preverjene podatke o lastnostih avtorja ter avtomatsko pripisane podatke o sentimentu tvita, kar obravnavamo v nadaljevanju razdelka.

3.1 Označevanje tipa in spola uporabnikov

Za poglobljene raziskave jezika tvitov so potrebni sociodemografski podatki, s katerimi tviti eksplicitno niso opremljeni in jih je treba pridobiti na drugačne načine. Glede na to, da je v slovenščini spol v prvoosebnih glagolskih oblikah eksplicitno izražen, smo ga na podlagi prevladujoče oblike uporabnikom najprej pripisali avtomatsko. Ker je avtorjev v korpusu še obvladljivo število (malo manj kot 7.600), smo nato prosili študentki medjezikovnega posredovanja, da pripisani spol pregledata in ga po potrebi popravita. Obenem smo ju prosili, da na podlagi profila uporabnika in pregleda objavljenih tvitov vsakemu uporabniku pripišeta še vrsto računa. Poleg moškega in ženskega spola smo z "nevtralnim" spolom označili tiste tvite, pri katerih niti iz uporabniškega imena niti iz besedil, ki so tipično poročevalska, ni mogoče ugotoviti spola pisca. Tip avtorja je bodisi "ustanova" bodisi "osebno", pri čemer so v prvo kategorijo uvrščeni računi medijskih hiš, javnih ustanov in podjetij, v drugo kategorijo pa računi posameznikov.

3.2 Označevanje sentimenta

Označevanje sentimenta (pozitiven, negativen ali nevtralen) na področju uporabniško ustvarjenih vsebin, še posebej tvitov, postaja vse popularnejše (Pak in Paroubek, 2010). S pripisom sentimenta tvitom o posamezni temi lahko namreč ugotovimo, ali je javnost neki temi (kot npr. predsedniškemu kandidatu, izdelku, vrednostnim papirjem) naklonjena ali ne, spremljamo pa lahko tudi trende v sentimentu na določeno temo. Program, opisan v Smailović et al. (2014), ki temelji na uporabi metode podpornih vektorjev (SVM), je bil kasneje nadgrajen, predvsem pa naučen označevanja besedil v slovenščini na večji ročno označeni zbirki raznovrstnih slovenskih tvitov. Ta program oz. model je bil nato uporabljen za označevanje sentimenta v korpusu tvitov Janes, s čimer imamo možnost preučevanja tvitov tudi po tem kriteriju.

Za evalvacijo označevanja sentimenta v tvitih smo s sentimentom ročno označili 1.977 tvitov s področja politike in športa, pri čemer sta vsak tvit označila dva anotatorja. To podatkovno množico smo uporabili za evalvacijo točnosti avtomatskega označevanja, ki jo podamo v Tabeli 2. Prva vrstica pokaže ujemanje v pripisovanju sentimenta med anotatorjema, kjer vidimo, da sta se skoraj v četrtini primerov razhajala glede pripisane ocene, kar gre pripisati predvsem dejstvu, da je določanje sentimenta v veliki meri subjektivno. Še posebej problematični so cinični in sarkastični tviti o aktualnem političnem dogajanju, pri katerih je pravi sentiment mogoče določiti šele s pomočjo širšega konteksta, s katerim sta anotatorja seznanjena v različni meri. Avtomatsko označevanje se s prvim anotatorjem ujema v 57,3 % primerov in za spoznanje bolje z drugim anotatorjem. Ujemanje se dvigne na 62,1 %, če se omejimo samo na tiste problematične tvite, ki sta jim oba anotatorja pripisala enak sentiment, na 67,1 % pa, če štejemo kot pravilne tiste oznake, ki se ujemajo z vsaj enim anotatorjem.

Primerjava		Ujemanje
Anotator 1	Anotator 2	76,5 %
Anotator 1	Avtomatsko	57,3 %
Anotator 2	Avtomatsko	57,4 %
Anotator 1 in 2	Avtomatsko	62,1 %
Anotator 1 ali 2	Avtomatsko	67,1 %

Tabela 2: Ujemanje ročno in avtomatsko pripisanih oznak sentimenta.

Anotator	Neg.	Nevt.	Poz.	Povpr.
Anotator 1	32,6 %	37,2 %	30,2 %	-2,4 %
Anotator 2	27,7 %	38,1 %	34,2 %	13,1 %
Oba anotatorja	30,1 %	37,7 %	32,2 %	2,1 %
Avtomatsko	22,2 %	45,8 %	32,1 %	9,9 %

Tabela 3: Razporeditev ročno in avtomatsko pripisanih oznak sentimenta.

Zanimivo je še pogledati, kako so tri vrednosti sentimenta razporejene, kar podamo v Tabeli 3, ki pokaže, koliko odstotkov tvitov je bilo ocenjenih kot negativnih (-1), nevtralnih (0) in pozitivnih (+1); zadnji stolpec prikazuje odstopanje povprečne ocene od nevtralne (če bi torej anotator vse tvite ocenil kot negativne, bi bila ta ocena -100 %). Tabela pokaže, da je sentiment razmeroma enakomerno porazdeljen med tri vrednosti ne glede na anotatorja, čeprav je anotator 1 nekoliko bolj nagnjen k pripisovanju negativnih, anotator 2 pa pozitivnih ocen. Rezultat avtomatskega označevanja delno razkrije tudi razmeroma slabe rezultate evalvacije iz Tabele 2, saj je avtomatska metoda bolj konzervativna, tj. mnogo več tvitom pripiše nevtralen sentiment, kar je s stališča uporabnosti aplikacije verjetno ustrezna odločitev. Tretja vrstica poda razporeditev sentimenta, če povprečimo oceni obeh anotatorjev, s čimer dobimo večinski razred, ki je nevtralen sentiment. S tem lahko tudi sklenemo evalvacijo: zelo enostaven, večinski označevalnik bi vsakemu tvitu pripisal vrednost sentimenta nevtralno, s čimer bi dosegel točnost 37,7 %. Avtomatsko označevanje doseže točnost 57,3 % kot najslabši rezultat glede na prvega anotatorja oz. 62,1 %, kjer se anotatorja ujemata, kar zabeležimo v 76,5 % primerov. Avtomatsko označevanje je tako po kvaliteti na sredini med naključnim in ročnim pripisovanjem sentimenta.

Kot zanimivost v Sliki 2 podamo še razporeditev relativnih deležev treh vrednosti sentimenta po besedilih glede

na čas objave, kjer je zanimivo manjšanje pozitivnega in enakovredno večanje negativnega sentimenta. Kot pa smo videli v Sliki 1, je v korpusu zelo malo tvitov izpred 2013, zato je ta opažanja trebna jemati z dobršno mero previdnosti.

3.3 Sestava

Tabela 4 poda velikost korpusa Janes Tviti v0.3.4 v celoti in glede na posamezne označene metapodatke. Korpus 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 %), za katere je zanimivo, da so v korpus prispevali najmanjši delež besedil, in sicer nekaj nad 17 % tvitov in le za odstotek večji delež pojavnic. Ti podatki kažejo, da moški in ženske tvitajo približno enako pogosto in v podobni dolžini, medtem ko uporabniki, ki jim spola nismo mogli določiti, objavijo precej manj sporočil, ki pa so nekoliko daljša.

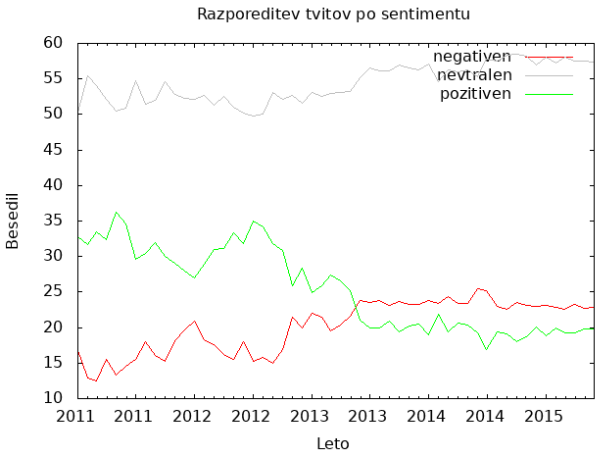
Dobre tri četrtine uporabnikov, zajetih v korpusu, tvita v osebnem imenu (76 %), medtem ko je slaba četrtina korporativnih računov oz. računov javnih ustanov (24 %). Zasebni uporabniki so v korpus prispevali 79 % tvitov oz. dobrih 80 % pojavnic, institucijski pa le 21 % tvitov oz. za odstotek manj pojavnic. Primerjava teh deležev pokaže, da zasebni uporabniki tvitajo več in objavljajo nekoliko daljša sporočila kot predstavniki ustanov. Zanimiva je tudi primerjava tipa uporabnika z njegovim spolom, saj bi pričakovali, da so tviti ustanov nevtralnega spola, kar sicer večinoma drži, ne pa vedno, saj je za 20 % institucionalnih uporabniških računov spol mogoče določiti, in sicer v 268 primerih moški, v 67 pa ženski.

		Besed	Tvitov	Avtorjev
Korpus	Tviti v0.3.4	58.311.996	4.337.767	7.570
Spol	ženski	15.417.064	1.151.300	1.858
	moški	32.718.987	2.399.365	4.011
	nevtralen	10.175.930	787.101	1.700
Vir	ustanova	11.629.005	908.454	1.782
	osebno	46.682.991	3.429.313	5.788
Senti.	negativen	15.964.247	1.006.123	-
	nevtralen	31.424.765	2.455.801	-
	pozitiven	10.922.984	875.843	-

Tabela 4: Velikost trenutnega korpusa Janes Tviti v0.3.4.

4 Označevanje nestandardnosti besedil

Ker so prve analize pokazale, da zgrajeni korpus vsebuje številna besedila podjetij (novice, reklame) in javnih ustanov (obvestila), ki tako po komunikacijskem namenu kot jezikovni podobi v ničemer ne odstopajo od klasičnih besedil na njihovih spletnih straneh, smo se odločili razviti postopek, ki bo vsakemu besedilu pripisal stopnji (ne)standardnosti, kar bo uporabniku korpusa omogočilo, da izbere samo besedila, ki ustrezajo tisti stopnji standardnosti, ki ga za konkretno raziskavo zanima. To je koristna informacija tudi za jezikovnotehnološka orodja, saj se na



Slika 2: Število tvitov po sentimentu glede na čas objave.

osnovi tega lahko odločijo, ali je normalizacija besednih oblik potrebna. Če to normalizacijo uporabimo nad standardnimi besedili, bo namreč naredila več škode kot koristi.

Razvili smo avtomatsko metodo (Ljubešić et al., 2015), ki besedilo opredeli glede na njegovo stopnjo standardnosti, pri čemer se izkaže, da je koristno ločiti med dvema vrstama (ne)standardnosti, in sicer tehnično in jezikovno. Tehnična (ne)standardnost se izrazi predvsem v (ne)uporabi velikih začetnic, ločil in presledkov, medtem ko jezikovna (ne)standardnost upošteva izbor in zapis besed, njihove oblikoslovne lastnosti ter besedni red. Za obe meri uporabljamo lestvico od 1 (povsem standardno) do 3 (zelo nestandardno). Za vtis, kakšne lastnosti besedil smo upoštevali, podamo dva primera:

- T=1 / L=3
  A gdo pozna koga ka bi zaceu trenirat pr 15 ih.

- T=3 / L=1
  komunistična ideologija ubijaj,kradi laži.....zelo primerna za aktualno vlado„„,

Postopek temelji na metodah nadzorovanega strojnega učenja, zato smo najprej ročno označili 1.200 tvitov, komentarjev in forumskih sporočil, nato pa definirali značilke, ki so pomembne za določanje stopnje standardnosti. Glede na izbrane značilke in s pomočjo učne množice se je program naučil pripisati vsakemu besedilu vrednost 1–3 za obe meri standardnosti. Rezultati evalvacije so pokazali, da je povprečna absolutna napaka najboljše od preizkušenih metod 0,38 za določanje tehnične in 0,42 za določanje jezikovne standardnosti, kar mdr. kaže na to, da je tehnična stopnja standardnosti lažje določljiva.

S tem programom smo nato določili obe stopnji standardnosti vsem besedilom v korpusu razen blogom, zato informacija o standardnosti tudi ni prisotna v celotnem korpusu Janes v0.3. Z informacijo o standardnosti so tako opremljeni korpusi Tviti v0.3.4, Forum v0.3 in Komentarji v0.3.

V Tabeli 5 podamo podatke o številu besedil, ki so v posameznih podkorpusih prejela različne oznake standardnosti, kjer pa se je treba zavedati, da pri pripisovanju obeh ocen prihaja do razmeroma velikih napak. Gledano v celoti

(Pod)korpus	Št. besedil	$\overline{T}$	T=1 %	T=2 %	T=3 %	$\overline{L}$	L=1 %	L=2 %	L=3 %
Skupaj	5.410.140	1,5	67,4	25,9	6,7	1,5	71,2	21,9	7,0
Tviti v0.3.4	4.337.767	1,5	70,1	24,8	5,1	1,4	75,0	19,2	5,7
Forumi v0.3	772.953	1,7	52,5	32,8	14,7	1,8	50,3	34,8	14,9
avtomobilizem	569.594	1,8	45,5	36,8	17,7	1,8	42,8	38,2	19,1
medovernet	122.613	1,5	66,8	24,7	8,6	1,6	66,6	29,5	3,9
kvarkadabra	80.746	1,4	80,5	16,4	3,1	1,4	78,5	19,0	2,4
Komentarji v0.3	299.420	1,5	66,2	25,1	8,7	1,5	68,7	26,8	4,4
rtvslo	267.932	1,5	65,6	25,2	9,1	1,5	68,0	27,3	4,7
mladina	26.084	1,4	72,4	22,8	4,9	1,5	74,5	23,3	2,1
reporter	5.404	1,5	66,1	26,8	7,1	1,4	76,6	20,7	2,7

Tabela 5: Standardnost jezika v posameznih (pod)korpusih Janes.

so besedila v korpusu precej bolj standardna, kot bi morda pričakovali, tako na tehničnem kot na lingvističnem nivoju (1,5). V povprečju so tehnično najbolj standardni tviti in komentarji na novice (1,5), najmanj pa forumi (1,7), po zaslugi velikega in najbolj nestandardnega Avtomobilizma (1,8). Najvišjo stopnjo jezikovne standardnosti v povprečju dosegajo tviti (1,4), najnižjo pa zopet forumi (1,8) z Avtomobilizmom (1,8) na čelu. Po drugi strani med vsemi viri, vključenimi v korpus, najvišjo stopnjo standardnosti dosega forum Kvarkadabra (1,4), kar je bilo glede na obravnavano tematiko tudi pričakovano.

5 Zaključek

V prispevku smo predstavili gradnjo, jezikoslovno označevanje in opremljanje z metapodatki trenutne različice korpusa spletne slovenščine Janes v0.3 in njegovih podkorpusov, posebej pa korpusa Tviti v0.3.4. Od klasičnih korpusnih projektov se predstavljeni razlikuje po tem, da smo pred oblikoskladenjskim označevanjem in lematizacijo nestandardni zapis besed standardizirali, besedilom v korpusih pa smo tudi dodali oznako za stopnjo standardnosti na tehnični in jezikovni ravni, s čimer smo omogočili bolj fokusirane jezikoslovne raziskave in razvoj orodij za procesiranje nestandardne slovenščine. Korpus tvitov smo še opremili s številnimi dragocenimi metapodatki, kot so oznaka tipa in spola uporabnika ter sentiment objavljenega tvita.

V nadaljevanju razvoja korpusa nameravamo evalvirati zanesljivost razvitih avtomatskih metod za dodajanje jezikoslovnih podatkov in vključenih besedilnih metapodatkov ter postopke nadgraditi in jih prilagoditi specifikam spletne slovenščine. Za izboljšanje standardizacije načrtujemo razvoj orodja za avtomatsko rediakritizacijo besed, ki so zapisane brez šumnikov, in ročno normalizacijo učnega korpusa za strojno učenje. V korpus tvitov na podlagi geolociranih tvitov načrtujemo dodati še podatke o prevladujoči regiji uporabnika, preostale metapodatke pa razširiti še na preostale podkorpuse korpusa Janes. Za diskurzivne in pragmatične raziskave pa bi bilo zelo koristno, če bi lahko v korpusu omogočili sledenje dialogom in pogovorom med uporabniki.

Korpus bo treba razširiti z zadnjimi zbranimi podatki in ga oblikovati v celoto, ob tem pa tudi določiti metode, ki bodo omogočile njegovo čim širšo uporabo.

Zahvala

Avtorji se zahvaljujejo Jasmini Smailović za označevanje sentimenta v korpusu Tvitov, Sašu Rutarju pa za izdelavo programske kode za to nalogo. Hvala tudi Jasmini, Marku Robniku Šikonji, Katji Zupan in anonimnima recenzentoma za koristne pripombe. Za vse preostale napake avtorji krivijo drug drugega. 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.

6 Literatura

Naomi S. Baron. 2008. *Always On: Language in an Online and Mobile World*. Oxford University Press.

Michael Beißwenger. 2013. Raumorientierung in der netzkommunikation. korpusgestützte untersuchungen zur lokalen deixis in chats. V: *Die Dynamik sozialer und sprachlicher Netzwerke*, str. 207–258. Springer.

David Crystal. 2011. *Internet Linguistics: A Student Guide*. Routledge, New York.

Helena Dobrovoljc in Nataša Jakop. 2012. *Sodobni pravopisni priročnik med normo in predpisom*. Založba ZRC.

Tomaž Erjavec, Camelia Ignat, Bruno Pouliquen in Ralf Steinberger. 2005. Massive Multilingual Corpus Compilation: Acquis Communautaire and ToTaLe. V: *2nd Language and Technology Conference*, str. 32–6, Poznan, Poland.

Tomaž Erjavec in Darja Fišer. 2013. Jezik slovenskih tvitov: korpusna raziskava. V: *Družbena funkcijskost jezika: (vidiki, merila, opredelitve)*, str. 109–116. Znanstvena založba Filozofske fakultete.

Tomaž Erjavec in Nikola Ljubešić. 2014. The slWaC 2.0 Corpus of the Slovene Web. V: *Language Technologies: Proceedings of the 17th International Multiconference Information Society IS2014*, Ljubljana, Slovenia.

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

Nikola Ljubešić, Tomaž Erjavec in Darja Fišer. 2014a. Standardizing Tweets with Character-Level Machine Translation. V: *CICLing: 15th International Conference on Intelligent Text Processing and Computational Linguistics*, Lecture notes in computer science, str. 164–75. Springer.

Nikola Ljubešić, Darja Fišer in Tomaž Erjavec. 2014b.

TweetCaT: a Tool for Building Twitter Corpora of Smaller Languages. V: *Ninth LREC*, Reykjavik. ELRA.

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: *RANLP - Recent Advances in Natural Language Processing*.

Mija Michelizza. 2008. Jezik SMS-jev in SMS-komunikacija. *Jezikoslovni zapiski*, 14:151–166.

Alexander Pak in Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. V: *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*.

Jasmina Smailović, Miha Grčar, Nada Lavrač in Martin Žnidaršič. 2014. Stream-based active learning for sentiment analysis in the financial domain. *Information sciences*, 285:181–203.

Velika in mala dilema pri imenih industrijskih izdelkov in znamk pri uradnih in zasebnih računih na družbenem omrežju Twitter

Teja Goli,* Damjan Popič,† Darja Fišer†

* Kropa
teja.goli@gmail.com

† Oddelek za prevajalstvo, Filozofska Fakulteta
Aškerčeva 2, 1000 Ljubljana
darja.fiser@ff.uni-lj.si
damjan.popic@ff.uni-lj.si

Povzetek

V prispevku obravnavamo rabo velike in male začetnice v imenih industrijskih izdelkov in znamk v slovenskem jeziku na družbenem omrežju Twitter, pri čemer se osredotočamo na to, ali so bili pisci besedil uporabniki uradnih ali zasebnih uporabniških računov. Z analizo podkorpusa tvitov v slovenščini želimo ugotoviti, kakšni so trendi rabe velike in male začetnice, kje se pojavljajo težave pri zapisovanju teh imen ter do kakšnih odstopanj od jezikovnega standarda prihaja pri obeh vrstah računov. Prav tako želimo ugotoviti, ali raba, ki jo lahko opazimo pri eni vrsti računov, vpliva tudi na rabo pri drugi.

Industrial product and brand names capitalisation in Slovene on corporate and private Twitter accounts

In this paper, we will analyse and discuss the use of capitalization of industrial products and brand names in Slovene on the social network Twitter. We will focus on the type of users that produced the texts, meaning that they will be from either corporate or private accounts. By analysing the Slovene Twitter subcorpus, we wish to understand how capitalization of industrial product and brand names is being used, where we can find any usage problems, and in which cases and how the Slovene language standard differentiates from the actual usage in both user types. Furthermore, we wish to find out if the usage in one account type influences the other.

1 Uvod

V informacijski dobi in potrošniški družbi se pojavljajo vedno novi trendi, tudi v jeziku. Pričakovali bi, da bodo tudi jezikovni priročniki in standard, ki ga predpisujejo, v koraku s časom. V tem prispevku bomo na podlagi analize korpusa spletne slovenščine preverili, ali je trenutni jezikovni standard aktualen, in sicer z vidika zapisovanja imen industrijskih izdelkov. Njihove proizvajalce ščitijo zakoni, ki med drugim pravijo, da je v slovarjih, enciklopedijah in podobnih delih reprodukcijo znamke, ki daje vtis generične rabe zaščitenih izdelkov, treba na zahtevo imetnikov znamk popraviti in jasno označiti, da gre za znamko (Dobrovoljc, 2009: 12).

S tem v slovenščino prihaja izjemno problematičen segment jezika, saj smo njegovi uporabniki izpostavljeni določeni rabi, ki jo zahtevajo proizvajalci, pravila pa za to rabo (morda) niso več aktualna oziroma niso dovolj jasna. Zaradi tega že na samem začetku prihaja do odstopanja od standarda.

Kaj to pomeni za naš standard, bomo poskušali izluščiti iz analize tvitov uporabnikov uradnih računov, kot so podjetja, radii, novice, revije itd., ki jih bomo primerjali s tviti uporabnikov zasebnih računov, ki predstavljajo povprečnega uporabnika slovenskega jezika.

Osredotočili se bomo na rabo velike in male začetnice v imenih industrijskih izdelkov, pri čemer bomo še posebej pozorni na to, kakšna so odstopanja med rabo pri uradnih računih, ki pogosto predstavljajo prav proizvajalce teh izdelkov, zaradi česar so pozorni na rabo, in pri zasebnih računih, za katere je značilno, da v neformalni komunikaciji uporabljali nestandardno slovenščino. Izvedeli bomo, če in kje prihaja do razlik pri rabi med prvim in drugim tipom uporabnikov in oboje tudi primerjali s trenutnim knjižnim standardom, ki ga predstavlja Slovenski pravopis 2001.[1]

2 Imena industrijskih izdelkov in znamk v slovenskem jeziku

Po najaktualnejšem kodifikacijskem priročniku (SP 2001: §77–109) imena industrijskih izdelkov spadajo med stvarna lastna imena, ki so razdeljena na devet skupin, in sicer:

1. imena, naslovi stvaritev,
2. imena organizacij in družbenih teles,
3. imena delovnih skupnosti,
4. imena oddelkov ustanov in nesamostojnih enot delovnih organizacij,
5. imena (samo)upravnih enot,
6. imena meddržavnih zvez,
7. imena posameznih vozil, npr. ladij, vesoljskih in zračnih plovil ali vlakov,
8. imenovalni prilastki k vrstnim imenom tehničnih izdelkov in trgovskih znamk,
9. pri mednarodnih naravoslovnih poimenovanjih latinska ali polatinjena imena živalskih in rastlinskih vrst.

Takoj lahko opazimo, da je osma kategorija, kamor so uvrščena imena industrijskih izdelkov, nekoliko drugačna

[1] V nadaljevanju SP 2001.

od ostalih, saj so v ospredju zgolj imenovalni prilastki k vrstnim imenom in ne samostojna imena industrijskih izdelkov. Primeri, ki jih navaja SP 2001, so: *cigarete Filter 57*, *otroška soba Boštjan*, *pisalni stroj Olivetti*, *gramofon Melodija* ter *zobna krema Kalodont* (ibid.: § 107).

Že Dobrovoljc (2009: 4) govori o potrebi po posodobitvi definiranja fonda stvarnih imen glede na aktualno jezikovno stanje in za osmo kategorijo predlaga imenovanje »imena znamk in industrijskih izdelkov«. Poraja pa se tudi vprašanje, ali je trenutna kategorizacija imen izdelkov in znamk sploh še primerna, saj nemalokrat pride do spreminjanja referenčnega razmerja, ki je značilno za lastna imena. Tako izgubljajo osnovne lastnoimenske funkcije in vzpostavljajo nanašalna razmerja z več denotati. Pri tem denotat lastnega imena lahko postane serija, skupina izdelkov, predmetov ali objektov, lahko pa stvarna lastna imena dobijo tudi predmetni pomen, čemur rečemo apelativizacija ali poobčnoimenjenje.

Do prve oblike preoblikovanja nanašalnega razmerja pogosto pride pri imenih sodobnih izdelkov, pri katerih se proizvajalci poslužujejo serijske proizvodnje vedno novih izdelkov zaradi potreb in povpraševanja na tržišču. Na trg torej pošiljajo nove modele iste serije, ki v imenih namenoma ohranjajo povezavo z osnovnim izdelkom (ibid.: 5–6). Tako sčasoma imenovalni prilastki k vrstnim imenom lahko postanejo samostojna stvarna lastna imena, saj uporabniki iz njih lahko razberejo pomen jedrnega dela izvirne besedne zveze (iz izvirne besedne zveze urejevalnik besedil Word lahko sklepajo sledeče: Word = urejevalnik besedil). S tem se spremeni skladenjski položaj nepregibnega prilastka, ki postane pregibna beseda. Vsemu temu lahko sledi tudi slovenjenje tujega zapisa in zapis z malo začetnico, ki pa ni dosleden (ibid. 6).

Ko taka serija ali skupina izdelkov dobi še predmetni pomen, pride do apelativizacije in denotat postane občno ime, ki pa že označuje nekaj vrstnega (*kolodont*, *pips*, *žiletka* itd.). Prav v takih primerih se lahko poraja vprašanje o smiselnosti kategorizacije imen industrijskih izdelkov in znamk, saj ne gre več za lastna imena, poleg tega pa se jim spremenijo tudi slovnične lastnosti (ibid. 7).

3 Velika in mala začetnica pri imenih industrijskih izdelkov in znamk v slovenskem jeziku

Pogorelec (1975) piše, da se je o problematiki zapisa stvarnih lastnih imen razpravljalo že pri pripravah načrta pravil za Slovenski pravopis 1962, kjer je bilo opredeljeno razlikovanje med enakopisnicama, pri katerih gre lahko za stvarno lastno ali občno ime. »Če bo šlo za stvarno lastno ime, ga bo pisec pisal z veliko začetnico, občna imena pa z malo, o izbiri začetnice pa naj bi pisec odločal sam na podlagi pomenske vloge posamezne besede oz. besedne zveze«. Izbira začetnice za industrijske izdelke in znamke je bila nato v Načrtu pravil za novi slovenski pravopis osnovana na skladnji, saj naj bi se v primeru, da se rabijo kot imenovalni prilastki k vrstnim imenom, pisali z veliko začetnico (*ure Omega*, *avto znamke Fiat*), če pa nastopajo samostojno, kot sklonljivi samostalniki, pa se pišejo z malo (*slikam s kodakom*, *vozim se s fiatom*).

Dobrovoljc (2009) ugotavlja, da aktualni SP 2001 pravil za zapis imen industrijskih izdelkov in znamk ni spreminjal, so pa bili dodani novi primeri za ponazoritev, ki vključujejo apelativizirana imena: *Zobe si umivam s kalodontom* in *Komarje uničujem s pipsom* (ibid.). Za ta imena SP 2001 navaja, da se lahko poobčnoimenijo na dva različna načina. Prvič lahko postanejo generična, torej da poimenujejo vse tovrstne izdelke, ne le ene vrste (*kalodont*, *superge*). V drugih primerih pa lahko postanejo vrstna in pomenijo le tip nekega izdelka (*oliveti, ford*). Pravila za merilo za razlikovanje med enim in drugim tipom imen določajo začetnico, ki je pri lastnih imenih velika, pri občnih pa mala. V nadaljnjih raziskavah Dobrovoljc (2012) govori o tem, da je »kriterij problematičen posebej zaradi tega, ker izpust skladenjskega jedra (in tako zapis z malo začetnico) besedne zveze ne implicira vedno izgube lastnoimenske funkcije, kar pomeni, da se lastnoimenskost imen, ki so po SP 2001 zapisana z malo začetnico, ohrani tudi v neprilastkovnih skladenjskih položajih«.

3.1 Kodifikacija in raba

Dobrovoljc (2009) predstavi trostopenjski model preoblikovanja znamk v občne besede, ki opisuje tudi štiri različne načine rabe. Pri tem so imena znamk lahko:

1. v vlogi nesklonljivih imenovalnih prilastkov ob občnoimenskem jedru. Ta položaj zahteva rabo velike začetnice.
2. v skladenjski vlogi jedra besedne zveze. Takrat se pregibajo in jih pišemo z veliko začetnico.
3. prešla v občna imena, ki označujejo vrsto. V tej situaciji jih pišemo z malo začetnico.
4. popolnoma samostojna in ne nastopajo več kot prilastki. Tega primera rabe imen industrijskih izdelkov in znamk pravila v SP 2001 ne opredeljujejo.

Problematično četrto skupino bi lahko po SP 2001 (§ 37 in § 148) opredelili na dva različna načina, in sicer kot okrajšana večbesedna lastna imena (*Microsoft Office Word* tako postane *Word*) ali pa kot t. i. poljudna poimenovanja (zapis *word* z malo začetnico). Tu je torej nejasnost glede statusa in zapisa besede, ki je SP 2001 ne pojasni ali opiše, poleg tega pa niti ne definira pomena poljudnosti, da bi uporabniki o tem lahko presodili sami (ibid.).

To skupino bi lahko poimenovali tudi »imena dvoživke«, saj omogočajo tako zapis z malo kot tudi z veliko začetnico. Poleg tega, da so samostojna, ta imena določa tudi dejstvo, da so do te samostojnosti prišla s pogosto rabo, ki je pripomogla k temu, da lastno ime vsebuje tudi pomen nelastnoimenskega dela izvirne besedne zveze. Zapis sčasoma zaradi redne rabe preide z velike na malo začetnico. Tu pa se pojavi problematična stran besed dvoživk, saj je potrebno definirati, kdaj ime izdelka postane občna beseda in ali se to zgodi tudi na pomenski ravni. Glede na številne raziskave različnih (in predvsem tujih) jezikoslovcev se je ta skupina imen izkazala za neprototipično. To so imena, ki ne izpolnjujejo več vseh kriterijev lastnoimenskosti (pragmatično,

pomensko in skladenjsko) in se zato pogosto uresničujejo kot občne besede. Zaradi tega jih lahko dojemamo kot imena izdelkov ali pa kot označitve posameznih primerkov določene serije izdelkov. Pri tem je potrebno upoštevati vse različne pomene, ki jih določena beseda lahko nosi in glede na to prilagoditi tudi zapis bodisi z veliko bodisi z malo začetnico (Dobrovoljc, 2012: 32–33).

Iz teh primerov in preteklih dognanj lahko torej vidimo, da je zapis imen industrijskih izdelkov in znamk dejansko problematična tema z vidika jezikovnega standarda, saj dosedanji poskusi kodificiranja niso uspeli zajeti vseh vidikov rabe. Pri tem moramo imeti v mislih tudi dejstvo, da so se že pojavile nove variante zapisa teh imen (npr. *iPhone*), ki jih obravnavamo v nadaljevanju.

4 Metodologija

4.1 Priprava podatkov

Analiza je bila izvedena s pomočjo korpusa Janes (Fišer et al., 2014). Najprej smo pregledali in razvrstili uporabnike družbenega omrežja Twitter po tipu računa na uradne in zasebne račune. Ali gre za uradne ali zasebne račune, smo preverili tako, da smo pregledali njihov profil, v primeru nejasnosti pa tudi nekaj njihovih objav.

V korpusu prevladujejo zasebni računi (5.806), ki so v korpus prispevali 55.407.416 pojavnic oz. 1.189.180 lem, medtem ko je uradnih računov (1.780), ki v korpusu predstavljajo 13.275.838 pojavnic oz. 866.352 različnih lem, precej manj. Zaradi različne velikosti podkorpusov smo pri analizi upoštevali odstotke in ne absolutnih števil pojavitev posameznega tipa zapisa, pri čemer smo odstotke izračunali za vsako lemo posebej.

4.2 Zasnova raziskave

V nadaljevanju smo s pomočjo regularnih izrazov poiskali vse leme lastnih samostalnikov v podkorpusu tvitov, pri čemer smo enkrat iskali samo po tvitih uporabnikov uradnih računov, drugič pa po tvitih zasebnih uporabnikov. Iz zadetkov smo izločili še vsa uporabniška imena, saj so za našo raziskavo nerelevantna. Iz dobljenih konkordančnih nizov smo nato izdelali in izvozili frekvenčni seznam za oba tipa uporabnikov. Oba seznama smo ročno pregledali in označili 130 imen industrijskih izdelkov, od katerih smo nato izbrali šestdeset lem, ki smo jih umestili v štiri različne kategorije:

- avtomobilizem,
- mobilna telefonija,
- računalništvo ter
- hrana in pijača.

Zatem smo za vsako kategorijo določili po pet lem (skupaj dvajset), ki smo jih izbrali glede na pogostost in zanimivost posamičnih primerov zapisa velike in male začetnice. Minimalni pogoj za vključitev v analizo je bil, da se lema tako pri uradnih kot pri zasebnih računih pojavi vsaj tridesetkrat. Tu velja opozoriti, da smo z izborom poskušali vsaj do neke mere izločiti morebiten šum v izpisanih rezultatih, ki so posledica napačne avtomatske lematizacije. V primerih, kjer smo to odkrili, v analizo nismo vključili pojavnic, ki niso predstavljale pravih lem.

Seznam izbranih industrijskih izdelkov za analizo:

- avtomobilizem: *audi, renault, volvo, mercedes, corsa*,
- mobilna telefonija: *samsung, galaxy, xperia, blackberry, iphone*,
- računalništvo: *twitter, google, windows, minecraft, itunes*,
- hrana in pijača: *union, radenska, poli, nutella, cedevita*.

V nadaljevanju smo za vse izbrane leme iskanja izdelali frekvenčne sezname za vse besedne oblike, pri čemer smo enkrat iskali po uradnih, drugič pa po zasebnih računih. S pomočjo teh seznamov smo dobili frekvenčne podatke o variantnosti zapisov iskanih lem (v analizo smo vključili vse atestirane besedne oblike, za večjo preglednost rezultatov pa v nadaljevanju prispevka podatke navajamo na nivoju lem). Za nazornejši prikaz rabe v kontekstu smo na koncu ročno izbrali po nekaj primerov osnovnih besednih oblik s pripadajočim kontekstom, ki jih predstavimo pri posameznih kategorijah.

5 Analiza rabe velike in male začetnice

5.1 Avtomobilizem

Kot kaže tabela 1, uporabniki uradnih računov avtomobilistične izdelke pogosteje pišejo z veliko začetnico. Zanemarljiv odstotek obeh tipov uporabnikov ta imena zapisuje s samimi velikimi črkami. Pri zasebnih računih lahko opazimo, da se za zapis z veliko začetnico odloča malenkost nižji odstotek uporabnikov v primerjavi z uradnimi računi, vendar ta še vedno presega 70 %. Pri tej kategoriji se manj običajni zapisi, kjer bi se neprve črke pisale z velikimi črkami, ne pojavljajo. Štiri izmed izbranih pojavnic predstavljajo znamke vozil, ki imajo lahko različne serije in se njihova imena lahko tudi prekrivajo z imeni podjetja (v primeru ford celo z osebnim lastnim imenom), ena izmed pojavnic pa je ime tipa vozila (corsa). Pri slednji se je izkazalo, da je nekoliko višji odstotek uradnih računov besedo pisalo z malo začetnico (35,3 %), vendar se je še vedno več kot polovica (62,7 %) uporabnikov odločila za uporabo velike začetnice.

Avtomobilizem	Primeri	Absolutna frek.		Odstotki	
		uradni	zasebni	uradni	zasebni
prva z veliko	*Audi*	1.270	1.026	80,7 %	70,6 %
vse z malo	*audi*	227	532	18,0 %	27,7 %
vse z veliko	*AUDI*	10	25	1,3 %	1,7 %
neprva z veliko	-	0	0	0,0 %	0,0 %
oba dela z veliko	-	0	0	0,0 %	0,0 %
variacije	-	0	0	0,0 %	0,0 %
skupaj	-	1.507	1.583	100,0 %	100,0 %

Tabela 1: Variante zapisov besed s področja avtomobilizma.

Rabo v kontekstu ponazarjamo še na štirih različnih primerih, ki prikazujejo nestandardno rabo velike in male začetnice pri obeh tipih računov:

[1] Uradni računi (469 pojavitev za besedno obliko *Audi*):

hja veš, če se ne pripelješ z ***Audijem*** *ali vsaj Passatom...*

[2] Zasebni računi (217 pojavitev za besedno obliko *Audi*):

Audi A6 Avant je zagotovo odlična izbira. Mimo grede, ljubitelji psov pogosteje posežejo po ***Audiju*** *(raziskava);)*

[3] Uradni računi (16 pojavitev za besedno obliko *audi*):

Prihaja nov kupe med robustnimi velikani - ***audi*** *Q8*

[4] Zasebni računi (58 pojavitev za besedno obliko *audi*):

jaz osebno bi raje preveril smart fortwo, ***audi*** *a2, kangoo ter mercedes A (stari jasno)*

Pri prvih dveh navedenih primerih opazimo, da tako uporabniki uradnih kot zasebnih računov besedo *Audi* pišejo z veliko, čeprav gre v obeh primerih za poimenovanje vrstnega, kar bi glede na pravopisni standard sicer zahtevalo rabo male začetnice. V drugem primeru lahko izpostavimo tudi rabo velike začetnice pri lastnem imenu točno določenega izdelka, ki ga dopolnjuje desni prilastek (*Audi A6 Avant*). Tretji in četrti primer pa predstavljata primer nestandardne rabe male začetnice. Obakrat gre namreč za lastno ime določnega modela vozila, saj ob njem stoji desni prilastek. Če primerjamo še števila pojavitev, izpisanih za posamično rabo velike in male začetnice, je razvidno, da raba velike začetnice močno prevladuje pri izbranem primeru, ne glede na to, ali račun pripada uradni ali fizični osebi.

5.2 Mobilna telefonija

Kot je razvidno iz Tabele 2, pri mobilni telefoniji prihaja do pogostejšega pisanja neprve črke z velikimi črkami (npr. *iPhone*). V tej kategoriji opažamo tudi pisanje posameznih delov izdelkov z veliko začetnico (npr. *BlackBerry*). Predvidevamo, da razlog za tak trend leži v politiki podjetja, ki zahteva in promovira tak zapis, ne glede na trenutno veljavni slovenski standard.

Na naslednjih 8 naključnih primerih ponazarjamo okoliščine rabe velikih in malih črk pri kategoriji mobilne telefonije.

Mobilna telefonija	Primeri	Absolutna frek.		Odstotki	
		uradni	zasebni	uradni	zasebni
prva z veliko	*Iphone*	1.524	1.681	65,9 %	54,0 %
vse z malo	*xperia*	96	986	4,1 %	24,7 %
vse z veliko	*GALAXY*	51	35	1,5 %	0,7 %
neprva z veliko	*iPhone*	974	1.640	18,1 %	13,9 %
oba dela z veliko	*BlackBerry*	46	90	10,4 %	6,7 %
variacije	*iPhOnE*	2	2	0,0 %	0,0 %
skupaj	-	2.693	4.434	100,0 %	100,0 %

Tabela 2: Variante zapisov besed s področja mobilne telefonije.

[5] Uradni računi (183 pojavitev za besedno obliko *Xperia*):

/.../ preverjamo ali je ***Xperia*** *uporabna za objavo fotk v reviji. Bo potrebno iskat dalje...*

[6] Zasebni računi (174 pojavitev za besedno obliko *Xperia*):

/.../ sicer sam 4mega piksli... Ampak dela boljs kot moja trenutna ***Xperia*** *z s 16 megapiksli... It's magic i guess...*

[7] Uradni računi (6 pojavitev za besedno obliko *xperia*):

Sonyjeva ***xperia*** *z je v našem uredništvu, prvi telefon z zaslonom full hd. Jutri jo v Londonu soočimo z novim HTCjem. /.../*

[8] Zasebni računi (56 pojavitev za besedno obliko *xperia*):

/.../ Js mam android 4.0, ce to misls.. Sicer je pa tole moja prva xperia pa se ne spoznam se lih kej

[9] Uradni računi (965 pojavitev za besedno obliko *iPhone*):

Twelve South, je po popularni paleti starinskih ovitkov za ***iPhone*** *in iPad predstavil še BookBook ovitek za iMac. /.../*

[10] Zasebni računi (1810 pojavitev za besedno obliko *iPhone*):

Ko se glih odlocam o tem ali prodat ***iPhone*** *in it nazaj na BlackBerry, Apple objavi update.. in iOS ima zopet star sijaj! #nebomprodal*

[11] Uradni računi (51 pojavitev za besedno obliko *iphone*):

super bi bilo ce bi vam uspelo dol dati baterijo, ker je ***iphone*** *se zmeraj pod napetostjo.*

[12] Zasebni računi (620 pojavitev za besedno obliko *iphone*):

Prodam ***iphone*** *4s črne barve, več info zs;)*

Iz navedenih primerov je razvidno, da se tudi tokrat raba nagiba k uporabi velikih oziroma »izvirnih« začetnic, ki jih promovirajo proizvajalci, ne glede na to, v kakšnem kontekstu se besede pojavijo. Pri primeru besedne oblike *Xperia* sicer lahko opazimo, da se je uporabnik odločil za nekoliko nekonsistenten pristop k zapisu, saj je *Xperia* zapisal z veliko, desni prilastek »z« pa z malo, kljub temu da je to del lastnega imena izdelka, za katerega bi jezikovni in pravni standard sicer zahteval pisavo z veliko začetnico. Ko se osredotočimo na male začetnice, opazimo, da je pri uradnih računih zgolj šest takih besednih oblik, kar priča o neuveljavljenosti male začetnice pri zapisu tega izdelka. Poleg tega je primer št. 7 nekoliko presenetljivo zapisan z malo začetnico, saj govori o specifičnem izdelku podjetja Sony, za katerega bi navadno pričakovali zapis z veliko začetnico. Pri osmem primeru lahko vidimo rabo male začetnice, ki je v skladu z jezikovnim standardom, saj uporabnik govori o seriji izdelkov – liniji modernih pametnih telefonov podjetja Sony.

Še posebej je zanimiva raba besednih oblik *iPhone* in *iphone*, saj število pojavitev pri inovativnem načinu zapisa daleč presega zapis samo z malimi črkami, konteksti, v katerih je pojavlja zapis *iPhone*, so različni, ker ta lahko predstavlja jedro besedne zveze, stoji samostojno, ali opisuje serijo izdelkov. Rabo s samimi malimi črkami pri uradnem računu bi lahko razumeli kot neke vrste generično poimenovanje za telefon, ki je v tem specifičnem primeru *iPhone*. Pri primeru, ki pripada zasebnemu računu, pa je raba nekoliko problematična, uporabnik namreč govori o določenem modelu telefona, pri katerem bi sicer predvidevali rabo velikih črk (*iPhone 4S*).

5.3 Računalništvo

V tretji tabeli, ki vsebuje podatke o industrijskih izdelkih s področja računalništva, opazimo podoben trend kot pri mobilni telefoniji. Raba pri obeh vrstah uporabnikov se očitno nagiba k uporabi velike začetnice, le da je ta težnja pri uradnih računih še toliko bolj izrazita. Ponovno opazimo rabo neprve črke z veliko začetnico (npr. *iTunes*). Slednja se obnaša podobno kot *iphone*, saj sta oba izdelka podjetja Apple, ki svoje izdelke poimenuje in oglašuje z značilno malo začetnico »*i*«, ki ji sledi velika neprva črka. To nam še posebej nazorno prikaže, kakšna je korporativna politika rabe velike in male začetnice. Rezultati pri zasebnih uporabnikih pa kažejo, da tudi oni upoštevajo zapis, kakršnega želijo proizvajalci, bodisi zaradi oglaševanja bodisi zaradi drugih razlogov.

Računalništvo	Primeri	Absolutna frek.		Odstotki	
		uradni	zasebni	uradni	zasebni
prva z veliko	*Windows*	3.219	8.272	70,2 %	54,6 %
vse z malo	*twitter*	450	5.835	7,9 %	27,7 %
vse z veliko	*GOOGLE*	69	130	2,3 %	1,3 %
neprva z veliko	*iTunes*	108	404	19,3 %	16,1 %
oba dela z veliko	*MineCraft*	2	10	0,3 %	0,3 %
variacije	*tWItter*	1	1	0,0 %	0,1 %
skupaj	-	3.848	14.654	100,0 %	100,0 %

Tabela 3: Variante zapisov besed s področja računalništva.

Za področje računalništva smo izbrali besedni obliki *Google* in *google*:

[13] Uradni računi (1280 pojavitev za besedno obliko *Google*):

Tudi ***Google*** *Prevajalnik gre v šolo. Lahko mu pomagaš /.../*

[14] Zasebni računi (2735 pojavitev za besedno obliko *Google*):

Kako naj jaz sestavim Rubikovo, če jo ***Google*** *vrti že ure, neuspešno /.../*

[15] Uradni računi (57 pojavitev za besedno obliko *google*):

Foto: Nenavadni, strašljivi in čudoviti svet ***google*** *street viewa /.../*

[16] Zasebni računi (1336 pojavitev za besedno obliko *google*):

/.../ šerlok bi dandanes uporabil ***google*** *. meni pa to ni izziv. :)*

Tudi pri tej kategoriji opazimo, da številke kažejo v prid rabe velike začetnice. Nastopa pa tako v besedni zvezi, kjer prestavlja lastno ime programa, kot tudi samostojno, kjer predstavlja ime podjetja oziroma iskalnika. Kar se tiče rabe male začetnice, je ta v 15. primeru zagotovo nestandardna, kajti pojavi se v lastnem imenu programa. V primeru št. 16 pa bi zapis lahko imeli za standarden, ker besedo *google* v tem (nekoliko skopem) kontekstu lahko dojemamo kot generično poimenovanje za internetni iskalnik. Dilema, ki se tu pojavi, je, kje in kako postaviti mejo za (ne)generičnost določenih imen.

5.4 Hrana in pijača

V Tabeli 4 opazimo najvišji delež pisanja z veliko začetnico pri obeh tipih računov, saj je ta pri uradnih uporabnikih skoraj 90-odstoten, pri zasebnih pa krepko presega polovico (62,5 %). Ta kategorija sicer malce izstopa v primerjavi z ostalimi tremi, ker ne predstavlja tehničnih izdelkov. Ravno zaradi tega je morda nekoliko presenetljivo, da se v njej kaže enaka težnja kot pri ostalih kategorijah. Uporabniki se namreč ravno tako pretežno odločajo za zapis imen z veliko začetnico. Kot zanimivost bi tu lahko dodali še dejstvo, da je to edina kategorija, v kateri poleg imen izdelkov tujih proizvajalcev najdemo tudi imena slovenskih industrijskih izdelkov in znamk.

Hrana in pijača	Primeri	Absolutna frek.		Odstotki	
		uradni	zasebni	uradni	zasebni
prva z veliko	*Cedevita*	1375	1635	88,4%	62,5%
vse z malo	*radenska*	26	852	7,8%	36,9%
vse z veliko	*POLI*	20	23	3,9%	0,6%
neprva z veliko	-	0	0	0,0%	0,0%
oba dela z veliko	-	0	0	0,0%	0,0%
variacije	*NUtella*	0	1	0,0%	0,0%
skupaj	-	1421	2511	100,0%	100,0%

Tabela 4: Variante zapisov besed s področja hrane in pijače.

Za ponazoritev konteksta rabe imena izdelkov iz zadnje kategorije predstavljamo sledeče primere:

[17] Uradni računi (101 pojavitev za besedno obliko *Radenska*):

/.../ Živjo! Še enkrat smo preverili stanje v trgovini in ***Radenska*** *je na polici.*

[18] Zasebni računi (110 pojavitev za besedno obliko *Radenska*):

Domač metin sirup + ***Radenska*** *= #win poletna pijača. #izum tedna:)*

[19] Uradni računi (2 pojavitvi za besedno obliko *radenska*):

*Na lastnike žal nimamo vpliva... Je pa še vedno "****radenska*** *" marsikje po ex YU generično ime za mineralno vodo. :)*

[20] Zasebni računi (59 pojavitev za besedno obliko *radenska*):

/.../ Meni je bolj ali manj vseeno, kako kdo reče špricerju, samo da je posredi dobro vino in ***radenska****.*

Ponovno se pri obeh tipih računov pogosteje pojavi raba velike začetnice. Tokrat bi pri rabi male začetnice še posebej radi izpostavili rabo pri uradnih računih (kjer sta sicer zgolj 2 pojavitvi za besedno obliko *radenska*). Uporabnik namreč že sam govori o generičnosti imena tega izdelka, ki ga ponazoritveno zapiše z malo začetnico. V primeru št. 20 prav tako najdemo zapis, ki bi lahko predstavljal generično rabo. Ne moremo pa povsem zagotovo vedeti, ali je uporabnik tu mislil katerokoli mineralno vodo ali vodo podjetja *Radenska*, saj govori o kvaliteti te vode, s čimer bi lahko v mislih imel prav specifično vodo. V slednjem primeru bi bila bolj ustrezna raba velike začetnice. Podobno velja za primer št. 18: če se uporabnik nanaša na mineralno vodo podjetja Radenska, raba povsem ustreza jezikovnemu standardu. Tu se torej pojavi problem t. i. imen dvoživk, za katera je v takih primerih potrebno ugotoviti, ali so že prešla v občna imena in na podlagi tega izbrati ustrezen zapis.

5.5 Diskusija

Iz korpusne analize je razvidno, da je tako pri uradnih kot zasebnih računih močan trend zapisovanja imen industrijskih izdelkov z veliko začetnico. Z izjemo kategorije mobilne telefonije je pri uradnih računih povsod odstotek zapisa z veliko začetnico vsaj 70-odstoten, pri zasebnih pa vsaj 54-odstoten. Razlog za nižji odstotek pri kategorijah mobilna telefonija in računalništvo leži v dejstvu, da je sta bili med drugim upoštevani lemi *iphone* in *itunes*. Prvo lemo 90,5 % uporabnikov uradnih računov in 69,6 % uporabnikov zasebnih računov zapisuje kot *iPhone*, drugo pa 96,4 % uradnih in 80,5 % zasebnih piše kot *iTunes*, torej tako, da je neprva črka zapisana z veliko. Med vsemi dvajsetimi obravnavanimi lemami je izstopala le lema *twitter*, ki jo zasebni uporabniki pogosteje pišejo z malo začetnico (54,9 %), medtem ko v zapisih uradnih računov še vedno prevladuje raba velike začetnice (76,0 %).

Glede na to, da sicer splošen trend zapisovanja imen industrijskih izdelkov z veliko začetnico odstotkovno izstopa predvsem pri uradnih računih, lahko sklepamo, da proizvajalci in podjetja stremijo k doslednemu zapisu takih izdelkov z veliko začetnico oziroma z morebitnimi izvirnimi zapisi tipa *iPhone*. Pri zasebnih uporabnikih se za enak zapis, kot ga uporabljajo uradni računi, odloča sicer nekoliko nižji delež, vendar ta še vedno predstavlja veliko večino uporabnikov. To pomeni, da se tudi zasebni uporabniki ravnajo po zapisu, ki ga uporabljajo proizvajalci in prodajalci, četudi ni vedno v skladu z rabo, ki jo predpisuje aktualna kodifikacija, še posebej pa je to razvidno iz trenda zapisovanja izvirnih imen izdelkov. V prihodnosti bi bilo smiselno raziskavo tudi razširiti, in sicer tako, da bi rabo na družabnem omrežju Twitter primerjali z rabo v referenčnih korpusih, kjer bi se lahko osredotočili na razliko v rabi med leposlovnimi oziroma strokovnimi ter publicističnimi besedili. Če bi tudi v teh besedilih odkrili podoben trend kot v spletnih besedilih, bi bilo to vsekakor zanimivo za nadaljnje diskusije o tem, kaj nam raba sporoča.

6 Sklep

Rezultati raziskave nam kažejo, da se raba ravna po taktirkah podjetij, ki želijo, da se njihovi izdelki zapisujejo po njihovih standardih, ne nujno po jezikovnem. Za jezikovni standard, ki ne predvideva vseh okoliščin, v katerih se imena industrijskih izdelkov in znamk pojavljajo, to pomeni, da ni več aktualen. Predvsem je to odraz potrebe po prevetritvi standarda, ki bi se moral prilagajati razvoju jezika, ki ga opažamo v rabi. Pri tem bi morali upoštevati tudi nove načine pristopa k uporabi velikih in malih začetnic, ki se odraža v izdelkih, kot je *iPhone*, saj gre tu z vidika slovenščine za praktično povsem nov koncept, ki še ni bil upoštevan v nobenem kodifikacijskem priročniku. V mislih je treba imeti, da je tudi spletna slovenščina del našega jezika in da odraža, kako jezik napreduje in v katero smer se govorci nagibajo s svojo rabo. To rabo bi bilo potrebno upoštevati pri bodočih priročnikih za slovenski jezik, saj naj bi ti predstavljali aktualno stanje jezika in naj ga ne bi omejevali po nepotrebnem, sploh ko gre za področja, ki v preteklosti še niso bila dovolj podrobno in premišljeno kodificirana. Sem zagotovo spada tudi področje imen industrijskih izdelkov, zlasti ko gre za izdelke moderne dobe.

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

Helena Dobrovoljc. 2009. Pravopisna obravnava imen znamk in industrijskih izdelkov ter posledice spreminjanja njihovih lastnoimenskih funkcij. Jezik in slovstvo, 54(6): 3–19.

Helena Dobrovoljc. 2012. Pisanje imen izdelkov in znamk. V: Nataša Jakop, Helena Dobrovoljc, ur., Pravopisna stikanja: razprave o pravopisnih vprašanjih. Založba ZRC, Ljubljana. 27–39.

Darja Fišer, Tomaž Erjavec, Ana Zwitter Vitez, Nikola Ljubešić. 2014. JANES se predstavi: metode, orodja in viri za nestandardno pisno spletno slovenščino. V: *Jezikovne tehnologije: zbornik 17. mednarodne multikonference Informacijska družba – IS 2014*. 56–61. Institut "Jožef Stefan", Ljubljana

Darja Fišer, Tomaž Erjavec, Jaka Čibej in Nikola Ljubešić. 2015. Gradnja in analiza korpusa spletne slovenščine JANES. V: *Slovnica in slovar – aktualni jezikovni opis. Obdobja 34*. Znanstvena založba Filozofske fakultete, Ljubljana. (v tisku)

Breda Pogorelec. 1975. Kako je z veliko in malo začetnico pri stvarnih lastnih imenih? Jezik in slovstvo 21(1). 30–31.

SP 2001 = Slovenski pravopis. Ljubljana: Znanstvenoraziskovalni center SAZU, Založba ZRC.

Rana ura, slovenskih fantov grob: analiza frazeoloških prenovitev v spletni slovenščini

Martin Justin,* Nejc Hirci,* Polona Gantar‡

* Ljubljana
martin1123581321@gmail.com
nhirci@gmail.com
‡ Oddelek za prevajalstvo, Filozofska fakulteta Univerze v Ljubljani
Aškerčeva 12, 1000 Ljubljana
apolonija.gantar@ff.uni-lj.si

Povzetek

V prispevku analiziramo osem izbranih frazemov glede na njihovo zastopanost v korpusu Janes kot specializiranem viru za preučevanje jezika spletno specifičnih besedil in v korpusu Kres kot uravnoteženem referenčnem korpusu pisne slovenščine. Poleg zastopanosti obravnavanih frazemov v obeh korpusih nas je zanimal tudi delež frazeoloških prenovitev v korpusu Janes glede na besedila referenčnega korpusa. V raziskavi ugotavljamo, da je vključevanje tako frazemov kot njihovih prenovitev ena od bistvenih značilnosti spletne komunikacije ter da pisci pri tem uporabljajo prepoznavne strategije, kot so ustvarjanje dobesednega ali nasprotnega pomena, vključevanje nepričakovanih besed ali vsebin, združevanje frazemov ter vključevanje aktualnih družbenih in političnih elementov.

Early hour, early grave of Slovene boys: analysis of phrasological innovations in the internet Slovene

The paper analyses eight selected phrasemes with regard to their frequency in the Janes corpus as a specialised resource for researching texts specific for communication on the web, and in the Kres corpus as a balanced representation of written Slovene. In addition to general frequency, we were interested in the proportion of phraseological innovations in the Janes corpus, in relation to their frequency in the reference corpus. Research shows that both phrasemes and their innovations one of the important characteristics of communication on the web, and that writers use standard techniques, such as generating literal or opposite meaning, the inclusion of unexpected words or substance, combining two or more phrasemes and the inclusion of topical social and political elements.

1 Uvod: Frazeologija in frazeološke enote

Frazeologijo je mogoče definirati kot jezikoslovno vedo, ki preučuje različne tipe večbesednih enot, njihove pomenske, izrazne, skladenjske in besedilne lastnosti. Vendar pa jezikoslovci in znotraj njih frazeologi, leksikologi in leksikografi postavljajo različna merila določanja večbesednih leksikalnih enot in to kljub temu, da je mogoče v strokovni literaturi prepoznati nekatere skupne definicijske lastnosti, kot so na primer: večbesednost, izrazna in skladenjska ustaljenost, pomenska neprozornost ter ekspresivnost, pogosto opredeljena kot metaforičnost oz. slikovitost v načinu izražanja (Gantar, 2007: 72).

Razpon, ki ga zajema definicija večbesednosti, postavlja pod vprašaj enote, ki jim je na izrazni ravni težko pripisati enobesednost, npr. pri fraznih glagolih, kot so *odstopiti od, računati na, izhajati iz*, ali izrazih, nastalih po frazeološki poti, npr. *piš-me-v-uh-ovstvo, nebodi-ga-treba* ipd. Stopnja pomenske prozornosti je problematična zaradi različnih pomenskih interpretacij posameznih besed v odnosu do pomena zveze kot celote, npr. koliko je pomen 'črna barva' prisoten v pomenu sestavine *črn* v zvezi *črni humor* ali *črna gradnja*? Izrazna in skladenjska ustaljenost je zrahljana s številnimi variantami in različnimi skladenjskimi vlogami istega frazema, npr. *občutiti, izkusiti, okusiti ... na lastni koži*, ali če navedemo primer iz korpusa Janes za izhodiščni frazem *dobiti/imeti kurjo kožo*: *kurja koža ol ouver; ti kr kurja koža rata ...; men gre kr kurja koža ...; kr kurjo kožo dobite; da se ti kurja koža po hrbtu dela; kurja koža v trenutku!*

Poleg omenjenega se kot merilo frazeološkosti postavlja tudi neterminološkost zveze kot celote, npr. *črna ovca* je navadno prepoznana kot frazeološka enota, ne pa tudi *črna skrinjica* ali *črna luknja*, ki se jima pripisuje terminološka vrednost – ta tip se v literaturi največkrat prepoznava kot stalne besedne zveze.

Problem vključenosti v širšo množico frazeoloških enot predstavljajo kolokacije kot tipične, za jezik specifične besedne sopojavitve, ki pa jih materni govorci pogosto težko prepoznavamo kot tipične in ustaljene, saj gre za kombinacije, ki se jih naučimo po naravni poti, tj. v procesu usvajanja jezika. Na primer v besedni zvezi *serijski morilec* bi katerakoli sopomenka ene od besed zvenela nenavadno, celo napačno, npr. *serijski ubijalec*. Podobno je za slovenščino običajno reči *gosta megla* in *trda tema*, ne pa tudi **trda megla* in **gosta tema*.

Delitev večbesednih enot na kolokacije, stalne zveze in frazeme, ki smo jo uporabili v raziskavi, temelji na razdelitvi, ki je bila uporabljena pri gradnji Leksikalne baze za slovenščino (Gantar, 2015) v okviru projekta Sporazumevanje v slovenskem jeziku,[1] in omogoča prepoznavanje večbesednih enot na podlagi analize besedilnega okolja v korpusu. Za raziskovanje kolokacij je korpusna analiza tako rekoč nujna, saj presega meje naše jezikovne intuicije. Posledično je šele s korpusnimi analizami tako raziskovanje končno docela omogočeno, hkrati pa je z opazovanjem besednega vzorčenja v realnih besedilih mogoče opazovati tako regularno kot tudi ustvarjalno in nepričakovano jezikovno rabo.

[1] Rezultati projekta so dostopni na http://www.slovenscina.eu/.

2 Izhodiščna predvidevanja

Izhajali smo s stališča, da pisci izvirnost, drugačnost ali veščino najlažje izkazujejo s slogom svojega pisanja, kamor sodi tudi besedni zaklad in uporaba različnih večbesednih enot, predvsem frazemov. Po našem mnenju zahteva bogat in izviren slog običajno tudi bolj zapletene stavčne strukture,[2] bolj poglobljeno razmišljanje o napisanem, torej več časa, ki pa ga splet kot mesto hitre, dvosmerne, hkrati pa bolj sproščene in manj formalne komunikacije, običajo ne daje oziroma si ga pisci niso pripravljeni vzeti. Zato predvidevamo, da pisci spletno specifičnih besedil, zlasti tvitov in komentarjev, pogosto posežejo po frazemih in predvsem po njihovih prenovitvah ter da posledično sproščena in neformalna komunikacija predvideva tudi večjo rabo frazeoloških enot in njihovih prenovitev na sploh.

V prispevku nas je torej zanimalo, kako pogosto se frazemi pojavljajo v spletno specifičnih besedilih, pri čemer smo za primerjavo vzeli besedilno heterogen referenčni korpus Kres. Nadalje, kako pogosto se v obeh korpusih pojavljajo frazeološke prenovitve,[3] torej inovativne spremembe ustaljenih frazemov, kot npr. *rana ura, zlata ura → rana ura, slovenskih fantov grob*. Poleg tega nas je zanimalo, katere so najopaznejše skupne značilnosti takih rab v spletnih besedilih.

3 Raziskava

V raziskavi[4] smo se osredotočili na frazeme, za katere poleg večbesednosti in frazeološkega pomena, ki ga določa pomenska neprozornost sestavin glede na celostni pomen, velja tudi slikovitost in ekspresivnost izražanja, pogosto z namenom ustvariti določeno opaznost ali izraziti bodisi kolektivni bodisi osebni pogled na svet. Preučevanje tovrstnih enot v besedilih spletne slovenščine daje po našem mnenju dobra izhodišča za preučevanje ustvarjalne in humorne jezikovne rabe.

3.1 Metodologija

Z uporabo korpusnih orodij, ki omogočajo prepoznavanje vzajemno odvisnih besednih kombinacij v določenih slovničnih relacijah, pa tudi z urejanjem besedilnega okolja v konkordančnem nizu je mogoče razmeroma enostavno ugotoviti ponavljajoče se jezikovne vzorce in njihove leksikalne zapolnitve. Hkrati je mogoče opazovati tudi širši besedilni kontekst, ki narekuje pomensko razpoznavnost pa tudi način rabe takih enot v določenem besedilnem tipu ali govornem položaju.

Za raziskavo smo uporabili korpusa Janes (pribl. 160 mil. besed) in Kres (pribl. 120 mil. besed). Korpus Janes (Fišer et al., 2014) je sestavljen izključno iz spletno specifičnih besedil, in sicer tvitov (38 %), komentarjev (9 %), blogov (24 %) in forumov (29 %), Kres pa kot uravnotežen referenčni korpus (Logar et al., 2012) iz približno enakih deležev revij (20 %), časopisov (20 %), leposlovja (17 %) ter stvarnih (18 %) in spletnih besedil (20 %). Ker vključuje korpus Kres tudi delež spletnih besedil, smo pri primerjalni kvantitativni analizi frazeme, ki so se v takih besedilih pojavljali v korpusu Kres, obravnavali kot spletno specifične in jih pripisali končnemu številu pojavitev v korpusu Janes. V Tabeli 1 je število pojavitev obravnavanega frazema v spletnih besedilih korpusa Kres prikazano v oklepajih. Omenjena primerjalna metoda nam služi predvsem za ugotavljanje lastnosti frazeološke rabe v spletno specifičnih besedilih, ne želimo pa ugotovitev posploševati na razmeroma heterogena in žanrsko različno opredeljena besedila, ki jih vsebuje korpus Kres. V tem smislu bi bilo ustrezneje primerjati rabo frazemov v spletno specifičnih besedilih npr. z besedili, ki so v korpusu Kres kategorizirana kot leposlovje, kar je lahko ena od prihodnjih raziskovalnih nalog pri preučevanju rabe frazemov.

Za analizo korpusov smo uporabljali orodje Sketch Engine (Kilgarriff et al., 2004), in sicer funkcijo iskanja v konkordančniku, urejanje besedilnega okolja in filtriranja besedil, deloma pa tudi funkcijo Besedne skice in frekvenčni prikaz obravnavanega frazema v določenem besedilnem tipu. Zadnje smo uporabili predvsem pri določitvi števila pojavitev frazema v spletnih besedilih, vključenih v korpus Kres.

Pri določanju iskalnih pogojev smo izhajali iz sedmih stavčnih frazemov oz. pregovorov in enega glagolskega, ki smo jih poznali sami in so se nam zdeli bodisi vsebinsko bodisi glede na svojo zgradbo možen predmet prenovitev. Spodaj jih navajamo v obliki, ki se je pokazala kot najbolj tipična:

- kdor drugemu jamo koplje, sam vanjo pade
- jabolko ne pade daleč od drevesa
- ni vse zlato, kar se sveti
- kdor visoko leta, nizko pade
- kdor čaka, dočaka
- brez muje se še čevelj ne obuje
- rana ura, zlata ura
- iz muhe delati slona

V iskalnik smo vpisali del frazema, za katerega smo predvidevali, da bo ostal nespremenjen (npr. *kdor čaka, rana ura* ipd.) ali pa smo ločeno iskali pojavitve za oba dela frazema, če smo sklepali, da je mogoče oba prenoviti (npr. *ni vse zlato ...* in *... kar se sveti*). Dobljeni konkordančni niz smo ročno pregledali in (a) izločili vse konkordance, ki niso vsebovale frazema oz. jih ni bilo mogoče opredeliti kot frazeološke, (b) prešteli vse rabe frazema v izhodiščni obliki, (c) prešteli vse ponovitve izhodiščnega frazema in (č) zabeležili vse različne in (d) ponovljene oz. enake prenovitve. Postopek smo izvedli na obeh korpusih in rezultate med seboj primerjali. Pri analizi frazemov v korpusu Kres smo, kot rečeno, upoštevali dejstvo, da ta vključuje tudi uravnotežen delež spletnih besedil.

3.2 Kvantitativna analiza

Kot je razvidno iz Tabele 1, se obravnavani frazemi – gledano v absolutnih vrednostih – pojavljajo v tipično spletnih besedilih (korpus Janes) v povprečju trikrat pogosteje kot v (nespletnih) besedilih korpusa Kres. Izjemo predstavlja frazem *brez muje ...*, ki se sicer v obeh korpusih pojavlja redko, vendar v korpusu Kres enkrat

[2] Izrazita prvina stila so seveda lahko tudi zelo kratke, celo enobesedne povedi.

[3] Izraz prenovitev uporabljamo v pomenu, kot ga je uvedla Kržišnik (Kržišnik, 1987, prim. tudi Kržišnik, 1990 in 1996: 140/141), in sicer gre za "inovativne spremembe oblike in/ali pomena frazema, izpeljane v besedilu in v njem prepoznavne, ki so namerne in (praviloma) enkratne".

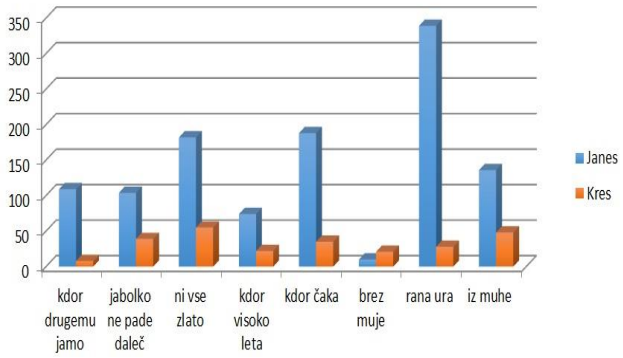
[4] Raziskava je nastala v okviru srednješolskega tabora Janes, ki je potekal na Filozofski fakulteti v Ljubljani od 24. do 28. avgusta 2015.

pogosteje. Prav tako je mogoče opaziti, da je v spletni slovenščini več neopredeljivih primerov, tj. takih, ki jim ni mogoče določiti frazeološke vrednosti (ker so bodisi nejasni ali pa je zveza rabljena dobesedno), kar je mogoče pripisati kratkosti in specifičnosti izražanja, značilnega zlasti za tvite in komentarje.

Iskalni pogoj	**Korpus Janes**				**Korpus Kres**			
	vse pojavitve	**izhodiščni frazemi**	**prenovitve**	**nejasno /dobesedno**	**vse pojavitve**	**izhodiščni frazemi**	**prenovitve**	**nejasno/ dobesedno**
kdor drugemu jamo	109	59	34	12	8 (3)	7	1	0
jabolko ne pade daleč	104	91	9	4	39 (2)	18	21	0
ni vse zlato	182	146	33	3	55 (13)	48	3 (1)	4 (1)
kdor visoko leta	74	38	32	4	22 (3)	17	5	0
kdor čaka	188	155	25	10	35 (5)	27	5	3 (1)
brez muje	10	3	2	5	21 (0)	16	1	4
rana ura	339	215	92	32	28 (5)	17	5 (2)	6
iz muhe	136	128	7	1	48 (11)	45	1	2
skupaj	1142	835	234	71	256 (42)	195	42 (3)	19 (2)

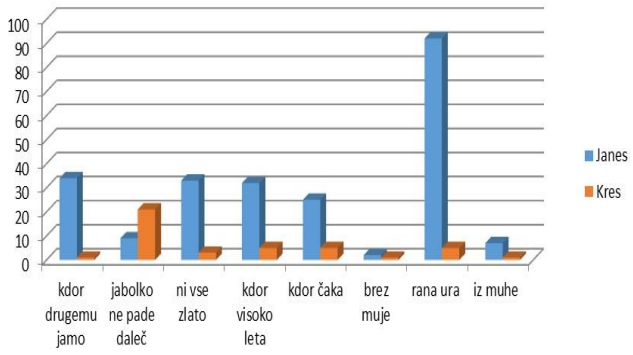
Tabela 1: Kvantitativna analiza frazemov v korpusih Janes in Kres.

Razmerja v številčni zastopanosti za vse obravnavane frazeme, torej tako za izhodiščne kot za prenovitvene, v obeh korpusih prikazuje Graf 1.



Graf 1: Zastopanost obravnavanih frazemov v korpusih Kres in Janes.

Podobne ugotovitve kot za zastopanost frazemov na splošno veljajo tudi za zastopanost frazeoloških prenovitev, kjer je mogoče ugotoviti, da je ustvarjalna raba za spletna besedila zelo značilna (Graf 2), medtem ko je v besedilih korpusa Kres sicer opazna, ne moremo pa govoriti o značilnostih frazeološke rabe na splošno, saj bi se v tem primeru morali osredotočiti na primerjalno analizo posameznih besedilnih tipov v korpusu Kres. Izjemo predstavlja frazem *jabolko ne pade daleč* ..., ki ima kljub pogostejši zastopanosti v spletnih besedilih (Graf 1), več prenovitev v korpusu Kres (Graf 2).



Graf 2: Zastopanost frazeoloških prenovitev pri obravnavanih frazemih v korpusih Kres in Janes.

Posledično je tudi razmerje med prenovljenimi in izhodiščnimi frazemi v korpusu Janes manjše, iz česar lahko sklepamo, da je verjetnost prenovitve glede na uporabljeni frazem v spletni slovenščini precej velika. Če upoštevamo primerjavo z besedili korpusa Kres, je izjema že omenjeni frazem *jabolko ne pade* ..., ki je v korpusu Kres večkrat rabljen prenovljeno, kot pa v izhodiščni obliki.

3.3 Vsebinska analiza

Podrobnejša analiza posameznih prenovitev je pokazala, da jih je glede na učinek ali namen avtorja ter glede na način, kako je prenovitev dosežena, mogoče razdeliti v nekaj prepoznavnih skupin.

Zelo pogoste so prenovitve, ki dajo izhodiščnemu frazemu **dobesedni pomen**, kar zaradi izvorne prenesenosti, metaforičnosti oz. ustaljenosti običajno deluje ironično ali kako drugače humorno. Spodaj navajamo nekaj primerov v obeh korpusih (v oklepaju je navedeno več kot enkratno število enake prenovitve v posameznem korpusu).

Janes
- kdor drugemu jamo koplje, ga bo sosed ovadil zaradi dela na črno
- kdor drugemu jamo koplje, je prvega že zasul
- kdor drugemu jamo koplje, je grobar (3)
- kdor visoko leta, je pilot (4)/ptič/stevardesa
- kdor visoko leta, ima lep razgled
- kdor visoko leta, daleč vidi (3)
- kdor čaka, ima dobre živce
- kdor čaka, zamudi/umre
- kdor čaka, se načaka/čaka
- kdor čaka, postane star
- kdor čaka, ima čas (3)
- ni vse zlato, kar je drago

Kres
- kdor drugemu jamo koplje, naj se ne čudi, če bo že zasedena
- Zidane ne pade daleč od očeta
- kdor visoko leta, je ptič (2)
- kdor čaka, se načaka

Prenovitve lahko učinkujejo, če vsebujejo glede na izvirni frazem **nepričakovano besedo**, ki vsebinsko ali izrazno zamenjuje izhodiščno, npr.

Janes
- iz muhe delati Airbus/Dumbota
- ne delat avijona iz muhe k ji je slon na nogo stopil
- jabolko ne pade daleč od hruške (2)
- brez muje, se še modrc ne odpne
- brez muje, se še papuč ne sezuje
- kdor visoko leta, iz drevesa pade
- rana ura, bronasta ura
- ni vse zlato, kar je nafta
- rana ura, vroča/črna kava
- rana ura, podočnjaki do kolen
- rana ura, veliki podočnjaki

Kres
- jabolko ne pade daleč od hruške (2)/jablane (2)
- žena ne pade daleč od drevesa
- jabolko ne pade daleč od – brata

Humorni učinek lahko avtor doseže tudi s prenovitvijo, ki da frazemu **nasprotni pomen** od pričakovanega, npr.

Janes
- rana ura, zaspana ura
- rana ura – srčna rana
- brez muje se čevelj z jezičkom zapne
- kdor visoko leta, daleč vidi
- rana ura, rano dan zjebe

Pogosto je humorni učinek dosežen z **združitvijo dveh frazemov**, npr.

Janes
- kdor drugemu jamo koplje, je sam svoje sreče kovač (3)
- kdor drugemu jamo koplje, je sam svoje ovadbe kovač
- kdor drugemu jamo koplje, volk iz gozda pride
- ni vse zlato, kar se Janezek nauči
- kdor visoko leta, se najslajše smeje
- kdor visoko leta, ne pade daleč od drevesa
- kdor čaka, ostanek dobi
- rana ura, slovenskih fantov grob (38)
- rana ura, sam vanjo pade (4)
- rana ura, zlata žila
- rana ura, zlata kura
- rana ura – brazilske kave grob

Kres
- kdor visoko leta, se najslajše smeje
- kdor visoko leta, daleč pride
- rana ura, slovenskih fantov grob (2)

Izstopa prenovitev *rana ura, slovenskih fantov grob*, sestavljena iz frazema *rana ura*, *zlata ura* in verza, ki ga poznamo iz znane slovenske ljudske vojaške pesmi Oj Doberdob,[5] ki se je tako ustalil (36 pojavitev v korpusu Janes in 4 v korpusu Kres, od tega 2 v spletnih besedilih), da bi ga lahko obravnavali kot samostojni frazem.[6]

Prenovitve se lahko nanašajo tudi na **konkretno politično** ali **družbeno dogajanje**, npr.

Janes
- kdor drugemu jamo koplje, vlada vanjo pade
- kdor drugemu jamo koplje... je jutri vabljen v Rovte
- ni vse zlato, kar prihaja iz Amerike/ Skandinavije
- ni vse zlato, kar zraste S od nas
- ni vse zlato, kar je nemško
- ni vse zlato, kar je slovensko
- iz muhe Patria se dela slon

Prenovitve se lahko nanašajo na **znane osebnosti** ali **blagovne znamke**, npr.

Janes
- ni vse Apple, kar se sveti
- ni vse zlato, kar je Žižek (2)
- ni vse zlato, kar je od forda
- kdor visoko leta, je Peter Prevc
- rana ura, zlati Rolex

4 Zaključek

V preučevanih spletnih besedilih, zlasti v tvitih in komentarjih, se pisci pogosto izražajo na način, ki želi biti drugačen, učinkovit in odmeven. Ena od možnosti, kako to doseči pri hitrem, impulzivnem načinu pisne komunikacije, ki se pogosto osredotoča na izražanje mnenj, nazorov ter na opisovanje odnosov in deljenje izkušenj – največkrat v povezavi s konkretnim družbenim dogajanjem, je uporaba frazemov. Frazemi namreč s svojo izhodiščno »metaforo«, ki je lastna (celotni) jezikovni

[5] Recenzentu se zahvaljujemo za opozorilo, da gre dejansko za verz in ne za Prežihov roman Doberdob, saj Prežih pesmi v svojem romanu nikjer izrecno ne navaja (Tucovič, 2010: 44).

[6] Npr. v zgledu iz korpusa Kres: *Čeprav slovenski pregovor pravi, da je rana ura slovenskih fantov grob, se bo zgodnje bujenje obrestovalo z obiskom ene najbolj zanimivih ribjih tržnic na svetu Tsukiji.*

skupnosti, predstavljajo dobro izhodišče za ustvarjanje novih metafor, asociacij in kulturnospecifičnih pomenskih vsebin. Na ta način se od bralca zahteva določen napor, da najprej prepozna izhodiščno metaforo, nato pa še lastnosti, ki mu omogočajo razumevanje njene nadgradnje. Učinek, kot kažejo analizirani primeri, želi biti predvsem humoren, prenovitveni postopki pa so vezani na izrabljanje pomenskih lastnosti frazemov (nasprotni ali dobesedni pomen, združevanje frazemov) in vpletanjem aktualnih družbenih in političnih pa tudi osebnih dogodkov.

Zgledi primerjave med korpusoma Janes in Kres potrjujejo predvidevanje, da je raba frazemov tako v izhodiščnem pomenu kot v prenovitvah za spletna besedila zelo značilna, posledično pa je mogoče sklepati tudi na specifične lastnosti spletne komunikacije (hitra odzivnost, drugačnost, učinkovitost, neformalnost), ki se tudi sicer kažejo na vseh ravneh njenega jezikovnega opisa.

5 Zahvala

Raziskava, opisana v prispevku, je bila opravljena v okviru nacionalnega temeljnega projekta “Viri, orodja in metode za raziskovanje nestandardne spletne slovenščine” (J6-6842, 2014-2017), ki ga financira ARRS.

6 Literatura

Darja Fišer, Tomaž Erjavec, Ana Zwitter Vitez in Nikola Ljubešić. 2014. JANES se predstavi: metode, orodja in viri za nestandardno pisno spletno slovenščino V: T. Erjavec in J. Žganec Gros, ur. Zbornik Devete konference Jezikovne tehnologije, str. 56–61. Ljubljana: Institut Jožef Stefan.

Polona Gantar. 2007. Stalne besedne zveze v slovenščini: korpusni prostop. Založba ZRC, ZRC SAZU.

Polona Gantar. 2015. Leksikografski opis slovenščine v digitalnem okolju. Ljubljana: Znanstvena založba Filozofske fakultete UL.

Adam Kilgarriff, Pavel Rychlý, Pavel Smrz in David Tugwell. The Sketch Engine. 2004. V: W. Geoffrey 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.

Erika Kržišnik. 1987. Prenovitev kot inovacijski postopek. Slava I(1): 49–56.

Erika Kržišnik. 1990. Tipologija frazeoloških prenovitev v Cankarjevih proznih besedilih. Slavistična revija 38(4): 400–420.

Erika Kržišnik. 1996. Norma v frazeologiji in odstopi od nje v besedilih. Slavistična revija 44(2): 133–154.

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

Vladka Tucovič. 2010. Ljudska pesem v romanih Prežihovega Voranca *Doberob* in *Jamnica*. Jezik in slovstvo 55 (3–4): 41–52.

Postavljanje vejic v slovenščini s pomočjo strojnega učenja in izboljšanega korpusa Šolar

Anja Krajnc, Marko Robnik-Šikonja

Fakulteta za računalništvo in informatiko, Univerza v Ljubljani
Večna pot 113, 1000 Ljubljana
anja.krajnc@gmail.com, marko.robnik@fri.uni-lj.si

Povzetek

Poskušamo izboljšati trenutne pristope strojnega učenja za postavljanje vejic v slovensko besedilo. Generiramo nove atribute na podlagi slovničnih pravil za slovenski jezik, ki z dodatno informacijo omogočijo boljše učenje. Za analizo uporabimo korpus Šolar, ki je bil uporabljen v že obstoječi raziskavi, in izboljšano verzijo tega korpusa. Uporabimo vzorčenje neuravnoteženih množic ter odstranimo neinformativne atribute. Z metodo ReliefF ocenimo kakovost množice atributov. Testiramo različne klasifikacijske algoritme: naključne gozdove, metodo podpornih vektorjev, naivni Bayesov klasifikator, RBF mrežo, alternirajoče odločitveno drevo, AdaBoostM1 ter odločitveno tabelo. Najboljše rezultate dosežemo z naključnimi gozdovi, alternirajočimi odločitvenimi drevesi in odločitveno tabelo. Ugotovimo, da trenutni korpusi niso primerni za izdelavo modelov, ki bi bili splošno uporabni.

Placing comma in Slovene using machine learning and updated corpus Šolar

We improved current machine learning approaches to comma placement in Slovene language. First we generate new features based on grammar rules. This additional information improves performance of machine learning approaches. We learn from corpus Šolar and its updated version. We use undersampling of imbalanced data sets and feature subset selection with ReliefF algorithm. The classification methods we test are random forests, support vector machines, naïve Bayesian classifier, RBF network, alternating decision trees, AdaBoost.M1, and decision table. The best performing methods are random forests, alternating decision trees, and decision table. We conclude that current learning corpuses do not allow construction of generally applicable models.

1 Uvod

Pravilno postavljene vejice v besedilu ne puščajo samo dobrega vtisa pri bralcih in odsevajo verodostojnost in strokovnost besedila, pač pa omogočajo tudi lažje in enolično razumevanje vsebine stavkov, ločijo stavke znotraj povedi in nakazujejo premor v govoru. Hkrati lahko napačno postavljene vejice spremenijo pomen stavkov. Programi, ki uspešno postavljajo vejice v besedilo, so pomembni tako za pisce, ki naletijo na zadrego pri pisanju, kot tudi pri računalniški razpoznavi in obdelavi govora. Postavljanja vejic s strojnim učenjem se lotimo, ker ima slovenščina zahtevno oblikoslovno podobo in so pravila za pisanje vejic velikokrat zahtevna za razumevanje, njihova programska implementacija pa težko uresničljiva. Prav tako je strojno postavljanje vejic del zahtevnejših jezikovnih tehnologij, katerih cilj je vzdrževanje večjezičnosti.

Za postavljanje vejic v slovensko besedilo obstajata dva programa, ki delujeta na podlagi ročno napisanih pravil: Besana (Amebis d.o.o., 2015) in LanguageTool (LanguageTool skupnost, 2015). Peter Holozan se je lotil postavljanja vejic s pomočjo strojnega učenja na seznamu primerov iz korpusa Šolar (Holozan, 2012; Holozan, 2013). Strojno učenje je uporabil za problem iskanja vseh vejic ter za problem popravljanja realnih napak v besedilu, torej za iskanje odvečnih vejic. Njegove obsežne študije, ki bistveno presegajo obseg tega dela, želimo izboljšati predvsem z izboljšano predstavitvijo informacije, ki je na voljo strojnemu učenju. Uvedemo več izboljšav množice atributov in predlagamo uporabo še netestiranih uspešnih učnih algoritmov, ki so se izkazali na podobnih nalogah. Zahtevnost računskih operacij, ki se je v dosedanjih raziskavah izkazala kot omejitveni dejavnik za več učnih algoritmov, želimo zmanjšati z ocenjevanjem kakovosti atributov in s podvzorčenjem. Podrobneje o naših poskusih poročamo v (Krajnc, 2015), tukaj pa povzamemo bistvene ugotovitve.

Bistveno za uspeh napovedovanja z algoritmi strojnega učenja je, da algoritmom priskrbimo kakovostno informacijo o problemu in to v takšni obliki, da so jo sposobni izkoristiti, glede na njihove predpostavke in formalizem, ki ga uporabljajo za predstavitev napovednega modela. Izhajali smo iz korpusa Šolar, ki vsebuje popravljena besedila šolskih esejev. Ker se je med analizo izkazal za neprimernega za strojno učenje, smo preizkusili tudi izboljšano verzijo tega korpusa, ki jo je sestavil in nam jo posredoval Peter Holozan (2015). Ta korpus vsebuje približno trikrat več primerov z vejicami, prav tako je bilo veliko primerov ročno pregledanih in popravljenih. Podrobneje o zgradbi podatkovne množice poročamo v razdelku 2

Dosedanji poskusi s strojnim učenjem so bili le delno uspešni. Menimo, da je mogoče algoritmom priskrbeti še dodatno informacijo in vključiti tudi znanje, skrito v pravilih za postavljanje vejic. Za generiranje novih atributov na podlagi slovničnih pravil smo razvili orodje, ki uporabi pravila, ki jih za postavljanje manjkajočih vejic uporablja LanguageTool (2015), in jim dodali še dodatna pravila, ki jih narekuje Slovenski pravopis (Jože Toporišič in sod., 2001). Dosedanji atributni opis korpusa ima več pomanjkljivosti, ki jih analiziramo v razdelku 3 Neinformativne atribute odstranimo, nekatere pa modificiramo.

Na izboljšani podatkovni množici uporabimo algoritme, ki so že bili uporabljeni na osnovni podatkovni množici v predhodni raziskavi, to so naivni Bayesov klasifikator,

RBF mreža, alternirajoče odločitveno drevo, AdaBoostM1 in odločitvena tabela, ter uporabimo še dva algoritma, ki sta se v večini primerjalnih študij, npr. (Caruana in Niculescu-Mizil, 2006) izkazala z robustnim delovanjem, to sta algoritem naključnih gozdov ter metoda podpornih vektorjev. Za zmanjšanje potrebnih računskih virov smo preizkusili dve tehniki, ocenjevanje atributov, ki bi ga lahko potencialno uporabili za izbiro podmnožice pomembnih atributov in vzorčenje. Kakovost atributov smo ocenili z algoritmom ReliefF, ki zna upoštevati tudi pogojne odvisnosti atributov glede na razred. Pri neuravnoteženih učnih množicah, kot je v našem primeru (razmerje med učnimi primeri, ko je vejica in ko je ni, je približno 1:10), se pogosto kot koristno izkaže podvzorčenje (ang. undersampling), le-to nam istočasno zmanjša velikost učne množice, kar pohitri učenje.

V nadaljevanju najprej v 2 razdelku podrobneje opišemo podatkovne množice, obstoječe atribute in nove atribute, ki jih generiramo na podlagi pravil. Izboljšave predstavitve že uporabljene množice atributov si ogledamo v razdelku 3 V 4 razdelku opišemo ocenjevanje atributov ter podvzorčenje. Rezultate klasifikacije si ogledamo v razdelku 5 V 6 razdelku povzamemo poglavitne ugotovitve študije in podamo nekaj idej za nadaljnje delo.

2 Opis podatkovnih množic

Študijo smo začeli na podatkovnih množicah, uporabljenih v Holozan (2013), ki izhajajo iz korpusa Šolar. Te podatkovne množice in korpus imenujemo *Šolar1*. V korpusu je 11.892 pravilno postavljenih vejic ter 11.399 manjkajočih. Korpus je bil oblikoskladenjsko označen z oblikoslovnim označevalnikom in lematizatorjem Obeliks (Grčar et al., 2012) ter lematiziran in skladenjsko razčlenjen s skladenjskim razčlenjevalnikom (Dobrovoljc et al., 2012). Vsaka beseda z okoliškim oknom, ki se pojavi v korpusu, je pretvorjena v seznam atributov, dodan pa je tudi atribut, ki pove, ali tej besedi sledi vejica (Holozan, 2013). Nastala podatkovna množica je v formatu ARFF, ki je sestavljen iz glave ter podatkovnega dela in je namenjen orodju WEKA (Witten in Frank, 2005). V glavi so opisani atributi, tako da je za vsak atribut podano ime atributa in vrednosti, ki jih lahko atribut zavzame, v podatkovnem delu pa je v vsaki vrstici ena beseda, opisana z vrednostmi atributov. Korpus vsebuje 23.282 primerov z vejico (pozitivnih) in 185.875 brez vejice (negativnih), skupaj torej 209.157 besed.

Ker so začetni rezultati testiranj pokazali, da je obstoječi korpus neprimeren za strojno učenje, smo uporabili izboljšano in posodobljeno verzijo tega korpusa, ki jo je sestavil Peter Holozan. Veliko stavkov v tej novi podatkovni množici, ki jo imenujemo *Šolar2*, je ročno pregledanih in popravljenih (Holozan, 2015). Izboljšan korpus vsebuje 728.927 učnih primerov, od tega 65.356 z vejico in 663.571 brez vejice.

2.1 Opis atributov

Prvotna podatkovna množica Šolar1 je vsebovala po 67 atributov za vsako besedo. Prvi atribut je pojavnica. Ta atribut lahko zavzame toliko vrednosti, kolikor je različnih besed in njenih oblik v besedilu, npr. pojavnici človek in človeka zasedeta dve mesti v zalogi vrednosti tega atributa. Naslednji atribut je lema trenutne besede. Tudi zaloga vrednosti tega atributa je zelo velika, sestavljajo pa jo vse leme besed v besedilu. Tukaj pojavnici človek in človeka zasedeta eno vrednost in sicer osnovno obliko besede, človek. Sledi zapis MSD kode besede, zaloga vrednosti pa zavzame vse možne oblikoskladenjske oznake po oblikoskladenjskih specifikacijah JOS (Erjavec et al., 2010), s katero besedi določimo besedno vrsto, sklon, spol, število itd. Naslednji trije atributi so binarni in nosijo informacijo, ki jo priskrbi skladenjski označevalnik (Dobrovoljc et al., 2012). Zasedejo lahko eno izmed vrednosti 0 ali 1 ter sporočajo, ali trenutna beseda kaže na veznike, del povedi (obstaja povezava med trenutno besedo in neko drugo besedo v povedi) ter ali obstaja povezava bodisi na osebke, predmete ali prislovna določila. Opisanih šest atributov se ponovi za vse besede znotraj okoliškega okna. Za analizo uporabljamo okoliško okno pet besed pred in pet besed za trenutno besedo. To pomeni, da imamo skupaj enajstkrat po šest zgoraj opisanih atributov. Tem atributom sledi še razred, ki pove, ali besedi sledi vejica. Zavzema lahko dve vrednosti [je-vejica, ni-vejica].

2.2 Predstavitev pravil v atributni obliki

Obstoječim atributom smo dodali 45 novih, od tega 41 z implementacijo pravil, ki jih za postavljanje vejic uporablja orodje LanguageTool ter 4 dodatna pravila, ki jih narekuje Slovenski pravopis in ki piscem pogosto povzročajo težave. Za vsako pravilo ustvarimo nov atribut.

Oglejmo si primer implementacije pravila za vezniško besedo *ker* po pravilih, ki jih za postavljanje vejic uporablja orodje LanguageTool: kadar naletimo na besedo *ker* in beseda pred njo ni eno izmed ločil *,(;-:* ali ena izmed besed *in*, *ali*, *ter*, *a* in *temveč*, potem trenutni besedi verjetno sledi vejica, zato atribut za trenutni veznik zavzame vrednost 1. Če beseda ni veznik, ki ga iščemo, ali pa je veznik, ki ga iščemo, in ne ustreza podanemu pogoju, potem atribut za ta veznik pri tej besedi zavzame vrednost 0.

Implementirali smo še nekaj dodatnih pravil, ki bodisi piscem povzročajo največ težav bodisi zahtevajo dodatno pozornost pri pisanju. Nekatera pravila smo implementirali tako, da smo dodali pogoj pri implementaciji obstoječih pravil, ostala pa smo ustvarili kot nove samostojne atribute.

2.2.1 Členek da

Kadar se *da* v stavku rabi kot podredni veznik, pred njim najverjetneje stoji vejica. Primere povzemamo po Toporišič in sod. (2001) . Primeri uporabe: *Zeblo ga je tako, da se je ves tresel. - Daj mu nageljček, da bo tudi on vedel, da je prvi maj. - Bral je knjige, da bi spoznal svet.* Posebnost nastopi, kadar se *da* v stavku rabi kot členek. V takšnem primeru pred njim ne pišemo vejice. Primeri uporabe: *Ti da tega ne znaš? - Baje da vozi. - Iz Pirana da ste?* Obstoječemu atributu, generiranemu s pomočjo pravil za veznik *da*, smo dodali nov pogoj, kjer smo preverili, ali je MSD koda trenutne besede enaka L (oblikoskladenjska oznaka členka). Če je pogoj izpolnjen, atribut brez nadaljnjih preverjanj dobi vrednost 0. Šele ob neizpolnjevanju pogoja so se izvršila preverjanja ostalih pravil za postavitev vejice, ki ustrezajo besedi *da*, kadar se rabi kot podredni veznik. Tako program, kadar naleti na besedo *da*, preveri

ali je MSD koda besede enaka L, kar pomeni, da je beseda členek. Če je pogoj izpolnjen, atributu za ta veznik vrednost nastavi na 0. Če pogoj ni izpolnjen, kar pomeni, da beseda ni členek (veliko upov polagamo v točnost oblikoskladenjskega označevalnika), se izvršijo pogoji za ta veznik in vrednost atributa nastavi v skladu z njimi. Za pogoj, ki preverja, ali je beseda členek, ne pa za pogoj, ki preverja, ali je beseda veznik, smo se odločili, ker lahko beseda *da* nastopa tudi v vlogi drugih besednih zvez, npr. glagol v sedanjiku in tretji osebi ednine. Primer: *Nikoli ga ne da iz rok.*

2.2.2 Večbesedni vezniki

Slovenski pravopis zapoveduje, da vejic med deli večbesednega veznika ne pišemo. Primeri uporabe: *Namesto da bi se učil, je lenaril. - Lenaril je, namesto da bi se učil. - Že ko sem ga prvič videl, sem ga spregledal. - Kljub temu da je bil bolan, je šel na delo. - Rad ga imam, zato ker je tako miren.* Taki vezniški izrazi so še: tako da, toliko da, potem ko, vtem ko, še ko, brž ko, šele ko, s tem da, posebno ko, zlasti če itd. (Jože Toporišič in sod., 2001). To smo implementirali tako, da smo vsem do sedaj generiranim atributom dodali dodaten pogoj, ki pravi: če je trenutna beseda veznik, ki ustreza vsem pravilom za postavljanje vejice za ta vezniški izraz, in pred njo ne stoji beseda z MSD kodo za veznik (Vp ali Vd), potem atribut za ta veznik prejšnji besedi nastavi na 1. V nasprotnem primeru dobi atribut vrednost 0.

2.2.3 Dodaten pogoj za večbesedne veznike

S prejšnjim načinom smo poskušali čim bolj natančno najti večbesedne veznike v besedilu z uporabo MSD kod ob predpostavki, da označevalnik povsem natančno določa besedne vrste. To smo storili zaradi enostavnosti te rešitve. Ker Obeliks besedne vrste pripisuje z 98,30 % natančnostjo (Grčar et al., 2012), morda z zgoraj opisanim postopkom v oblikoslovni označevalnik polagamo preveč zaupanja. Zato implementiramo tudi izvedbo pravila brez uporabe MSD kod in brez zanašanja na označevalnik za primere, ko veznik *in* nastopa v kombinaciji z besedami *sicer*, *to*, *če* in *ko* ter tvori vezniški izraz, pred katerim verjetno stoji vejica. To smo implementirali tako, da smo za trenutno besedo preverili, ali je enaka besedi *in*, nato preverili, ali je naslednja beseda enaka eni izmed zgoraj naštetih besed, in če sta pogoja izpolnjena, prejšnji besedi (besedi na položaju pred trenutno besedo) atribut za ta veznik nastavili na 1, v nasprotnem primeru pa na 0.

3 Izboljšave predstavitve atributov

Predstavitev nekaterih atributov smo spremenili, da bi algoritmi strojnega učenja bolje izkoristili v njih vsebovano informacijo. Nekatere atribute, ki strojnemu učenju ne koristijo, smo izločili iz podatkovne množice.

3.1 Obravnava pojavnic in lem

V dosedanjih poskusih strojnega učenja na problemu vejic so bili atributi sestavljeni tako, da sta dve besedi, ki imata isto oblikoskladenjsko oznako, obravnavani kot dve različni vrednosti atributa. To ni smiselno, saj v slovenščini (ne)obstoj vejice ni odvisen od besede same kot oznake pojma, pač pa od njene besedne vrste (ponekod tudi od nekaterih oblikoskladenjskih lastnosti besede) in povezav z drugimi besedami v povedi. Zato je bolj smiselno, da posameznih besed ne obravnavamo kot ločenih vrednosti atributov. S tem se izognemo drugačni obravnavi besed iste besedne vrste (ali celo iste oblikoskladenjske oznake), ki določajo enako vez z drugimi besedami in v stavku nosijo isti pomen. S tem premislekom atribute, ki določajo posamezne besede, odstranimo. Enako kot je nesmiselno definirati atribut glede na samo besedo, velja tudi za leme. Tudi atribute, ki opisujejo leme besed, odstranimo iz množice atributov.

3.2 Obravnava oblikoskladenjskih oznak

Vsaka beseda iz korpusa Šolar1 je bila do sedaj v atributih predstavljena z eno od 930 različnih oblikoskladenjskih oznak. Samostalnik lahko zavzame eno izmed 78 vrednosti, zaimek eno izmed 451 vrednosti, pridevnik ima 168 oznak in glagol 127 različnih oznak. Z bogatejšim in večjim korpusom bi bile te številke še večje, saj bi nastopalo še več možnih oblik. Pri tem se pojavi vprašanje, ali je to smiselno. V slovenščini vejice postavljamo glede na besedne vrste in povezave med besedami v povedi, ne pa glede na vse možne kombinacije lastnosti, vidov in oblik teh besednih vrst. Povezave in pravila za postavljanje vejic je težko razbrati, če za stvari, ki bi nam morale podati isto informacijo, uporabljamo več deset različnih oznak. Po tem premisleku smo neinformativne atribute, ki opisujejo MSD kode in so lahko zavzeli eno izmed 930 vrednosti, spremenili na dva načina. Tako dobimo dve podatkovni množici, ki se razlikujeta v številu in vrednosti tistih atributov, ki nam povedo oblikoskladenjske lastnosti trenutne besede in ostalih besed znotraj okna.

3.2.1 Zapis MSD kode s po enim atributom za vsako besedo znotraj okna

Prvi način spremembe atributov, ki nosijo podatek o MSD kodi, vpliva na enajst atributov znotraj okna. Na primer, vse različne oznake samostalnika, ki so kombinacije vrste, spola, števila, sklona in živosti, ki pri samostalniku tvorijo mnogo različnih oblikoskladenjskih oznak, nadomestimo z oznako S, ki bo strojnemu učenju povedala najpomembneje - naleteli smo na samostalnik. S tem se izognemo situacijam, kjer algoritem strojnega učenja nima dovolj podatkov o neki oznaki, ki se redko pojavi (ali pa se npr. pojavi samo v testni množici), lahko pa bi se naučil ukrepanja pri isti besedni vrsti v drugem spolu (in/ali številu, sklonu,...) in s tem z drugačno oznako. Tudi pri pridevniku vse možne oznake, ki določajo, da je beseda pridevnik in hkrati njegovo vrsto, stopnjo, spol, število, sklon in določnost (po oblikoskladenjskih oznakah JOS jih je 279, v našem korpusu pa 168), zamenjamo s skupno oznako za pridevnike (P). Enako storimo, kadar naletimo na prislov (R), zaimek (Z), številnik (K), predlog (D) in glagol (G). MSD oznako za kategorijo Neuvrščeno, kamor spadajo razne neznane besede, tipkarske napake in besede, ki jih je program napačno tokeniziral, nadomestimo z N.

Obdržimo oznake za veznik. Po oblikoskladenjskih oznakah JOS (Erjavec et al., 2010) veznik označimo s črko V ter dodatno črko, ki določa, ali je veznik podredni ali

priredni. Tako besedo, kadar ugotovimo, da je veznik, označimo z oznako Vd ali Vp. Ker je stavljenje vejic velikokrat odvisno od vrste veznika, obdržimo obe oznaki. Prav tako obdržimo oznake za medmet (M), okrajšavo (O) in členek (L), saj so pri njih oznake dolge en znak.

3.2.2 Zapis MSD kode s po 9 atributi za vsako besedo znotraj okna

S prvim načinom izboljšanja MSD oznak smo atribute, ki opisujejo oblikoskladenjske oznake, zamenjali s prvo črko oznake (razen pri veznikih, kjer smo obdržali prvi dve črki). S tem smo obdržali samo podatke o besedni vrsti. Z drugim načinom vključimo vse podatke, ki jih nosi posamezna MSD koda. S tem želimo zagotoviti potencialno več informacije za strojno učenje, tvegamo pa morebitno pretirano prilagajanje podatkom.

Najdaljšo oblikoskladenjsko oznako ima zaimek, kjer je MSD koda sestavljena iz 9 znakov, zato iz osnovne MSD kode generiramo devet atributov. V dosedanji predstavitvi MSD koda določa en atribut za vsako besedo, to je 11 atributov za vse besede znotraj okna, posamezna vrednost vsakega izmed atributov je dolga največ 9 znakov, vseh vrednosti pa je 930. Z novo implementacijo vsak atribut, ki nosi podatek o MSD kodi, spremenimo v devet novih, tako da vsak izmed novih atributov dobi vrednost enega od znakov MSD kode. Kadar je dolžina MSD kode manjša od 9 znakov, atributom, ki določajo preostale znake, nastavimo vrednost *, ki je znak za manjkajočo vrednost. Ker spreminjamo 11 atributov, torej po en atribut za vsako besedo znotraj okna, to pomeni, da enajst osnovnih atributov nadomestimo z 99 novimi. Spremenjena podatkovna množica se tako poveča iz 90 na 178 atributov za opis trenutne besede.

3.2.3 Zapis MSD kode s po 38 atributi za vsako besedo znotraj okna

MSD oznake bi lahko opisali še na tretji način, ki bi strojnemu učenju morebiti podal nekaj več informacije o povezavah med deli MSD kode za posamezno besedno vrsto. Kot prvi atribut bi obdržali prvo črko MSD oznake, ki nam pove besedno vrsto. Ta atribut bi bil edini skupen vsem besednim vrstam in bi zasedal eno izmed dvanajstih vrednosti: S za samostalnik, G za glagol, P za pridevnik, R za prislov, Z za zaimek, K za števnik, D za predlog, V za veznik, L za členek, M za medmet, O za okrajšavo ter N za neuvrščeno. Temu atributu bi sledili ločeni atributi za vsako posamezno besedno vrsto in sicer za vsako besedno vrsto toliko atributov, kolikor dolge so njene MSD oznake. Če bi začeli s samostalnikom, pri katerem lahko njegova MSD oznaka zasede šest mest, od tega prvi znak določa besedno vrsto in smo njegovo vrednost že zapisali v prvi atribut, bi to pomenilo, da bi bili atributi na mestih od drugega do šestega rezervirani za podatke o samostalniku. Pri vseh ostalih besednih vrstah bi bili atributi na teh mestih nastavljeni na vrednost *. Temu bi sledilo 7 atributov za oznake glagola, 6 za pridevnik, 2 za oznake prislova, 8 za zaimek, 6 za števnik, po 1 za predlog, veznik in neuvrščeno. MSD oznake medmetov in okrajšav so dolge po en znak, ki nam pove besedno vrsto, to pa že vsebuje prvi atribut, zato za ti dve besedni vrsti dodatnih atributov ne bi definirali. MSD oznake bi torej opisali s po 38 novimi atributi za vsako besedo znotraj okna. To bi skupaj pomenilo 418 atributov, ki bi nosili informacije o MSD oznakah. Implementacijo in testiranje tretjega načina puščamo za nadaljnje delo, saj zahteva več računskih kapacitet.

4 Ocenjevanje atributov in podvzorčenje

Zanima nas pomembnost atributov, tj. koliko informacije posamezen atribut vsebuje pri klasifikaciji za dani učni problem. Kvaliteto atributov ocenjujemo z mero za ocenjevanje atributov ReliefF (Robnik-Šikonja in Kononenko, 2003). ReliefF je široko uporabljan algoritem za ocenjevanje atributov, ki zaznava tudi pogojne odvisnosti med atributi, uspešno obravnava šumne in neznane vrednosti ter deluje tudi za večrazredne probleme.

Ocenjevanje atributov smo pognali na podatkovnih množicah, ki izhajajo iz korpusa Šolar1 in Šolar2, na vsakem po dve množici. Par podatkovnih množic se med seboj razlikuje pri atributih, ki podajajo informacijo o oblikoskladenjski oznaki. Prva množica iz para vsebuje 90 atributov in ima vsako oblikoskladenjsko oznako opisano z enim atributom. Znotraj okna je tako 11 atributov, ki opisujejo oblikoskladenjske oznake, zato ti množici imenujemo Šolar1-11 in Šolar2-11. Drugi dve množici na vsakem korpusu vsebujeta 178 atributov in imata vsako oblikoskladenjsko oznako opisano s po devetimi atributi (99 atributov za opis oblikoskladenjskih oznak in zato poimenovanje Šolar1-99 in Šolar2-99). Zanima nas, kako pomembni so atributi in koliko informacije nosijo oblikoskladenjske oznake, če jih opišemo bodisi z enim bodisi z devetimi atributi. Ker se zaključki glede ocen atributov z mero ReliefF ne razlikujejo med korpusom Šolar1 in Šolar2, predstavimo v tabeli 1 le rezultate za korpus Šolar2 in sicer po 15 najbolje ocenjenih atributov. Rezultati množice Šolar2-11 so na levi strani in množice Šolar2-99 na desni strani tabele 1.

Za Šolar2-11 so najbolje ocenjeni atributi o MSD kodah besed na mestih 0, 1, -1, -2, in 2. Na desni strani tabele 1 se za Šolar2-99 pokaže, da je najpomembnejša beseda na mestu 0, ki z različnimi oznakami zavzame kar štiri od petih najvišjih mest (msd0-5, msd0-3, msd0-4, msd0-1). Iz tega sklepamo, da nosi drugi način opisa MSD kod več koristne informacije. Kljub temu pa je treba upoštevati, da sta korpusa Šolar1 in Šolar2 zelo specifična in da vsebujeta več med seboj podobnih besedil, tako da rezultatov ne moremo preveč posploševati.

Rezultate algoritma ReliefF bi lahko uporabili za izbor podmnožice pomembnih atributov, s čimer bi lahko omejili trenutno dokaj visoko porabo računskih virov. Zaradi pomanjkanja prostora poskusov s to možnostjo, ki koristi predvsem manj zmogljivim algoritmom strojnega učenja, nismo vključili v članek.

Za neuravnotežene učne množice, kot je naš korpus Šolar, kjer je razmerje med mesti z vejico in brez nje približno 1:10, pogosto za izboljšanje delovanja učnih algoritmov koristi uravnoteženje učne množice. Ena najbolj uveljavljenih tehnik uravnoteženja, ki istočasno tudi zmanjša zahtevnost računanja, je podvzorčenje (ang. undersampling). Le-to smo preizkusili tudi na našem problemu, a ker rezultati bistveno ne odstopajo od tistih brez uravnoteženja, o tem v nadaljevanju ne poročamo, bi pa bilo tehniko smiselno uporabiti za določanje optimalnih parametrov, ki zahtevajo mnogo ciklov učenja, in na bodočih, večjih korpusih.

	Šolar2-11		Šolar2-99	
rang	ocena	atribut	ocena	atribut
1	0,355	msd0	0,357	msd0-5
2	0,271	msd1	0,353	msd0-3
3	0,237	msd-1	0,340	msd0-4
4	0,161	msd-2	0,337	msd-1-1
5	0,132	msd2	0,320	msd0-1
6	0,132	msd-3	0,292	msd-1-2
7	0,120	msd-4	0,261	msd0-2
8	0,106	msd-5	0,257	je_vez1
9	0,099	zac_modrega0	0,233	msd-1-4
10	0,099	je_vez1	0,215	msd-1-5
11	0,082	zac_modrega1	0,214	msd-1-3
12	0,082	msd3	0,195	msd1-1
13	0,063	zac_rdecega1	0,189	zac_modrega1
14	0,061	je_vez0	0,187	msd-2-1
15	0,060	msd4	0,165	msd1-2

Tabela 1: Rezultat ocenjevanja atributov z mero ReliefF na množicah Šolar2. Na levi strani imamo za vsako MSD oznako po en atribut (Šolar2-11), na desni pa po 9 atributov (Šolar2-99).

5 Rezultati klasifikacije

Testirali smo več različnih klasifikacijskih algoritmov, ki so že bili uporabljeni v prejšnjih raziskavah. Naivni Bayesov klasifikator (NaiveBayes) je enostaven verjetnostni klasifikator, ki predpostavlja, da so atributi med seboj pogojno neodvisni. Odločitvena tabela (DecisionTable) izbere pomembne atribute, potem pa primere klasificira glede na njim najbolj podoben primer iz učne množice. RBF mreža (ang. Radial Basis Function Network, RBFnetwork) je nevronska mreža, ki klasificira glede na razdaljo primerov do središč Gaussovskih jeder, ki jih vsebujejo nevroni. Alternirajoče odločitveno drevo (ang. Alternating Decision Tree, ADTree) je posebne vrste odločitveno drevo, ki implementira boosting. AdaBoostM1 prav tako uporablja idejo boostinga, kar pomeni, da med učenjem napačno klasificiranim učnim primerom povečuje uteži, zato postanejo ti pomembnejši in se učenje osredotoča nanje. Dodatno smo uporabili dva nova algoritma, ki pri večini primerjalnih študij, npr. (Caruana in Niculescu-Mizil, 2006), dosegata visoko klasifikacijsko točnost, to sta metoda podpornih vektorjev (SMO) in metoda naključnih gozdov (RandomForest). Pri SMO lahko, za razliko od večine drugih algoritmov, vključimo vse razpoložljive atribute, saj metoda sama kombinira atribute, vendar je tudi zato časovno zahtevna. Metoda RandomForest vrača skupinsko napoved množice randomiziranih odločitvenih dreves in s tem zmanjša napako zaradi variance napovedi.

Testiranje smo najprej opravili na originalnem korpusu Šolar1. Rezultati testiranj so potrdili, da je korpus neprimeren za strojno učenje. Odvzem neinformativnih atributov je poslabšal rezultate in najboljše rezultate smo dobili z algoritmom odločitvena tabela, kjer ne gre toliko za učenje v smislu posploševanja, ampak si zapomnimo le natančno okolico besed, kjer stoji vejica. Po natančnejšem pregledu korpusa smo ugotovili, da je poleg veliko stavkov, ki ne vsebujejo vejice, vsebuje tudi veliko napačno postavljenih vejic. Sestavlja ga mnogo podobnih besedil. Prav tako je poln nepravilno tvorjenih stavkov z mnogimi pogovornimi in tujimi besedami.

Testiranje smo zato osredotočili na podatkovne množice izboljšanega korpusa Šolar2, kjer je primerov z vejicami skoraj trikrat več. V tabeli 2 vidimo rezultate za različne statistike: natančnost, priklic in F1 posebej za učne primere, ko vejica je in ko je ni, ter dodatno še klasifikacijska točnost in AUC, ki sporoča informacijo o pravilnosti rangiranja oz. napovedanih verjetnosti. Zaradi velike časovne zahtevnosti metode podpornih vektorjev in odločitvena tabela smo morali pri teh algoritmih uporabiti zmanjšane podatkovne množice. V tem primeru smo uporabili naključno izbrano desetino množice Šolar2, kar pomeni, da smo uporabili 72.892 primerov. Pri vseh ostalih algoritmih smo uporabili Šolar2 originalne velikosti, to je 728.927 primerov. Pri metodi podpornih vektorjev smo testirali linearno in kvadratno jedro. Pri vseh algoritmih smo uporabili privzete parametre in 10-kratno prečno preverjanje. Zaradi velikih množic so razlike med metodami skoraj povsod statistično značilne.

V tabeli 1 vidimo, da za primere, kadar ni vejice, najboljše rezultate pri meri F1, ki je harmonična sredina med natančnostjo in priklicem, dobimo z odločitveno tabelo (F1=0,977), z naključnimi gozdovi in obema inačicama boostinga (AdaBoostM1 in ADTree). Razlik v rezultatih med množicami, ki različno definirajo atribute za opis MSD kod, praktično ni.

Za primer, kadar je vejica, se z F1=0,71 najbolje odreže odločitvena tabela, sledi ji algoritem naključnih gozdov. Podobno velja za klasifikacijsko točnost, kjer ponovno izstopata odločitvena tabela (95,7 %) in metoda naključnih gozdov. Pri AUC najboljši rezultat doseže metoda naključnih gozdov (97,3 %) na podatkovni množici Šolar2-99, ostale metode precej zaostajajo.

6 Zaključki

Testirali smo več pristopov strojnega učenja za postavljanje vejic v slovenščini. Modificirali smo podatkovne množice, ki so bile uporabljene v preteklih raziskavah, in uporabili nekatere nove algoritme strojnega učenja. Originalne podatkovne množice smo spremenili tako, da smo jim najprej odstranili neinformativne atribute, ki so oteževali strojno učenje in povečevali časovno zahtevnost ter porabo virov. Nato smo dodali nove atribute, generirane na podlagi pravil, ki jih za postavljanje vejic uporablja orodje LanguageTool, in na podlagi pravopisnih pravil za slovenski jezik. Na dva načina smo preoblikovali atribute za zapis oblikoskladenjskih oznak. Pri prvem načinu smo za zapis atributa uporabili le besedno vrsto besed znotraj okoliškega okna. Pri drugem načinu smo obdržali celotno oblikoskladenjsko oznako tako, da smo jo razdelili na 9 novih atributov, kjer je vsak atribut nosil po en znak oznake. Poskusili smo izboljšati rezultate s prilagojenim podvzorčenjem, vendar za zdaj še ne dosežemo enake klasifikacijske točnosti.

Najbolj uspešne metode strojnega učenja so algoritmi naključna drevesa, alternirajoče odločitveno drevo ter odločitvena tabela. Predvidevamo, da dober rezultat odločitvene tabele kaže, da izboljšan in posodobljen korpus Šolar2 še vedno ni najbolj primeren za strojno učenje, saj je stopnja posplošitve pri tem algoritmu nizka. Rezul-

		Ni vejice			Je vejica			Točnost [%]	AUC
		Natančnost	Priklic	F1	Natančnost	Priklic	F1		
Šolar2-99	NaiveBayes	0,979	0,840	0,904	0,333	0,813	0,473	83,746	0,905
	RBFNetwork	0,927	0,985	0,955	0,591	0,217	0,318	91,634	0,868
	ADTree	0,951	0,995	0,972	0,898	0,482	0,638	94,866	0,925
	AdaBoostM1	0,951	0,970	0,961	0,620	0,489	0,547	92,735	0,882
	RandomForest	0,956	0,996	0,976	0,925	0,536	0,579	95,455	0,973
Šolar2-11	NaiveBayes	0,983	0,769	0,863	0,269	0,861	0,410	77,754	0,903
	RBFNetwork	0,922	0,992	0,956	0,654	0,147	0,240	91,655	0,870
	ADTree	0,946	0,996	0,971	0,916	0,426	0,581	94,502	0,913
	AdaBoostM1	0,952	0,970	0,961	0,621	0,499	0,553	92,775	0,867
	RandomForest	0,957	0,995	0,975	0,913	0,542	0,680	95,428	0,943
1/10	DecisionTable	0,958	0,995	0,976	0,918	0,653	0,698	95,589	0,939
Šolar2-99	SMO-linearno	0,955	0,996	0,975	0,928	0,529	0,674	95,366	0,762
	SMO-kvadratno	0,927	1,000	0,962	0,996	0,210	0,347	92,848	0,605
1/10	DecisionTable	0,959	0,995	0,977	0,920	0,577	0,709	95,720	0,940
Šolar2-11	SMO-linearno	0,954	0,997	0,975	0,942	0,520	0,670	95,367	0,758
	SMO-kvadratno	0,943	0,992	0,967	0,834	0,392	0,534	93,799	0,692

Tabela 2: Rezultati klasifikacije na korpusu Šolar2.

tati pri najboljših algoritmih so podobni pri obeh podatkovnih množicah, kjer smo različno generirali atribute o MSD oznakah. To zelo verjetno pomeni, da ob uporabi primernega korpusa zadostuje, če za opis oblikoskladenjskih oznak uporabimo le besedno vrsto (razen pri veznikih, kjer dodamo še informacijo ali gre za priredni ali podredni veznik).

Rezultati za priklic, natančnost in mero F1 na korpusu Šolar2 z izboljšanimi in za strojno učenje primernejšimi atributi so podobni rezultatom na korpusu Šolar1 z atributi, ki opisujejo konkretne besede, leme in celotne MSD oznake (Holozan, 2013). Menimo, da je s tem postavljen dober temelj za uporabo strojnega učenja tudi na morebitnih prihodnjih izboljšanih korpusih in v orodjih za avtomatsko postavljanje vejice.

Zanimivo bi bilo videti, kako se strojno učenje odreže, če podatkovnim množicam dodamo atribute, generirane glede na vsa pravila za postavljanje vejic v slovenščini, a je to zaradi včasih ne dovolj jasnih in nedvoumnih definicij težko implementirati. Vejica poleg tega nosi tudi semantično informacijo, zato brez semantičnih atributov ne moremo pričakovati bistvenih izboljšav za ta del problema. Zanimalo bi nas tudi, kako bi se obnesla implementacija tretjega načina definiranja MSD oznak, ki smo ga opisali v razdelku 3.2.3 in zahteva več računskih zmogljivosti.

Za kvalitetno procesiranje naravnega jezika so izjemno pomembne jezikovne tehnologije, kot so lematizator, označevalnik in skladenjski razčlenjevalnik. Bistven je tudi kvaliteten korpus, ki mora biti sestavljen iz dobrih in (večkrat) lektoriranih besedil. Pri teh komponentah so izboljšave mogoče in potrebne.

7 Zahvala

Zahvaljujemo se Petru Holozanu za nasvete, podatkovne množice in dostop do izboljšanega korpusa Šolar.

8 Literatura

Amebis d.o.o., 2015. *BesAna - slovnični pregledovalnik*. http://besana.amebis.si/ [Dostop: 06/06/2015].

Rich Caruana in Alexandru Niculescu-Mizil. 2006. An empirical comparison of supervised learning algorithms. V: *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, str. 161–168, New York, NY, USA. ACM.

Kaja Dobrovoljc, Simon Krek in Jan Rupnik. 2012. Skladenjski razčlenjevalnik za slovenščino. V: *Zbornik 8. konference Jezikovne tehnologije*, Institut Jožef Stefan, Ljubljana.

Tomaž Erjavec, Darja Fišer, Simon Krek in Nina Ledinek. 2010. The JOS linguistically tagged corpus of Slovene. V: *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*, Malta.

Miha Grčar, Simon Krek in Kaja Dobrovoljc. 2012. Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. V: *Zbornik 8. konference Jezikovne tehnologije*, Institut Jožef Stefan, Ljubljana.

Peter Holozan. 2012. Kako dobro programi popravljajo vejice v slovenščini. V: *Zbornik 8. konference Jezikovne tehnologije*, Institut Jožef Stefan, Ljubljana.

Peter Holozan. 2013. Uporaba strojnega učenja za postavljanje vejic v slovenščini. *Uporabna informatika*, 21(4).

Peter Holozan. 2015. Izboljšani korpus Šolar. Osebna komunikacija.

Jože Toporišič in sod. 2001. *Slovenski pravopis*. Slovenska akademija znanosti in umetnosti.

Anja Krajnc. 2015. Postavljanje vejic v slovenščini s pomočjo strojnega učenja. Univerza v Ljubljani, Fakulteta za računalništvo in informatiko. Diplomsko delo.

LanguageTool skupnost, 2015. *LanguageTool Style and Grammar Check*. http://www.languagetool.org [Dostop: 06/06/2015].

Marko Robnik-Šikonja in Igor Kononenko. 2003. Theoretical and empirical analysis of ReliefF and RReliefF. *Machine Learning Journal*, 53:23–69.

Ian H Witten in Eibe Frank. 2005. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Leksika na spletu (in kje jo iskati)

Mija Michelizza

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

Povzetek

V prispevku primerjamo korpus besedil blogov in Wikipedije s korpusom Nova beseda (ki spletnih besedil ne vsebuje) in ugotavljamo, kateri tipi novejše leksike se pojavljajo na spletu oz. konkretno v obeh omenjenih zbirkah besedil. Uporabniki zaradi nepoznavanja jezikovnih pripomočkov na spletu pa tudi zaradi neažurnosti slovarjev in besedilnih korpusov za razna jezikovna in jezikoslovna vprašanja pogosto uporabljajo spletne iskalnike, zato v prispevku prikažemo, kaj vpliva na razvrstitev zadetkov v spletnih iskalnikih, ter se vprašamo, kaj nam lahko pove podatek o številu zadetkov. Na konkretnih primerih si ogledamo, na kakšne težave lahko s tovrstnim iskanjem naletimo, in opozarjamo, na kaj naj bodo (predvsem jezikoslovci v svojih raziskavah) pozorni.

Lexicon on the Web (and Where to Search for it)

In the article, we compare a corpus of blog and Wikipedia texts to a corpus Nova beseda (which does not contain web texts). Moreover, we determine which types of modern lexicon appear on the Web, namely in both text collections mentioned above. Users often use web search engines to answer language and linguistic questions either due to their lack of familiarity with online language tools or unupdated dictionaries and corpora. We therefore show what affects the ranking of hits in web serach engines, and explore what the information on number of hits can indicate. We use actual examples to examine what kind of problems can be encountered in searching with web search engines, and point out what (especially linguists in their research) to pay attention to.

1 Uvod

Mnoge jezikoslovne raziskave danes nastanejo s pomočjo besedilnih korpusov in trend, ki se kaže, je, da se v referenčnih korpusih besedilom pisnega in govornega prenosnika vse bolj dodaja tudi besedila elektronskega prenosnika. Največji korpus pisnih besedil za slovenščino Gigafida vključuje 16 % internetnih besedil, in sicer besedila novičarskih portalov in predstavitvene strani podjetij ter državnih, pedagoških, raziskovalnih, kulturnih ipd. ustanov. Pred tem je imel manjši delež besedil z interneta (1,24 %) tudi referenčni korpus FidaPLUS, ki je sedaj skoraj v celoti vključen v Gigafido. Poskusno je bilo v besedilni korpus FIDA vključenih enajst dokumentov (ali 20.999 pojavnic) spletnih besedil že leta 1998 (Logar Berginc et al., 2012). Vključevanje spletnih in internetnih besedil v besedilne korpuse je običajna praksa tudi v tujini (prim. Michelizza, 2011; Logar Berginc in Ljubešić, 2013), mnogi so tudi korpusi spletnih besedil (npr. korpusi WaC (ang. Web as Corpus), za slovenščino slWaC). Za jezikoslovne raziskave (kot tudi za običajne poizvedbe o jeziku) pa so pogosto v uporabi tudi spletni iskalniki.

2 Namen članka

V članku želimo s primerjavo korpusa besedil blogov in Wikipedije s korpusom Nova beseda (ki spletnih besedil ne vsebuje) ugotoviti, kateri tipi novejše leksike se pojavljajo na spletu oz. konkretno v obeh omenjenih skupinah besedil. Ker uporabniki za jezikovna in jezikoslovna iskanja pogosto uporabljajo spletne iskalnike, v prispevku prikažemo, kaj vpliva na razvrstitev zadetkov in kako nezanesljiva je informacija o številu zadetkov. S pomočjo konkretnih primerov večinoma iz primerjave korpusov dobljene leksike opozorimo na težave, s katerimi se lahko srečamo. Čeprav je iskanje s spletnimi iskalniki zelo enostavno in hitro, pa so informacije, ki jih na ta način dobimo, pogosto zelo omejene, česar se moramo uporabniki, zlasti pa jezikoslovci v svojih raziskavah, zavedati.

3 Leksika na spletu

Z namenom preveriti leksikalne razlike med obema zbirkama besedil, smo analizirali manjši korpus blogov in Wikipedije (skupaj okrog 500.000 pojavnic) ter ga primerjali s besedilnim korpusom Nova beseda, ki spletnih besedil ne vsebuje. V raziskavi Michelizza (2011) je bilo s primerjavo besed omenjenih korpusov dobljenih 1058 novih leksemov (545 iz blogov in 513 iz Wikipedije), torej takih, ki se v Novi besedi niso pojavili. Te lekseme smo razdelili v devet skupin[1] ter nekatere izmed njih v nadaljevanju uporabili tudi za ponazoritev težav pri iskanju na spletnih iskalnikih.

(1) Nov leksem za novo predmetnost: V podkorpusu besedil z Wikipedije takih primerov ni bilo, v podkorpusu blogov pa zasledimo le manjši del tovrstnih leksemov (npr. *flipanje*, *skrolanje*, *tvit*, *bail-out*).

(2) Novoopomenjeni leksem: Novoopomenjenih leksemov s primerjavo pojavnic dveh korpusov ni

[1] Delitev temelji na skupinah novejše slovenske leksike z vidika njenega generiranja, kot jih predstavlja Gložančev (2009), na koncu pa sta dodani še dve skupini.

mogoče iskati – vseeno je bilo med analizo najdenih nekaj primerov (*izgubljenček* 'zguba', *vsipan* 'pijan' ipd.).

(3) Nov leksem kot slovenska ustreznica za novejšo prevzeto: V analiziranem gradivu najdemo izraz *čivkač* v pomenu družabnega omrežja Twitter.

(4) Nov leksem kot posledica determinologizacije: Tovrstni novi izrazi so bili najdeni samo pri analizi besedil Wikipedije (npr. *melanopsin*, *dimetrodon*, *pleziozaver* ipd.).

(5) Novotvorjenke:

a) z lastnoimensko podstavo (npr. *janšegrad*, *murkovalec*, *barbivic*, *zdrnovškati se*, *nebondovsko*, *barbibutelj* ipd.),

b) z občnoimensko podstavo, ki je zaznamovana glede na kvalifikator v SSKJ (npr. *fejstnost*, *aufbiksati*, *šlatalec* ipd.),

c) novotvorjenke, ki izkazujejo zaznamovanost zaradi nenavadnih, manj običajnih obrazil (npr. *floskularjenje*, *premožnjakar* ipd.),

č) novotvorjenke, ki so zaznamovane tako zaradi podstave kot zaradi obrazila (npr. *futrač*, *glupača*, *frajerišenje*, *pifling* ipd.),

d) pomanjševalnice (npr. *blogec*, *zdrsek*, *lišajček* ipd.),

e) novotvorjenke, ki so poenobesedeni leksemi besednih zvez (iz polnopomenskih besed so nastale zloženke, iz predložnih zvez pa izpeljanke iz predložne zveze), nekatere izmed njih imajo frazeološki pomen (npr. *navrstnež*, *prvožogaški*, *pralnomožganski*, *polnoriten*, *vžepljivost* ipd.).

(6) (Fonetično)-oblikoslovna prevzetost tujih leksemov: Mnogi novi leksemi te skupine so posledica želje po stilnem učinkovanju v besedilih, nepoznavanja citatnega zapisa ali pa gre za primere, ko izraz v slovenščini (še) ne obstaja oz. ni vsesplošno uveljavljen (npr. *autlet*, *goodi*, *rumor*, *inboks* ipd.).

(7) Poobčnobesedenje lastnoimenskega izhodišča: Do poobčnobesedenja lastnoimenskih izhodišč prihaja pri imenih znamk in industrijskih izdelkov (npr. *skajpanje*, *uggice*, *pofotošopati*, *salomonke*, *martenske*) ali pa gre za poobčnobesedenje osebnih imen (npr. *potrč*, *golubić*, *anderlić* ipd.).

(8) Lastnoimenske novosti: V besedilih Wikipedije je delež novosti s področja lastnoimenskih poimenovanj precejšen. Navajamo nekaj primerov lastnoimenskih poimenovanj za zemljepisna imena: *Bernissart* (valonska občina v Belgiji), *Vajots Dzor* (armenska provinca); znane osebnosti: *Bellincione* (Dantejev oče), *Biondetti* (dirkač formule 1); viskije: *Balvenie*, *Bunnahabhain*, *Glengoyne* ipd. Hkrati pa je treba opozoriti, da so zaradi prevajanja geselskih člankov Wikipedije iz drugih jezikov ta imena pogosto zapisana citatno ali z mednarodno prečrkovalno različico, npr. za *Djalalabad* (ali *Jalalabad*) v slovenščini uporabljamo tudi zapis *Džalalabad*. Posebej je treba omeniti še skupino leksemov, ki jih uvrščamo med besedne igre in grafološke inovacije (npr. *pešhonda* (*...čaka me še cca. 20min. "pešhonde" do doma :()*, *Stovodiček* (*Tale arhitekt Stovodiček (Hundertwasser je Čeh drugače) mi je zelo všeč zaradi njegovega rekla, da če je lepo krivo, je tudi lepo.*) ipd.).

Analiza korpusa izbranih besedil z blogov in Wikipedije v primerjavi s korpusom Nova beseda pokaže, da se največ razlik na področju leksike (izvzemši stalne besedne zveze in frazeologijo) kaže pri blogih v obliki novotvorjenk, ki so pogosto priložnostnice,[2] pa tudi kot poobčnobesedenja lastnoimenskih izhodišč, s čimer se izkazuje želja po inovativnosti in ekspresivnosti ter stilnem učinkovanju v izražanju. Na Wikipediji je že zaradi same narave besedil izpostavljen vidik terminološkosti oz. determinologizacije ter lastnoimenskih novosti. Večine teh besed v obstoječih slovarjih ne bi našli (izjema je beseda *tvit*, ki jo najdemo v SNB in SSKJ[2]), če pa bi nas zanimalo kaj več o kaki od teh besed, bi si verjetno marsikdo pomagal tudi s spletnimi iskalniki.

4 Jezikovno in jezikoslovno iskanje s spletnimi iskalniki

Slovarji (vsaj za jezike z manjšim številom govorcev) niso vedno najbolj ažurni z objavo nove leksike, zato moramo uporabniki večkrat uporabiti druge pripomočke. Poleg besedilnih korpusov (ki pa imajo podobno težavo kot slovarji, saj je bila npr. zadnja posodobitev referenčnega korpusa za slovenščino leta 2012, zadnja besedila v njem pa so iz leta 2011) za jezikovne in jezikoslovne težave, analize in raziskave tako običajni uporabniki kot tudi jezikoslovci pogosto iščejo s pomočjo spletnih iskalnikov (Michelizza, 2011). Tu pa je nujno opozoriti na posebno previdnost pri interpretaciji rezultatov.

Interpretacija števila zadetkov in razvrstitev

Spletni iskalnik deluje na podlagi t. i. dvofaznega algoritma: ko vpišemo iskani izraz,

[2] V tej raziskavi dobljeno leksiko smo iskali še s pomočjo spletnih iskalnikov in mnoge izmed novotvorjenk se v spletnem iskalniku Google pojavijo samo enkrat (gre za ista besedila, ki so bila zajeta v korpus besedil blogov in Wikipedije).

najprej pajki poiščejo vse spletne strani, ki ta izraz vsebujejo, v drugi fazi pa program te zadetke razvrsti (Oblak in Petrič, 2005). Razvrstitev spletnih iskalnikov ni naključna in je odvisna od različnih dejavnikov (Gatto, 2009; Lana, 2004): (1) priljubljenost strani (merjena s številom drugih strani, ki se povezujejo na to stran); (2) pojavnost iskane besede ali besedne zveze na strani (kolikokrat se pojavlja, kje se pojavlja: če je beseda v naslovu, podnaslovu, v oznakah (ang. *Tags*) ter v hiperpovezavah, bo imela stran višjo razvrstitev); (3) geografski izvor poizvedovanja (višje bodo razvrščene strani, ki so geografsko gledano bližje uporabniku); (4) komercialni vidik (sponzorirane spletne strani so vedno na vrhu oz. na drugače izpostavljenem mestu iskalnika); (5) omejen čas iskanja (par stotink ali tisočink sekunde za obdelavo; ko iskalniki dosežejo omejeni čas, se poizvedba zaključi in rezultati so poslani uporabniku). Gre zgolj za nekatere izmed dejavnikov, ki vplivajo na razvrstitev, saj so ti algoritmi bolj kot ne skrivnost. Popolno razkritje algoritmov pri enem iskalniku bi lahko povzročilo vzpon drugega. Pri Googlu je ena pomembnejših tehnologij za razvrščanje zadetkov PageRank,[3] ki ne prešteva neposrednih povezav, ampak povezavo s strani A na stran B interpretira kot glas strani A za stran B. Nato oceni pomembnost strani glede na število prejetih glasov.[4] Zaradi omejenega časa iskanja in zaradi vse večjega števila podatkov na spletu, je tudi vse več strani, ki jih spletni iskalniki ne prikažejo. Starejše, neažurirane strani izpadejo, vse pomembnejša za razvrstitev v spletnih iskalnikih je tudi prilagoditev spletnih strani za mobilne naprave. Kot bomo videli v nadaljevanju, lahko v spletnem iskalniku Google npr. leta 2015 za identično iskanje najdemo manj pojavitev kot leta 2011, zato se je pri analizi spletnih besedil nujno zavedati relativnosti frekvenčnih podatkov na spletnih iskalnikih.

Omenili smo že, da je razvrščanje zadetkov v spletnih iskalnikih pogosto sponzorirano. Pojavili so se ponudniki t. i. optimizacije spletnih strani, ki naročniku pomagajo, da se uvrsti čim višje pri razvrščanju zadetkov v spletnih iskalnikih. Znano pa je, da uporabniki pogosto po liniji najmanjšega odpora pogledajo le prvih nekaj zadetkov v iskalniku. Pri spletnem korpusu projekta CUCWeb npr. opozarjajo na tovrstni šum, ki ga je mogoče opaziti v spletnih iskalnikih, saj so nekateri uporabniki spleta razvili posebne programe, ki pretentajo iskalnike, da so nekatere spletne strani višje razvrščene pri rezultatih iskanja, kot si dejansko zaslužijo. Predvideva se, da 8 % vsega, kar najdejo spletni iskalniki, spada med tovrstni šum. Uporabniki naredijo osnovno spletno stran in zraven še veliko drugih strani, s katerih naredijo povezavo na osnovno stran, in si na ta način povišajo razvrstitev na iskalniku (Fetterly et al., 2004; v Boleda et al., 2006).[5]

24. oktobra 2009 je takratni predsednik vlade RS Borut Pahor v enem izmed javnih nastopov uporabil besedo *krucefiks* in logična posledica je bila, da se je ta oblika na spletu razširila, čeprav so novičarski spletni portali navajali v SSKJ in SP 2001 uslovarjeno različico *krucifiks*. Še v začetku leta 2011 je iskalnik *Google* z omejenim iskanjem na slovenščino našel približno 4.500 zadetkov za *krucifiks*, medtem ko se je *krucefiks* pojavil kar 34.700-krat. Iskalnik pa je ob rezultatih vseeno priporočal zapis *krucifiks*.

Slika 1: Iskanje besede *krucefiks* v spletnem iskalniku Google leta 2011.

Stanje v začetku leta 2015 je bilo diametralno nasprotno, saj smo v Googlu (ponovno z omejitvijo na slovenščino) našli 5.280 pojavitev *krucefiksa*, *krucifiks* pa je imel kar 369.000 pojavitev. Konec leta 2015 z enakim iskanjem spet dobimo drugačno sliko. *Krucifiks* ima 3.540 zadetkov, *krucefiks* pa 6.670, pri čemer nas pri iskanju *krucifiks* vpraša, če smo morda mislili *krucefiks*. Google pogosto menja iskalne algoritme in kot kaže, je tudi v vmesnem času med temi iskanji prišlo do večjih sprememb, za jezikoslovca, ki bi želel ugotoviti prevladujočo

[3] Poimenovana je po ustanovitelju Googla in izumitelju algoritmov Larryju Pageu.

[4] Pred nastankom Googla so iskalniki delovali s pomočjo preprostih algoritmov, ki so temeljili na ključnih besedah – to pa je bilo precej enostavno zlorabiti. Še posebej pornografska industrija je začela izkoriščati to pomanjkljivost iskalnikov. Pogoste iskane besede so skrili po vsej svoji strani, npr. v drobnem tisku na beli podlagi. Leta 1998 so bili rezultati iskanja za poizvedbo avtomobil na takrat priljubljenem spletnem iskalniku Lycos večinoma pornografske spletne strani (Battelle, 2010).

[5] Meja med tovrstnim legalnim in nelegalnim početjem je pogosto nejasna. V okviru optimizacije spletnih strani sta se uveljavili poimenovanji črni klobuki in beli klobuki (ang. *Black Hats* in *White Hats*). Gre za izraza, ki izhajata iz žargona Divjega zahoda oz. kavbojskih filmov, kjer so bele klobuke nosili dobri, črne pa slabi kavboji (Battelle, 2010). V sklop nedovoljenih postopkov optimizacije npr. štejemo, (1) ko je v kodo HTML dodano besedilo, ki ga zaznajo iskalniki, na zaslonu pa ni vidno; (2) ko obstajata različni strani za iskalnike in za uporabnike in (3) ko se za optimizacijo spletnih strani uporablja program, ki sam proizvaja vsebino, zanimivo za iskalnike (<http://www.seo-blog.com/hats.php>, 22. november 2010).

obliko v rabi na spletu, pa bi lahko bila informacija zavajujoča.[6]

	krucifiks	*krucefiks*
začetek leta 2011	4.500	34.700
začetek leta 2015	369.000	5.280
konec leta 2015	3.540	6.670

Tabela 1: Pojavitve za *krucifiks* in *krucefiks* v iskalniku Google.

V času nogometnega svetovnega prvenstva v Južnoafriški republiki leta 2010 se je veliko pisalo o navijaškem glasbilu, t. i. *vuvuzeli*.[7] Za primer smo vzeli iskanje v spletnem iskalniku Google (z omejitvijo na slovenščino), kjer smo izbrali časovno obdobje po meri. Število pojavitev besede v času svetovnega prvenstva, torej med 11. junijem in 11. julijem 2010, in v istem obdobju leto poprej (med 11. junijem in 11. julijem 2009) smo primerjali konec leta 2010 in konec leta 2015 (gl. tabelo 2). Tako pri iskanju leta 2010, kot tudi leta 2015 je frekvenca višja v mesecu, ko je potekalo svetovno prvenstvo, vendar glede na izrazito manjše število zadetkov v obeh izbranih časovnih obdobjih, sklepamo, da iskalnik v letu 2015 med prikazom zadetkov upošteva le še redke spletne strani iz leta 2010.

vuvuzela	11. 6.–11. 7. 2010	11. 6.–11. 7. 2009
konec leta 2010	21.300	301
konec leta 2015	45	3

Tabela 2: Pojavitve za *vuvuzelo* v iskalniku Google.

Nelematiziranost spletnih iskalnikov

Težave z nelematiziranostjo spletnih iskalnikov omenja že Kilgarriff (2006), konkretno pa jih pokažemo z iskanjem primerov novih leksemov, ki smo jih obravnavali v 3. poglavju. Če v iskalnik vpišemo[8] osnovno obliko nekaterih besed, ki smo jih našli v korpusu blogov in Wikipedije, spletni iskalnik Google ne najde zadetkov. Taki primeri so: *murkovalec* (v korpusu se pojavi v obliki *murkovalca*), *barbivic* (*barbivice*), *zdrnovškati se* (*se zdrnovška*), *barbibutelj* (*barbibutlja*), *premožnjakar* (*premožnjakarjev*), *vžepljivost* (*vžepljivostjo*) in *pofotošopati* (*bi pofotošopal*). Čeprav je iskalnik Google splošno razširjen in med uporabniki najbolj priljubljen tudi za jezikovne poizvedbe (Michelizza, 2011), smo preverili zgornje primere še v iskalniku Najdi.si, ki prav tako ne najde zadetkov za osnovne oblike zgoraj obravnavanih iskanih besed.

Zapis z veliko in malo začetnico

Iskalniki tudi ne ločujejo zadetkov, zapisanih z veliko in malo začetnico. V preteklosti je tako razlikovanje omogočal iskalnik AltaVista, ki se je kasneje združil z Yahoojem in zato izgubil prenekatere prednosti, ki jih je prej nudil jezikoslovcem: poleg že omenjenega razlikovanja male in velike začetnice še iskanje posebnih znakov, vseboval je tudi določeno jezikoslovno znanje, kot je npr. enačenje nemškega ß in ss. Kasneje se je iskanje združilo z oglaševanjem, kar je nedvomno vplivalo tudi na mnoge odločitve o možnostih iskanja (Fletcher, 2007). Če želimo v Googlu iskati prevzeta tuja leksema *goodi* in *rumor*, se med zadetki pojavi tako zapis z veliko kot z malo začetnico, ki pa jih uporabniki ne moremo ločiti in moramo zato pregledati vse zadetke oz. toliko, kolikor se nam zdi potrebno. Podoben primer so zgledi poobčnobesedenja osebnih imen (*potrč*, *golubić*, *anderlić*), ki jih je v spletnih iskalnikih, ki nimajo možnosti iskanja z ločevanjem male in velike začetnice, praktično nemogoče iskati.

Zapis skupaj, narazen in z vezajem

V podobno težavnem položaju se znajdemo, če želimo izvedeti, kako se piše določena beseda – skupaj, narazen ali z vezajem. V rezultatih iskanja dobimo zadetke, ki zanemarjajo presledek (npr. pri iskanju *peshonda* dobimo zadetke *peshonda*, *peš honda* in *peš-honda*). Tiste, ki so pisani skupaj, sicer lahko najdemo s pomočjo t. i. Boolovega operatorja. V iskalnik vpišemo poizvedbo »*peshonda -peš -honda*«, kar pomeni, da bo iskalnik izločil vse zadetke, ki vsebujejo *peš* in *honda*, torej bo poiskal samo tiste, zapisane skupaj.

Tujejezični elementi

Čeprav lahko v iskalniku Google nastavimo iskanje po slovenskih spletnih straneh in po spletnih straneh v slovenščini, pa so pri iskanju tujejezičnih leksemov na tak način precejšnje težave. Pri primeru *rumor* najde Google z omejitvijo na slovenščino 20.900 zadetkov, ki pa že na prvi strani niso vsi v slovenščini. Podobno je pri primeru *bail-out* (22.600 zadetkov), kjer se pojavi še težava zapisa skupaj, narazen in z vezajem. Če v iskalniku Google iščemo besedo *tvit*, sicer najde kar 30.900 rezultatov, vendar nas hkrati vpraša »*Ste morda mislili tweet*«, kar bi lahko kakega uporabnika usmerilo v tujejezični zapis. Podoben predlog dobimo, ko v iskalno okence vpišemo *martenske* (1.330 zadetkov). Google nam predlaga *martinske*

[6] Katera od obeh prevladuje, težko rečemo, v Gigafidi je npr. razmerje med njima *krucifiks* 129 : *krucefiks* 70, za leto 2010 *krucifiks* 49 : *krucefiks* 48.

[7] Avgusta 2010 se je beseda *vuvuzela* prvič pojavila v oxfordskem slovarju (angl. *Oxford Dictionary of English*), v slovenščini je bila prvič vključena v SNB, ki je izšel leta 2012, čez dve leti pa je bila dodana tudi v SSKJ[2].

[8] Vsa nadaljnja iskanja so bila izvedena konec leta 2015.

(10.500 zadetkov). Čeprav je poimenovanje nastalo iz znamke Dr. Martens, bi lahko iz rezultatov sklepali, da se je v slovenščini uveljavilo poimenovanje *martinske* in ne *martenske*. Vendar pa podrobnejši pregled pokaže, da so med zadetki pri iskanju *martinske* tudi spletne strani, ki omenjajo slovaško smučišče Martinské Hole, martinske jedi ('jedi, ki jih jemo za martinovo'), martinske peči,[9] Martinsko jamo ipd. Kateri izmed izrazov (*martinske* ali *martenske*) se je v slovenščini (bolj) uveljavil, bi bilo treba natančneje preučiti in verjetno bi za to potrebovali drugo orodje (uravnoteženi korpus). Na primeru iskanja *martinske*, smo videli, da je iskalnik zanemaril tudi zahtevo po iskanju brez diakritičnega znamenja in med zadetke uvrstil *Martinské Hole*. Konec leta 2010 in v začetku leta 2011 so bile pri Googlu podobne težave z iskanjem besed s šumevci oz. brez njih v slovenščini (npr. pri iskanju besednih oblik *mizi* in *miži*). Konec leta 2015 pa iskalnik v omenjenem primeru že razmeroma dobro ločuje med zadetki, ki so zapisani s šumevcem oz. brez njega.

Seveda pa so nam spletni iskalniki lahko v veliko pomoč tako pri iskanju osnovnih jezikovnih podatkov, kot tudi pri zahtevnejšem jezikoslovnem raziskovanju. Med zadetki (navadno precej visoko uvrščeni) so pogosto različni spletni slovarji, ki lahko uporabniku pomagajo posredno (prek iskalnika) pri razreševanju jezikovnih težav (Lorentzen in Theilgaard, 2012). Za primer vzemimo v skupino novoopomenjenih izrazov uvrščeni leksem *vsipan*, ki ga lahko v enakem pomenu kot v podkorpusu blogov najdemo tudi med množico izrazov za pomen 'pijanost' na Prostem slovarju žive slovenščine Razvezani jezik (<http://razvezanijezik.org/?page=pijanost>, 24. oktober 2015), do katerega nas usmeri prav spletni iskalnik. Iskalnik nas lahko privede tudi do informacije o zapisu v lastnoimenske novosti uvrščenega *Djalalabada*. Na drugi strani zadetkov lahko najdemo zapis *Džalalabad*, ki je v Slovarju slovenskih eksonimov na portalu Termania.

Navkljub enostavnosti uporabe in vsem drugim prednostim, ki jih spletni iskalniki omogočajo, pa je nujno, da se uporabniki, predvsem pa jezikoslovci pri svojem delu omejitev in možnosti teh pripomočkov zavedamo in jih upoštevamo pri interpretaciji informacij. Pri spremljanju novejše leksike na spletu bo oblikovanje (in sprotno dopolnjevanje) specializiranega korpusa spletnih besedil, ki se nam obeta v sklopu projekta Janes, več kot dobrodošlo, za potrebe uslovarjanja novih, pogosto čez noč uveljavljenih gesel (tudi na račun spletne rabe jezika) pa je lahko uporaben Sprotni slovar slovenskega jezika, ki nastaja v sklopu portala Fran <www.fran.si>. S pomočjo neuspešnih poizvedb po slovarjih na spletni strani Inštituta za slovenski jezik Frana Ramovša ZRC SAZU <http://bos.zrc-sazu.si>, vključuje besedje, ki v drugih splošnih slovarjih še ni vključeno. Gre za metodo »log-files«.[10]

5 Zaključek

S primerjavo besedilnega korpusa Nova beseda in korpusov besedil blogov in Wikipedije smo pokazali, da na ta način pridobljeno besedje s spleta predstavlja predvsem novotvorjenke, ki so pogosto priložnostnice in poobčnobesedenja lastnoimenskih izhodišč. Pogoste so še lastnoimenske novosti in determinologizirani leksemi. Za iskanje tudi takih leksemov pogosto uporabljamo spletne iskalnike, ki pa za jezikoslovca prinašajo zelo omejene, včasih tudi zavajajoče informacije, česar se moramo zavedati. Prav zaradi tega je nujno oblikovanje večjega korpusa spletnega jezika, novosti pa je treba v slovarski obliki tudi sproti beležiti.

6 Literatura

Battelle, John. 2010. Iskanje. Kako so Google in njegovi tekmeci na novo napisali pravila poslovanja. Ljubljana: Pasadena.

Boleda, Gemma et al. 2006. CUCWeb: a catalan corpus built from the Web. Proceedings of the 2nd International Workshop on Web as Corpus. Trento. 19–26.

Fletcher, William H. 2007. Concordancing the web: promise and problems, tools and techniques. Corpus Linguistics and the Web (ur. M. Hundt et al.). Amsterdam: Rodopi. 25–45.

Gatto, Maristella. 2009. From Body to Web. An Introduction to the Web as Corpus. Bari: Editori Laterza.

Gložančev, Alenka, Jakopin, Primož, Michelizza, Mija, Uršič, Lučka, Žele, Andreja. 2009. Novejša slovenska leksika (v povezavi s spletnimi jezikovnimi viri). Ljubljana: Založba ZRC, ZRC SAZU.

Kilgarriff, Adam. 2006. Googleology is bad science. Computational linguistics, Volume 1, number 1. 147–151.

Lana, Maurizio. 2004. Il testo nel computer. Dal web all' analisi dei testi. Torino: Bollati Boringhieri.

Lorentzen, Henrik, Theilgaard, Liisa. 2012. Online dictionaries – how do users find them and what do they do once they have? Proceedings of the 15th EURALEX International Congress. 7-11

[9] Pridevnik *martinski* najdemo že v SSKJ (*martínski* -a -o prid. (î) metal., v zvezah: martinski postopek *postopek za pridobivanje jekla v martinovki*; martinska peč *martinovka*; martinsko jeklo *jeklo, ki se pridobiva v martinovki*).

[10] Tema je bila 25. maja 2015 obravnavana na Lingvističnem krožku Filozofske fakultete v Ljubljani. Predstavili so jo Primož Jakopin, Helena Dobrovoljc in Aleksandra Bizjak Končar.

August 2012. Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo. 654–660.

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

Logar Berginc, Nataša, Ljubešić, Nikola. 2013. Gigafida in slWaC: Tematska primerjava. Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave, št. 1. 78–110.

Michelizza, Mija. 2011. Vloga in pomen spletnih besedil v slovenščini. Doktorska disertacija. Ljubljana: Univerza v Ljubljani, Filozofska fakulteta.

Oblak, Tanja, Petrič, Gregor. 2005. Splet kot medij in mediji na spletu. Ljubljana: Univerza v Ljubljani, Fakulteta za družbene vede.

Primerjava rabe ekspresivnih ločil v tvitih slovenskih uporabnikov in uporabnic

Eneja Osrajnik,* Darja Fišer,† Damjan Popič†

*Maribor
eneja.osrajnik@gmail.com

†Filozofska fakulteta Univerze v Ljubljani
darja.fiser@ff.uni-lj.si, damjan.popic@ff.uni-lj.si

Povzetek

V pričujočem prispevku obravnavamo rabo ekspresivnih ločil, ki jih razdelimo na pravopisna ločila in emotikone, v tvitih slovenskih uporabnikov in uporabnic. Na podlagi podatkov, zbranih v podkorpusu tvitov korpusa Janes (Fišer et al., 2014), analiziramo razlike v rabi ekspresivnih ločil v tvitih moških in žensk ter njihov sentiment. Kot osnova za določanje sentimenta nam služijo avtomatsko pripisane oznake sentimenta tvitov v korpusu, upoštevamo pa tudi kontekst posameznih tvitov. Raziskava je pokazala, da uporabniki obeh spolov uporabljajo več emotikonov kot pravopisnih ločil, ženske emotikone uporabljajo pogosteje kot moški, ti pa pogosteje uporabljajo pravopisna ločila. Raziskava je pokazala tudi, da je glede na kontekst prevladujoč sentiment tako emotikonov kot pravopisnih ločil pri obeh spolih negativen.

Comparing the use of expressive punctuation marks in Slovenian tweets written by male and female users

The following article deals with the use of expressive punctuation marks, which we divide into ortographic punctuation marks and emoticons, in the tweets of Slovenian male and female users. Based on the data gathered in the corpus Janes (Fišer et al., 2014), specifically in its subcorpus of tweets, we analyse the differences in the usage of expressive punctuation marks and their sentiment in the tweets of Slovenian male and female users. The automatically appointed sentiment designations and the context of each analysed tweet served as the basis for our analysis. We concluded that both genders use emoticons more frequently than ortographic punctuation marks and that female users use emoticons more frequently than male users. On the other hand, male users use orthographic punctuation more frequently than female users. We also established that the sentiment of the overwhelming majority of tweets – both male and female – is negative.

1 Uvod

Ena od značilnosti človeške komunikacije je izražanje čustev, načini njihovega izražanja pa se razlikujejo glede na medij, preko katerega sporazumevanje poteka. »V govorjeni komunikaciji čustva izražamo s pomočjo telesne govorice, obrazne mimike, tona glasu, glasnosti, tempa, poudarkov in s pomočjo pavz, v standardni pisni komunikaciji pa jih verbaliziramo« (Crystal, 2001). Upoštevati je treba, da je v tradicionalnem smislu narava javnega pisnega diskurza taka, da se čustva pretežno ne izražajo – vsaj ne odkrito – saj naj bi bil pisni diskurz bil objektiven in nezaznamovan. Kadar pa pisci svoje počutje oz. čustva kljub temu želijo izraziti, v diskurz zavestno vnašajo zaznamovane ekspresivne prvine. Komunikacija, ki poteka preko družbenih omrežij, predstavlja svojevrstno kategorijo pisnega diskurza, saj uporabniki svoja čustva sicer lahko verbalizirajo, vendar jim družbena omrežja omogočajo tudi neverbalno izražanje »temeljnih družbenih drž in osebnih občutij, čustev in čustvenih stanj, odnosov do drugih« (Ule in Kline, 1996: 43). Izražajo jih lahko z emotikoni, torej kombinacijami »različnih znakov, ki jih najdemo na tipkovnici, namenjeni pa so izražanju čustev in nadomeščanju izrazov obraza« (Crystal, 2001: 36), z ločili in s ponavljanjem ločil, ponavljanjem črk v besedah, velikimi tiskanimi črkami, gifi,[1] heštegi[2] ipd.

V pričujoči raziskavi smo se omejili na ločila, s katerimi uporabniki izražajo svoja čustva in jih v nadaljevanju imenujemo ekspresivna ločila. Neekspresivna ločila (vejica, pika, vprašaj, klicaj, dvopičje, podpičje, narekovaj, vezaj, pomišljaj ipd. v primerih, ko gre za običajno skladenjsko ali neskladenjsko rabo, ki ne izraža uporabnikovih čustev), smo iz analize izločili. Ekspresivna ločila smo razdelili na emotikone in »pravopisna ločila«. Med pravopisna ločila v prispevku prištevamo tri pike oz. zamolk, vprašaj in klicaj, s katerimi uporabnik izraža svoja čustva, pa tudi kopičenje ene vrste ločila za izražanje večje intenzivnosti čustev. Prvotno smo med pravopisna ločila uvrstili tudi piko (primer ekspresivne rabe pike je *Ha. Ha. Ha. Zelo smešno.*),[3] vendar v korpusu nismo našli primerov ekspresivne rabe pike (treba je upoštevati, da je ekspresivno rabo ločil težko analizirati iz konkordanc), zato je nismo vključili v raziskavo.

V analizo smo vključili tudi emotikone, torej ustaljene kombinacije ločil z ekspresivno funkcijo, ki izražajo zadržek do dobesednega pomena, do izbrane formulacije ipd. Kot ugotavlja Praprotnik (2003: 14), so »tovrstne simbolne reakcije pravzaprav nujne, saj izjave same na sebi ne pomenijo dovolj in ne kažejo čustvenega stanja govorca/govorke, tako da bi brez fizične ekspresije, ki razkodira specifični kontekst izjave, obstajala velika možnost napačne interpretacije.« Emotikone obravnava tudi Crystal (2008), ki preučuje značilnosti sms-sporočil, (te lahko prenesemo tudi na jezik tvitov), in sicer identificira šest glavnih značilnosti sms-sporočil – uporabo piktogramov, logogramov, inicializmov, izpust

[1] Prevzeto iz angleškega izraza *gif*, gre pa za vrsto animiranih slik.

[2] Slovenjen angleški izraz *hashtag*. Gre za znak #, ki stoji pred ključnimi pojmi sporočila v tvitu, s pomočjo katerih lahko iščemo tvite s podobno vsebino, na primer:
Papa cokolado in je najsrecnejsi na svetu :-) #lumpi#funny#dog#chocolate.

[3] Primer je za večjo nazornost dodala avtorica, saj v korpusu primerov ekspresivne pike nismo našli.

črk in nestandardno črkovanje. Trdi, da je »ena najvidnejših značilnosti pisanja sms-sporočil uporaba posebnega pravopisa – uporaba posameznih črk, številk in tipografskih simbolov, ki predstavljajo besede ali celo /.../ zvoke, ki jih asociiramo z določenimi dejanji. /.../ Kadar grafične simbole uporabljamo v te namene, gre za *logograme* ali *logografe*. Logograme lahko v sms-sporočilih uporabljamo posamično ali jih medsebojno kombiniramo: *s5* spet ali *ju3* jutri.[4]

Pri logogramih je najpomembnejša njihova izgovorjava, ne pa vizualna podoba. Prav v tem se najbolj razlikujejo od grafičnih znakov, imenovanih *emotikoni*[5]/.../, kjer pomen v celoti temelji na obliki simbola (kadar jih prebiramo s strani, z glavo, nagnjeno v levo stran): *:-)* smeško, *;-)* pomežik, *:-@* krik (ali držimo glavo pokonci, na primer v japonskem ali nekaterih drugih vzhodnoazijskih sistemih): *(*o*)* presenečenje, *(^_^)* ljubko. Kadar vizualne podobe ali slike predstavljajo predmete ali koncepte, jih imenujemo *piktogrami*, med katere sodijo tudi emotikoni.«[6] Emotikoni kot dogovorjene kombinacije grafičnih simbolov torej izražajo koncept posameznih čustev. Po drugi strani ekspresivna pravopisna ločila sicer omogočajo izražanje čustev, vendar se uporabljajo posamično, če jih uporabljamo več skupaj (v zaporedni ponovitvi ali kombinaciji različnih ločil), pa gre za govorčevo odločitev, in ne za družbeno dogovorjeno kombinacijo znakov, ki bi predstavljala določeno čustvo.

V raziskavi opazujemo ekspresivno rabo ločil v jeziku moških in žensk, ki je tudi sicer pogosto obravnavana tema v sociolingvističnih raziskavah. Tannen (1990) na primer ugotavlja, da ženski stil komunikacije izraža več podpore in bolje vzpostavlja odnose, medtem ko je za moško komunikacijo značilno podajanje poročil in visoka stopnja informativnosti, Schler et al. (2006) , ki so analizirali 300-milijonski korpus blogov, pa so ugotovili, da blogerke veliko pogosteje pišejo o svojem zasebnem življenju in uporabljajo bolj osebnen stil pisanja kot blogerji. Razlike v jeziku moških in žensk je na podlagi korpusa norveških sms-sporočil preučeval tudi Ling (2015) in ugotovil, da še posebej mlajše ženske v sms-sporočilih uporabljajo širši register, pišejo daljša in kompleksnejša sporočila kot moški ter pogosteje uporabljajo ekspresivne oz. čustvene elemente. Ekspresivno rabljena ločila v povezavi s spolom pogosto obravnavajo kot »označevalce razburjenosti« (Waseleski, 2006). Kivran-Swaine et al. (2013) pa raziskujejo povezavo med jezikom, spolom in socialnimi razmerji, ki se kažejo v komunikacijskih vzorcih družbenih medijev, in ugotavljajo, da »ženske uporabljajo več jakostnih prislovov, zaimkov in emotikonov, še posebej v komunikaciji z drugimi ženskami«. Med ločili v moškem in ženskem jeziku pa najbolj izstopa klicaj, saj se ugotavlja, da ga mnogo pogosteje uporabljajo ženske kot moški (Colley in Todd, 2002; Rubin in Greene, 1992; Scates, 1981)«.

Analiza ekspresivne rabe ločil v slovenskih uporabniških spletnih vsebinah še ni bila opravljena, prav tako ne poznamo primerjav ekspresivne rabe ločil pri moških in ženskah nasploh, zato se jim posvečamo v pričujočem prispevku, v katerem se omejimo na družbeno omrežje Twitter. S pomočjo podatkov, zbranih v podkorpusu tvitov korpusa Janes (Fišer et al., 2014), ugotavljamo, kakšne so razlike v rabi ekspresivnih ločil v tvitih moških in žensk ter kakšen je prevladujoč sentiment – pozitiven ali negativen –[7] identificiranih ekspresivnih ločil. Glede na pregled literature pričakujemo, da slovenske uporabnice Twitterja uporabljajo širši nabor ekspresivnih ločil in da jih uporabljajo pogosteje kot moški, hkrati pa želimo potrditi domnevo, da je prevladujoč sentiment ekspresivnih ločil pozitiven in da se emotikoni v tvitih uporabnikov obeh spolov pojavljajo pogosteje kot pravopisna ločila, saj v nasprotju s pravopisnimi ločili emotikoni poleg veselja in jeze oz. žalosti izražajo tudi sarkazem in ironijo (to lahko izraža tudi na primer navaden smeško), včasih pa le omilijo negativno sporočilo, ki ga spremljajo.

2 Priprava podatkov

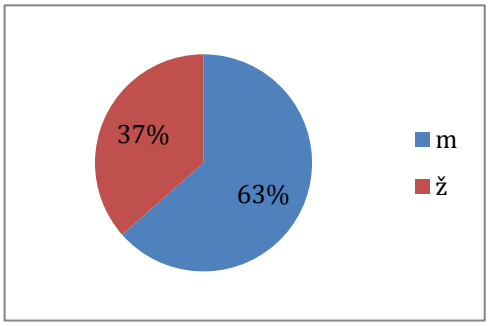
Za potrebe raziskave smo pregledali seznam uporabnikov družbenega omrežja Twitter, zajetih v korpus Janes (Fišer et al., 2014), in jim določili spol (moški, ženski, nevtralno). Najprej smo pregledali seznam uporabnikov, pri katerih je bilo mogoče zaznati rabo vsaj 5 prvoosebnih glagolskih oblik v pretekliku ali prihodnjiku, ki v slovenščini eksplicitno izražajo spol uporabnika in je zato oznako spola uporabnikom mogoče pripisati avtomatsko glede na prevladujočo uporabljeno glagolsko obliko. Avtomatsko pripisane spole smo nato preverili še ročno, in sicer s pregledom uporabniških profilov uporabnikov na Twitterju, torej njihovega uporabniškega imena, morebine rabe osebnih imen in opis profila, ter tako preverili, ali pripisana oznaka spola drži. Na ta način smo določili spol 5.883 uporabnikom omrežja Twitter. Z avtomatskim postopkom je bil spol pravilno pripisan 5.320 uporabnikom (90 %), pri 563 (10 %) pa smo oznako spola morali popraviti. Med slednjimi je 12 (2 %) takšnih, ki so dobili pripisano napačno oznako zaradi navajanja tvitov drugih uporabnikov Twitterja, v 551 primerih (98 %) pa je šlo za uporabniške račune javnih ustanov (na primer potovalnih agencij, društev, podjetij), ki jim ni mogoče določiti spola, zato smo jim zaradi večje preglednosti ob predhodnem posvetu z avtorji korpusa pripisali oznako nevtralno.

Nato smo pregledali še datoteko s seznamom 1.708 preostalih uporabnikov Twitterja v korpusu Janes, ki niso uporabili dovolj osebnih glagolskih oblik, s pomočjo katerih bi jim bilo mogoče avtomatsko pripisati spol, in jim ga pripisali ročno. 561 uporabnikom (33 %) smo na ta način spol lahko določili, 1.147 uporabnikom (67 %) pa ne, saj so pisali v izrazito neosebnem stilu, iz katerega ni bilo mogoče razbrati spola, ali pa je šlo za uporabniške račune javnih ustanov. Tem smo zato pripisali oznako nevtralno. Skupno smo torej določili spol 7.591 uporabnikom, med katerimi močno prevladujejo moški

[4] Avtorjeve primere smo za večjo nazornost nadomestili z ustreznimi slovenskimi primeri.
[5] ang. *emoticons*
[6] Iz angleščine prevedla avtorica članka.
[7] Prevzeto iz angleškega izraza *sentiment*, ki pomeni počutje, občutek, razpoloženje (gl. Merriam-Webster).

(53 %), preostali uporabniki pa so enakomerno razporejeni na ženske (24 %) in nevtralne uporabniške račune (23 %).

Na podlagi pripisanih oznak smo izdelali podkorpus tvitov moških in žensk, ki ob upoštevanju zgolj tistih, ki so glede na avtomatsko pripisani oznaki L2 in L3 (Ljubešić et al. 2015) napisani v nestandaradni slovenščini, vsebuje 18.207.584 pojavnic. Od tega so jih 11.561.770 (63 %) prispevali moški, 6.645.814 (37 %) pa ženske (glej graf 1).



Graf 1: Delež pojavnic uporabnikov in uporabnic, napisanih v nestandardni slovenščini.

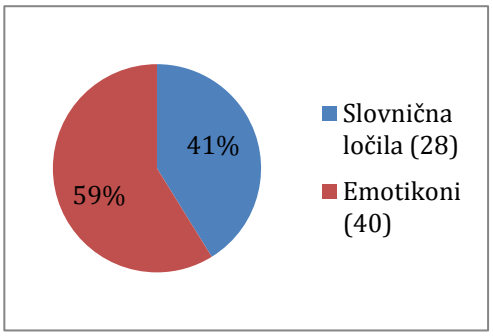
Tvitov moških je v podkorpusu torej dvakrat več kot tvitov žensk. Ker razmerje med številom pojavnic moških in ženskih uporabnikov ni uravnoteženo, pri analizi podatkov upoštevamo relativno frekvenco (normalizirano na milijon pojavnic).

3 Analiza podatkov

V podkorpusih tvitov moških in žensk z avtomatsko pripisanima stopnjama nestandardnosti L2 in L3 (gl. Ljubešić et al., 2015) smo identificirali vsa ločila, ki se v tvitih pojavijo vsaj desetkrat. Ob upoštevanju širšega konteksta, v katerem se ločilo pojavi, smo iz nadaljnje analize izločili neekspresivna ločila, torej ločila v skladenjski in neskladenjski rabi, s katerimi uporabnik ne izraža svojih čustev. Tako smo npr. med naslednjima primeroma prvega izločili iz analize, saj gre za neekspresivno rabo klicaja, drugega pa vključili, saj klicaj izraža negativno čustvo uporabnika:

Pridi v četrtek na Medijsko ogledalo!
Ampak ti si povrhu še ženska, bogtenimarad!

Iz relevantnih zadetkov smo izdelali seznam 68 ekspresivnih ločil uporabnikov in uporabnic, ki smo jih nadalje razdelili na 28 pravopisnih ločil (od katerih smo jih 27 identificirali v moških tvitih, 22 pa v ženskih) in 40 emotikonov (34 se jih pojavlja v moških tvitih, 33 pa v ženskih), kar prikazuje Graf 2:



Graf 2: Delež pravopisnih ločil in emotikonov med identificiranimi ekspresivnimi ločili.

Pri tem zaporedne ponovitve istega ločila upoštevamo kot različna ločila, saj izražajo različno stopnjo čustev. Tako na primer pravopisno ločilo !!!! izraža večjo stopnjo razburjenosti kot !, enako emotikon :))) izraža večjo intenzivnost čustev kot :). Kot različna ločila upoštevamo tudi različne kombinacije vprašaja in klicaja s ponovitvami enega ali drugega ločila, saj te izražajo različno stopnjo čustev – ?!?! na primer izraža intenzivnejše čustvo kot ?! (seveda ob upoštevanju širšega konteksta, v katerem se ločilo pojavi), kar lahko razberemo iz sledečih primerov iz korpusa:

*Z busom le, če je nujno. Ha**!** Kdo pravi, da so novinarji nemočni?*
*hahaha.. dobra **!!!!***
*al je lepo, e? **:)***
*Voje ti si res ena #faca **:)))***
*A s' kej Slovenca al nč **?!***
*A niste mogl it na fotra a **?!?!***

Zavedamo se, da nekateri identificirani emotikoni (na primer :/ in :-/) izražajo isto vrsto čustva, vendar nas zanima tudi nabor različnih ekspresivnih ločil uporabnikov in uporabnic, zato smo na seznam uvrstili vse identificirane različice emotikonov.

3.1 Razlike v rabi ekspresivnih ločil pri uporabnikih in uporabnicah

Kot prikazuje tabela 1, smo izmed vseh 68 identificiranih ekspresivnih ločil v podkorpusu moških tvitov identificirali 61 ločil – od tega se uporablja dvakrat več emotikonov (34 oz. 56 %) kot pravopisnih ločil (27 oz. 44 %). V podkorpusu ženskih tvitov smo identificirali 55 različnih ločil, od tega prav tako več emotikonov (33 oz. 60 %) kot pravopisnih ločil (22 oz. 40 %). Pri tem naj spomnimo, da je podkorpus moških tvitov dvakrat večji od podkorpusa ženskih tvitov, zato nekoliko večjega nabora ločil pri moških brez podrobnejše raziskave na enako velikih naključnih vzorcih iz obeh podkorpusov ne gre interpretirati kot znak večje nagnjenosti moških k uporabi bolj raznolikega nabora ekspresivnih ločil.

	Ekspresivna ločila	Pravopisna ločila	Emotikoni
m ločila	61	27 (44 %)	34 (56 %)
ž ločila	55	22 (40 %)	33 (60 %)

Tabela 1: Delež pravopisnih ločil in emotikonov v tvitih uporabnikov in uporabnic.

Tako moški kot ženske uporabljajo več emotikonov kot pravopisnih ločil, razlika med obojimi pa je pri obeh spolih primerljiva (razmerje je okoli 60:40). Ena izmed možnih razlag za večje število emotikonov je, da emotikoni izražajo večji nabor čustev kot pravopisna ločila, lahko so tudi le dodatek k sporočilu, ki omili njegov ton. V nadaljevanju podrobneje predstavimo analizo rabe emotikonov in pravopisnih ločil v moških in ženskih tvitih, popoln popis katerih smo vključili v tabelo 3 na koncu članka.

3.1.1 Emotikoni

Moški najpogosteje uporabljajo naslednjih pet emotikonov: :-), :)), :))), :-)), ;). Ženske pa :)), :))), ^^, :)))), :-))). Zanimivo je, da sta obema spoloma skupna le :)) in :))). Čeprav se ostali emotikoni razlikujejo, oba spola najpogosteje uporabljata emotikone, ki izražajo pozitivna čustva, na primer smeške in pomežike. Glede na relativno frekvenco ženske pogosteje od moških uporabljajo 21 (78 %) emotikonov (na primer (:, :))), :-)))), moški pogosteje uporabljajo 6 (22 %) emotikonov (na primer :-)), :-)))), 13 emotikonov, ki pa jih uporabljajo le moški ali le ženske, pri tem nismo upoštevali. Zaključimo lahko, da uporabnice emotikone uporabljajo pogosteje od uporabnikov.

Tako pri rabi pravopisnih ločil kot emotikonov je opaziti pogosto uporabo zaporedne ponovitve istega ločila. Med emotikoni je več takih z zaporedno ponovitvijo istega ločila – teh je 21 (52 %), nekaj manj pa je emotikonov brez ponovitev – teh je 19 (48 %). Emotikoni s ponovitvami so variacije sledečih emotikonov: :), :-), :(, :-(in ;), glede na intenzivnost izražene ekspresivnosti pa se ponavlja število oklepajev oz. zaklepajev. Med emotikoni vsebuje največje število ponovitev smeško z 9 ponovitvami zaklepaja, ki smo ga identificirali v podkorpusu ženskih tvitov.

V izboru najpogosteje rabljenih emotikonov obeh spolov sta emotikona :)) in :))).

3.1.2 Pravopisna ločila

Moški najpogosteje uporabljajo naslednja pravopisna ločila: ..., ?, !, !!!, ????, ženske pa: !, ?, ..., !!!, ??? . Zanimivo je, da se izbor petih najpogosteje rabljenih pravopisnih ločil pri obeh spolih prekriva, razlikujejo se le po pogostosti rabe. Glede na relativno frekvenco ima 7 (33 %) pravopisnih ločil več pojavitev v tvitih uporabnic (na primer !, ?!), 14 (67 %) ločil ima več pojavitev v tvitih uporabnikov (na primer !?!, ??), 7 pravopisna ločil pa uporabljajo le moški, zato smo jih iz tega dela analize izključili. Analiza kaže, da uporabniki pravopisna ločila uporabljajo pogosteje od uporabnic.

Pravopisna ločila lahko razdelimo v tri kategorije:

- pravopisna ločila brez zaporednih ponovitev istega ločila (v korpusu smo identificirali 2 taki ločili, in sicer vprašaj in klicaj),
- zaporedne ponovitve pike, vprašaja in klicaja (v korpusu smo identificirali 21 takih ločil) in
- kombinacije vprašaja in klicaja (v korpusu smo identificirali 5 takih ločil).

Ugotovili smo, da vprašaj in klicaj brez zaporednih ponovitev pogosteje uporabljajo uporabniki moškega kot ženskega spola, vendar smo identificirali več pravopisnih ločil z zaporednimi ponovitvami istega ločila kot brez njih, in sicer gre za ponovitve pik, vprašajev in klicajev. Med ločili z zaporednimi ponovitvami istega ločila tako moški kot ženske najpogosteje uporabljajo večpičja (..., ..., ...), in sicer smo identificirali 12 večpičij z več kot tremi zaporednimi pikami, pri čemer moški in ženske najpogosteje uporabljajo večpičje s štirimi zaporednimi pikami, zaporedje z devetimi pikami uporabljajo le ženske, zaporedja z dvanajstimi, s trinajstimi in štirinajstimi zaporednimi pikami pa le moški. Najdaljše večpičje torej vsebuje štirinajst ponovitev in se pojavi v podkorpusu moških tvitov.

Identificirali smo tudi 4 različne zaporedne ponovitve vprašaja (tako moški kot ženske najpogosteje uporabljajo tri zaporedne ponovitve vprašaja, največje identificirano število ponovitev je pet, to ločilo pa uporabljajo le moški) in 4 različne zaporedne ponovitve klicaja (tako moški kot ženske najpogosteje uporabljajo tri zaporedne ponovitve, največje identificirano število ponovitev klicaja je pet, ločilo pa moški uporabljajo pogosteje od žensk). Pri tem se zavedamo, da se je (še posebej pri več kot treh zaporednih ponovitvah pike) uporabnik morda zatipkal, vendar bi v takšnem primeru dodal le eno ali dve »odvečni« ločili. Zaporedna ponovitev na primer sedem pik pa kaže, da jih je uporabnik napisal namenoma, da poudari intenziteto svojih čustev.

Analizirali smo tudi kombinacije vprašaja in klicaja ter ugotovili, da oba spola najpogosteje uporabljata kombinaciji ?! in !?, moški redko uporabljajo tudi ?!? in ?!?!, ženske pa sploh ne.

3.1 Sentiment ločil

Sledi analiza ekspresivnosti pravopisnih ločil in emotikonov, kjer smo klasificirali 60 naključnih pojavitev posameznega ločila. Če se je ločilo v korpusu pojavilo manj kot 60-krat, smo pregledali vse pojavitve.

Avtomatsko pripisane oznake sentimenta tvitov, v katerih se pojavijo identificirana ekspresivna ločila, so nam služile kot smernica za določanje sentimenta ločil. Za končno določitev pozitivnega ali negativnega sentimenta ločil smo ročno pregledali tvite, v katerih se identificirana ekspresivna ločila pojavijo, saj je njihov sentiment odvisen od širšega konteksta. Klicaju ali vprašaju lahko na primer določimo tako pozitiven kot negativen sentiment, odvisno od konteksta, v katerem se pojavljata.

Posebej smo analizirali moška in ženska pravopisna ločila ter emotikone. Za ponazoritev navajamo primer pravopisnega ločila, ki smo mu glede na avtomatsko pripisano oznako sentimenta tvita in širšega konteksta besedila določili pozitiven sentiment (prvi primer), in primer emotikona, ki smo mu določili negativen sentiment (drugi primer):

*Košiiiiiiiir**!!!** Srebrna medalja, osma za Slovenijo! Sramota! **:/***

Rezultate analize sentimenta ekspresivnih ločil prikazuje tabela 3 (v prilogi). Največjo razliko med deležema negativnih in pozitivnih sentimentov je opaziti pri pravopisnih ločilih, saj jih ima med 27 pravopisnimi ločili v tvitih moških kar 24 (89 %) negativen sentiment, le 3 (11 %) pa pozitivnega. Podobno ima med 22 pravopisnimi ločili v tvitih žensk kar 17 (77 %) ločil negativen in le 5 (23 %) pozitiven sentiment. Razlika med deležem pozitivnih in negativnih sentimentov je nekoliko manjša pri emotikonih – med 34 emotikoni v tvitih moških jih ima 21 (62 %) negativen sentiment, 13 (38 %) pa pozitivnega. Med 33 ženskimi emotikoni jih ima 18 (55 %) negativen sentiment, 15 (54 %) pa pozitivnega, kar tudi prikazuje tudi tabela 2:

	Pravopisna ločila		Emotikoni	
Spol	-	+	-	+
m	24 (89 %)	3 (11 %)	21 (62 %)	13 (38 %)
ž	17 (77 %)	5 (23 %)	18 (55 %)	15 (54 %)

Tabela 2: Deleži ekspresivnih ločil glede na sentiment in spol uporabnikov.

Zdi se presenetljivo, da se največ ekspresivnih ločil pojavlja v negativnem kontekstu, vendar je treba upoštevati, da ekspresivna ločila v negativnih tvitih velikokrat ne izražajo le jeze ali žalosti, temveč tudi sarkazem in ironijo (tudi na primer navaden smeško), včasih pa tudi le omilijo negativno sporočilo, ki ga spremljajo.

4 Sklep

Za komunikacijo na družbenih omrežjih je značilno, da osebe, s katero komuniciramo, ne vidimo ali slišimo. Pomanjkanje telesne govorice in obrazne mimike ter nadsegmentnih zvočnih lastnosti diskurza za izražanje čustev uporabniki družbenih omrežij nadomestijo z uporabo ekspresivnih ločil – emotikonov in pravopisnih ločil. Naša analiza je pokazala, da med identificiranimi ekspresivnimi ločili prevladujejo emotikoni (59 % vseh identificiranih ekspresivnih ločil), manj pa je pravopisnih ločil (41 %). Tudi med ločili, ki jih uporabljajo le moški ali le ženske, večino predstavljajo emotikoni (56 % v moških in 60 % v ženskih tvitih), manj pa je pravopisnih ločil (44 % v moških in 40 % v ženskih tvitih). Moški uporabljajo nekoliko večji nabor ekspresivnih ločil kot ženske, vendar je to morda rezultat neuravnoteženosti korpusov moških in ženskih tvitov. Za potrditev tega zaključka bi torej morali raziskavo ponovno izvesti na enako velikih vzorcih iz obeh korpusov.

Dokazali smo tudi, da uporabniki ne glede na spol pogosteje uporabljajo emotikone kot pa pravopisna ločila. Večjo pogostost emotikonov lahko razložimo z dejstvom, da lahko s pravopisnimi ločili uporabniki družbenih omrežij izražajo precej omejen izbor čustev – izražajo lahko le osnovna pozitivna in negativna čustva (kot sta veselje ali jeza), ne morejo pa izražati žalosti, ironije, sarkazma ipd. Izražanje te vrste kompleksnih čustev jim omogočajo emotikoni, ki pogosto igrajo zgolj vlogo dodatka k besedilu, saj podkrepijo njegovo sporočijo ali ga omilijo. Analiza rabe emotikonov v tvitih uporabnikov in uporabnic je pokazala, da uporabniki pogosteje uporabljajo emotikone z zaporedno ponovitvijo istega ločila (52 %) kot pa emotikone brez ponovitev ločila (48 %). Najpogosteje rabljeni emotikoni s ponovitvami so variacije sledečih emotikonov: :), :-), :(, :-(in ;), glede na intenzivnost izražene ekspresivnosti pa se ponavlja število njihovih oklepajev oz. zaklepajev.

Podrobna analiza pravopisnih ločil pa je pokazala, da pravopisna ločila brez zaporednih ponovitev istega ločila (torej enojna vprašaj in klicaj) predstavljajo 7 % vseh identificiranih pravopisnih ločil, 18 % pravopisnih ločil predstavljajo kombinacije vprašaja in klicaja, uporabniki pa najpogosteje uporabljajo pravopisna ločila z zaporednimi ponovitvami istega ločila – ta predstavljajo 75 % vseh pravopisnih ločil.

Po pričakovanjih smo glede na število ekspresivnih ločil v tvitih uporabnikov in uporabnic ob upoštevanjeu neuravnoteženosti števila pojavnic v podkorpusu moših in ženskih tvitov ugotovili, da uporabnice emotikone uporabljajo bistveno pogosteje od uporabnikov (med emotikoni, skupnimi obema spoloma, se jih 78 % pogosteje pojavi v tvitih uporabnic, le 22 % pa pogosteje v tvitih uporabnikov). Presenetljivo pa je, da se pravopisna ločila pogosteje pojavljajo v tvitih moških (67 %) Zaključimo lahko, da slovenski uporabniki Twitterja za izražanje čustev pogosteje uporabljajo pravopisna ločila, uporabnice pa emotikone.

V nasprotju z našimi prvotnimi pričakovanji pa je analiza sentimenta pri emotikonih in pravopisnih ločilih pokazala, da s pravopisnimi ločili uporabniki obeh spolov najpogosteje izražajo negativna čustva. To velja tudi za emotikone, čeprav je razlika med deleži pozitivnih in negativnih sentimentov pri njih precej manjša kot pri pravopisnih ločilih. To se morda na prvi pogled zdi nekoliko presenetjjivo, vendar je treba upoštevati, da ekspresivna ločila v negativnih tvitih velikokrat ne izražajo le jeze ali žalosti, temveč tudi kompleksna čustva, kot sta sarkazem in ironija, včasih pa le omilijo ton sporočila, ki ga spremljajo.

Zaradi razlike v velikosti obeh analiziranih podkorpusov bi za potrditev teh zaključkov morali izvesti podrobnejšo raziskavo na enako velikih naključnih vzorcih iz obeh podkorpusov. Prav tako bi bilo zanimivo raziskavo razširiti še na forume, komentarje, spletne novice in bloge.

5 Zahvala

Raziskava, opisana v prispevku, je bila opravljena v okviru nacionalnega temeljnega projekta "Viri, orodja in metode za raziskovanje nestandardne spletne slovenščine" (J6-6842, 2014-2017), ki ga financira ARRS.

6 Literatura

Teja Antončič. 2006. *Jezikovne in stilne posebnosti novih medijev: primer spletnih klepetalnic*. Diplomsko delo, FDV, Ljubljana.

Michael Argyle. 1988. Bodily Communication. London: Methuen, druga izdaja.

Brittney G. Chenault. 1998. Computer-Mediated Communication and Emotion: Developing Personal Relationship Via CMC. *Computer-Mediated*

Communication Magazine. Dostopno na: http://www.december.com/cmc/mag/1998/may/chenault.html (3. julij 2015).

Ann Colley in Zazie Todd. 2002. Gender-linked differences in the style and content of e-mails to friends. *Journal of Language and Social Psychology*, 21(4): 380–392.

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

David Crystal. 2008. *Txtng: The Gr8 Db8*. Oxford University Press.

Darja Fišer et al., 2015: Gradnja in analiza korpusa spletne slovenščine JANES (Zbornik simpozija Obdobja – v tisku).

Darja Fišer, Tomaž Erjavec, Ana Zwitter Vitez, Nikola Ljubešić. 2014. JANES se predstavi: metode, orodja in viri za nestandardno pisno spletno slovenščino. Zbornik Devete konference Jezikovne tehnologije. Ljubljana: Institut Jožef Stefan.

Tomaž Erjavec in Darja Fišer. 2013. Jezik slovenskih tvitov: korpusna raziskava. V: A. Žele, ur., *Družbena funkcijskost jezika: (vidiki, merila, opredelitve)*, Obdobja 32, str. 109–116. Ljubljana. Filozofska fakulteta.

Seth Grimes. 2008. *Sentiment analysis: A Focus on Applications.* BeyeNETWORK – Global coverage of the business Intelligence Ecosystem. Dostopno na: http://www.b-eye-network.com/view/6897 (27. september 2015).

Funda Kivran-Swaine, Sam Brody in Mor Naaman. 2013. *Effects of gender and tie strength on twitter interactions*, 18(9). Dostopno na: http://firstmonday.org/ojs/index.php/fm/article/view/4633/3746 (27. september 2015).

Simona Kranjc. 2003. Jezik v elektronskih medijih. V: A. Vidovič Muha, ur., *Slovenski knjižni jezik – aktualna vprašanja in zgodovinske izkušnje*, Obdobja 20, str. 435–446. Ljubljana. Filozofska fakulteta.

Robin Lakoff. 1975. Language and Woman's Place. *Language in Society*, 2: 45–80.

Rich Ling. 2015. The Sociolinguistics of SMS: An Analysis of SMS Use by a Random Sample of Norwegians. V: R. Ling in P. Pedersen, ur., *Mobile communications: Renegotiation of the social sphere*, str. 335–349, London: Springer.

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. *1oth International Conference on Recent Advances in Natural Language Processing: Proceedings of RANLP 2015 Conference*, 7.–9. September 2015, Hissar, Bolgarija: 371-378.

Merriam-Webster. Dostopno na: http://www.merriam-webster.com/dictionary/sentiment (25. oktober 2016).

Mija Michelizza. 2014. Slovenščina v elektronskih medijih. *Razpotja*, 15. Dostopno na: http://www.razpotja.si/slovenscina-v-elektronskih-medijih/ (26. septembr 2015).

Róisín Parkins. 2012. Gender and Emotional Expressiveness: An Analysis of Prosodic Features in Emotional Expression. *Griffith Working Papers in Pragmatics and Intercultral Communication* 5(1): 46–54.

Sarah Pedersen in Caroline Macafee. 2007. Gender differences in British blogging. *Journal of Computer-Mediated Communication*, 12(4): 1472–1492.

Tadej Praprotnik. 2003. Skupnost, identiteta in komunikacija v virtualnih skupnostih. ISH, Ljubljana.

Donald Rubin in Kathryn Greene. 1992. Gender-typical style in written language. *Research in the Teaching of English*, 26(1): 7–40.

Carol Scates. 1981. *A Sociolinguistic Study of Male/Female Language in Freshman Composition.* Neobjavljena doktorska disertacija. University of Southern Mississippi, Hattiesburg, Mississippi.

Jonathan Schler, Moshe Koppel, Shlomo Argamon in James Pennebaker. 2006. Effects of age and gender on blogging. V: N. Nicolov, F. Salvetti, M. Liberman in J. H. Martin, ur., Zbornik konference *AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, str. 199–206, AAAI Press.

SP 2001 – Slovenski pravopis. Ljubljana: SAZU – ZRC SAZU – Založba ZRC. § 226.

Deborah Tannen. 1990. *You Just Don't Understand*. New York, Ballantine.

Mirjana Ule in Miro Kline. 1996. *Psihologija tržnega komuniciranja*. FDV, Ljubljana.

Tina Verovnik. 2003. Analiza jezikovne kakovosti besedil v vladnem spletu. V: A. Lukšič in T. Oblak, ur., *S poti v digitalno demokracijo*, 147–158. Dostopno na: http://odkw.fdv.uni-lj.si/eknjige (3. julij 2015).

Carol Waseleski. 2006. Gender and the Use of Exclamation Points in Computer-Mediated Communication: An Analysis of Exclamations Posted to Two Electronic Discussion Lists. V: G. Gay, ur., *Journal of Computer-Mediated Communication*, 11, str. 1012–1024.

Diane F. Witmer in Sandra Lee Katzman. 2006. On-line smiles: Does gender make a difference in the use of graphic accents? *Journal of Computer-Mediated Communication*, 2 (4). Dostopno na: http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.1997.tb00192.x/full (26. september 2015).

7 PRILOGE

Priloga 1: Tabela 3

Emotikon	Št. pojavitev	Št. pojavitev	Sentiment med naključnimi 60 zadetki				Pravopisno ločilo	Št. pojavitev	Št. pojavitev	Sentiment med naključnimi 60 zadetki			
			Negativen		Pozitiven					Negativen		Pozitiven	
	m	ž	m	ž	m	ž		m	ž	m	ž	m	ž
(:	30	51	19	32	11	19	...	122.251	59.983	48	48	12	12
:))	5.591	5.585	27	31	33	29	....	10.349	6.318	44	31	16	29
:)))	2.027	2.705	21	17	26	43	.....	1.903	1.273	42	16	18	44
(((:	0	12	0	7	0	5	.....	792	351	39	42	21	18
:))))	584	574	38	21	22	39	.....	306	247	45	38	15	22
:))))))	252	162	27	45	33	15	.....	165	103	47	38	13	22
:))))))	80	56	32	21	28	35	.....	136	59	46	41	14	18
:))))))	36	19	15	4	21	15	.....	0	32	0	26	0	6
:))))))	18	18	7	12	11	6	.....	37	16	28	10	9	6
:))))))	0	11	0	5	0	6	.....	25	16	16	12	12	4
:-)	13.906	225	12	24	48	36	.....	24	0	20	0	4	0
:-))	1.946	239	35	18	25	42	.....	15	0	6	0	9	0
:-)))	308	517	25	32	35	18	.....	14	0	8	0	6	0
:-))))	107	49	38	13	22	26	?	141.724	75.542	42	50	16	10
:-)))))	47	22	25	26	22	34	??	397	121	55	53	5	7
:-)))))	15	0	10	0	5	0	???	427	146	59	45	1	15
:((	148	158	56	55	4	5	????	80	31	49	29	11	2
:(((	46	155	43	59	3	1	?????	22	0	16	0	6	0
:((((	14	38	14	31	0	7	!	110.498	77.202	44	34	17	26
:((((((	11	0	11	0	0	0	!!	400	168	32	16	28	44
:-(	551	0	57	0	3	0	!!!	696	191	27	21	33	39
:-((	50	12	44	12	6	0	!!!!	158	48	25	23	35	25
:-(((	10	0	10	0	0	0	!!!!!	38	21	22	6	17	15
:/	26	18	23	16	3	2	?!	152	129	52	47	8	13
:-/	322	0	56	0	4	0	!?	125	18	48	17	12	1
:'(	33	44	31	42	2	2	?!?	39	22	35	22	4	0
:')	0	52	0	19	0	33	!?!	17	0	15	0	2	0
(=	21	47	15	31	6	16	?!?!	15	0	15	0	0	0
=)	13	0	5	0	8	0							
;)	684	461	23	22	37	38							
;))	0	10	0	4	0	6							
(;	20	14	9	8	11	6							
^^	364	1.170	26	16	34	44							
^.^	11	34	4	23	7	11							
>_<	0	23	0	14	0	9							
>.<	55	125	39	48	16	12							
^_^	46	13	27	3	19	10							
.	11	21	5	8	6	13							
;*	10	0	3	0	7	0							
>=|	0	15	0	11	0	4							

Tabela 3: Podatki o ekspresivnih ločilih, številu njihovih pojavitev in njihov sentiment glede na spol uporabnikov.

Identifikacija spletno specifičnih kolokacij pogostega besedišča

Senja Pollak

Odsek za tehnologije znanja, Institut »Jožef Stefan«
Jamova cesta 39, 1000 Ljubljana
senja.pollak@ijs.si

Povzetek

V prispevku predstavimo pristop k luščenju spletno specifičnih kolokacij pogostega slovenskega besedišča. Izbrali smo več kot 150 najpogostejših besed, ki se pojavljajo tako v korpusu uporabniških vsebin Janes kot v vzorčenem referenčnem korpusu slovenščine Kres. Za te leme smo izluščili kolokatorje, ki se pojavljajo tik pred izbrano besedo in izluščili sezname za oba korpusa. Nato smo identificirali tiste kolokacije, ki se pojavljajo izključno v korpusu uporabniških vsebin. Predstavljena metodologija omogoča hitro luščenje in modularne nastavitve parametrov, primernih za luščenje podobnih specifičnih seznamov kolokacijskih kandidatov.

Identification of Web-specific collocations of common lexis

In this article we present an approach to extracting Web-specific collocations for standard Slovene lexis. We selected more than 150 most frequent Slovene words in Janes, the corpus of user-generated content, and Kres, the reference balanced corpus of Slovene. In each of the two corpora we extracted collocators which immediately preceded the selected words. We then identified the collocations that were specific to user-generated content, i.e. those appearing only in the Janes corpus. The presented methodology enables a rapid extraction of collocation candidates and modular parameter settings, suitable for extracting similar lists of specific collocation candidates.

1 Uvod

Kolokacije so tipične sopojavitve besed. Pristopi k preučevanju kolokaciji se v grobem delijo na frekvenčne in frazeološke (Nesselhauf, 2005). *Frekvenčni pristopi* razumejo kolokacije kot statistične sopojavitve besed znotraj določenega okna[1] (prim. Firth, 1957; Halliday, 1961; Sinclair, 1966), tipičnost oz. statistična signifikantnost sopojavitev pa je definirana s pogostostjo in različnimi merami statistične povezanosti. Iz te tradicije, ki razume kolokacije v širšem pomenu, izhajajo tudi korpusni pristopi in računalniški pristopi k luščenju kolokacij, sama dolžina okna (definicija sopojavitve), frekvence in statistične mere povezanosti pa se pojavljajo v več različicah. *Frazeološki pristopi* pa obravnavajo kolokacije v ožjem pomenu. Poudarjajo sintaktične (Sinclair, 1991; Kjellmer, 1987), predvsem pa semantične aspekte kolokacij, ki jih definirajo kot besedne zveze, ki se nahajajo med »prostimi« besednimi kombinacijami in idiomi (npr. Cowie, 1994; Benson, 1989). Razlikovalni kriteriji pa se opirajo predvsem na pomensko razstavljivost in transparentnost zveze. Avtorji se tudi razlikujejo po tem, ali se osredotočajo le na leksikalne besede ali tudi na slovnične besede (npr. Bartsch, 2004; Siepmann, 2005).

V prispevku se pri postopku luščenja naslanjamo na strogo statistične kriterije. Z začetnim izborom lem, kjer upoštevamo izključno samostalnike, ter pri podajanju primerov, kjer diskutiramo tako o slovničnih kot o pomenskih strukturah, pa se odmikamo od strogo frekvenčnega pristopa.

Kolokacije lahko razumemo kot pare besed (predvsem pri frekvenčnih pristopih) ali pa izbrano besedo (lemo) interpretiramo kot bazo, ki ga kolokator pobliže določa (Hausmann,1989 po Gorjanc in Jurko, 2004). V tem prispevku luščimo kolokacije kot pare samostalniške leme *l* in kolokatorja *k*.

Kolokacije so pomembno področje preučevanja jezika in tako tudi za analizo spletne slovenščine. Termin *spletni jezik* oz. *jezik uporabniških vsebin* uporabljamo za jezik komunikacije na forumih, blogih in družbenih omrežjih, kot je Twitter. Omejujemo se na preučevanje točno določenega korpusa, korpusa Janes (Fišer et al., 2014), tako da ne pokrivamo celotnega spektra spletnega jezika (npr. spletnega časopisja, komunikacije preko elektronske pošte). Prav tako v naši obravnavi izpuščamo besedila nekaterih pomembnih družbenih omrežij (npr. Facebook) in marsikatere uporabniške vsebine (npr. Wikipedija). Zato mora bralec izraz *spletni jezik*, ki ga uporabljamo v nadaljevanju, primerno interpretirati.

Spletni jezik ima svoje značilnosti. Pisno spletno komunikacijo določajo okoliščine, kot so (ne)interaktivnost, (a)sinhronost, fizična (ne)prisotnost sogovornika in drugi situacijski dejavniki (Noblia, 1998; Fišer et al., 2014). Bolj kot je izbrana oblika komuniciranja interaktivna, poteka v realnem času in ima na drugi strani prisotnega sogovornika, več prvin spontanega govorjenega jezika vsebuje, vključno s prozodičnimi elementi kot tudi s paralingvističnimi elementi, prilagojenimi za računalniško komunikacijo (Crystal, 2001). Specifike spletnega jezika lahko preučujemo z različnih vidikov, kot je ortografija (npr. bolj fonetičen zapis besed, opuščanje ločil in ponavljanje črk za čustveno poudarjanje zapisane izjave), skladnja (npr. nestandardni vrstni red, stavčne strukture), diskurzivna raven (interaktivnost) in leksikalna raven, kjer se najhitreje in najpogosteje kaže inovativna raba jezika, značilna za uporabniške vsebine (neologizmi, aktualne tematike). V tem prispevku nas zanima predvsem leksikalna specifika, ki jo preučujemo na nivoju večbesednih zvez oz. kolokacij.

V članku predstavimo metodo identifikacije kolokacij, ki so specifične za spletne vsebine, ali natančneje, za identifikacijo tistih kolokacij, ki se pojavljajo le v jeziku

[1] Za razliko od besednih nizov oz. n-gramov, ki so zaporedja besed, okno upošteva sopojavitev besed, četudi ne gre za pravo zaporedje. Npr. v nizu besed b_1 b_2 b_3 je pri pri preučevanju kolokacije besede b_1 in oknu 2 upoštevana tudi beseda b_3. Pri oknu 1 oz. bigramih $b_1 b_2$ pa ni razlike.

(oz. korpusu) spletnih uporabniških vsebin in ne v bolj splošni, »standardni« rabi jezika, ki jo predstavlja referenčni korpus. Za izhodiščne leme (tj. leme, za katere nas zanimajo njihova kolokacijska okolja) izberemo najpogostejše splošno (torej nespecifično) slovensko besedišče.

S področja kolokacij je na voljo vrsta tujih študij. Poleg osrednjih leksikografskih vidikov (Castro in Faber, 2014) so kolokacije pomembne z vidika uporabe v številnih aplikacijah, kot so poučevanje tujega jezika (Orenha-Ottaiano, 2012), luščenje informacij (Lin, 1998), strojno prevajanje (Gerber in Yang, 1997) itn. Na uporabniške vsebine se osredotočajo npr. Rösiger et al. (2015), ki luščijo terminologijo iz korpusa uporabniških vsebin strani tipa »naredi si sam«), Seretan (2015) se posveča prevodom večbesednih izrazov iz uporabniških spletnih vsebinah z vidika strojnega prevajanja, analizo sentimenta v uporabniških spletnih vsebinah obravnava npr. Yu (2014).

V slovenščini se je z luščenjem kolokacij ukvarjalo že več avtorjev. Za leksikografske namene so podatke iz referenčnega korpusa Gigafida luščili Gantar in Krek (2011), Kosem et al. (2013), s terminološkega vidika so bile kolokacije obravnavane v Vintar (2010) in Logar Berginc et al. (2014).

V dosedanjem delu smo že obravnavali tematiko iskanja kolokacij, specifičnih za korpus spletnih vsebin (Pollak, 2015), vendar se od preteklih raziskav pričujoči prispevek razlikuje v več pogledih:

- luščimo spletno specifične kolokacije lem najpogostejšega splošnega besedišča,
- nova metodologija vključuje orodje za hitri izvoz kolokacijskih seznamov (API),
- v prejšnjem eksperimentu smo se osredotočili na analizo okoli 30 lem, v tem eksperimentu pa gre za mnogo večji izbor, ki obsega 150 lem,
- modularna nastavitev parametrov (frekvenca, mera kolokabilnosti, mera razpršenosti kolokatorja) omogočajo enostavno uporabo orodja glede na specifične cilje.

2 Metodologija luščenja

Za poljubno lemo *l,* ki se pojavlja v seznamu *n* najpogostejših besed slovenskega besedišča (tako spletnega kot splošnega) nas zanimajo tiste kolokacije oz. pari kolokatorjev in lem <*k*, *l*>, ki se pojavljajo izključno v uporabniških spletnih vsebinah, ne pa v referenčnem korpusu slovenščine. Z dodatnimi parametri definiramo ustrezno mero kolokabilnosti, frekvence v korpusu, okno iskanja ter razpršenost kolokatorja.

2.1 Korpusa

Za iskanje razlik med kolokacijami slovenščine uporabniških spletnih vsebin in »splošno« slovenščino uporabimo dva korpusa. Spletno slovenščino predstavlja korpus Janes v0.3 (Fišer et al., 2014), ki zajema besedila forumov, tvitov in blogov (okoli 161 milijonov pojavnic).

Kot referenčni korpus »splošne« slovenščine izbremo korpus Kres (Logar et al., 2012). Kres je iz korpusa Gigafida (ibid.) vzorčeni uravnoteženi podkorpus in ga uporabljamo kot referenčni korpus. Ima 121 milijonov pojavnic in vsebuje stvarna besedila, leposlovje, časopisje, revije in internetne vsebine.

Metodologija je neodvisna od izbora korpusov, vendar korpus definira same rezultate. Če bi vzeli drug referenčni korpus (npr. Gigafida), bi bile tudi izluščene spletno specifične kolokacije druge. V primeru križanja seznamov z govornim korpusom (Verdonik et al., 2014) pa bi dobili le tiste kandidate, ki so specifični za uporabniške vsebine, ne pa za govor. V primeru sproti posodabljanega korpusa »splošne« oz. standardne slovenščine pa bi z metodo lahko natančneje luščili spletno oz. nestandardno besedišče.

2.2 Izbor lem

V predstavljeni raziskavi nas zanimajo spletno specifične kolokacije za leme splošnega slovenskega besedišča. Izdelali smo frekvenčne sezname lem po posameznih besednih vrstah. Za opisano raziskavo smo se osredotočili na 250 najpogostejših samostalnikov vsakega korpusa. Seznama iz obeh korpusov smo med seboj križali in ohranili 150 občnih samostalnikov, ki se pojavljajo na obeh seznamih.

Nekaj samostalnikov iz seznama (izbor je naključen in zajema primer vsake črke abecede): *avto*, *beseda*, *cerkev*, *človek*, *dogodek*, *energija*, *fant*, *glava*, *hiša*, *igralec*, *jezik*, *knjiga*, *ljubezen*, *mati*, *namen*, *oče*, *pesem*, *roka*, *sistem*, *šola*, *telo*, *ura*, *vas*, *zgodba*, *želja.* Primeri tistih lem, ki so bili na seznamu le enega od obeh korpusov in jih zato nismo vključili, so npr. leme iz pravnih besedilih korpusa Kres (*ministrstvo, komisija, člen, besedilo*) ter npr. leme korpusa Janes, vezane na spletni medij (*novica, video, forum, komentar*).

Luščenje spletno specifičnih kolokacij lem pogostega splošnega slovenskega besedišča, ki je opisano v tem prispevku, je predvsem koristno za preučevanje pomenskih premikov, iskanje aktualnih kolokatorjev splošnega slovenskega besedišča ipd. Za razliko od te naloge bi lahko preučevali tudi leme, ki se pojavljajo izključno v korpusu Janes in ne v referenčnem korpusu, vendar v tem primeru ni treba primerjati kolokatorjev z referenčnim korpusom. Prav tako zanimiva je naloga preučevanja zelo pogostih besed v korpusu Janes, ki so v referenčnem korpusu manj pogoste, vendar tu najverjetneje ne bi želeli izključiti kolokacij, ki se pojavljajo v referenčnem korpusu (npr. *pisanje bloga* se kot kolokacija leme *blog* pojavlja v obeh korpusih), temveč bi rajši uporabili mero, ki primerja moč kolokacij (prim. Pollak in Arhar Holdt, 2015; Pollak, 2015).

Izbor lem je torej usklajen z motivacijo članka (iskanje nestandardnih in novih kolokacij zelo pogostega splošnega besedišča) in z metodologijo iskanja kolokacij (izključnost pojavljanja v korpusu uporabniških spletnih vsebin). Druge vidike bomo podrobneje preučili v nadaljnjem delu, možno metodologijo za njihovo obravnavo pa smo že predstavili v Pollak in Arhar Holdt (2015) in Pollak (2015).

2.3 Luščenje kolokacij iz posameznih korpusov

Na ravni izbora lem smo upoštevali leme, ki se pojavljajo v obeh korpusih, pri luščenju kolokacij pa je cilj izluščiti samo tiste kolokacije (kombinacije lem in kolokatorjev), ki so izključno v spletnih uporabniških vsebinah (glej 2.4), vendar moramo zato najprej izvoziti kolokacijske sezname iz obeh korpusov.

Za luščenje kolokacij uporabljamo funkcijo *kolokacije* orodja SketchEngine (Kilgariff et al., 2004). Hitro

luščenje za veliko število izbranih lem ter poljubno nastavljanje paramerov smo omogočili z API-ja, ki kliče lokalno instalacijo korpusov.

Nastavitve parametrov luščenja kolokacij v orodju Sketch Engine zajemajo razpon okna, tj. okolico besed izhodiščne leme, mero za izračun kolokacijske vrednosti ter minimalni frekvenci leme in kolokacijskega niza v izbranem korpusu.

V predstavljenih eksperimentih smo se omejili le na besede tik pred izbrano lemo, torej na okno -1, 0, vendar metoda sama ni odvisna od nastavitve okna.[2]

Za izračun kolokacijske vrednosti kolokacij za posamezni korpus uporabimo mero logDice (Rychlý, 2008), ki je priporočena mera za izračun kolokacij v orodju SketchEngine:

$$logDice = 14 + \log_2 D = 14 + \log_2 \frac{2f_{xy}}{f_x + f_y}$$

D se v formuli nanaša na originalni Diceov koeficient (Dice, 1945), medtem ko se f_x, f_y in f_{xy} nanašajo na relativne frekvence besed *x* in *y* ter njunih skupnih pojavitev. Ta mera ima več prednosti (Rychlý, 2008), kot so lahka interpretacija vrednosti (med 0 in 14) ter neobčutljivost na velikosti korpusov, saj temelji na relativnih frekvencah in omogoča primerljivost med različnimi korpusi.

Za minimalne frekvence luščenja smo pri luščenju kolokacij posameznega korpusa določili nastavitev 5 pojavitev za kolokator in 3 za kolokacijski niz v izbranem oknu, vendar se v koraku izbora spletno specifičnih kolokacij (glej 2.4) omejimo na kandidate z bolj pogostimi pojavitvami (min. frekvenca 10).

Za vsako lemo izvozimo kolokatorje in pripadajoče logDice vrednosti ter relativne frekvence.

V Tabeli 1 prikažemo najmočnejše kolokatorje leme *družina*, ki jo izberemo za ponazoritev postopka luščenja tudi v nadaljevanju prispevka.

Kolokacije (lema: *družina*)	
Kres	**Janes**
član družine	enostarševska družina
ribiška družina	član družine
kraljeva družina	primarna družina
lovska družina	kraljeva družina
ustvariti družino	mlada družina
rejniška družina	ogrožena družina
mlada družina	cela družina
kmečka družina	ustvariti družino
njegova družina	(*n*) članska družina
svoja družina	romska družina

Tabela 1: Najpogostejših deset kolokacij z besedo družina v korpusih Kres in Janes (nastavitve: -1, 0, mera logDice).

Vidimo, da je več kolokacij presečnih, vendar se v kolokacijah korpusa Janes bolj odražajo aktualne tematike (npr. kolokacija *enostarševska družina* se pojavlja tudi v korpusu Kres, vendar je logDice vrednost bistveno nižja (5,7 za Kres; 8 za Janes), kakor tudi število pojavnic (63 v korpusu Kres oz. 0,52/1M, 199 oz. 1,23/M v korpusu Janes). Podobno velja za nekatere druge tematsko aktualne primere, npr. *primarna družina* ima logDice vrednosti 7,6 v korpusu Janes (156 pojavnic) in 4,8 v korpusu Kres (34 pojavnic), kar pomeni, da je tematika veliko bolj zastopana v spletnih vsebinah, kljub temu da kolokacija obstaja v obeh korpusih. V našem prispevku se za razliko od Pollak (2015) osredotočimo na kolokacije, ki se pojavljajo izključno v enem od dveh korpusov, in tovrstnih primerov, kjer gre le za razliko med vrednostma kolokatorjev, podrobneje ne obravnavamo, kljub temu da predstavljajo zanimiv material za analizo.

2.4 Identifikacija spletno specifičnih kolokacij

Spletno specifične kolokacije iščemo s pomočjo programa, ki ga izdelamo v ta namen. Ta na podlagi izvoženih kolokacijskih seznamov dveh korpusov izlušči seznam tistih kolokacij oz. kolokatorjev dane leme, ki se pojavljajo izključno[3] v izbranem korpusu (v našem primeru v korpusu Janes).[4] Poleg ključnega pogoja iskanja kolokacij, ki se pojavljajo izključno v spletnem korpusu, omogočamo tudi nastavitev najnižje dopuščene frekvence kolokacije v korpusu, najnižje vrednosti logDice ter razpršenost kolokatorja po lemah (v nadaljevanju *razp*).

Razp je odstotek lem, ki vsebujejo določeni kolokator.[5] *Razp* izračunamo tako, da število izdelanih kolokacijskih seznamov specifičnega korpusa, ki vsebujejo določeni kolokator, delimo s številom vseh preučevanih seznamov kolokacij tega korpusa (tj. s številom obravnavanih lem). Torej, če se nek kolokator pojavlja na kolokacijskih seznamih vseh preučevanih lem, je njegova vrednost 1 in čim redkeje se pojavlja, tem bolj je specifičen za določeno lemo in vrednost *razp* je bližje 0. S to mero lahko določimo zgornjo mejno vrednost in izločimo funkcijske besede z visoko *razp* vrednostjo, ki se pojavljajo kot kolokatorji velikega števila lem (npr. *vsak, kakšen, moj*)[6]. Tem je sicer pripisana praviloma nizka vrednost logDice. Prav tako imajo relativno visoko *razp* vrednost pogoste besede, ki so specifika korpusa, npr. besede z opuščenimi strešicami (*vec*), pogoste okrajšave v nestandardnem zapisu (*slo*), pogosti tematski izrazi (*političen*), razlike med lematizaroji dveh korpusov (edini vs. edin). itd.

[2] V Pollak (2015) smo npr. vzeli okno -3 +3.

[3] To pomeni, da se v referenčnem korpusu pojavljajo manj kot trikrat, saj je bil ta pogoj upoštevan pri luščenju iz posamičnih korpusov (glej sekcijo 2.3).

[4] Alternativni pristop k iskanju korpusnospecifičnih kolokacij je, da se ne omejimo na kolokatorje, ki se pojavljajo izključno v izbranem korpusu, temveč dopuščamo njihove pojavitve v obeh korpusih, vendar določimo razliko v vrednosti kolokabilnosti (mero CorpDiff smo vpeljali v Pollak in Arhar Holdt (2015) in jo uporabili v Pollak (2015)). V pričujoči raziskavi specifičnost definiramo kot izključnost.

[5] *Razp* ima enako motivacijo kot mera *inverse document frequency* oz. *idf*, ki jo je uvedla K. Spärck Jones (1972) v kontekstu iskanja dokumentov (angl. *document retrieval*). Mera *idf* uteži termine tako, da zmanjša vrednost terminov, ki se pojavljajo v več dokumentih neke zbirke, in poveča vrednost tistim, ki so prisotni v manj dokumentih. V našem primeru uporabimo različico *frekvence po dokumentih* (angl. *document frequency*) in ne *idf,* zato je manjša vrednost bolj informativna.

[6] Kolokatorji z vrednostjo *razp*, ki presega 0,9, so: *a, ampak, brez, dober, en, glede, isti, kak, kako, kakšen, moj, morati*, *nek*, *nekaj*, *niti*, *njihov*, *noben*, *oz.*, *reči*, *svoj*, *tak*, *tisti*, *torej*, *tvoj*, *vaš*, *veliko*, *vsak*, *zaradi.*

Kolokator	Lema	Frekvenca	Rel. frek.	LogDice	Razp.
festival	družina	81	0,632	6,313	0,093
celje	družina	44	0,344	5,073	0,258
homoseksualen	družina	21	0,164	4,783	0,066
ožji	družina	19	0,148	4,638	0,040
narcisističen	družina	16	0,125	4,431	0,007
janšev	družina	18	0,141	4,314	0,205
prazničen	družina	14	0,109	4,088	0,152
razdreti	družina	10	0,078	3,715	0,033
razpasti	družina	10	0,078	3,611	0,046
priloga	družina	8	0,062	3,313	0,033

Tabela 2: Izpis 10 najvišje rangiranih specifičnih kolokacijskih kandidatov za lemo *družina* (mera logDice).

Čeprav je preučevanje kolokatorjev z relativno visoko vrednostjo *razp* zanimivo z vidika razumevanja spletnega diskurza (npr. kolokatorji *sploh*, *pač, tale*) in nestandardne uporabe leksike (*rabiti* v pomenu *potrebovati*), pomenskih premikov na ravni besed (*hud* v pomenu *dober*) ali zapisa (npr. kratica *slo*), so bolj zanimivi za samostojno preučevanje ali v kombinaciji z njihovi kolokatorji kot z vidika preučevanja kolokacij drugih lem.

Izdelani končni seznami kolokacij uporabniških vsebin vsebujejo par <kolokator, lema>, ki mu pripišemo še dodatne informacije, kot so frekvenca kolokacijskega niza, relativna frekvenca (glede na milijon besed v korpusu Janes), logDice vrednost, kolokatorja z vrednostjo *razp* ter povezavo do konkordanc v korpusu. Končna oblika izpisa je prikazana v Tabeli 2, dodane pa so ji še povezave na konkordance iz korpusa, kar pa zaradi predolgih povezav v prispevku izpuščamo.

3 Rezultati raziskave

Predstavljena metodologija omogoča izdelavo različnih seznamov s poljubnimi nastavitvami (okno, mera kolokabilnosti, minimalna frekvenca, minimalna vrednost mere kolokabilnosti, uporaba statistike *razp*, izključnost[7]). Kot je podrobneje opisano v prejšnjem poglavju v naši raziskavi izberemo parameter za dolžino okna -1, 0, mero logDice in parameter izključnosti. Od nastavitev je odvisna tudi količina izluščenih kandidatov. V Tabeli 3 prikažemo razliko v številu izluščenih kandidatov glede na uporabo različnih nastavitev mejne vrednosti logDice, minimalne frekvence ter uporabe filtriranja z mero *razp.*

Nastavitve	Št. izluščenih kolokacijskih kandidatov
brez omejitev (logDice>0)	11.148
logDice>3	4.183
logDice>3, frek>=10	2.928
logDice>3, frek>=10, *razp*<0.1	1.661

Tabela 3: Število izluščenih kolokacijskih kandidatov, ki se pojavljajo v korpusu Janes in ne v korpusu Kres, za 151 izbranih lem, z različnimi nastavitvami parametrov.

V sorodni študiji (Pollak, 2015) smo glavne kategorije izluščenih kolokacij ločili na aktualno tematiko, spletno tematiko, frazeologijo, lastna imena in specifične kolokatorje. V tem prispevku se osredotočimo predvsem na metodologijo luščenja, za razliko od Pollak (2015) pa iščemo izključno spletne kolokacije splošnega besedišča. Za ponazoritev različnih nastavitev pa se osredotočimo na lemo *družina.*

Strožji kot so kriteriji (glej Tabelo 3), manj kandidatov dobimo v pregled. Na primeru *družina* nam z najstrožjimi nastavitvami iz Tabele 3 (zadnja vrstica) ostanejo naslednje kolokacije:

- *homoseksualna*, *narcisistična*, *ožja družina*
- *razdreti družino*
- *Festival družin*

Še en kandidat, in sicer par <razpasti, družina>, je le delno relevanten, saj polovica konkordanc izhaja iz samostalniške besedne zveze *razpadla družina,* ki pa se dejansko pojavlja v obeh korpusih. Ob iskanju zaporedne sopojavitve lem *razpadel* (pridevnik) in *družina*, najdemo kolokacijo *razpadla družina* v korpusu Kres štirikrat, v korpusu Janes pa trikrat. Kolokacijski par, ki je izluščen kot specifičen za korpus Janes, pa izhaja iz glagola *razpasti* in besede *družina,* vendar je od desetih primerov le v štirih dejansko uporabljen glagol *razpasti,* v šestih pa

[7] Izključnost pomeni, da se kolokacija pojavlja le v enem od dveh korpusov, kar je tudi izbrani pristop v tem članku. V primeru, da ne izberemo parametra izključnosti, dopuščamo presečnost kolokatorjev, določimo pa željeno razliko v vrednosti ključnosti kolokacije med korpusoma.

gre za pridevnik *razpadel*, ki mu je napačno dodeljena glagolska oznaka.

Zaradi filtriranja na podlagi mere *razp* smo izpustili *Praznično Družino* (posebna izdaja revije Družina), *Janševo družino* pa tudi <celje, družina>, ki izhaja iz napačne lematizacije kolokatorja *cel* v kolokaciji *cela družina*. Predvsem posamezni akterji, kot je *Janša*, se pojavljajo v različnih kolokacijah korpusa in jih je v našem primeru smiselno izpustiti.

Z ohlapnejšimi pogoji pojavitev dobimo veliko večji nabor kolokacij, npr. pri edinem pogoju logDice>3 so poleg že omenjenih kolokacij izluščene tudi:

- *raznospolna družina, partnerjeva družina*
- *črkovna družina* (grafika), *modelna družina* (avtomobilizem), *imenska družina*
- *razdirati družino*
- *časopis Družina*, *priloga Družina*, *Prehrana družine*
- *grožnja družini*

Več tovrstnih kandidatov se nahaja v enem samem viru ali pa so deli večbesednih zvez (npr. *grožnja družini* je del niza *grožnja družini Janša*). V nadaljevanju bi bilo smiselno nadgraditi pristop še z možnostjo filtriranja zadetkov, ki se pojavljajo le pri enem samem viru oz. domeni.

V primeru, da opustimo vse omejitve (logDice=0, min. frekvenca=1), je kolokacij, v katerih nastopa lema *družina* in se pojavljajo izključno v korpusu Janes, kar 180. Nekateri izmed naštetih pridevnikov, ki določajo družino so *homoseksualna*, *oligarhična*, *čefurska*, *gejevska*, zelo splošni kolokatorji so *super* in *ok,* najdemo pa tudi veliko pogostih besed, ki z lemo ne tvorijo sintaktičnih in semantičnih kolokacij (npr. *potem*, *kdo*, *pač* in *sploh*).

Podrobneje smo si ogledali primer kolokacij besede *družina.* V Tabeli 4 navajamo še kolokatorje petih drugih lem, ki so bili izluščeni z najstrožjimi pogoji (nastavitve v zadnji vrstici Tabele 3). Kot vidimo, je veliko neformalnega izrazja, vezanega na sodobne tematike (*zajebati*, *pokrasti državo*), vidimo pa tudi probleme avtomatskega luščenja kolokacij iz uporabniških spletnih besedil, saj je npr. odsotnost strešic pri zapisu vzrok za vrsto izluščenih kolokacij (npr. *smetišče zgodovine* je pogost frazem tudi v korpusu Kres, vendar ga zapis brez strešic prikaže kot specifičnega za spletno slovenščino). Med primeri vidimo tudi nestandardni zapis okrajšav brez ločil (*ang jezik*), luščenje lem na podlagi napačne lematizacije[8] (*tuja* namesto *tuj* v kolokaciji *tuj jezik*; <foto, mama> iz napačne lematizacije besede *foot,* ki ga vsebuje ime *Big foot mama*, <kurac, mama> pa iz ekspresivnega izraza *poln kurac mam* oz. *koji kurac mamo*). Korpus uporabniških spletnih vsebin je tudi bogat vir za preučevanje idiomatike (par <jezik, muca> iz Tabele 4 izvira iz izraza *muca jezik papala* (oz. *popapala, popapcala* ali celo *muca jezik papne*).

Nekatere pogoste napake predprocesiranja bi bilo mogoče odpraviti z izboljšanjem orodij za lematizacijo (učenje nad večjim označenim spletnim korpusom, boljša standardizacija), postopek rediakritizacije za pripis izpuščenih strešic.

Lema	Kolokatorji
država	vitek, zadolžiti, rušiti, pravično, ugrabiti, koruptiven, zavoziti, gnili, pokrasti, ugrabljen, zajebati
jezik	muca, eksotičen, tuja, ang, venetski, šparati
mama	foto, yo, platišče, mreža, odkar, doječ, vreme, kurac, feltna
moški	hetera, napasti, feminizacija, kastracija
zgodovina	servisen, preverljiv, smetisce, spisati, retuširanje, brisanje, odpasti

Tabela 4: Izluščeni kolokatorji za izbor petih lem.

4 Zaključek in nadaljnje delo

V prispevku smo prikazali metodo luščenja kolokacij, ki se pojavljajo le v izbranem korpusu (v našem primeru korpusu uporabniških spletnih vsebin) in ne v referenčnem korpusu. Z izdelanim orodjem za hiter dostop do kolokacij posameznega korpusa (API za dostop do korpusa na lokalni instalaciji orodja SketchEngine) lahko izvozimo kolokacije poljubnih lem. Nato seznama kolokatorjev vsake leme v različnih korpusih med seboj križamo in ohranimo le tiste kolokacijske kandidate, ki se pojavljajo izključno v specifičnem korpusu.

Z različnimi strožjimi nastavitvami parametrov frekvence besed, mere kolokabilnosti in mere specifičnosti kolokatorja na podlagi razpršenosti po seznamih (*razp*) smo iz prvotnega seznama z nad 11.000 kolokacijskimi kandidati nabor skrčili na cca. 1.600 specifičnih kolokacij. Metoda je uporabna za analizo diskurza, leksikografske naloge (sprotno dopolnjevanje drugih s specifičnimi vsebinami) ali pri poučevanju slovenščine kot tujega jezika, kjer je potrebno poznavanje in ločevanje formalnih in neformalnih načinov izražanja.

V nadaljnjem delu je potrebno predvsem izboljšati sama orodja za predprocesiranje (lematizacija), rediakritizacija pa bi pomagala luščiti bolj relevantne leksikalne specifike. S predstavljeno metodo bomo izluščili sezname kolokatorjev tudi za druge besedne vrste, zanimiva pa bi bila tudi primerjava z govornim korpusom.

5 Zahvala

Za implementacijo API vmesnika se zahvaljujem Borutu Lesjaku. Raziskava je bila opravljena v okviru projekta »Viri, orodja in metode za raziskovanje nestandardne spletne slovenščine« (J6-6842, 2014-2017), ki ga financira ARRS.

[8] Evalvacije v tem prispevku nismo naredili, smo pa v prispevku Pollak (2015) ocenili, da je več kot 35 % izluščenih kandidatov neprimernih zaradi predprocesiranja in sestave korpusa.

6 Literatura

Morton Benson. 1989. The structure of collocational dictionary. *The International Journal of Lexicography*, 2: 1–14.

Sabine Bartsch. 2004. *Structural and functional properties of collocations in English. A corpus study of lexical and pragmatic constraints on lexical co-occurrence.* Tübingen, Verlag Gunter Narr.

Miriam Buendía Castro in Pamela Faber. 2014. Collocation Dictionaries: A Comparative Analysis. *MonTI. Monografías de Traducción e Interpretación* 6: 203–235.

Anthony. P. Cowie. 1994. *Phraseology. Encyclopedia of Language and Linguistics* (6. zv). Oxford in New York. 3168–3171.

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

Lee R. Dice. 1945. Measures of the Amount of Ecologic Association Between Species. *Ecology* 26(3): 297–302.

John R. Firth. 1957. Modes of Meaning. Frank R. Palmer (ur.): *Papers in Linguistics 1934-51*, str. 190–215. London: Oxford University Press.

Darja Fišer, Tomaž Erjavec, Ana Zwitter Vitez in Nikola Ljubešič. 2014. JANES se predstavi: metode, orodja in viri za nestandardno pisno spletno slovenščino. V: *Zbornik 9. konference Jezikovne tehnologije*, str. 56–61.

Polona Gantar in Simon Krek. 2011. Slovene Lexical Database. V: *Zbornik 6. konference Natural language processing, multilinguality*, str. 72–80.

Laurie Gerber in Jin Yang. 1997. Systran MT dictionary development. V: *Proceedings of Past, Present, and Future: Machine Translation Summit 6,* str. 211–218.

Vojko Gorjanc in Primož Jurko. 2004. Kolokacije in učenje tujega jezika. *Jezik in slovstvo* 49(3/4): 49–62 .

Michael Alexander Kirkwood Halliday. 1966. Lexis as a Linguistic Level. *In Memory of F. R. Firth*. Longman.

Kjellmer, Göran. 1987. Aspects of English Collocations. *Corpus linguistics and Beyond*. Amsterdam, Atlanta: Rodopi.

Franz Josef Hausmann. 1989. Le dictionnaire de collocations. V: Franz Josef Hausmann, Oskar Reichmann, Herbert Ernst Wiegand in Ladislav Zgusta (ur.): Wörterbücher (3 zvezki). Berlin: Walter de Gruyter.

Iztok Kosem, Polona Gantar in Simon Krek. 2013. Avtomatizacija leksikografskih postopkov. *Slovenščina 2.0* 1(2): 139–164.

Adam Kilgarriff, Pavel Rychly, Pavel Smrz in David Tugwell. 2004. The Sketch Engine. V: *Proceedings of EURALEX* 2004, str. 105–116.

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

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

Nadja Nesselhauf. 2005. *Collocations in a Learner Corpus*. Amsterdam: John Benjamins Publishing.

Maria V. Noblia. 1998. The Computer-Mediated Communication: A New Way of Understanding The Language. V: *Proceedings of Internet Research and Information for Social Scientists Conference*, str. 10–12.

Adriane Orenha-Ottaiano. 2012. English collocations extracted from a corpus of university learners and its contribution to a language teaching pedagogy. *Language and Culture* 34 (2): 241–251.

Senja Pollak in Špela Arhar Holdt. 2015. Identifying Corpus-specific Collocations: The Case of Spoken Slovene. V: *Proceedings of 8th International Conference on Natural Language Processing, Corpus Linguistics, Lexicography* (Slovko 2015). RAM-Verlag, str. 117–125.

Senja Pollak. 2015. Luščenje kolokacij iz korpusa uporabniških spletnih vsebin. *Zbornik 34. simpozija Obdobja. Slovnica in slovar – aktualni jezikovni opis.*

Ina Rösiger, Johannes Schäfer, Tanja George, Simon Tannert, Ulrich Heid in Michael Dorna. 2015. Extracting terms and their relations from German texts: NLP tools for the preparation of raw material for specialized e-dictionaries. V: *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference*, str. 486-503

Pavel Rychlý. 2008. A Lexicographer-Friendly Association Score. V: *Proceedings of Recent Advances in Slavonic Natural Language Processing*, str. 6–9.

Violeta Seretan. 2015. Multi-Word Expressions in User-Generated Content: How Many and How Well Translated? Evidence from a Post-editing Experiment. V: *Proceedings of the Second Workshop on Multi-word Units in Machine Translation and Translation Technology (*MUMTTT 2015), Malaga, Spain.

Dirk Siepmann. 2005. Collocation, Colligation and Encoding Dictionaries. Part I: Lexicological Aspects. *International Journal of Lexicography* 18: 409–443.

John Sinclair. 1966. Beginning the Study of Lexis. *In Memory of F. R. Firth*. London: Longman.

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

Karen Spärck Jones. 1972 (2004). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28(1): 11–22. Ponatis: *Journal of Documentation* 60 (5): 493–502.

Darinka Verdonik, Iztok Kosem, Ana Zwitter Vitez, Simon Krek in Marko Stabej. 2014. Compilation, transcription and usage of a reference speech corpus: The case of the Slovene corpus GOS. *Language resources and evaluation* 47(4): 1031-1048

Špela Vintar. 2010. Bilingual term recognition revisited. *Terminology* 16(2): 141–158.

Ning Yu. 2014. Sentiment analysis in UGC. Marie-Francine Moens, Juanzi Li, Tat-Seng Chua (ur.): *Mining User Generated Content*, str. 43–66. CRC Press. Taylor and Francis book.

Slovenščina pod palcem interneta: vezajne in dvozačetniške e-tvorjenke

Teja Rebernik

Groningen, Nizozemska
teja.rebernik@gmail.com

Povzetek

Internet in nove tehnologije so v slovenščini spodbudili pojavitev novega nesistemskega besedotvornega postopka, s pomočjo katerega je v zadnjem desetletju nastalo več kot tisoč novih tvorjenk z enoglasovnim krnom, in sicer vezajne e-tvorjenke, zapisane z vezajem, ter dvozačetniške e-tvorjenke, zapisane s medbesedno veliko začetnico. Morfemu *e-*, ki je najpogostejši in katerega podstava nosi v e-tvorjenkah pomen »spletni«, »elektronski«, »električni«, »ekološki« ali »energijski«, se pridružujejo tudi drugi morfemi, najpogosteje *i* v dvozačetniških e-tvorjenkah. Podan bo premislek o sistemskosti teh tvorjenk ter narejen pregled najpogostejših vezajnih in dvozačetniških e-tvorjenk, prav tako pa bodo omenjena lastna imena istega zapisa.

The Slovene Language Under the Internet's Thumb: Hyphenated and Bicapitalized E-words

The Internet and new technologies encouraged the emergence of a new, unsystematic pattern of word-formation. In the last decade, more than a thousand new words with one-sound truncation have been formed through this pattern, namely the so-called »vezajne e-tvorjenke«, written with a hyphen, and »dvozačetniške e-tvorjenke«, written with a capital letter mid-word. The morpheme e- is the most frequent, its meaning derived either from »Internet-related«, »electronic«, »electric«, »ecological« or »energy-related«, but other morphemes appear as well, most often the morpheme i. The systematic character of these newly formed words will be considered and an overview of the most frequent hyphenated and bicapitalized »e-words« will be done. Proper names of the same notation will be mentioned as well.

1 Uvod

Z razvojem novih tehnologij in prevlado interneta so se v zadnjih letih pojavili tudi novi (nesistemski) besedotvorni postopki ter zapis novonastalih tvorjenk, predvsem kadar gre za besede, ki se nanašajo na splet oziroma elektroniko in se najpogosteje uporabljajo prav v spletnih zapisih. Vendar pa nesistemski besedotvorni postopki in t. i. e-tvorjenke, ki jih prispevek obravnava, niso vezani le na besede z enoglasovnimi krni e-, m- in i- (kot je o tem že pisala Logar, 2004), temveč so se pojavili tudi zapisi, pri katerih vezaj zamenja velika začetnica podstave (tipa *eDavki* namesto *e-davki*).

V prispevku predstavljamo nesistemske tvorjenke z enoglasovnim krnom, in sicer vezajne e-tvorjenke (tipa *e-pošta*, *e-račun*) in dvozačetniške e-tvorjenke (tipa *eDnevnik*, *mCobiss*), prav tako pa omenjamo lastna imena izdelkov istega zapisa (tipa *iPhone*, *e-HDi*). Na začetku naredimo kratek pregled sistemske tvorjenosti in netvorjenosti, nadaljujemo s predstavitvijo naše analize različnih nesistemskih tvorjenk z enoglasovnim krnom in pregledom lastnih imen izdelkov istega zapisa, prav tako pa se povprašamo o problemih uporabnikov in tvorcev novih besed ter o pomanjkanju aktualiziranih kodifikacijskih priročnikov.

2 E-tvorjenke in njihova (ne)sistemskost

Slovenščina je bila v zadnjih letih podvržena veliko spremembam, predvsem v besedišču, ki se nanaša na internet in elektroniko. Na eni strani uporabniki to besedišče namreč izrazito uporabljajo, na drugi pa se soočajo s težavami pri zapisu in uporabi t. i. e-tvorjenk.

Novonastale vezajne in dvozačetniške e-tvorjenke so tako relativno neraziskan del slovenske leksike, zato smo želeli narediti jasen pregled tistih, ki se pojavljajo najpogosteje, ter jih razvrstiti v kategorije, ki bodo v pomoč bodočim raziskovalcem na tem področju.

2.1 E-tvorjenke

V prispevku obravnavamo nesistemske tvorjenke z enoglasovnim krnom oziroma (krajše imenovane) e-tvorjenke dveh tipov, in sicer č-podstava (vezajne e-tvorjenke) ter čPodstava (dvozačetniške e-tvorjenke). Gre za dvobesedne levoprilastkovne besedne zveze, pri katerih pa pomen enoglasovnega krna, ki predstavlja začetno črko prve besede (pri nas *č*), ni več nujno transparenten, temveč lahko o njem sklepamo le po drugi besedi (pri nas *podstava*) in kontekstu. To, kar imenujemo e-tvorjenke, so pravzaprav okrajšane besedne zveze, saj lahko z razširjenostjo interneta opazimo tudi novo tendenco, da je tisto, kar je zapisano krajše, posledično primerneje in bolje prilagojeno spletnemu mediju in načinu komunikacije.

2.2 Sistemska tvorjenost

Toporišič (2004) tvorjene besede razdeli na štiri besedotvorne vrste, vendar pa nesistemskih tvorjenk z enoglasovnim krnom ne moremo uvrstiti v nobeno od njih. Ker so v svojem bistvu okrajšane besedne zveze, ne morejo biti niti izpeljave niti zloženke (manjka jim obrazilo). Morda so e-tvorjenke, kot pravi Logar (2004), zaradi neobstoječega obrazila podobne kraticam, vendar je pri e-tvorjenkah podstava krnjena predvidljivo, dodano pa je jedro podstavne besede ali besedne zveze (tako *elektronska pošta* postane *e-pošta*, *elektronska knjiga* postane *e-knjiga* itd.), kar se razlikuje od kratičnih zloženk, kjer je podstavna beseda krnjena in krni združeni (npr. spletne kratice, kot sta *LP* za *lep pozdrav* ali *RTM* za *rad/a te (i)mam*).

E-tvorjenke tudi niso sestavljenke, saj se slednje ne udejanjajo z vezajem ali medbesedno veliko začetnico, ohranjajo pa svoje slovnične značilnosti in pomensko transparentnost (medtem ko za e-tvorjenke tega ne moremo zanesljivo trditi). Zadnja kategorija po Toporišiču so sklopi, od katerih pa se e-tvorjenke razlikujejo, saj se dela besedne zveze ne združita, temveč se prva beseda okrajša ter na drugo besedo veže z vezajem.

Vidovič Muha (2011: 331) skladenjsko podstavo definira kot »nestavčno podredno besedno zvezo (s predvidljivo izjemo), katere predmetno- in slovničnopomenske sestavine so pretvorljive v tvorjenko«, kar nam, kot izpostavi tudi Logar (2004), pove, da pri e-

tvorjenkah ne moremo govoriti niti o skladenjski podstavi niti o besedotvorni podstavi in obrazilu.

2.3 Nesistemska tvorjenost

Logar (2004) prva na slovenskem področju definira in se posveti tvorjenkam »z enoglasovnim krnom, ki bi jih lahko pogojno imenovali e-tvorjenke«, četudi se že takrat pojavljajo tudi take z morfemom *m-* (danes bolj kot ne zastarel) in *i-*. V prvem delu prispevka tako e-tvorjenke definira kot tvorjenke z enoglasovnim krnom, v katerih je »podstava samostalniška stalna besedna zveza s pridevniškim prilastkom«. Posebej poudari, da vsekakor gre za *nesistemski* besedotvorni postopek, saj *e-* sam po sebi nima pomena (tj. ni simbol s tem pomenom), v zloženki nima pomena razreda (tipa *C-vitamin*), hierarhičnosti (tipa *D-dur*) ali razločevalne vrednosti (tipa *C-disk*). Logar (2004) primerja e-tvorjenke z nemškim tipom tvorjenja zloženk tipa *U-bahn*, s katerimi jih ne moremo enačiti, lahko pa vsekakor vidimo podobnost v tem, da se oba tipa vedno krnita do prve črke.

Od časov zapisa Logar (leto 2004) do danes (leto 2015) se je povečal doseg interneta in elektronskih tehnologij, posledično pa je to prineslo tudi veliko novih besed tipa *č-podstava* in *čPodstava.* Za razliko od približno 70 primerov vezajnih e-tvorjenk, ki jih našteva Logar (2004), smo jih mi v korpusu Janes (več o korpusu spletne slovenščine Janes v Fišer et al. (2014)) našli več kot 1000, od tega skoraj 700 takih, ki se pojavijo dvakrat ali več. Besed, ki se začnejo z enoglasovnim krnom *e-*, je največ, in sicer približno 600 takih, ki se pojavijo več kot dvakrat, kar je skoraj desetkrat več kot 65 primerov, ki jih je leta 2004 našla Logar.

Tudi pomeni teh e-tvorjenk so se razširili, saj so se pomenom krna *e-*, ki jih je našla Logar (tj. elektronski, informacijski, internetni) pridružili tudi drugi (kot npr. električni, energijski, ekološki). Na drugi strani se je pojavil tudi alternativni zapis e-tvorjenk (tj. dvozačetniške e-tvorjenke), prav tako pa lahko najdemo skoraj vse morfeme, ne le *e-, m-* in *i-*.

3 Nesistemske tvorjenke z enoglasovnim krnom

3.1 Metodologija

Analizirali smo nesistemske tvorjenke z enoglasovnim krnom v korpusu spletne slovenščine Janes v0.3, in sicer smo v Corpus Query Language (CQL) iskali po naslednjih parametrih:

- [word="(?i)[a-zčšž]-.*"] (vezajne e-tvorjenke tipa *č-podstava* in *Č-podstava*)
- [word="[a-zčšž][A-ZČŠŽ].*"] (dvozačetniške e-tvorjenke tipa *čPodstava*)

Pri tem je bilo v prvem iskalnem nizu 28.301 zadetkov in smo analizirali vse vezajne e-tvorjenke, ki so se pojavile dvakrat ali več (680), v prispevku pa podrobneje obravnavamo take s pet ali več pojavitvami (318). V drugem iskalnem nizu je bilo 27.914 zadetkov, vendar pa od tega le 157 dvozačetniških e-tvorjenk s pet ali več pojavitvami.

Ker smo za analizo uporabili korpus spletnih uporabniških vsebin, se vsi rezultati nanašajo na slovenščino, uporabljeno v spletnih medijih (npr. forumih, socialnih omrežjih), kjer imajo uporabniki pri izbiri jezika pogosto prostejše roke kot drugod. Če bi želeli ugotoviti, kako pogosto se vezajne in dvozačetniške e-tvorjenke pojavljajo v slovenščini nasploh, bi morali analizirati tudi druge vire, kot je na primer korpus GigaFida[1] (več o korpusih slovenskega jezika v Logar et al. (2012)).

3.2 Vezajne e-tvorjenke

134 e-tvorjenk z enoglasovnim krnom *e-* se pojavi petkrat ali več. Med analizo smo izločili tiste, pri katerih enoglasovni krn *e-* nakazuje kategorijo, znamko ali razred (npr. avto *Audi e-tron*, fotoaparat *E-M5*, vrsta avtomobilov *E-class* itd.) pa tudi tiste tvorjenke, kjer je *e-* del tujega lastnega imena (najpogosteje *e-Bay* in *e-Leclerc*). Ostalo nam je 117 e-tvorjenk, ki bi jih lahko razdelili na samostojne besede (tipa *e-pošta*), imena spletnih strani, portalov in forumov (tipa *e-fotograf*) ter imena izdelkov, dogodkov, projektov itd. (tipa *E-dem*). Prav tako smo določili pomen enoglasovnega krna, in sicer smo našli 5 prevladujočih pomenov, tj. *elektronski*, *spletni*, *električni*, *ekološki* in *energijski.*

Pojavitve	Beseda
1699	e-mail, e-mailing, e-majlirati
722	e-pošta, e-poštni
420	e-knjiga, e-knjigica
301	e-naslov
248	e-novice, e-novičke
242	e-volitve
181	e-račun
151	e-trgovina, e-trgovinica
140	e-uprava, e-upraven, e-upravljanje
114	e-fotograf, e-fotografija

Tabela 1: 10 najpogostejših vezajnih e-tvorjenk v korpusu Janes.

3.2.1 Pomen *elektronski* in *spletni*

E-tvorjenk z morfemom *e-* v pomenu *elektronski* ali *spletni*[2] je bilo največ, moramo pa poudariti, da razlika med pomenoma pri številnih e-tvorjenkah ni očitna. Le pri redkih e-tvorjenkah lahko namreč govorimo samo o enem ali drugem pomenu, saj se prepletata in je tisto, kar je elektronsko, velikokrat tudi spletno in obratno.

Veliko pomembneje je, da poznamo kontekst posameznih e-tvorjenk in razlikujemo med e-tvorjenkami, ki predstavljajo imena portalov, spletnih strani in dogodkov na eni strani ter e-tvorjenkami, ki so samostojne besede na drugi. Razlika je pomembna, ker so prve govorci prisiljeni uporabljati, saj nekaj poimenujejo, druge pa so tiste besede, ki so jih govorci sprejeli in jih uporabljajo, ker jim ustrezajo. Zato so v prispevku lastna imena portalov, spletnih strani, dogodkov in podobno označena z nadpisano črko l (npr. *e-študentski*[l]), lastna imena izdelkov pa predstavljamo v posebnem razdelku.

V korpusu Janes v0.3 se pojavlja okoli 600 besed z enoglasovnim krnom *e-*, ki nosi pomen *elektronski* ali *spletni*, mi pa bomo našteli le tiste, ki imajo 5 ali več

[1] Dostop do korpusa: www.gigafida.net.

[2] Pri Logar (2004) »internetni«.

pojavitev (109) ter najbolj zanimive od e-tvorjenk s 4 ali manj pojavitvami. Besede in njihove izpeljanke (npr. *e-pošta* in *e-poštni*) smo združili in njihove pojavitve sešteli. To, da ne nastajajo le besede, temveč tudi izpeljanke iz prvotnih e-tvorjenk, nam kaže tako živost jezika kot tudi uveljavljanje in vse pogostejšo uporabo novega nesistemskega besedotvornega postopka.

Pomen *elektronski* in *spletni* z več kot 100 pojavitvami (11 besed in izpeljank): *e-mail* (*e-mailing*, *e-majlirati*, *e-mail*), *e-pošta* (*e-posta*, *e-poštni*),[3] *e-knjiga* (*e-knjigica*), *e-naslov*, *e-novice* (*e-novičke*, *e-novičnik*), *e-volitve*, *e-račun*, *e-trgovina* (*e-trgovinica*, *e-trgovinčka*), *e-uprava*[l] (*e-upraven*, *e-upravljanje*), *e-bančništvo* (*e-bančni*, *e-banka*, *e-banka*), *e-fotograf*[l] (*e-fotografija*[l]).

Pri tem je potrebno poudariti, da se največkrat pojavita prav *e-mail* (1699) in *e-pošta* (722) ter njune izpeljanke. Medtem ko Logar (2004) govori o slovenskem prevodu *e-pošta*, lahko sedaj opazimo, da se je anglicizem uveljavil, slovenska ustreznica pa je, kljub nezanemarljivi pojavnosti, manj pogosta, vsaj v korpusu Janes. Omeniti moramo tudi, da ima alternativni zapis *email* (tudi *eMail*) v korpusu še dodatnih 2232 pojavitev, medtem ko ima zapis *epošta* le dodatnih 40, celotna izpisana besedna zveza *elektronska pošta* pa dodatnih 58.

Seveda pa je treba pripomniti, da morda ne gre le za postopno uveljavitev anglicizma, temveč je do razlike prišlo tudi zaradi besedil in virov, ki smo jih uporabili. Korpus Janes je, kot že omenjeno, večinoma sestavljen iz besedil v spletni slovenščini, medtem ko je Logar (2004) uporabila gradivo iz časopisov in revij, kot so Delo, Mladina, Mobinet in Joker. Če zgoraj omenjena primera analiziramo v referenčnem korpusu Gigafida, lahko opazimo veliko razliko, saj ima beseda *elektronska pošta* 26.838 pojavitev, medtem ko se *e-mail* pojavi v 21.294 primerih, pa še to le v kontekstu naštevanja osebnih in kontaktnih podatkov.

Pomen *elektronski* in *spletni* z 20 do 100 pojavitvami (23 besed in izpeljank): *e-bralec* (*e-bralnik*, *e-branje*), *e-poslovanje*, *e-oblika*, *e-storitev*, *e-revija*, *e-sporočilo*, *e-gradivo*, *e-davek*[l], *e-demokracija* (*e-demokratičen*), *e-participacija*, *e-cigareta (e-cigaretni)*, *e-študentski*[l,4], *e-učbenik*, *e-izobraževanje*, *e-obrazec*, *e-obvestilo* (*e-obveščanje*, *e-obveščevalec*), *e-šolstvo*[l], *e-glasovanje* (*e-glasovati*, *e-glasovnica*), *e-ink*, *e-časopis* (*e-časnik*), *e-vem*[l], *e-dnevnik* (*e-Dnevnik*[l,5] *e-dnevnikar*), *e-družba*.

Pomen *elektronski* in *spletni* z 10 do 20 pojavitvami (25 besed in izpeljank): *e-streznik*[l], *e-verzija*, *e-voščilnica* (*e-voščilo*), *e-prijava*, *e-vsebina*, *e-book*,[6] *e-kartica*, *e-prostor*[l], *e-stave*, *e-tečaj*, *e-kozmetika*[l], *e-opomnik*, *e-knjigarna*, *e-komunikacija* (*e-komuniciranje*), *e-svetovalnica*[l], *e-učenje*, *e-commerce*, *e-članek*, *e-izdaja*, *e-listovnik*, *e-pismo*, *e-seminar*, *e-asistent*, *E-dem*[l], *e-zdravje*.

Pomen *elektronski* in *spletni* s 5 do 9 pojavitvami (49 besed in izpeljank): *e-informator*, *e-knjižnica*, *e-kompetenten*, *e-medij*, *e-mesečnik*, *e-nabiralnik*, *e-pismen* (*e-pismenost*), *e-vloga, e-zemljiška*[l,7] *e-brošura*, *e-krog*[l], *e-naročanje* (*e-naročila*, *e-naročnik*), *e-papir*, *e-predal*, *e-priročnik*, *e-redovalnica*, *e-sistem*, *e-veščina*, *e-klic*, *e-podpis*, *e-priprava*, *e-recept*, *e-učilnica*, *e-zavarovanje*, *e-bonton*, *e-demokracija*[l], *e-gold*[l], *e-learning*, *e-lista*, *e-naprava*, *e-portal*, *e-pravo* (*e-pravosodje*), *e-različica*, *e-trgovec*, *e-arhiv*, *e-čestitka*, *e-črnilni*, *e-dokument*, *e-indeks* (alternativno: *e-index*), *e-kompetentnost*, *e-Leader*[l], *e-marketing*, *e-parlament*, *e-razred*, *e-reader*, *e-študent*, *e-študij*, *e-zloraba.*

Če pogledamo e-tvorjenke, ki se pojavijo 4-krat ali manj, takoj opazimo več imen spletnih strani in portalov (npr. *e-Delo*, *e-Večer*, *e-FotoPOTEP*, *e-nefiks*, *e-kamini*, *e-Rodna*, *e-justice*, *e-kupon* itd.) ter anglicizmov (npr. *e-card*, *e-pen*, *e-tax* itd.) in različnih izdelkov (npr. *e-buddy*, *e-light*, *e-studio* itd.). Vendar prevladujejo druge, samostojne besede, ki bodisi kažejo ustvarjalnost enega uporabnika bodisi jo uporablja več ljudi. Tako imamo besede, kot so npr. *e-pisanje*, *e-država*, *e-državljan*, *e-izvod*, *e-poročilo*, *e-položnica*, *e-kompetenca*, *e-omrežje*, *e-lekarna*, *e-publikacija*, *e-smetišče*, *e-šport, e-slama*,[8] *e-identiteta* in na stotine drugih.

Pojavljajo se tudi slovenski prevodi uveljavljenih anglicizmov (npr. *e-črnilo* kot manj uveljavljena alternativa anglicizma *e-ink*). Na drugi strani pa so zanimive tudi pogovorne besede, ki se niso izmaknile temu novemu besedotvornemu postopku (npr. *e-klapa*, *e-bukve*, *e-kas(t)lc*, *e-cajteng*, *e-biznis*), in žaljivke, kot so npr. *e-bebavost*, *e-bedak*, *e-tič*.[9]

3.2.2 Drugi pomeni enoglasovnega krna *e-*

Razširjenost obravnavanega besedotvornega postopka kaže tudi pojavitev drugih pomenov morfema *e-*. V pomenu *električni* se morfem *e-* v e-tvorjenkah pojavi v besedah *e-avto* (12), *e-golf* (9), *e-kolo* in *e-bike* (5 in 4), *e-skuter* (5), *e-mobilnost* in *e-odpadek*[10] (4); v pomenu *energijski* se pojavi v *e-svet* (alternativni zapis eSvet)[11] (14) in *e-ples*[12] (6); v pomenu *ekološki* se pojavi predvsem v izdelkih, kot sta npr. motor *e-Hdi* (34) ali *e-gas* (10); v eni besedi pa se pojavi tudi v pomenu *eksperiment* (*e-hiša*[l]).

Pri tem lahko zaznamo, da je nekaterim uporabnikom krn *e-* postal dobrodošel način okrajšave besednih zvez, saj Evrosong postane *E-song*, Evropska unija se spremeni v *E-Unijo*, elektronska doba v *e-dobo* itd.

3.3 Vezajne e-tvorjenke z drugimi enoglasovnimi krni

3.3.1 M-tvorjenke

Za razliko od Logar (2004), ki je našla 6 primerov e-tvorjenk z morfemom *m-*, smo mi ugotovili, da ga je v

[3] Najpogosteje v kolokaciji »e-poštni naslov«.
[4] Govor je o »e-študentskem servisu«.
[5] Govor je o spletni verziji časnika Dnevnik.
[6] Anglicizem z manjšo pogostostjo kot njegova slovenska verzija e-knjiga.
[7] Gre za e-zemljiško knjigo.
[8] V pomenu angleške besede »spam«.
[9] V kontekstu »Ne mislim e-tiča primerjat«.
[10] Tvorjenka »e-odpadek« se nanaša na odpadno električno opremo, baterijo itd.
[11] Spletna stran, ki odgovarja na vprašanja s področja energije in energetike.
[12] Ime sproščujočega plesnega programa.

veliki meri nadomestil morfem *e-* (tako je npr. Logar pisala o *m-bančništvu*, *m-poslovanju* in *m-prodaji*, mi pa o *e-bančništvu, e-poslovanju* in *e-prodaji*). Kljub temu pa je nekaj e-tvorjenk z morfemom *m-*, ki se pojavijo 4-krat ali več, in sicer: *M-teorija*,[13] *M-Tech*, *M-tuning* in *M-styling* (*m-* v treh primerih v pomenu moto) ter *m-vstopnice* (*m-* v pomenu Moneta).

3.3.2 I-tvorjenke

Logar (2004) je odkrila le eno e-tvorjenko z morfemom *i-* (in sicer *iprogramček*, ki se v naši analizi ni pojavil), mi pa smo v korpusu Janes našli 37 takih, ki se pojavijo 2-krat ali več, pri čemer je 6 samostojnih besed, 31 pa lastnih imen izdelkov. Od tega jih je kar 10 povezanih z izdelki podjetja Apple, pri čemer gre pri petih primerih za napačno črkovanje izdelka (zapisi *i-phone*, *i-pod*, *i-pad*, *i-tunes* in *i-fon*), pri petih pa za tvorjenke, kjer se morfem *i-* pojavi v pomenu Apple izdelek (in sicer *i-naprava*, *i-dioti*, *i-igračka*, *i-produkt* in *i-zadeva*).

Morfem *i-* se pojavi tudi v pomenu *internetni*, najpogosteje v naslednjih e-tvorjenkah: *i-net* (okrajšava za *internet*), *i-volitev*, *i-učbenik*, *i-Računi*, *i-Slovar*[1] in *i-tabla*.

3.3.3 Drugi enoglasovni krni

V primerjavi z enoglasovnimi krni *e-, m-* in *i-* imajo preostale črke veliko manj pojavitev, predvsem pa je malo takšnih, ki niso povezane z znamkami vozil (tipa *X-mini* ali *C-Airdream*), kategorijami (tipa *C-liga* ali *R-klasa*) ter izdelki (tipa *X-porter*) in lastnimi imeni, v katerih črka ne nosi posebnega pomena (tipa *A-kanal* ali *C-span*). Izločili smo tudi angleška lastna imena (tipa *D-pad* ali *s-video*).

Besed, ki se pojavijo 4-krat ali več in ustrezajo tem parametrom, smo tako našli 16. Več kot 10-krat se pojavijo naslednje tvorjenke: *n-ta* (nešteta), *n-let* (neznano), *V-racing* (Velenje), *n-tem* (nevemkolikem); manj kot 10-krat pa naslednje: *A-kanalizacija* (leti na A-kanal), *A-cosmos* (avto), *F-trendi* (povezano s časnikom Finance), G-Fart[1] (gledališki), *G-sila* (gravitacijska), *C-forum*[1] (Citroen forum), *n-mesecev* (neznano), *P-garaža* (parkirna hiša), *a-c* (avtocesta), *A-infoshop*[1] (anarhistični), *v-izrez* (izrez majic v obliki črke v), *x-firma* (poljubna).

3.3.4 Zanimivi primeri s 3 ali manj pojavitvami

Veliko tvorjenk z enoglasovnim krnom, ki se v korpusu Janes v3.0 pojavijo 3-krat ali manj, je odvisnih od osebne odločitve posameznika, kar pripelje do zanimivih e-tvorjenk, ki kažejo tendenco ustvarjanja domiselnih novotvorjenk in okrajšav tam, kjer je to le mogoče. Tako lahko najdemo tvorjenke, kot so: *A-Maze* (alternativni zapis *aMaze*), *p-hiša* (parkirna hiša), *g-pogovor*, *g-račun* in *g-maps* (vse troje Google) pa tudi *ž-klub* in *ž-reprezentanca* (ženski/ženska).

3.4 Dvozačetniške e-tvorjenke

3.4.1 Dvozačetniške e-tvorjenke z morfemom *e*

Zaradi nestandardnosti e-tvorjenk in pomanjkanja obravnave v kodifikacijskih priročnikih, pa tudi zato, ker se v spletnih naslovih ustvarjalci radi izognejo vezajem, lahko najdemo tudi dvozačetniške e-tvorjenke, pri katerih je podstava zapisana z veliko začetnico. Razen izdelkov podjetja Apple (tipa *iNekaj*) je največ dvozačetniških e-tvorjenk prav takih, ki se začnejo z enoglasovnim krnom *e*.

E-tvorjenke z enoglasovnim krnom *e* in veliko začetnico, ki imajo 10 ali več pojavitev: *eRačun*, *ePonedeljek*, *eZdravje*, *ePlay*[1], *eTrgovina*, *eVolitev*, *ePub*, *eTri*[1], *ePolicist*[1].

E-tvorjenke z enoglasovnim krnom *e* in veliko začetnico, ki imajo 9 ali manj pojavitev: *eVino*[1], *ePoslovanje*, *eSlog*, *eUprava*[1],[14] *eSATA*[1], *ePER*[1], *eRedovalnica*, *eTwinning*[1], *eVEM*, *eZPIZ*[1], *ePosavje*[1], *eReader*, *eRegion*, *eRepublika*, *eSkiing*, *eSkills*, *eŠport*, *eVročanje*.

3.4.2 E-tvorjenke z drugimi morfemi

Kakor hitro izločimo imena izdelkov (tipa *iRobot*, *nVidia*), aplikacij (tipa *iO*, *iQpon*), Applovih izdelkov, znamk avtomobilov in z avtomobili povezane tehnologije (tipa *xDrive*, *iX20*), podjetij (tipa *iPROM*, *bHIP*) in anglicizmov (*sRGB*, *mIRC*), nam ostane le 25 dvozačetniških e-tvorjenk.

Največkrat se pojavi slovenski turistični zapis *sLOVEnija* (tudi *sLOVEnia*, *sLOVEnščina*, *sLOVEnec*, pa tudi priredbe, kot sta *sLOLvenija* in ime *sLOVErotika*).[15] Podobno poznamo tudi navijaški slogan *aMaze* (tudi *aMaze* ter negativni *zMaze*) za slovensko športnico Tino Maze. Slednje besede sicer nimajo pomenonosnega krna, vendar pa so zaradi zapisa bile uvrščene med e-tvorjenke.

Nadalje imamo tvorjenke z enoglasovnim krnom *i* v pomenu *internetni*, in sicer *iTivi* (*iTV*), *iBanka*, *iConcert*, *iMDB*, *iRazglednica*, *iBar*, *iGorenje*, *iPoker* in *iUčbenik* pa tudi v pomenu *informacijski* v tvorjenki *iCenter* in *instant* v tvorjenki *iM*. Končamo lahko z besedami z enoglasovnim krnom *m* v pomenu *mobilni*, od katerih smo našli le tri, in sicer *mCOBISS*, *mTerminal* ter *mVzajemna.*

3.4.3 Izdelki podjetja Apple

Največ dvozačetniških e-tvorjenk je izdelkov podjetja Apple, in sicer se jih je v našem iskalnem nizu pojavilo kar 56, ki jih lahko razdelimo v uradna imena izdelkov in aplikacij na eni strani ter nove tvorjenke na drugi. Četudi je krn *i* v imenih izdelkov Apple izvorno pomenil *internetni*, je to dejstvo v veliki meri utonilo v pozabo. Tako pri uradnih imenih izdelkov in aplikacij enoglasovni krn *i* danes označuje ime izdelka, pri novih tvorjenkah pa pomeni Applove izdelke nasploh.

Tako imamo med uradnimi imeni izdelkov naslednje i-tvorjenke: *iPhone* (tudi *iP4*, *iP5* in *iP6)*, *iPad*, *iOS*, *iTunes*, *iPod*, *iMac*, *iCloud*, *iStore*, *iMessage*, *iWatch*, *iStyle*, *iSpot*, *iHelp*, *iBooks*, *iWork*, *iMovie*, *iDrive*, *iPhoto*, *iP*, *iPlayer*, *iLife*, *iCal*, *iLoop*, *iBookstore*, *iGlove*, *iStvar*, *iBand*, *iCalendar*, *iSight*, *iTouch*, *iWeb*, *iFlicks*, *iMaps*, *iDevice*, *iUser*, *iRadar*, *iSteve*.[16]

Na drugi strani pa imamo 14 primerov i-tvorjenk, ki so specifično slovenske, in sicer: *iNaprava*, *iTelefon*, *iNapravica*, *iZadeva*, *iPriročnik*, *iTrgovina*, *iAplikacija* pa tudi *iCrap*, *iDiot*, *iGrača*, *iFanatik*, *iFan*, *iNeki* in »šaljivo«

[13] Prevod iz angleške tvorjenke M-theory (ki se prav tako pojavi 5-krat), v kateri črka M predstavlja »magic, mystery or membrane«, torej čarovnijo, skrivnost ali membrano.

[14] Alternativni zapis portala e-uprava.

[15] Slovenski sejem erotike.

[16] Naslov filma.

(lahko le upamo) sklanjanje besede iPhone kot *iPhoneta*/*iPhonetov*.

3.5 Navidezne e-tvorjenke

Veliko tvorjenk bi na prvi pogled uvrstili med nesistemske tvorjenke z enoglasovnim krnom, vendar so pravzaprav lastna imena, ki so jih svojim izdelkom nadela razna podjetja, pa tudi sistemske tvorjenke.[17] V tem razdelku obravnavamo le lastna imena, in sicer imena avtomobilov ter nekaterih drugih izdelkov, ki so v korpusu Janes najpogostejša. Četudi teh navideznih e-tvorjenk ne obravnavamo podrobno, jih je pomembno omeniti, saj razlika med e-tvorjenkami in lastnimi imeni, ki sledijo zapisu *č/Č-podstava* oz. *čPodstava*, ni vedno očitna na prvi pogled.

Naštevamo 10 izdelkov iz avtomobilske industrije (tabela 2) ter 10 raznovrstnih izdelkov (tabela 3), ki so se v korpusu Janes pojavili najpogosteje. Lastna imena so zapisana tako, kot se pojavljajo izvorno na uradnih straneh podjetij, kar pa ni vedno enako zapisom, ki se pojavljajo v korpusu, saj le redki uporabniki preverijo pravilno črkovanje posameznih lastnih imen in se ga držijo. V razpredelnici, ki sledita, smo vključili tako vezajne kot tudi dvozačetniške tvorjenke, ne pa Applovih izdelkov, ki smo jih že obravnavali v razdelku 4.4.3.

Pojavitve	Izdelek
470	dCi
375	S-Max
197	C-Max
184	S-line
146	T-Jet
98	S-Class / s-klasa
85	E-Class / e-klasa
84	X-Trail
73	C-Elysée
68	B-Max

Tabela 2: 10 izdelkov avtomobilske industrije z največ pojavitvami v korpusu Janes.

Pojavitve	Izdelek
78	B-complex / B-kompleks
68	iGo
57	X-treme
34	e-HDi
32	uTorrent
32	V-lube
29	iRobot
24	nVidia
16	x-box / xBox[18]
13	X-TRM

Tabela 3: 10 izdelkov z največ pojavitvami v korpusu Janes.

Kljub nekaterim pravilnim zapisom lastnih imen v korpusu slednji ne prevladujejo, temveč obstaja veliko raznolikih verzij, ki imajo pogosto skupne le črke, ne pa tudi njihovega zaporedja.

4 Sklep

Predstavili smo relativno nov segment leksike, ki se je v zadnjem desetletju z vse večjo prisotnostjo interneta in razvojem tehnologij močno razširil in postal del vsakdanjega življenja. Ukvarjali smo se s t. i. vezajnimi in dvozačetniškimi e-tvorjenkami oz. nesistemskimi tvorjenkami z enoglasovnim krnom tipov *č-beseda* in *čBeseda*, ki so nam pokazali, kako priljubljen in razširjen je ta postal. Naredili smo pregled najpogostejših tvorjenk obeh zapisov, pri čemer smo ugotovili, da je še vedno največ e-tvorjenk z enoglasovnim krnom *e-*, pri vseh črkah pa obstajajo inovativni zapisi, ki kažejo, da se tvorci besedil tega novega besedotvornega vzorca zavedajo in ga, vsaj na spletu, tudi uporabljajo.

V besedotvorni analizi je bilo mogoče opaziti tako nihanja v zapisu e-tvorjenk kot tudi probleme pri njihovem sklanjanju, kar nakazuje jasno potrebo po aktualizaciji kodifikacije na podlagi gradiva, predstavljenega v prispevku. Raznolikost zapisov in negotovost uporabnikov na eni strani ter njihova navdušenost nad uporabo novega besedotvornega postopka na drugi nam namreč kaže, kako pomembno je, da osnovni jezikovni priročniki vsebujejo tudi ažurne informacije o spreminjajočih se segmentih jezika.

5 Zahvala

Zahvaljujem se asist. dr. Damjanu Popiču in doc. dr. Darji Fišer z Oddelka za prevajalstvo Filozofske fakultete UL za spodbudo in dragocene napotke pri pripravi tega prispevka. Zahvala gre tudi trem anonimnim recenzentom, katerih premisleki in pripombe so prispevek pomembno izboljšale.

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.

6 Literatura

Helena Dobrovoljc in Nataša Jakop. 2012. Sodobni pravopisni priročnik med normo in predpisom. Založba ZRC, Ljubljana.

Darja Fišer ... [et al.]. 2014. JANES se predstavi. V: Zbornik 17. mednarodne multikonference Informacijska družba – IS 2014 : zvezek G = Jezikovne tehnologije, str. 56–61, Ljubljana, Slovenija. http://is.ijs.si/zborniki/2014_IS_CP_Volume-G_%28LT%29.pdf.

Nataša Logar. 2004. Nove tehnologije in nekateri nesistemski besedotvorni postopki. V: Mednarodni znanstveni simpozij Obdobja - metode in zvrsti; Obdobja 22 – Aktualizacija jezikovnozvrstne teorije na Slovenskem: členitev jezikovne resničnosti, str. 121–

[17] Tipa *B-kategorija* in v podobnih primerih, kjer gre za črkovne simbole, ki nakazujejo hierarhičnost, razločevalno vrednost itd.

[18] Gre za napačen zapis imena igralne konzole *Xbox*, ki se sicer v korpusu pojavi 583-krat.

132, Ljubljana, Slovenija. http://www.centerslo.net/files/File/simpozij/sim22/Logar.pdf.

Nataša Logar Berginc ... [et al.]. 2012. Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba. Trojina, zavod za uporabno slovenistiko: Fakulteta za družbene vede, Ljubljana.

Mija Michelizza. 2008. Nove tvorjenke v spletnih besedilih (primer Wikipedije). V: Slovenščina med kulturami, str. 328–338, Ljubljana, Slovenija. Slavistično društvo Slovenije.

Damjan Popič. 2013. Korpusnojezikoslovni mo(nu)menti: Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKres: gradnja in uporaba. V: Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave, str. 176–180. Trojina, zavod za uporabno slovenistiko, Škofja Loka. http://www.trojina.org/slovenscina2.0/arhiv/2013/1/Slo2.0_2013_1_09.pdf.

Irena Stramljič Breznik. 2003. Besedotvorna tipologija novonastalega besedišča s področja mobilne telefonije. V: Slavistična revija 51, str. 105–118, Ljubljana, Slovenija. Slavistično društvo Slovenije. http://www.srl.si/arhiv/2003-kongr/pdf/stramljic-breznik.pdf.

Jože Toporišič. 1984. Slovenska Slovnica. Založba "Obzorja", Maribor.

Ada Vidovič-Muha. 2011. Slovensko skladenjsko besedotvorje. Znanstvena založba Filozofske fakultete, Ljubljana.

Terminologija v spletnih forumih

Špela Vintar

Oddelek za prevajalstvo, Filozofska fakulteta
Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
spela.vintar@ff.uni-lj.si

Povzetek

V prispevku raziskujemo rabo terminologije v treh slovenskih spletnih forumih: avtomobilizem.com, med.over.net in kvarkadabra.net. Z uporabo luščilnika terminologije in ročne validacije terminoloških kandidatov skušamo ugotoviti delež specializiranih enot v posameznih forumih in podforumih, posebej pa se posvetimo nestandardnim elementom pri rabi terminologije. Rezultati kažejo precejšnje razlike v specializiranosti forumov, ta pa ne sovpada z ravnijo standardiziranosti. Prikažemo tudi razlike v rabi terminologije glede na tip foruma (moderiran ali nemoderiran) in raven znanja razpravljalca (specialist ali laik).

Terminology in Internet Forums

The paper explores the use of terminology in three Slovene internet forums: avtomobilizem.com, med.over.net and kvarkadabra.net. We employ automatic term recognition and manual validation of term candidates in an attempt to establish the term density of individual subforums, A set of methods is proposed for the recognition of nonstandard terms. Our results reveal large differences between forums in terms of specialisation level, which does not necessarily correlate to the level of standardisation. Finally we explore some properties of terminology depending on the type of forum (active or passive moderation) and on the level of expertise of the forum user (expert vs. lay).

1 Uvod

Spletni forum je spletni prostor, kjer lahko uporabniki v obliki objavljenih sporočil sodelujejo v različnih razpravah (Wikipedija, 17. 8. 2015). Kot eden od precej razširjenih spletnih žanrov ima forum določene značilnosti (prim. Varga, 2011: 32–34):

- Spletni forum običajno združuje vrsto razprav z določenega področja in tako uporabnike povezuje v virtualno skupnost (npr. mama.si, mojpes.net, hribi.net itd.).
- Spletni forumi imajo drevesno strukturo, pri kateri se vrhnje področje deli na posamezne podforume, znotraj njih pa na teme, sestavljene iz niza objav.
- Razprave v spletnih forumih imajo pretežno značilnosti pisne komunikacije in se tipično vsaj nekaj časa arhivirajo.

Zaradi teh značilnosti so spletni forumi – vsaj tisti, ki obravnavajo določeno področje – oblika specializirane komunikacije in so kot taki zanimivi za terminološko analizo, prek njihove hierarhične strukture pa jih lahko uredimo v bolj ali manj specializirane podkorpuse.

Spletni forum lahko s komunikološkega vidika opredelimo kot interaktivno okolje, kjer poteka množično komuniciranje v razmerju eden z mnogimi in/ali kot skupinsko razmerje mnogih z mnogimi, v njem pa potekata javna in zasebna komunikacija (Petrovčič, 2005: 13). Predvsem zaradi anonimnosti in odsotnosti neposrednega vizuelnega stika lahko uporabnik spletnega foruma kommunicira drugače kot pri neposredni, fizični komunikaciji, po drugi strani pa forum ni povsem neregulirano spletno okolje, saj danes praktično vsa forumska spletišča eksplicitno določajo pravila vedenja, za njihovo upoštevanje pa skrbi moderator (Plant, 2004: 60).

Najbolj obiskani forumi pri nas (in podobno v tujini) so mesta za izmenjavo izkušenj in znanja o določeni temi, zato jim v nadaljevanju lahko rečemo strokovni spletni forumi. Namen prispevka je raziskati značilnosti rabe terminologije v takšnih forumih. Raziskava temelji na šestih tematskih podkorpusih s treh slovenskih forumov, in sicer med.over.net, avtomobilizem.com in kvarkadabra.net, ki so zajeti v korpusu Janes v0.3. Osrednja vprašanja raziskave so:

- Kolikšen je delež terminoloških enot v posameznem podkorpusu in ali iz tega lahko sklepamo o ravni specializiranosti posameznega foruma?
- Kakšno je razmerje med standardnim in nestandardnim terminološkim besediščem v posameznih forumih?
- Ali so obstoječe metode samodejnega luščenja terminologije uspešne tudi pri luščenju terminologije iz spletnih forumov?
- Kako se raven standardnosti spletnih besedil kaže pri rabi terminologije?

Pričakujemo, da bodo rezultati raziskave uporabni pri razvoju orodij za računalniško obdelavo spletnih besedil, obenem pa bodo prinesli nova spoznanja o jezikoslovnih značilnostih specializirane komunikacije na spletu.

2 Sorodne raziskave

S sociološko-komunikološkega vidika so spletni forumi zanimivi kot družbena okolja, v katerih se med člani izmenjujejo izkušnje, prepričanja in znanje, ob tem pa se gradi tudi posameznikova pripadnost skupnosti. Kimmerle et al. (2011) tako raziskujejo konstruiranje znanja v spletnem forumu alternativne medicine in ugotavljajo, da pri opazovanem spletnem forumu prepričanja uporabnikov tvorijo holistično ideologijo, ki prerašča okvire dejstev in iz virtualne skupnosti ustvarja parareligiozno bratovščino.

Podrobno se s procesi prenašanja znanja v spletnih forumih ukvarja Varga (2011) v svoji doktorski disertaciji, pri čemer s postopki analize diskurza raziskuje strategije prepričevanja, razlaganja in podajanja navodil v različnih strokovnih spletnih forumih. Skozi analizo pokaže, kako se

identiteta razpravljalcev samodejno vzpostavi skladno z njihovo stopnjo obveščenosti in s skupnim ciljem izmenjati znanje.

Petrovčič (2005) v svojem diplomskem delu raziskuje deliberativnost v razpravljalnih forumih, pri tem pa izhaja iz dveh nasprotujočih si izhodišč. Nekateri avtorji namreč (Pavlik, 1994; citirano po Petrovčič, 2005: 5) zagovarjajo stališče, da »odsotnost normativne strukture in relativna anonimnost posameznikov razpravljalcem ponujata večjo svobodo za javno izražanje argumentiranih mnenj, medtem ko so drugi avtorji [...] nasprotnega mnenja, saj trdijo, da sta vizualna anonimnost razpravljalcev in pomanjkljiva normativna strukturiranost pogovorov glavni vzrok za kratkotrajnost in plitkost stikov med razpravljalci ter podajanje emocionalno nabitih nekritičnih mnenj.« To izhodišče je posredno relevantno tudi za jezikoslovno analizo forumov, saj se uporabnikova virtualna identiteta kaže primarno skozi njegov način izražanja.

Bolj jezikoslovno usmerjenih raziskav forumov je več, a se nobena nam znana raziskava ne usmerja prav v jezikoslovne značilnosti terminološkega besedišča. Montero et al. (2007) denimo ugotavljajo, da visoka pojavnost modalnih glagolov v forumskih besedilih ta približuje govorjenemu diskurzu. Za naš prostor sta relevantni še raziskava Nataše Jakop (2008), ki opazuje (ne)spoštovanje pravopisnih pravil v slovenskih spletnih forumih, čeprav brez kvantitativne analize, in magistrsko delo Nataše Zupančič (2009), ki izvede celovito korpusno analizo jezikovnih značilnosti šestih spletnih forumov.

Zupančič (ibid.) analizira različne ravni diskurza na spletnih forumih in predstavi kvantitativne rezultate v zvezi z ortografskimi posebnostmi, frazeologijo, skladenjskimi značilnostmi in pragmatiko, posebej se posveti tudi vprašanju jezikovnega izbiranja na spletnih forumih. Sklepne ugotovitve skladenjske in ortografske analize za izbrano gradivo pokažejo, da jezik spletnih forumov sicer v marsičem odstopa od standardne slovenščine, vendar v njem še vedno prevladujejo značilnosti zapisane besede. Za našo raziskavo so izsledki Nataše Zupančič uporabni zgolj posredno, saj je v njej uporabila gradivo z zelo splošnih forumov (diva.si, lunin.net, mobisux.com, mojtrener.com, razprave.com in siol.net), pri izboru objav pa se je omejila na teme, ki se ukvarjajo z resničnostnim šovom Bar.

V omenjeni raziskavi predstavljene značilnosti spletne slovenščine so bile torej pridobljene na podlagi korpusa splošnih, ne pa strokovnih forumov, kar ne pomeni, da v določeni meri ne držijo tudi za naše gradivo. Predvsem se ugotovitve o leksikalnih, skladenjskih in frazeoloških značilnostih forumskih besedil nikjer ne dotikajo terminologije, poleg tega je komunikacijski okvir precej drugačen: če je pri strokovnih forumih temeljni namen razpravljalcev izmenjava znanja, je pri splošnih forumih – še posebej pri razpravah o resničnostnem šovu – v ospredju izmenjava mnenj.

Pri naši raziskavi tako ohranjamo metodologijo korpusnoterminološke analize, ki temelji na samodejnem luščenju terminologije iz oblikoskladenjsko označenih besedil, dodali pa smo nekaj novih metod za raziskovanje nestandardnosti terminologije.

3 Predstavitev gradiva

Vir gradiva, uporabljenega v raziskavi, je korpus spletne slovenščine Janes v0.3 (Fišer et al., 2015), ki vsebuje 29 odstotkov besedil s spletnih forumov, kar je dobrih 47 milijonov pojavnic. V korpusu so zajeti trije spletni forumi, in sicer avtomobilizem.com, med.over.net in kvarkadabra.net, ki se precej razlikujejo po tematskih področjih, številu uporabnikov in objav, skupno pa jim je, da gre pri vseh treh v najširšem smislu za področno specializirane forume. Tako se forum avtomobilizem.com s skoraj 40.000 uporabniki razglaša za »največjo avtomobilistično skupnost v Sloveniji«, spletišče med.over.net se je izvorno vzpostavilo kot vrsta moderiranih zdravstvenih posvetovalnic, sčasoma pa so se jim pridružile še manj specializirane razpravljalske skupine, forum kvarkadabra.net pa je prostor za razprave o znanosti, predvsem astronomiji, fiziki, biologiji in filozofiji, katerega skupnost šteje nekaj čez 4 tisoč uporabnikov.

Potencialno pomemben parameter pri različnih (pod)forumih je aktivno oz. pasivno moderiranje. Čeprav imajo vsa tri forumska spletišča moderatorje, ki domnevno bdijo nad uporabniki in njihovim spoštovanjem pravil uporabe foruma, pa so le zdravstvene posvetovalnice na med.over.net forumi z aktivnim moderiranjem, se pravi razprave med nestrokovnjaki in strokovnjakom, običajno zdravnikom specialistom. Tudi med njimi so glede dinamike moderatorja razlike, ki jih lahko ugotovimo zgolj s prebiranjem posameznega podforuma. Pri nekaterih zdravstvenih posvetovalnicah je namreč v naslovu podforuma naveden moderator, ki pa se oglaša redko in razpravo prepušča laičnim uporabnikom, pri drugih pa so nizi objav strukturirani strogo v obliki vprašanj in odgovorov in torej potekajo v obliki dialogov med pacientom in zdravnikom, laični komentarji na »tujo« temo pa niso dovoljeni.

Podkorpus	Število objav	Število pojavnic	Aktivno moderiran
Avtomobilizem.Styling	21.634	866.974	ne
Avtomobilizem.SamSvojMojster	23.030	953.037	ne
Med.Over.Net.Ginekologija	1724	259.991	da
Med.Over.Net.Kontracepcija	5470	403.991	ne
Med.Over.Net.Hujšanje	2410	255.034	ne
Kvarkadabra.Astronomija	1601	107.365	ne

Tabela 1: Osnovni podatki o izbranih podkorpusih.

Za namene naše terminološke raziskave smo iz omenjenih forumskih spletišč izbrali 6 podforumov, ki smo jih iz korpusa Janes izluščili s pomočjo metabesedilnih oznak; pri vseh korpusnih besedilih je namreč v oznaki besedila zapisano ime foruma, podforuma, teme in avtorja, kar močno olajša ustvarjanje podkorpusov. S foruma avtomobilizem.com smo tako izbrali podforuma Styling o lepotnih posegih na vozilih ter Sam svoj mojster o popravilih, s spletišča med.over.net smo vključili zdravstveno posvetovalnico ABC ginekologije in porodništva, ki jo aktivno moderira zdravnik specialist, forum Kontracepcija, kjer uporabnice in uporabniki o tej temi razpravljajo pretežno brez moderatorja, in forum Hujšanje, ki je prav tako moderiran zgolj pasivno in je po naših predvidevanjih tudi najmanj terminološki. S spletišča Kvarkadabra smo izbrali le podforum O svetu za Luno, kjer so zbrani prispevki s področja astronomije. Tabela 1 predstavlja izbrane podkorpuse z osnovnimi korpusnimi podatki, dodan je še podatek o aktivnem moderiranju.

4 Delež in značilnosti terminologije v podkorpusih

V nadaljevanju smo iz vseh šestih podkorpusov izluščili ključne besede ter eno- in večbesedne terminološke kandidate, pridobljene sezname pa smo ročno pregledali in med prvimi 500 izluščenimi kandidati označili termine. Nato smo se posebej posvetili nestandardnim terminološkim elementom, nazadnje pa nas je zanimala še raba terminologije v moderiranem forumu, kjer lahko primerjamo objave, ki jih je napisal strokovnjak, s tistimi, ki so jih prispevali laični uporabniki.

4.1 Luščenje terminoloških kandidatov

Korpus Janes je oblikoskladenjsko označen (Fišer et al., 2015), kar pomeni, da lahko za samodejno luščenje terminologije uporabimo obstoječe orodje LUIZ, ki je bilo doslej uporabljeno že na številnih področjih (Vintar, 2010; Vintar in Fišer, 2011). Luščenje namreč temelji na oblikoskladenjskih vzorcih, ki so ponavadi samostalniške zveze, ti pa se v nadaljnji obdelavi razvrščajo na bolj ali manj terminološke. Na rezultate luščenja močno vpliva kakovost oblikoskladenjskih oznak, v korpusu Janes pa je označevalnih napak precej.

S programom LUIZ smo iz podkorpusov luščili izključno samostalniške besedne zveze, dolge od ene do štirih besed. Izračun terminološkosti temelji na primerjavi pogostosti med specializiranim in referenčnim korpusom; kot referenčni korpus smo uporabili Gigafido (Logar Berginc et al., 2012). Za izračun natančnosti smo pri vseh seznamih izluščenih kandidatov pregledali prvih 500 enot in ročno označili terminološko relevantne izraze. Ob tem je treba poudariti, da je bila presoja nujno subjektivna, saj se pri označevanju nismo posvetovali s področnimi strokovnjaki. Prav tako med termine nismo uvrščali splošnega besedišča, ki je se v določenem podforumu pogosto pojavlja zaradi tematskega okvira, a brez strokovno specifične reference. Tako se denimo pri podforumu Hujšanje kot terminološki kandidati pojavijo izrazi v zvezi s prehrano (*jogurt, zajtrk, obrok*), ki jih ne moremo obravnavati kot specializirane enote, čeprav so tematsko ključne za obravnavani forum. V kontekstu naše raziskave lahko natančnost razumemo kot izmerjeno raven specializiranosti posameznega podforuma.

Pri večini seznamov se pokaže, da je približno polovica izluščenih enot večbesednih, najproduktivnejši vzorec pa je pridevnik in samostalnik. Ostalih besednovrstnih kombinacij je bistveno manj, tri- in večbesedne enote so redke. To opažanje ni presenetljivo, saj se je dvobesedni vzorec pridevnik in samostalnik pokazal za najproduktivnejšega tudi v drugih raziskavah (Logar Berginc in dr., 2013).

Med prvimi 500 enotami ima največ terminoloških enot podforum Ginekologija (edini z aktivno moderacijo), ta forum izstopa tudi po visokem številu večbesednih, se pravi višje specializiranih enot pri vrhu seznama (*nosečnost, plod, porod, UZ, vstava; nuhalna svetlina, medenična vstava, morfologija ploda, maternični vrat*). Po deležu terminoloških enot sledijo Astronomija, vendar z manj večbesednimi (*planet, zvezda, sonce, vesolje, gravitacija, galaksija, teleskop, asteroid; črna luknja, spektralna črta, nevtronska zvezda*), Sam svoj mojster (*motor, volan, feltna, odbijač, vijak, platišče; prestavna ročica, sprednji odbijač, zadnja hauba, sesalni kolektor*), Kontracepcija (*tabletka, menstruacija, krvavitev, izcedek, kondom, obroček; spolni odnos, kontracepcijska tabletka, urgentna kontracepcija*), Styling (*homologacija, spojler, ledica, blatnik, števec; dimenzija gum, homologacijski kartonček, jekleno platišče*) in nazadnje Hujšanje, kjer je bilo terminološko relevantnih izrazov zgolj 15 odstotkov (*dieta, prehrana, GI, jedilnik, post; ogljikovi hidrati, način prehranjevanja, ločevalna dieta*).

Podkorpus	Št. izluščenih	Št. večbesednih	Natančnost (N=500)
Avtomobilizem.Styling	12.551	5.591	0,39
Avtomobilizem.SamSvojMojster	14.042	6.366	0,58
Med.Over.Net.Ginekologija	3.768	1.720	0,68
Med.Over.Net.Kontracepcija	5.145	3.280	0,52
Med.Over.Net.Hujšanje	4.037	1.464	0,15
Kvarkadabra.Astronomija	3.145	1.064	0,60

Tabela 2: Luščenje terminologije iz podkorpusov.

Posebna značilnost spletnih forumov so kratice, ki se uveljavijo znotraj določenega foruma oz. podforuma. Precej kratic smo opazili v podforumih med.over.net (*M – menstruacija, ZM – zadnja menstruacija, G – ginekolog, CR – carski rez, UZ – ultrazvok, KT – kontracepcijske tablete, GK – ginekološka klinika* itd.), manj v avtomobilizmu (*CZ – centralno zaklepanje*) in še manj na kvarkadabra.net. Kratice so povečini terminološke in predstavljajo učinkovito sredstvo za večjo ekonomičnost komunikacije, obenem pa igrajo vlogo kot povezovalni členi forumske skupnosti, ki z uporabo kratic izraža pripadnost.

4.2 Nestandardni elementi

V kontekstu terminologije je pojem nestandardnosti problematičen, kajti na mnogih specializiranih področjih se uporabljajo izrazi, ki odstopajo od besedotvornih načel splošne slovenščine, so prevzeti, vsebujejo nebesedne elemente ali simbole, pa zato niso nestandardni. V raziskavi nas je predvsem zanimalo, v kolikšni meri uporabniki forumov pri posredovanju znanja uporabljajo žargonsko terminologijo, sicer tipično za govorjeni diskurz, in v kolikšni meri pri zapisovanju izrazov prihaja do namernih ali nenamernih odstopanj.

Za ugotavljanje nestandardnosti smo uporabili tri metode:

- ročno označevanje nestandardnih terminoloških izrazov med prvimi 500 izluščenimi kandidati; pri tem moramo poudariti, da smo se omejili na termine in nestandardnih splošnih izrazov nismo upoštevali. V tabeli je zapisan odstotek nestandardnih enot od vseh terminološko relevantnih enot med prvimi 500 izluščenimi kandidati (ns_MAN),
- luščenje samostalniških enot, ki se z visoko relativno pogostostjo pojavljajo v podkorpusu, a imajo zelo majhno (<10) pogostost v referenčnem korpusu Gigafida (ns_GF), v tabeli je zapisano število takšnih enot,
- luščenje nestandardnih terminoloških kandidatov s pomočjo leksikona besednih oblik Sloleks (Arhar, 2009); iz prvotno izluščenega seznama kandidatov smo izluščili tiste, pri katerih se vsaj ena od besed ne pojavi v Sloleksu (ns_SL), v tabeli je zapisan odstotek od vseh izluščenih terminoloških kandidatov.

Rezultate na koncu primerjamo še z mero tehnične (std_tech) in jezikovne (std_ling) nestandardnosti, ki sta jo razvila Erjavec in Ljubešić (Erjavec, 2015) za namene projekta Janes, v tabeli podajamo povprečno vrednost za podkorpus in standardno deviacijo.

Podkorpus	ns_MAN	ns_GF	ns_SL	std_ling	stdev	std_tech	stdev
Avtomobilizem.Styling	31,1	184	27,0	1,7	0,6	1,8	0,5
Avtomobilizem.SamSvojMojster	20,7	165	21,8	1,7	0,5	1,8	0,5
Med.Over.Net.Ginekologija	3,0	21	10,2	1,5	0,2	1,5	0,3
Med.Over.Net.Kontracepcija	5,7	32	11,5	1,5	0,3	1,4	0,4
Med.Over.Net.Hujšanje	8,0	18	10,6	1,7	0,4	1,6	0,5
Kvarkadabra.Astronomija	6,6	9	11,6	1,5	0,4	1,4	0,4

Tabela 3: Mere nestandardnosti.

Iz Tabele 3 je razvidno, da pri vseh uporabljenih metodah po nestandardnosti izstopata oba podforuma avtomobilizma, najbliže standardu pa je aktivno moderirani forum ginekologije. Za forum avtomobilizma je predvsem značilna raba žargonskih poimenovanj in/ali nestandardnih zapisov za dele vozila (*auspuh/avspuh, felga/feltna, špegu/špegl, števc, spojlerček, žmigovec/žmigavec, radkapa/ratkapa, žarnca, lampa, akumolator*), orodja in postopke (*šrauf/šravf, šprickit, šmirgl, valar, strokovna ugradnja, pedenanje, ličarija*). To pa ne pomeni, da ljubitelji avtomobilizma ne uporabljajo tudi kanonične terminologije, saj je pri mnogih od omenjenih izrazov standardna različica pogostejša od pogovorne (*šipa*: 312, *šajba*: 23). Za avtomobilizem je značilno tudi prepletanje registrov, ko avtor v objavi uporabi nekaj pogovornih in nekaj kanoničnih izrazov:

»Torej, ali se čuje turbina (ker jo počas hudič jemlje), ali je prežgalo dihtungo med turbino in sesalno cevjo (oranžna dihtunga, ki stane cca 12€ – dokaj pogost fail...«

Naši rezultati se skladajo tudi s splošnima merama nestandardnosti, razen pri forumu Hujšanje, kjer pa moramo upoštevati, da se naša analiza usmerja v ugotavljanje nestandardnosti terminologije in ne tudi drugih jezikovnih značilnosti. Ker je v forumu Hujšanje specializiranega izrazja zelo malo, to tudi ne izstopa po nestandardnosti.

Pri bolj specializiranih podforumih se izkaže, da metoda primerjave s splošnimi referenčnimi viri, kot sta Gigafida in Sloleks, ni učinkovita za ugotavljanje nestandardnosti, saj se med žargonskimi, napačno zapisanimi in pogovornimi izrazi znajdejo tudi visokospecializirani izrazi, ki jih v splošnih referenčnih virih ni. To smo opazili predvsem pri medicinskem izrazju, kjer bi bilo za natančnejše luščenje nestandardnih izrazov poleg splošnih virov nujno uporabiti še medicinski leksikon.

Med nestandardnimi izrazi s foruma kvarkadabra.net je večina, če odštejemo izrazje brez šumnikov, angleških poimenovanj za fizikalne in druge pojave, nekaj pa je tudi zatipkank in hudomušnih tvorjenk (*nafitana krivulja, hobistična pamet, sključitev, kvazinaravoslovec, complex lajf*).

4.3 Primerjava objav strokovnjaka in nestrokovnjakov

Ker se je iz doslej predstavljenih rezultatov potrdilo, da aktivno moderirani forum Ginekologija izstopa tako po deležu terminologije, ravni specializiranosti in standardnosti, nas je zanimalo, ali je v interakciji med laično skupnostjo in strokovnjakom razlika v rabi terminologije med enimi in drugim. Forum ABC Ginekologije in porodništva že vsa leta njegovega obstoja moderira mag. Stanko Pušenjak, specialist ginekolog in porodničar, ki je v našem podkorpusu avtor 850 objav. Iz podkorpusa smo tako v eno skupino uvrstili zgolj njegove objave, v drugo pa vse ostale, in ju primerjali.

	Ginek.SP	Ginek.Drugi
Št. objav	850	874
Št. pojavnic	130.616	129.388
Št. izluščenih	1.835	2.388
ns_SL	5,0	12,7

Tabela 4: Strokovnjak in nestrokovnjaki.

Delež nestandardnih enot po metodi primerjave s Sloleksom je v objavah specialista precej nižji (5,0) kot v objavah laikov (12,7), ročni pregled pa pokaže, da je v izluščenih terminih večina specializiranih medicinskih izrazov, ki se v Sloleksu ne pojavijo (*spermiogram, luteinska cista, Metronidazol, endometritis*). V objavah specialista zasledimo le tri enote, ki bi jih lahko označili za nestandardne zaradi napačnega zapisa (*lumbrikant, stegnjenica, ovuilacija*). Nasprotno je v objavah laikov nestandardnih zapisov prek 300, uvrstimo pa jih lahko v tehnično nestandardne (pretežno zapisane brez šumnikov) in jezikovno nestandardne, napačno zapisane terminološke enote (*Dobroston/Dabresten, amnoreja, mestruacija, aborativna tabletka, agluntinacija semenčic, iscedek, endometiroza, gljivična okužba*). Nasploh je med laičnimi objavami opaziti veliko variabilnost pri zapisovanju specializiranih enot (*anastetik / anstetik / anastazistka; celarblue / clerblu; ejakuacija / ejakuiral; endometiroza / endometrioza / ednometrioza; epizotomija / epiziotomija; genikologinja / ginekloginja / ginekolginnja* itd.).

Žal je bil vzorec premajhen, da bi izvedli še analizo kolokacij; predvidevamo namreč, da nestrokovnjaki določene termine uporabljajo v drugačnih sintagmah kot strokovnjaki. Tak primer smo zabeležili pri laični rabi izrazov *tableta/tabletka/zdravilo* v povezavi z glagolom *jesti*; med objavami zdravnika specialista te kolokacije ni najti, temveč se *zdravila/tablete uživa, jemlje* ali *uporablja*.

5 Sklep

V raziskavi smo želeli raziskati značilnosti rabe terminologije v spletnih forumih, pri tem pa smo uporabili luščilnik terminologije kot pripomoček za ugotavljanje ravni specializiranosti posameznega foruma. Analiza je pokazala, da so med obravnavanimi forumi glede ravni specializiranosti precejšnje razlike, najbolj specializiran pa je bil aktivno moderirani forum zdravstvene posvetovalnice, pri katerem zdravnik specialist odgovarja na vprašanja pacientov. Terminološko bogata sta tudi oba podforuma s področja avtomobilizma, hkrati pa sta ta podforuma tudi najbolj nestandardna z uporabo številnih pogovornih izrazov in področno specifičnih okrajšav.

Ker je pri strokovnih forumih osnovni cilj udeležencev izmenjava znanja, ni presenetljivo, da je na splošno ta besedilni žanr s terminološkega vidika zanimiv, jezikovni register določenega foruma pa se oblikuje glede na različne zunajjezikovne parametre, kot so predznanje/izobrazba razpravljalcev, reguliranost foruma in strokovnost tematskega področja.

V nadaljnjih raziskavah se želimo podrobneje posvetiti pojmovnim razmerjem, ki se pojavljajo znotraj določene strokovne razprave, in odkrivanju tipičnih struktur, ki so značilne za bolj ali manj regulirane komunikacijske situacije posredovanja znanja (npr. pojmovne strukture vprašanj in pozvedb, odgovorov in razlag, navodil in napotkov itd.). Prav tako želimo sistematično raziskati terminološko variabilnost.

Jezikoslovne značilnosti rabe terminologije v spletnih forumih, pa tudi drugih spletnih žanrih, so relevantne z več vidikov. Z računalniško obdelavo bi iz njih lahko pridobivali znanje in izkušnje spletnih uporabnikov in jih uporabili za razvoj inteligentnih aplikacij, Z odkrivanjem terminoloških variacij bi lahko izboljšali spletne iskalnike in razpoznavalnike govora. V nekoliko širšem kontekstu foruma kot virtualne skupnosti za posredovanje znanja pa bodo vsekakor še naprej pomembne tudi komunikološke, socio- in psiholingvistične raziskave, ki nam pomagajo razumeti oblikovanje spletnih identitet in njihovega vpliva na komunikacijo.

6 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.

7 Literatura

Špela Arhar. 2009. Učni korpus SSJ in leksikon besednih oblik za slovenščino. Jezik in slovstvo, 54/3–4, 43–56.

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

Nataša Logar Berginc, Špela Vintar in Špela Arhar Holdt. 2013. Terminologija odnosov z javnostmi: korpus – luščenje – terminološka podatkovna zbirka. Slovenščina 2.0, 1 (2): 113–138.

Darja Fišer, Tomaž Erjavec, Jaka Čibej in Nikola Ljubešić. 2015. Gradnja in analiza korpusa spletne slovenščine JANES. Obdobja 2015 (v tisku).

Nataša Jakop. 2008. Pravopis in spletni forumi – kva dogaja? *Slovenščina med kulturami, Zbornik Slavističnega društva Slovenije 19*, 315–327.

Joachim Kimmerle, Kristina Abels, Katharina Becher, Anna Beckers, Annette Haussmann, Ansgar Thiel in U. Cress. 2011. Construction of health knowledge in an alternative medical community of practice: Hermeneutic analysis of a web forum.V Connecting research to policy and practice: Proceedings of the 9th computer supported collaborative learning conference, vol. 1. 1–8.

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 RANLP – Recent Advances in Natural Language Processing, Hissar, Bulgaria.
Begoña Montero, Frances Watts, and Amparo García-Carbonell. Discussion forum interactions: Text and context. System 35.4 (2007): 566–582.
Andraž Petrovčič. 2005. Deliberativnost komuniciranja v spletnih forumih. Diplomsko delo. Univerza v Ljubljani, Fakulteta za družbene vede.
Robert Plant. 2004. Online Communities. Technology in Society, 26 (1), 51–65.
Christina Varga. 2011. Knowledge Transmission in Cyberspace. Discourse Analysis of Professional Web Forums as Internet Subgenre. Doktorska disertacija, UPF Barcelona.
Špela Vintar. 2010. Bilingual term recognition revisited: the bag-of-equivalents term alignment approach and its evaluation. Terminology, 16 (2), 141–158.
Špela Vintar in Darja Fišer. 2011. Enriching Slovene WordNet with domain-specific terms. Translation: computation, corpora, cognition, vol. 1, no. 1, str. 29–44. http://www.t-c3.org/index.php/t-c3/article/view/4.
Nataša Zupančič. 2009. Korpusna analiza slovenskega jezika na spletnih forumih. Magistrsko delo, Univerza v Ljubljani, Filozofska fakulteta.

Prava frekvenca – analiza žaljivega govora v spletnih komentarjih

Urška Vranjek Ošlak,* Ajda Centa†

* Inštitut za slovenski jezik Frana Ramovša ZRC SAZU
Novi trg 2, Ljubljana
urska.vranjek@zrc-sazu.si

† Ljubljana
centajda@gmail.com

Povzetek

Svetovni splet omogoča uporabnikom nove načine komunikacije – eden takih je komentiranje spletnih objav novičarskih portalov. Nepregledno število komentarjev onemogoča njihovo hitro in zanesljivo pregledovanje ter morebitno posredovanje v primeru žaljivosti ali neprimernosti vsebine. Problematika je zlasti pereča v času, ko se je zaradi težav z njegovim zaznavanjem močno razširil žaljivi govor, ki ga omogočata anonimnost in množičnost spletne komunikacije. Pričujoči prispevek predstavlja interdisciplinarni študentski projekt, ki se je ukvarjal z analizo spletnih komentarjev in gradnjo korpusa žaljivih prispevkov.

The Right Frequency – Analysis of Offensive Speech in Online Comments

The Internet offers its users new ways of communication, one of them being the possibility to comment online articles. The endless number of comments makes it impossible to ensure their efficient screening as well as the possibility to intervene in the event of impropriety or offensive content. The problem is particularly acute at a time when offensive speech has greatly expanded due to problems with its detection; offensive speech is made possible by the anonymity and the plurality of online communication. This paper presents an interdisciplinary student project, which dealt with the analysis of online comments and the construction of offensive speech corpus.

1 Uvod

Internet je v zadnjih dvajsetih letih postal vodilni svetovni medij, ki močno vpliva na komunikacijo med različnimi skupinami ljudi in posamezniki. Načini sporazumevanja so se ne le spremenili, svetovni splet je uvedel tudi nove tipe besedil, ki so zaradi njegovih značilnosti specifični za to vrsto diskurza (Crystal, 2001). Primer posebnega tipa besedila so spletni komentarji.

Zaradi lahke dostopnosti in množične razširjenosti interneta je vse bolj razširjena in množična tudi spletna komunikacija (predvsem na spletnih forumih in v klepetalnicah, pa tudi na novičarskih straneh). Internet kot medij omogoča takojšnjo objavo sporočil in precejšnjo mero (četudi navidezne) anonimnosti pri sporočanju, zaradi česar se je v zadnjih letih v spletnih komentarjih razširil tudi žaljivi govor, ki ga je zaradi velike dinamike in količine objav velikokrat težko nadzorovati.

Raziskava, predstavljena v pričujočem prispevku, je bila opravljena v okviru projekta Prava frekvenca – korak k nenasilni komunikaciji, katerega cilj je bila zasnova orodja za avtomatsko zaznavanje neprimerne in žaljive vsebine spletnih komentarjev. V prispevku najprej predstavimo obravnavo žaljivega govora v literaturi in pregled preteklih raziskav, nato opišemo postopek zajema in obdelave besedil, na koncu pa predstavimo še ročno analizo spletnih komentarjev in označevanje žaljivega govora.

2 Žaljivi govor na spletu

Prvotni cilj projekta je bila detekcija *sovražnega govora*, ki je definiran kot »izražanje, ki zmerja, žali, ustrahuje in/ali spodbuja k nasilju, sovraštvu ali diskriminaciji, in sicer na podlagi rase, etničnega izvora, religije, spola, fizičnega stanja, invalidnosti, spolne usmerjenosti, političnega prepričanja itd.« (Matsuda et al., 1993). Sovražni govor je »sporna oblika komuniciranja proti določenim družbenim skupinam, ki je opredeljena v posebnem členu Kazenskega zakonika (trenutno 297. člen KZ-1) ter se preganja po uradni dolžnosti« (Vehovar et al., 2012, 173). Vendar je pregled komentarjev in primerov sovražnega govora na spletu, ki jih obravnavata varuh človekovih pravic in informacijski pooblaščenec,[1] pokazal, da je primerov, za katere odgovorne osebe presodijo, da spadajo na področje sovražnega govora, zelo malo, saj ga je izjemno težko dokazovati. Kot širša in primernejša alternativa se je tako izkazalo preučevanje *nesprejemljivega govora*, pri katerem »obstaja pravna osnova za pregon izven člena o sovražnem govoru, predvsem so to drugi členi iz KZ-1 (npr. čast in dobro ime, ogrožanje varnosti) ter drugi zakoni (npr. Zakon o medijih)« (Vehovar et. al., 2012, 173).

Neprimerni oz. *žaljivi govor* je opredeljen kot govor, ki »nima pravne osnove za sodno obravnavo in se regulira zgolj na osnovi internih pravil (formalnih ali neformalnih)« (Vehovar et al., 2012, 173). Glavna in razlikovalna lastnost žaljivega govora je, da gre pri njem za žaljivo komunikacijo, ki je označena kot neposredna kritika, žaljivka, žaljivo govorjenje, žaljivo in provokativno pošiljanje sporočil ali žaljiv izraz močnih čustev. Praprotnik (2003, 517) žaljivi govor definira kot »verbalno agresijo«, nekonformno vedenje, »antisocialno interakcijo« ali obliko socialne agresije.

Kljub temu, da je definicija žaljivega govora precej jasna, se je zaradi njene širine že med pregledom literature

[1] http://www.varuh-rs.si/, https://www.ip-rs.si/.

pokazalo, da je žaljivi govor v realnosti zelo izmuzljiv, razlog za to je subjektivni vidik govora na splošno. »Obstaja splošni konsenz, da žalitve na internetu sestojijo iz agresivne ali škodoželjne (sovražne) komunikacije. Natančnejši vpogled v problematiko žaljivk pa kaže, da definicije, ki jih uporabljajo v analizah žaljivega govorjenja, niso niti natančne niti konsistentne« (O'Sullivan in Flanagin, 2003). Kot meni Praprotnik (2003), do teh nejasnosti prihaja, ker se daje prevelik poudarek na vsebino sporočila, spregleda pa se pomen konteksta, v katerem se sporočilo pošilja. Kontekst ima pomemben vpliv na žaljivo v govoru. Bralec sicer zazna številne pomenske odtenke v besedilu (npr. sarkazem), od njegovih življenjskih izkušenj, vzgoje, moralnih vrednot in še česa pa je odvisno, ali oz. kako bo potencialno žaljiv komentar razumel. Zato je pri analizi žaljivosti bistvena interpretacija.

2.1 Sociolingvistične raziskave žaljivega govora

Čeprav je žaljivi govor v svetovnem merilu zaradi razširjenosti interneta zelo aktualen, pa je število raziskav tega področja omejeno. Za zgled smo vzeli nekatere slovenske in tuje sociolingvistične raziskave žaljivega govora.

V slovenskem prostoru se je z žaljivim govorom ukvarjal Vojko Gorjanc (2005), ki je obravnaval žaljivost v štirih slovenskih jezikovnih priročnikih: v *Slovarju tujk*, *Slovarju slovenskega knjižnega jezika*, *Slovenskem pravopisnem slovarju* in *Velikem slovarju tujk*. Ugotovil je, da se v obravnavanih priročnikih pogosto pojavljajo elementi žaljivega govora in da bodo morali biti slovenistični jezikoslovci v prihodnje pri svojem delu družbeno občutljivejši.

S prepoznavanjem pravno določenega sovražnega govora v besedilih so se ukvarjali v novinarskih krogih – raziskovalci so s pomočjo uveljavljenih definicij poskušali prepoznati elemente sovražnosti v izbranem besedilu. Ugotovili so, da obravnavano besedilo vsebuje sovražni govor, katerega bistvena značilnost je namen govorca, da bi nekoga ponižal ali prestrašil. Na (potencialno zelo hude) posledice sovražnega govora je po mnenju raziskovalcev treba čim pogosteje opozarjati (Campos Ferreira et al., 2012).

Z namenom kot nujno sestavino žaljivega govora so se ukvarjali tudi v raziskavah tožb proti novinarjem zaradi t. i. razžalitev. Raziskovalci so se ukvarjali s primeri medijskih razžalitev, in sicer z jezikovnostilskega, pravnega, etičnega, sociološkega in kognitivnega vidika. Za obsodbo zaradi razžalitve je poleg objektivnega žaljivega besedila treba dokazati tudi namen zaničevanja. Raziskovalci izpostavijo, da je žaljivi stil, ki ga izbere novinar, jasen znak njegovega zaničevanja neke osebe, s tem pa tudi namena, da bi osebo javnosti predstavil kot manjvredno. Žaljivost brez namena zaničevanja naj ne bi bila mogoča, saj ima novinar vedno na izbiro tudi nežaljiva jezikovna sredstva (Korošec et al., 2002).

2.2 Prepoznavanje žaljivega govora na novičarskih portalih

Da bi bolje razumeli razloge za izbris nekaterih komentarjev in jih upoštevali pri raziskavi, predstavljeni v pričujočem prispevku, smo se sestali z moderatorji novičarskih portalov MMC RTV Slovenija in Planet Siol.net ter se seznanili z njihovim delom.

Ekipo moderatorjev MMC RTV Slovenija sestavlja manjše število ljudi, izmensko delata dva na dan. Nimajo programske opreme za odločanje o komentarjih, imajo pa določene kategorije (sovražni govor, neprimerno, izven tematike, drugo), v katere lahko uporabnik s klikom na gumb na portalu prijavi žaljivi govor. Moderatorji po lastni oceni izbrišejo okoli 5 % vseh objavljenih komentarjev. Komentarje ročno pregledujejo in jih po potrebi brišejo po kriterijih, ki so vnaprej določeni z internimi pravili in pogoji uporabe spletne strani. V interesu spletnega portala je, da je pod novicami čim več konstruktivnih komentarjev, zato brišejo le res sporne.

Moderatorji Planeta pri svojem delu uporabljajo program, ki jim omogoča pregled komentarjev. V programu se komentarji razvrščajo na odobrene in na čakajoče, moderator mora pri problematičnih komentarjih (zaradi vsebine ali avtorja) potrditi objavo. Program uporablja listo prepovedanih besed, ki jih zaznava ter jih je mogoče spreminjati in dodajati. Politika objavljanja komentarjev je strožja kot na MMC. Na portalu želijo imeti samo dobre komentarje, zato jih brišejo pogosteje kot na MMC. Uporabniki lahko moderatorje opozorijo na sporne komentarje, čeprav ugotavljajo, da se ta možnost pogosto zlorablja za izražanje nestrinjanja z obravnavano tematiko, tudi kadar komentar sam po sebi ni žaljiv.

3 Potek raziskave

Raziskava je potekala v treh fazah, ki jih predstavljamo v tem razdelku: najprej smo izdelali korpus komentarjev na spletne novice, del korpusa smo ročno označili, na koncu pa rezultate označevanja še analizirali in ovrednotili.

3.1 Zajem in obdelava besedil

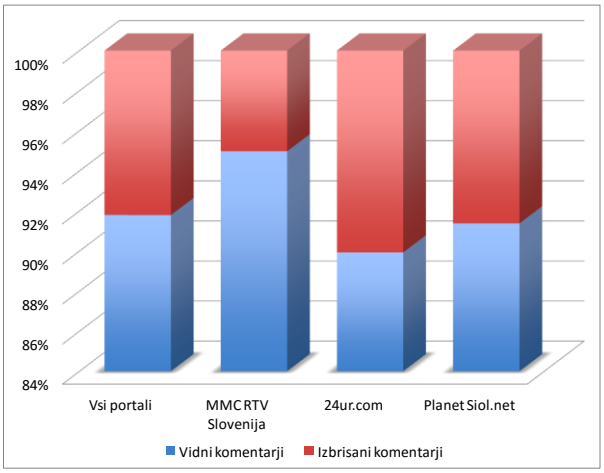
V raziskavi smo zgradili korpus komentarjev na novice s spletnih portalov treh glavnih medijskih hiš v Sloveniji: MMC RTV Slovenija, 24ur.com in Planet Siol.net. Za zajem besedil smo uporabili orodje, ki smo ga razvili v okviru projekta Prava frekvenca 2014 (Pogačnik, 2015), ki uporabniku omogoča celosten pregled nad odzivi na novice in zbira podatke o najpogosteje všečkanih, deljenih in komentiranih novicah.

Orodje omogoča pridobivanje in posodabljanje podatkov; z njegovo pomočjo smo od februarja do julija 2015 vsakodnevno in v realnem času pridobivali novice in komentarje nanje z obravnavanih novičarskih portalov ter nato s primerjavo starih in osveženih zajetih podatkov preverjali, katere izmed objavljenih komentarjev so moderatorji s portala izbrisali. Tako smo dobili bazo primernih in neprimernih komentarjev, ki smo jih razvrstili še po tematiki novice, na katero se nanašajo.

Zajem in obdelava podatkov sta potekala vsak dan ob 4. uri zjutraj, ko je bila obremenitev strežnika najmanjša. Ura je bila izbrana tudi glede na oceno moderatorjev spletnih portalov, da uporabniki največ žaljivih komentarjev objavijo po koncu delovnega časa moderatorjev, torej zvečer oz. ponoči.

Slika 1: Modularna zgradba računalniškega sistema (Blatnik in Jarm, 2015).

S treh izbranih spletnih portalov smo tako zbrali podatke o 29.904 novicah, za katere je bilo objavljenih 981.068 komentarjev. Moderatorji so izbrisali 80.515, torej 8,21 % komentarjev. Najvišji odstotek izbrisanih komentarjev je imel portal 24ur.com (10,07 %), najnižji pa MMC RTV Slovenija (5,02 %). Povprečna dolžina komentarjev je bila 191 znakov, pri čemer je imel najdaljšo povprečno dolžino komentarjev portal MMC RTV Slovenija (284 znakov), najkrajšo pa portal 24ur.com (133 znakov).



Graf 1: Zajeti komentarji – razmerje med vidnimi in pozneje izbrisanimi komentarji.

Zbrani korpus komentarjev smo lematizirali z orodjem LemmaGen (Juršič et al., 2010), kar ni neproblematično, saj je bilo orodje naučeno na standardni slovenščini, ne pa tudi na spletni oz. nestandardni slovenščini. Tako so posameznim besednim oblikam pripisane različne leme, kar vnaša šum pri nadaljnji obdelavi in uporabi korpusa.

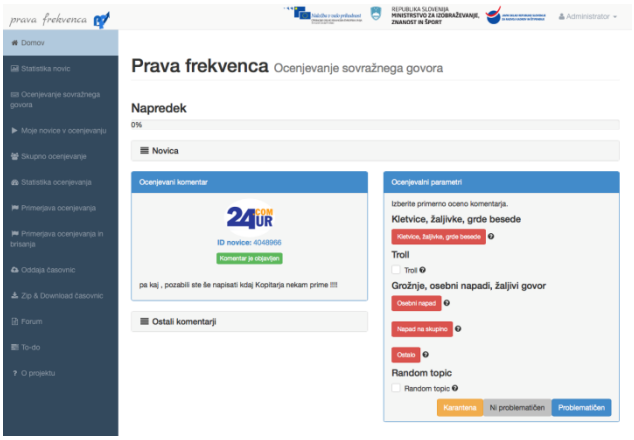
4 Ročno razvrščanje komentarjev v kategorije žaljivosti

Ročno smo pregledali in v kategorije, ki so opisane v nadaljevanju, razvrstili 32.563 komentarjev (tako objavljenih kot izbrisanih) na novice o istospolno usmerjenih in spremembi zakona o zakonskih razmerjih (največ ocenjenih komentarjev), Rimskokatoliški cerkvi, pedofiliji in migracijah.

Razvrščanje komentarjev je potekalo v uporabniškem vmesniku, ki je prikazal izbrano novico in komentarje pod njo. Komentarji, ki so bili s portala izbrisani s strani moderatorjev, so bili posebej označeni. Vsak komentar je v kategorijo razvrstila le ena od treh označevalk.

Na podlagi pregleda literature o žaljivem govoru smo za analizo in razvrščanje komentarjev določili štiri kategorije, pri zadnji pa še tri podkategorije:

1) **nepovezani komentarji** – komentarji, ki niso v skladu s tematiko novice, naključne teme, komentiranje administratorjev in storitev, smetenje, kršenje avtorskih pravic, reklame itd.; ni možnosti označitve problematičnega dela besedila;
2) **trol** – trolanje, norčevanje, posnemanje, izzivanje; ni možnosti označitve problematičnega dela besedila;
3) **kletvice, žaljivke, vulgarizmi** – vsi neprimerni izrazi v obliki besed ali besednih zvez; možnost označitve problematičnega dela besedila;
4) **grožnje, napadi, žaljivi govor** – kategorija je razdeljena na tri podkategorije:
   – osebni napad – grožnje, napadi in žaljivi govor, usmerjeni na posameznika; možnost označitve problematičnega dela besedila;
   – napad na skupino – grožnje, napadi in žaljivi govor, usmerjeni na skupino; možnost označitve problematičnega dela besedila;
   – drugo – druge grožnje, napadi in žaljivi govor; možnost označitve problematičnega dela besedila.



Slika 2: Videz uporabniškega vmesnika za ročno razvrščanje in označevanje komentarjev.

Protokol označevanja je bil sestavljen iz štirih korakov. Najprej je bilo treba presoditi, ali je komentar primeren ali neprimeren. Pri neprimernih komentarjih je sledilo izločanje tistih z nerelevantno vsebino – torej prispevkov trolov (namerno provokativni komentarji)[2] in komentarjev, ki niso v skladu s tematiko komentirane novice. Ti kategoriji smo vključili, ker imata pomembno vlogo v pravilih objavljanja komentarjev na spletnih portalih – nerelevantna vsebina je zelo pogost razlog za izbris komentarja s strani moderatorja.[3] Temu je sledila še

[2] Trolanje razumemo kot tukaj: http://www.nytimes.com/2010/11/30/opinion/30zhuo.html?_r=0.

[3] Pravila komuniciranja so za obravnavane spletne portale dostopna na naslovih: http://www.rtvslo.si/strani/moj-splet/669#pravila, http://www.siol.net/subsites/pravila_komentiranja.aspx in http://image.24ur.com/media/document/61374531.pdf.

odločitev, ali komentar vsebuje neprimerne besede (kletvice, žaljivke ali vulgarizme) ali dejanski žaljivi govor. Ta je nadalje razdeljen na tri podkategorije, in sicer glede na to, na koga je žalitev usmerjena. Vsak komentar je bil lahko uvrščen v več kategorij. Tako je npr. neki komentar lahko hkrati vseboval kletvice in žalil določeno skupino ljudi, pri razvrstitvi takega komentarja pa je bilo treba natančno označiti del besedila, ki je bil sporen iz enega ali drugega razloga.

5 Analiza rezultatov

Ročno razvrščenih 32.563 komentarjev smo v 74,14 % primerov uvrstili med primerne, v 25,86 % primerov pa med neprimerne komentarje. Zastopanost kategorij neprimernih komentarjev v označenem učnem korpusu je predstavljena v tabeli 1.

Kategorija	Odstotek komentarjev
Nepovezani komentarji	24 %
Trol	8 %
Kletvice, žaljivke, vulgarizmi	13 %
Grožnje, napadi, žaljivi govor:	55 %
– osebni napad	19 %
– napad na skupino	59 %
– drugo	22 %

Tabela 1: Zastopanost kategorij neprimernih komentarjev po ročnem razvrščanju.

Sledi nekaj primerov komentarjev po kategorijah. Krepko je zapisan tisti del komentarja, ki so ga označevalke označile kot problematičnega. To je bilo mogoče le pri dveh kategorijah, in sicer pri kletvicah, žaljivkah in vulgarizmih ter pri grožnjah, napadih in žaljivem govoru.

Primeri nepovezanih komentarjev:

> so me zbrisali cenzorji fašistični, zato sem napisal še enkrat

> Kar brišite. To je VAŠ kodeks. Vaš domet razumevanja.... Potem pa se čudite? A imate potem o čem pisati?

> Primca zopet kar nekaj časa ne bo v službo, kjer dobiva plačo iz državne blagajne. Evo, admin, lahko zopet brišeš.

Primeri trolanja:

> Treba je spoštovati pravico staršev homoseksualcev do vnukov.

> Fantek, kako se kliče tvoj očka? Marjan. Kaj pa mamica? Vinko.

> Otrok = Mama + Oče
> Otrok je tudi mama + poštar ali pa duhovnik + deklica

Primeri kletvic, žaljivk oz. vulgarizmov:

> Pa kaj tem **kretenom** ni jasno?!!

> Do tega pride ko te začnejo **pedri** posiljevati.

> Sm pa zihr da vsi tipi, ki so tako proti homoseksualcem, **drkajo** na lezbične porniče...

Primeri groženj, napadov oz. žaljivega govora:

> Še nikjer pa nisem zasledil, da bi se homoseksualci komu maščevali. Tudi to kaže kakšen **lažljivec je Primc**.

> Cerkev je proti, ker **duhovnikom ni tako zabavno in vznemrljivo z s polnoletnimi ministranti in farani moškega spola**, če se legalizira. Hipokrati.

> Imajo MOŽNOST, da se postavijo v dolgo čakalno vrsto, nič drugega kot to! BRAVO G.Bruno, **tako pač dela SDS, tudi podpise ponareja**!

Ujemanje med označevalkami in posledično zanesljivost pripisanih kategorij in oznak smo preverili na podlagi testnega podkorpusa 553 komentarjev na eno izmed novic. Te komentarje so v kategorije razvrščale vse tri označevalke. Kljub natančno določenim kategorijam se je izkazalo, da se označevalke pri razvrščanju niso ujemale v želeni meri (želeli smo doseči ujemanje v vsaj polovici primerov). V tabeli 2 so predstavljeni rezultati razvrščanja testne množice.

	Število objavljenih oz. sprejemljivih komentarjev	Število izbrisanih oz. nesprejemljivih komentarjev
Moderatorji	410	143
Označevalka 1	495	58
Označevalka 2	483	70
Označevalka 3	456	97

Tabela 2: Razvrstitev komentarjev iz testne množice.

Iz tabele je razvidno, da so označevalke bistveno manj komentarjev razvrstile med neprimerne kot moderatorji. Razlog za to je najverjetneje v strogih pravilih objavljanja na spletnih forumih, kjer moderatorji pogosto brišejo tudi odgovore na sporne komentarje, ki pa sami niso nujno neprimerni. Bolj problematičen je podatek o razlikovanju med označevalkami – označevalka 1 je v primerjavi z ocenjevalko 3 med neprimerne uvrstila skoraj 60 % manj komentarjev.

Še večje razlike so se pojavile pri pregledu strinjanja označevalk med seboj, kar je razvidno iz tabele 3 – že pri odločitvi, ali gre za neprimeren komentar, in pri uvrstitvi v kategorijo znotraj žaljivega govora je strinjanje označevalk ponekod nižje od 20 %. Stanje je podobno v vseh kategorijah.[4]

[4] Primerov nestrinjanja med označevalkami v prispevku nismo navedli, ker s spletnimi portali nismo dosegli dogovora o objavi izbrisanih komentarjev. Vsi navedeni komentarji so objavljeni ali pa so bili opaženi na spletnih straneh pred izbrisom.

	Neprimerno	Žaljivi govor
Moderatorji	143	99
Strinjanje O1 + O2 + O3	27	9
Strinjanje O1 + O2	29	17
Strinjanje O1 + O3	36	10
Strinjanje O2 + O3	44	23
Strinjanje samo O1 + O2	2	8
Strinjanje samo O1 + O3	19	1
Strinjanje samo O2 + O3	17	14
Ni strinjanja	78	67

Tabela 3: Strinjanje ocenjevalk pri odločanju o neprimernosti komentarjev in razvrščanju v kategorije žaljivosti.

Sklepamo, da je razlog za nestrinjanje v subjektivnosti odločanja – kaj je za nekoga žaljivo, je odvisno od številnih dejavnikov, npr. od spola, izobrazbe, vzgoje itd. Pomembna sta tudi kontekst, ki iz posameznega komentarja ni nujno razviden, in namen žaljivosti, ki ga lahko posameznik v besedilu prepozna ali pa ne.

6 Sklep

Zaradi anonimnosti se je v zadnjih letih v spletni komunikaciji razširil žaljivi govor, ki je s svojimi specifičnimi značilnostmi zanimiv za sociolingvistične raziskave. Sovražni govor, ki je pravno določena različica žaljivega govora, je težji za analizo, saj je primerov zaradi zapletenosti dokazovanja bistveno manj.

V prispevku smo predstavili analizo žaljivega govora v spletnih komentarjih. Rezultat projekta je korpus ročno razvrščenih komentarjev, ki je del večje baze zajetih novic s pripadajočimi komentarji. Korpus in baza omogočata številne nadaljnje raziskave žaljivega govora. Ob nadgradnji lematizacije in ponovnem ovrednotenju ročno razvrščenih komentarjev je lahko korpus primeren za morebitne raziskave avtomatskega zaznavanja žaljivega govora, uporabiti pa bi ga bilo mogoče tudi v raziskavah leksikalnih in skladenjskih značilnosti tako žaljivega govora kot nestandardne slovenščine.

Problematika žaljivega govora je v času prostega dostopa do spletne komunikacije zelo aktualna, zato je nanjo nujno treba opozarjati – tudi z raziskavami, ki bi lahko pripomogle k učinkovitejšemu nadzoru nad žaljivim govorom v spletnih komentarjih. Nekateri spletni portali se zaradi nezmožnosti učinkovitega nadzora nad spletnimi vsebinami zatekajo k drastičnim ukrepom, kot sta npr. ukinjanje možnosti komentiranja in uvajanje drugih besedilnih tipov (npr. pisma bralcev). Menimo, da so taki ukrepi pretirani, saj siromašijo besedilnotipsko raznovrstnost spletne komunikacije. Raziskave žaljivosti in razvoj korpusov žaljivega govora lahko pripomorejo k ohranjanju pestre spletne komunikacije in preprečevanju posledic neprimernega izražanja na spletu.

7 Zahvala

Interdisciplinarni študentski projekt, pri katerem so sodelovali Fakulteta za elektrotehniko in Filozofska fakulteta Univerze v Ljubljani ter podjetje Percipio, je potekal od februarja do julija 2015 in je bil financiran s sredstvi iz razpisa Po kreativni poti do praktičnega znanja Javnega sklada RS za razvoj kadrov in štipendije. Ob pomoči pedagoških mentorjev Matevža Pogačnika in Dana Podjeda ter delovnega mentorja Andreja Duha so pri projektu poleg avtoric sodelovali še naslednji študenti: Bernarda Baš, Aljaž Blatnik, Kaja Jarm, Jan Vidic in Gaja Naja Rojec. Za njihov doprinos se jim avtorici prispevka iskreno zahvaljujeva.

8 Literatura

Aljaž Blatnik in Kaja Jarm. 2015. Priloga k poročilu. Interno gradivo. Fakulteta za elektrotehniko UL.

Raphael Campos Ferreira, Petra Košič, Noemi Mavrič in Nina Mihalič. 2012. Kriteriji za prepoznavanje sovražnega govora v jeziku: študija primera. Teorija in praksa, 49(1): 204–215. http://dk.fdv.uni-lj.si/db/pdfs/TiP2012_1_Campos-Ferreira_idr.pdf.

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

Vojko Gorjanc. 2005. Neposredno in posredno žaljiv govor v jezikovnih priročnikih: diskurz slovarjev slovenskega jezika. Družboslovne razprave, 21(48): 197–209. http://dk.fdv.uni-lj.si/dr/dr48Gorjanc.PDF.

Matjaž Juršič, Igor Mozetič, Tomaž Erjavec in Nada Lavrač. 2010. LemmaGen: multilingual lemmatisation with induced Ripple-Down rules. Journal of Universal Computer Science, 16(9): 1190–1214. http://lemmatise.ijs.si/Download/File/Documentation%23JournalPaper.pdf.

Tomo Korošec, Monika Kalin Golob, Simona Zatler, Melita Poler, Maca Jogan in Gregor Tomc. 2002. Razžalitve v tiskanih medijih. Ljubljana: Založba FDV.

Mari J. Matsuda, Charles R. Lawrence, Richard Delgado in Kimberle Williams Crenshaw. 1993. Words That Wound. Boulder: Westview Press.

Patrick B. O'Sullivan in Andrew J. Flanagin. 2003. An Interactional Reconceptualization of "Flaming" and Other Problematic Messages. SAGE Publications, 5(1): 67−93.

Matevž Pogačnik. 2015. Končno poročilo o izvedenih projektnih aktivnostih. Interno gradivo. Fakulteta za elektrotehniko UL.

Tadej Praprotnik. 2003. Pragmatični vidiki žaljive komunikacije v računalniško posredovani komunikaciji – multipla perspektiva. Teorija in praksa 40(3): 515−540.

Vasja Vehovar, Andrej Motl, Lija Mihelič, Boštjan Berčič in Andraž Petrovčič. 2012. Zaznava sovražnega govora na slovenskem spletu. Teorija in praksa, 49(1): 171−189. http://safe.si/sites/safe.si/files/vehovar_et_al._-_zaznava_sovraznega_govora_na_slovenskem_spletu.pdf.

Arheologija začetnice pri stvarnih lastnih imenih

Iza Škrjanec,* Damjan Popič,† Darja Fišer†

*Ljubljana
skrjanec.iza@gmail.com

†Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
damjan.popic@ff.uni-lj.si, darja.fiser@ff.uni-lj.si

Povzetek

Prispevek obravnava zapis stvarnih lastnih imen, in sicer se osredotoča na rabo začetnice pri industrijskih imenih. Najprej je na kratko predstavljena njihova obravnava v slovenskem pravopisju, ki jo je usmerjala tudi težnja po odpravljanju tujih zgledov, nato pa na vzorcu štiridesetih najpogostejših industrijskih imen iz korpusa spletne slovenščine Janes poskušamo ugotoviti, kakšno rabo odražajo besedila novih medijev in kako se ta razlikuje od rabe v korpusu Kres.

The Archaeology of Proper Names Capitalization

The paper deals with the orthography of proper names and focuses on the usage of product names capitalization. Firstly, the rules for capitalization in Slovene orthography guides are discussed, as they have been subject to the tendency of restricting foreign influences. Then, a pattern of forty most frequent product names from the Janes corpus of internet Slovene is analysed in terms of its capitalization usage in new media texts and in terms of potential differences between the Janes corpus and the Kres corpus.

1 Uvod

V nasprotju z nekaterimi drugimi skupinami stvarnih lastnih imen gre pri imenih industrijskih oz. tehničnih izdelkov[1] za relativno novo kategorijo: šele od leta 1981 so samostojna podskupina stvarnih lastnih imen v slovenskem pravopisju, zato se v prispevku podrobneje posvetimo njihovim vzorcem v rabi. Kot pri nekaterih drugih stvarnih lastnih imenih je tudi pri industrijskih imenih zapis začetnice problematičen zaradi teženj po odpravljanju »tujih« zgledov v slovenski kodifikaciji, zato nas zanima ravno zapis začetnice v večjezičnem kontekstu spletne komunikacije.

Po slovenski pravopisni tradiciji ločujemo lastna in občna imena, meja, ki jih razločuje, pa je pogosto zabrisana. V vmesni prostor spadajo tudi tista lastna imena, ki jih v določenih vzorcih rabe lahko razumemo občno, saj naj bi šlo za izlastnoimenska vrstna poimenovanja. Kadar uporabnik ne misli na dejansko ime, temveč na njihovo vrsto, poimenovanja piše z malo začetnico in to »velja za vse vrste stvarnih lastnih imen« (SP 2001, § 104). Zapis izlastnoimenskih vrstnih poimenovanj je precej težaven, na kar kažejo tudi pogosti »popravki zapisov (začetnice, op. a.), ki so tudi v slovenskem pravopisju obravnavani nekonsistentno oz. vsiljujejo koncept ločevanja vrstnosti/lastnoimenskosti« (Popič, 2014). Vrstnost izlastnoimenskih poimenovanj tako služi kot poskus, da bi preprečili preobilje velike začetnice (Pogorelec, 1975).

Med stvarna lastna imena, pri katerih je zapis začetnice posebno problematičen, spadajo tudi industrijska imena. Po trenutno referenčni kodifikaciji jih zapisujemo z veliko začetnico, kadar se pojavljajo kot imenovalni prilastki k vrstnim imenom. V nasprotnem primeru so zapisana z malo začetnico in se uporabljajo kot sklonljivi samostalniki, saj jih razumemo kot stvarna imena, ki so prešla med vrstna, torej so se apelativizirala. Obravnava industrijskih imen po skladenjskem kriteriju ne predvideva, da lahko beseda ohrani lastnoimensko funkcijo, obenem pa je zapisana z malo začetnico, kar pomeni, da pri tem ni upoštevana pomenska vloga posamezne besede.

2 Namen članka

Na primeru štiridesetih imen industrijskih izdelkov in znamk s področja informacijske tehnologije in avtomobilizma poskušamo ugotoviti vzorce v rabi začetnice. Primere smo zajeli v korpusu spletne slovenščine Janes,[2] ki obsega tvite, forumska sporočila, blogovske zapise in komentarje na spletne novice (Fišer et al., 2015). Pri tovrstnih besedilih gre povečini za neformalno komunikacijo v t. i. internetnem jeziku (*netspeak*), kar med drugim pomeni, da »je značilna pogosta raba nestandardnih jezikovnih oblik, kot je nestandarden (bolj fonetičen) zapis besed (npr. izključno male tiskane črke, opuščanje večine ločil in večkratno ponavljanje črk za čustveno poudarjanje zapisane izjave), in pogoste specifične okrajšave« (Fišer et al., 2014). Za namen primerjave rabe začetnice v drugih besedilnih tipih smo uporabili vzorčni podkorpus Gigafide Kres,[3] v katerem najdemo približno enak delež leposlovnih, stvarnih, časopisnih, revijalnih in internetnih besedil, torej povečini jezikovno pregledana besedila. S sopostavitvijo korpusnih podatkov želimo ugotoviti, ali raba začetnice variira glede na besedilni tip in ali sledi aktualnim pravopisnim pravilom.

[1] V prispevku besedno zvezo 'industrijska imena' razumemo kot imena industrijskih izdelkov.

[2] Korpus je nastal v okviru nacionalnega raziskovalnega projekta Janes: http://nl.ijs.si/janes/ (dostop 19. 8. 2015).

[3] Korpus Kres: http://www.korpus-kres.net/ (dostop 25. 8. 2015).

3 Pregled kodifikacije začetnice pri stvarnih lastnih imenih

Analiza kodifikacije pravil za začetnico pri stvarnih lastnih imenih nam ponudi vpogled v različne, nasprotujoče si tendence v slovenski normativistiki, ki zajema tudi zapis začetnice. V pričujočem poglavju obravnavamo slovenske pravopise, ki so izšli med letoma 1899 in 1950, Načrt pravil za novi slovenski pravopis, SP 2001 in ob koncu na kratko kontrastivno predstavimo nekatere razlike s sodobno angloameriško kodifikacijo.

3.1 Kodifikacija pravil v prvih petih pravopisih

Vse do izida Načrta pravil za novi slovenski pravopis lahko v kodifikacijskih priročnikih opazimo spremembe stališč do zapisa začetnice pri stvarnih lastnih imenih. Za primer vzemimo pravila za zapis sestavljenih imen ustanov, uradov in časopisov, pri katerih je prvi del pridevnik in drugi samostalnik. V SP 1899 in SP 1920 so vsi deli tovrstnega imena zapisani z veliko začetnico, če so se v rabi ustalili kot lastna imena (*Slovenska Matica*, *Slovenski Pravnik*, *Ljubljanski Zvon*). SP 1920 za imena ustanov, ki so sestavljena iz občnih imen, predvideva zapis prvega dela z veliko le, kadar je rabljeno celotno ime (*Gospodarska zveza*). Zanimivo je, da SP 1935 ne obravnava imen, kot je *Slovenski Pravnik*, temveč zgolj našteje primere iz preteklih pravopisov.

Povojna pravopisa (SP 1950 in 1962) prikazujeta nasprotujočo si kodifikacijo pravil. Po SP 1950 pri sestavljenih imenih časopisov, knjig, društev in podjetij uporabimo veliko začetnico tudi za neprve besede, če gre za lastna imena. Priročnik opozarja, da je pri nekaterih imenih v rabi pisava z veliko začetnico v obeh delih (*Slovenska Matica*, *Slovenski Narod*, *Ljubljanski Zvon*). Imena uradov in ustanov po SP 1950 zapisujemo z malo, kadar so rabljena občno (*ministrstvo*, *fakulteta*). Če je ime rabljeno kot določeni naslov, postane lastno in je zapisano z veliko začetnico (*Ministrstvo za prosveto v Ljubljani*, *Filozofska fakulteta v Ljubljani*). Slednje pravilo je naletelo na kritike strokovne javnosti, saj naj bi šlo v tem primeru za posnemanje tujih zgledov, pri katerih se z veliko začetnico pišejo vsa lastna imena (Tomšič, 1955). Tomšič (1955/56) opozarja celo, da bi razmah velike začetnice v slovenščini lahko pomenil, da bi z veliko začetnico zapisovali tudi vse samostalnike, kakor je navada v nemščini.

SP 1962 glede na predhodne pravopise postavi pravila, ki za zgoraj obravnavane skupine stvarnih lastnih imen uvajajo zapis z malo začetnico. Kot novost predstavi tudi imena, ki »imajo poleg lastnosti lastnih imen /…/ še močan pomen občnih imen, tako da pomenijo poleg enkratnosti določene stvari hkrati tudi vrsto tiste stvari« (SP 1962, § 42). Tovrstna imena opredeli kot vrstna lastna imena, pišejo pa se z malo začetnico (*železniška postaja Celje*, *okrožno sodišče v Ljubljani*); v nadaljevanju pravopis eksplicitno pojasni, da jih od ostalih imen ne ločuje začetnica, temveč kraj njihovega sedeža, zato naj bi bil zapis z malo začetnico pomensko jasen in nedvoumen. V nasprotju s SP 1950 v SP 1962 ni pravila za zapis sestavljenih stvarnih imen, v katerih je sicer občno ime zapisano z veliko (*Slovenska Matica* v SP 1950), najdemo pa zapis *Slovenska matica*. Tudi imena za različne upravne organe, ustanove, šole itd. uvršča med vrstna imena in zanje priporoča zapis z malo začetnico (*državni sekretariat za pravosodno upravo*, *filozofska fakulteta univerze v Ljubljani*).

Povojna pravopisa v pravilih sicer ne določata, da gre pri imenih izdelkov za stvarna lastna imena, v poglavjih za nekatera druga pravila pa najdemo zglede za zapis (*motorji Diesel*, *žarnice Phillips*, *baterije Zmaj*; *Dieslovi motorji*, *Phillipsove žarnice*). V slovarskem delu obeh pravopisov je konsistenca pri rabi začetnice manj jasna (SP 1950: *cigareta Drava*, *z Drava mi postrezi*; SP 1962: *imam še eno drava/dravo*, *postregel mi je z drava*).

3.2 Načrt pravil za novi slovenski pravopis in aktualna kodifikacija

Načrt pravil za novi slovenski pravopis[4] je pri stvarnih lastnih imenih natančneje razdelal obstoječe podskupine. Videti je, da se je pri tem opiral na pravila in tendenco po mali začetnici iz SP 1962 – kadar namreč mislimo le na vrsto podjetij, ustanov in organov, se ta zapisujejo z malo (*študira na univerzi v Mariboru*) – vendar za isto skupino stvarnih lastnih imen predvidi tudi zapis z veliko začetnico, kadar gre za naslove (*Univerza v Mariboru*, *Železniška postaja Celje*). Tudi odzivi na Načrt izpostavljajo »neutemeljeno širjenje rabe velike začetnice« pri zapisu družbenopolitičnih in oblastvenih organov ter zakonov (Majdič, 1983).

Načrt predstavi nekatere nove skupine stvarnih lastnih imen, med katerimi so imena »tehničnih izdelkov« in nazivi »serijskih tehničnih izdelkov«. V primerjavi z denimo imeni podjetij, ustanov ali organov začetnica pri tehničnih izdelkih ni utemeljena s tem, kaj uporabnik misli (vrstna ali lastna raba), temveč se opira na skladnjo – kadar je ime proizvajalcev izdelka rabljeno kot imenovalni prilastek, ima veliko začetnico. V nasprotnem primeru lastno ime stoji brez občnoimenskega vrstnega poimenovanja in je rabljeno kot sklonljivi samostalnik, zapisano pa z malo začetnico. V času priprave Načrta so avtorji zagotovili, da bo imel novi pravopis »glede pisave velike in male začetnice natančna določila, utemeljena s pomensko vlogo posamezne besede ali besedne zveze v konkretnem besedilu« (Pogorelec, 1975), iz Načrta pa lahko razberemo, da uporabnik pravzaprav ne odloča sam o zapisu začetnice, saj je ta pogojena s tem, da je pri neprilastkovni rabi možen zgolj vrstni pomen, pri prilastkovni pa lastnoimenski (*avto znamke Fiat*, *vozim se s fiatom*), kar je problematično predvsem z vidika (ne)prilagodljivosti kodifikacije jezikovni ekonomičnosti.

Med pravili za zapis imen tehničnih izdelkov v Načrtu in SP 2001 ni prišlo do sprememb, v SP 2001 se pojavijo dodatni zgledi, med katerimi izstopata dva (*zobe si umivam s kalodontom*, *komarje uničujem s pipsom*). Za razliko od ostalih ponazoritvenih primerov gre pri imenih *kalodont* in *pips* za apelativizirani imeni, ki sta izgubili lastnoimenske lastnosti (gl. Dobrovoljc, 2009), česar priročnik ne izpostavi.

SP 2001 torej predvideva sledeči proces apelativizacije: lastna imena izdelkov serijske proizvodnje prehajajo med občna, in sicer lahko postanejo generična, kar pomeni, da poimenujejo vse tovrstne izdelke, ne le ene vrste (*kalodont*, *superge*),[5] ali pa postanejo vrstna in se nanašajo le na tip nekega izdelka (*oliveti*, *ford*). Za kriterij razlikovanja med

[4] V nadaljevanju Načrt.

[5] Za seznam izbranih apelativiziranih imen glej Dobrovoljc (2009).

tremi kategorijami imen SP 2001 določi začetnico, ki je pri lastnih imenih velika, pri občnih pa mala. Kriterij je problematičen zlasti zato, ker »izpust skladenjskega jedra (in tako zapis z malo začetnico, op. a.) besedne zveze vedno ne implicira izgube lastnoimenske funkcije« (Dobrovoljc, 2012), kar pomeni, da se lastnoimenskost imen, ki so po SP 2001 zapisana z malo začetnico, lahko ohrani tudi v neprilastkovnih skladenjskih položajih.

4 Korpusna analiza najpogostejših industrijskih imen

V korpusu spletne slovenščine Janes smo izdelali seznam štiridesetih najpogostejših industrijskih imen, za katera smo preverili ustaljenost zapisa začetnice v korpusu Janes in korpusu Kres ter primerjali rezultate, pri tem pa skušali ugotoviti, ali raba odseva ločevanje med vrstnimi in lastnimi imeni z začetnico.

Izhajamo iz predpostavke, da bo analiza pokazala zanimive razlike ali podobnosti med rabo začetnice, saj se korpusa razlikujeta po zajemu tipov besedil glede na okoliščine nastanka. Korpus Janes v celoti sestoji iz uporabniških spletnih vsebin, ki jih tvorijo uporabniki v neformalni računalniško posredovani komunikaciji. Korpus skupno šteje 161.289.153 pojavnic. V korpusu Kres najdemo uravnotežene deleže leposlovnih, stvarnih, revijalnih, časopisnih, spletnih in drugih besedil. S spleta izvira le 20 % celotnega korpusa ali približno 24.895.514 pojavnic, od tega je 12 % besedil s spletnih strani ustanov in podjetij, 8 % besedil pa izvira z novičarskih portalov (8.000.131 pojavnic). Korpus Janes tako v celoti sestoji iz besedil novih medijev, v korpusu Kres pa je takšnih besedil za 0,4 %.

S pomočjo regularnih izrazov smo iz celotnega korpusa Janes izluščili seznam lastnih imen, iz katerih smo ročno izbrali le industrijska imena.[6] Seznam obsega štirideset imen, ki se delijo v tri tematska področja: a) elektronske naprave, b) programska oprema in spletne aplikacije ter c) avtomobilizem. V nadaljevanju bomo v korpusih Janes in Kres preverjali ustaljenost zapisa dobljenih imen, in sicer se bomo osredotočili na rabo začetnice v imenovalniški obliki imen, saj je pogostnost imen v ostalih sklonih zanemarljiva.

4.1 Imena elektronskih naprav

Elektronske naprave so serijsko proizvedeni izdelki, kar pomeni, da se njihovo ime nanaša na celotno serijo primerkov istega modela.

Z izjemo imena Bosch se vsa imena v imenovalniški obliki v korpusu Janes pojavljajo pogosteje kot v korpusu Kres, kar prikazuje tabela 1. Velika začetnica v Kresu vidno prevladuje, medtem ko so razlike v deležu med začetnicama v korpusu Janes manjše. Kljub različni besedilnovrstni sestavi korpusov v obeh prevladuje zapis imen z veliko začetnico, in sicer je v korpusu Janes zapis z veliko začetnico vsaj sedemkrat pogostejši, v korpusu Kres pa vsaj petkrat.

Ime elektronske naprave	Janes				Kres			
	Velika	Velika (%)	Mala	Mala (%)	Velika	Velika (%)	Mala	Mala (%)
Sony	1.821	88,6 %	235	11,4 %	532	97 %	15	3 %
Apple	2.771	85 %	483	15 %	405	91 %	38	9 %
Samsung	2.599	90,7 %	267	9,3 %	344	98 %	7	2 %
Nokia	1.775	85,7 %	297	14,3 %	394	95 %	20	5 %
Bosch	13	93 %	1	7 %	148	98,7 %	2	1,3 %
Galaxy	1.867	78,1 %	262	21, 9 %	95	84,8 %	17	15,2 %
iPhone	Iphone (301)	6,3 %	iPhone (3.588) iphone (857)	75,6 % 18,1 %	Iphone (6)	4,3 %	iPhone (122) iphone (13)	86,5 % 9,2 %

Tabela 1: Pogostnost in delež začetnic v imenih elektronskih naprav v korpusih Janes in Kres.

V korpusu Kres je večje število konkordanc, v katerih se ime izdelka uporablja kot imenovalni prilastek. Z izpustom skladenjskega jedra se pomen zveze ne spremeni. V spodnjih primerih je ime Samsung zapisano z veliko začetnico, v prvi povedi je uporabljeno kot imenovalni prilastek, v drugi kot sklonljivi samostalnik.

Si.mobil je svoji ponudbi Re.misli, ki vsebuje solarne in varčne polnilnike, dodal še z okoljem prijazen ***telefon Samsung SGH-E200 ECO****.* [Kres, besedilni tip: revija]

Danes sem poteskala ***Samsung Galaxy SIII*** *- joj, kako je lahek! Skor prelahek zame.* [Janes, besedilni tip: tvit]

Primera se razlikujeta v kontekstu (ne)formalne komunikacije: v prvem primeru gre za besedilo oglaševalske narave, drugo besedilo pa je veliko bolj osebno. V obeh je poleg imena naveden tudi tip izdelka. V korpusu Janes so pogosti tudi zapisi imen izdelkov kot neujemalnih levih prilastkov, kar bi lahko pripisali tudi vplivu angleščine, kjer so samostalniški sklopi pogosta struktura. V korpusu Janes se kot neujemalni levi prilastki pojavijo tako lastna (*Sony izdelki*, *Power point predstavitev*) kot občna imena (*mobi številka*).

[6] Pri večini imen gre za večdenotatne pojave, saj lahko pomenijo podjetje, znamko ali izdelek, kar je potrebno upoštevati tudi pri interpretaciji statističnih podatkov.

4.2 Imena programske opreme ali spletnih aplikacij

Jezikovna raba slovenščine kaže na nekatere neevidentirane kategorije, med katerimi so tudi imena računalniških programov, operacijskih sistemov, spletnih aplikacij in spletnih strani ter omrežij (Dobrovoljc, 2009). Imena programske opreme in spletnih aplikacij se nekoliko razlikujejo od preostalih industrijskih imen, saj ne gre za fizične izdelke, temveč za spletne storitve (*Google*, *Twitter*, *Instagram*, *YouTube*, *Facebook*, *SiOL*) ali operacijske sisteme (*Android*, *iOS*, *Windows*).

Tudi pri poimenovanjih operacijskih sistemov si težko pomagamo s konceptoma občnosti in vrstnosti, kot sta prikazana v SP 2001, saj tudi programska oprema ni fizični izdelek, temveč lastniško zaščitena računalniška koda.

Kot lahko vidimo v tabeli 2, v obeh korpusih prevladuje zapis z veliko začetnico, saj je tako v korpusu Janes kot v korpusu Kres velika začetnica vsaj enkrat pogostejša. Zanimivo je, da je v obeh korpusih najmanjša razlika med deležem velike in male začetnice imen *YouTube* in *Google*. V korpusu Janes pri imenu *YouTube* je razlog za to lahko šum v obliki podvojenih zapisov ali napačno označevanje, v korpusu Kres pa se mala začetnica pri tem imenu pojavi po večini v spletnih besedilih.

Ime spletnega iskalnika Google je v korpusu Janes z veliko začetnico zapisano dvakrat pogosteje, v korpusu Kres pa kar štirikrat, kar je relativno malo glede na vrednosti za ostala imena v obeh korpusih Ker je iskalnik Google med najbolj znanimi in priljubljenimi iskalniki, lahko predvidimo možno apelativizacijo – ta se namreč »izvede le pri imenih, ki so bila v daljšem časovnem obdobju pogosteje rabljena in pogosto monopolna« (Dobrovoljc, 2012).

Komentar si zaslužijo dvozačetniške e-tvorjenke,[7] pri katerih je z veliko začetnico zapisana druga črka imena, medtem ko je prva črka zapisana z malo (*iPhone*, *iOS*), in imena, v katerih se velike črke pojavijo znotraj besede (*YouTube*, *SiOL*). Pri poimenovanju mobilnega telefona iPhone, operacijskega sistema iOS ter spletnih portalov YouTube in SiOL v obeh korpusih prevladuje originalni zapis. Z vprašanjem zapisa e-tvorjenk v imenih se bo morala kodifikacije slovenščine še spoprijeti, kakor so to storili že nekateri tuji slogovni priročniki (gl. CMOS).

Ime programske opreme/spletne aplikacije	Janes				Kres			
	Velika	Velika (%)	Mala	Mala (%)	Velika	Velika (%)	Mala	Mala (%)
Android	2.735	80,8 %	651	19,2 %	228	86,4 %	36	13,6 %
iOS	/	/	iOS (90)	100 %	/	/	iOS (35) ios (1) iOs (1)	94,6 % 2,7 % 2,7 %
Windows	1.906	79,5 %	492	20,5 %	2.783	97,6 %	67	2,4 %
Google	5.659	69,6 %	2.467	30,4 %	420	80,2 %	103	19,8 %
Twitter	2.536	54 %	2.165	46 %	118	92,2 %	10	7,8 %
Instagram	520	67,1 %	255	32,9 %	/	/	/	/
YouTube	YouTube (1.264) Youtube (575)	29 % 13,2 %	2.514	57,8 %	YouTube (130) Youtube (28)	49,2 % 10,6 %	106	40,2 %
Facebook	2.934	75,3 %	960	24,7 %	328	85,6 %	55	14,4 %
SiOL	SiOL (2.024) Siol (1.269)	52,2 % 32,7 %	587	15,1 %	SiOL (369) Siol (107)	69,2 % 20,1 %	57	10,7 %

Tabela 2: Pogostnost in delež začetnic v imenih programske opreme in spletnih aplikacij v korpusih Janes in Kres.

4.3 Imena avtomobilov

Imena avtomobilov in avtomobilskih znamk so edina od kategorij, obravnavanih v tem prispevku, za katera je zapis določil že Načrt in kasneje SP 2001, in predstavlja tudi najštevilčnejšo kategorijo v pričujočem prispevku. SP 2001 ločuje med imeni posameznih vozil in vozili serijske proizvodnje, kamor spadajo tudi avtomobili. Velika začetnica imena avtomobila naj bi bila utemeljena, kadar je ime imenovalni prilastek, v nasprotnem primeru pa naj bi ime poimenovalo vrsto avtomobila oz. primerek iz serije izdelkov. V tabeli 3 sta predstavljena število pojavnic in delež začetnic za 24 imen avtomobilov v korpusih Janes in Kres.

Iz tabele lahko razberemo dve skupini imen avtomobilov, in sicer imena avtomobilskih znamk in vrste modelov, ki so deležni enake pravopisne obravnave.[8]

[7] Za razlikovanje med dvozačetniškimi in vezajnimi e-tvorjenkami glej Rebernik (2015).

[8] V slovarskem delu SP 2001 tako najdemo geslo vólkswágen - gna [folksvagən] m, tudi živ. (o͡á); gl. folksvagen in geslo gólf 2 -a m, tudi živ. (o͡) |avtomobil|.

Imena vrst modelov lahko nastopajo skupaj z imenom znamke ali samostojno, kakor prikazujeta spodnja primera.

*Zanima me predvsem, če obstaja kakšen , ki ima vgrajene ledice, približno tako, kot jih ima **Renault Megane RS***. [Janes, besedilni tip: forumsko sporočilo]

*Verjetno se bo slika malo obrnila ko pride novi **Twingo**, naslednje leto pa še novi **Megane***. [Janes, besedilni tip: forumsko sporočilo]

Pri štirih imenih iz korpusa Janes (*Audi, Honda, Megane, Mustang*) in le petih imenih (*Megane, Octavia, Jaguar, Passat, Mustang*) iz korpusa Kres več kot tretjina pojavnic pogosteje zapisana z malo začetnico, sicer prevladuje velika začetnica. V povprečju je v korpusu Janes z veliko začetnico zapisanih 72,1 % imen avtomobilov, v korpusu Kres pa 72,3 %, kar pomeni, da je raba začetnice v splošnem precej podobna.

Ime avtomobila	Janes				Kres			
	Velika	Velika (%)	Mala	Mala (%)	Velika	Velika (%)	Mala	Mala (%)
Toyota	1.952	80,9 %	460	19,1 %	1.003	88,8 %	126	11,2 %
Fiat	2.757	69,5 %	1.209	30,5 %	486	68 %	229	32 %
Opel	2.976	70,1 %	1.272	29,9 %	452	71,5 %	180	28,5 %
Ford	3.441	76,4 %	1.060	23,6 %	907	83,9 %	174	16,1 %
Nissan	1.562	77,8 %	447	22,2 %	332	88,8 %	42	11,2 %
Mercedes	1.713	70 %	733	30 %	1.020	71,7 %	403	28,3 %
Ferrari	2.333	71,7 %	920	28,3 %	596	81,1 %	139	18,9 %
Audi	4.282	64,7 %	2.341	35,3 %	536	70,3 %	226	29,7 %
Renault	5.098	73,3 %	1.857	26,7 %	1.047	73,6 %	375	26,4 %
Škoda	10.133	44,3 %	12.762	55,7 %	2.036	36,9 %	3.486	63,1 %
Suzuki	678	76,2 %	212	23,8 %	184	84 %	35	16 %
Peugeot	2.087	74,2 %	725	25,8 %	621	80 %	156	20 %
Hyundai	1.673	75 %	560	25 %	296	87,1 %	44	12,9 %
Citroen	1.633	70,3 %	690	29,7 %	212	69,1 %	95	30,9 %
Honda	2.188	64,8 %	1.187	35,2 %	523	83,5 %	103	16,5 %
Mazda	1.238	67,7 %	590	32,3 %	333	77,8 %	95	22,2 %
Jaguar	395	76,3 %	123	23,7 %	159	62 %	97	38 %
Volkswagen	693	87,1 %	103	12,9 %	410	87,4 %	59	12,6 %
Fiesta	520	67,1 %	255	32,9 %	108	69,2 %	48	30,8 %
Passat	930	80,2 %	730	19,8 %	83	39,5 %	127	60,5 %
Mustang	430	60,6 %	280	39,4 %	45	63,4 %	26	36,6 %
Octavia	829	73,2 %	304	26,8 %	88	55 %	72	45 %
Megane	1.174	58 %	851	42 %	50	34,7 %	94	65,3 %
Golf	3.472	49,6 %	3.528	50,4 %	416	27,2 %	1.114	72,8 %

Tabela 3: Pogostnost in delež začetnic v imenih avtomobilov v korpusih Janes in Kres.

Ker SP 2001 veliko začetnico imena dopušča le v prilastkovni rabi, smo želeli za pet najpogostejših imen avtomobilov v korpusu Janes preveriti pogostnost tovrstne rabe, in sicer za besedne zveze 'avto/avtomobil/vozilo XY'. V tabeli 4 so podatki za število vseh pojavnic in delež prilastkovne rabe teh imen v korpusih Janes in Kres. Iz tabele je razvidno, da je delež prilastkovne rabe pri vseh imenih razmeroma nizek in da prevladuje neprilastkovna raba. Odstotek prilastkovno rabljenih imen je v korpusu Kres nekoliko višji. Če združimo ugotovitve iz tabele 3 in tabele 4, vidimo, da je v obeh korpusih večji delež imen, ki so rabljena neprilastkovno in zapisana z veliko začetnico.

Ime avtomobila	Janes			Kres		
	Prilastkovna raba	Št. vseh pojavnic imena	Prilastkovna raba (%)	Prilastkovna raba	Št. vseh pojavnic imena	Prilastkovna raba (%)
Renault	166	9.158	1,8 %	66	2.163	3,1 %
Audi	138	10.896	1,27 %	26	1.449	1,8 %
Ford	69	6.312	1,1 %	29	1.717	1,7 %
Opel	60	5.741	1 %	35	1.018	3,4 %
Fiat	79	7.367	1,1 %	40	1.207	3,3 %

Tabela 4: Pogostnost in delež prilastkovne rabe imen avtomobilov v korpusih Janes in Kres.

5 Sklep

V pričujočem prispevku smo obravnavali štirideset najpogostejših industrijskih imen v korpusu spletne slovenščine Janes in poskušali ugotoviti, v kakšnem odnosu sta raba začetnice in referenčna kodifikacija. Analiza je pokazala, da je pri veliki večini imen v obeh korpusih bolj ustaljen zapis z veliko začetnico, kar je v nasprotju z referenčno kodifikacijo. V SP 2001 je prehod iz lastnih med vrstna imena pojasnjen na primeru imena avtomobila, zato smo analizirali pogostnost prilastkovne rabe v korpusih Janes in Kres za pet najpogostejših imen avtomobilov. Analiza je pokazala, da se ta imena v obeh korpusih pogosteje pojavljajo v neprilastkovni rabi in z veliko začetnico, torej v kombinaciji, ki je SP 2001 ne dopušča.

Kodifikacijo male začetnice pri neprilastkovni rabi industrijskih imen je potrebno brati v kontekstu vrste poskusov slovenskega pravopisja, iz katerih lahko razberemo težnjo po ustalitvi male začetnice pri nekaterih stvarnih imenih, vse z željo po odpravljanju tujih zgledov.

V pravopisnih pravilih je začetnica večkrat obravnavana nekonsistentno, saj dopušča variantna zapisa brez ustreznega komentarja. Vendar pa vprašanje začetnice pri industrijskih imenih ni naslovljeno samo na tradicijo zapisovanja tovrstnih imen v slovenščini, temveč se problem neposredno dotika tudi vpliva tujih jezikov v smislu ortografije posameznih besed

Avtorji prispevka se zavedamo omejitev, ki jih prinaša obravnava statističnih podatkov o jezikovni rabi. Za postavljanje novih napotkov glede začetnice pri industrijskih imenih bi bila potrebna poglobljena kontekstualna analiza rabe v različnih medijih in okoliščinah. Korpus Janes se je izkazal za relevanten vir pri ugotavljanju zapisa industrijskih imen, saj se v vsebovanih besedilih izkazuje nerevidirana raba, po drugi strani pa v korpusu najdemo nezanemarljivo število pojavnic industrijskih imen.

V prispevku smo dobljena imena razdelili v tri vsebinske sklope in obravnavali posamezni sklop glede na posebnosti rabe, potrebno pa je poudariti, da je se v rabi izrisujejo še druge neevidentirane kategorije, kot so imena vremenskih pojavov, radijskih in televizijskih postaj, borznih indeksov in investicijskih skladov (Dobrovoljc, 2009). V nadaljnje šteje v raziskave vključiti tudi proces apelativizacije lastnih imen in imena, pri katerih je v rabi zaznati pomenske premike.

6 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.

7 Literatura

Helena Dobrovoljc. 2009: Pravopisna obravnava imen znamk in industrijskih izdelkov ter posledice spreminjanja njihovih lastnoimenskih funkcij. *Jezik in slovstvo*, 54(6): 3–19.

Helena Dobrovoljc. 2012: Pisanje imen izdelkov in znamk. V: Nataša Jakop, Helena Dobrovoljc, ur., *Pravopisna stikanja: razprave o pravopisnih vprašanjih*. Založba ZRC, Ljubljana.

Darja Fišer, Tomaž Erjavec, Jaka Čibej in Nikola Ljubešić. 2015: Gradnja in analiza korpusa spletne slovenščine JANES. V: *Slovnica in slovar – aktualni jezikovni opis. Obdobja 34*. Znanstvena založba Filozofske fakultete, Ljubljana. 149–155.

Darja Fišer, Tomaž Erjavec, Ana Zwitter Vitez in Nikola Ljubešić. 2014. JANES se predstavi: metode, orodja in viri za nestandardno pisno spletno slovenščino. V: *Jezikovne tehnologije: zbornik 17. mednarodne multikonference Informacijska družba – IS 2014*. 56–61. Institut "Jožef Stefan", Ljubljana.

Janes = http://nl.ijs.si/janes/. Dostop: 19. 8. 2015.

Kres = http://www.korpus-kres.net/. Dostop: 25. 8. 2015.

Viktor Majdič. 1983. Na rob Načrtu pravil za novi slovenski pravopis. *Jezik in slovstvo*, 28(6). 190–200.

Načrt pravil za novi slovenski pravopis. 1981. Državna založba Slovenije, Ljubljana.

Breda Pogorelec. 1975: Kako je z veliko in malo začetnico pri stvarnih lastnih imenih? *Jezik in slovstvo*, 21(1). 30–31.

Damjan Popič. 2014: *Korpusnojezikoslovna analiza vplivov na slovenska prevodna besedila*. Doktorsko delo. Univerza v Ljubljani.

Teja Rebernik. 2015: Slovenščina pod palcem interneta: vezajne in dvozačetniške tvorjenke. *Zbornik konference Slovenščina na spletu in v novih medijih*. V tisku.

SP 1899 = *Slovenski pravopis*. Cesarska kraljeva zaloga šolskih knjig, Dunaj.

SP 1920 = *Slovenski pravopis*. Jugoslovanska knjigarna, Ljubljana.

SP 1935 = *Slovenski pravopis*. Znanstveno društvo, Ljubljana.

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

SP 1962 = *Slovenski pravopis*. Državna založba Slovenije, Ljubljana.

SP 2001 = *Slovenski pravopis*. Ljubljana: Znanstvenoraziskovalni center SAZU, Založba ZRC.

CMOS = *The Chicago Manual of Style* 16th edition http://www.chicagomanualofstyle.org/home.html. Dostop: 29. 8. 2015.

France Tomšič. 1955. Za pravice male začetnice. *Jezik in slovstvo*, 1(4/5). 114–116.

France Tomšič. 1955/56. Še o veliki začetnici. *Jezik in slovstvo*, 1(8/9). 241–245.

Elementi interakcije v govorjenih in spletnih besedilih

Ana Zwitter Vitez,*† Darja Fišer†

*Oddelek za uporabno jezikoslovje, Fakulteta za humanistične študije
Titov trg 5, 6000 Koper
ana.zwitter@guest.arnes.si
† Oddelek za prevajalstvo, Filozofska Fakulteta
Aškerčeva 2, 1000 Ljubljana
darja.fiser@ff.uni-lj.si

Povzetek

Na družbenih omrežjih, spletnih forumih in komentarjih avtorji pri interakciji z drugimi udeleženci uporabljajo posebne jezikovne vzorce, ki se zdijo precej podobni govorjenemu diskurzu. Zato smo izvedli primerjavo korpusov Gos in Janes (s podkorpusi Tviti, Komentarji in Forumi). Kvantitativna analiza je pokazala visoko stopnjo ujemanja med kategorijami interakcije v korpusu Gos in podkorpusih Janes, kvalitativna analiza pa je razkrila pomembne razlike, ki so posledica prostorske oddaljenosti udeležencev, sočasnosti načrtovanja in tvorjenja govorjenih besedil in (ne)identifikacije z dejansko glasovno podobo nekaterih pogostih izrazov.

Interaction structures in spoken discourse and computer-mediated communication

On social networks, web forums and news comments, the authors interact with other participants and use specific language patterns which seem to be similar to spoken discourse. In this study, we present a comparison of the Gos and the Janes corpora (with subcorpora Tweets, Comments and Forums). Quantitative analysis showed a high degree of interactional correlation between the analysed corpora, while qualitative analysis revealed significant differences which seem to be related to spatial and temporal distance between CMC participants, to simultaneous planning and producing of spoken texts and to authors' perception of the acoustic image of frequent linguistic structures.

1 Uvod

V besedilih družbenih omrežij, spletnih forumov in komentarjev na novičarskih portalih lahko zaznamo jezikovne vzorce, precej drugačne od tistih, ki so tipični za tradicionalna pisna besedila. Taki jezikovni vzorci se kažejo skozi neposredno naslavljanje drugih udeležencev, sredstva sprotnega tvorjenja besedil in rabo neformalnih besed, saj avtorji skozi nenehno interakcijo z drugimi udeleženci utrjujejo svojo identiteto, gradijo samopodobo in razkrivajo svoj odnos do zunajjezikovne realnosti. Tovrstni jezikovni vzorci so tradicionalno značilni za govorjeni diskurz, zato se je besedil računalniško posredovane pisne komunikacije (CMC) prijela stereotipna oznaka hibrida med govorjenimi in pisnimi besedili. Ugotavljanje dejanskih zakonitosti besedil CMC lahko omogočijo empirične raziskave, ki temeljijo na njihovi primerjavi z govorjenim diskurzom in tradicionalnimi pisnimi besedili.

Cilj pričujoče analize je ugotoviti, kateri tipični elementi interakcije v govornih izmenjavah so prisotni v besedilih računalniško posredovane komunikacije in kakšna je njihova vloga v vsakodnevni komunikaciji na spletu. To bomo izvedli s pomočjo primerjave seznamov ključnih besednih oblik v korpusih GOS (Verdonik in Zwitter Vitez, 2011) glede na posamezne podkorpuse korpusa spletnih uporabniških vsebin Janes (Fišer et al., 2014). Predvidevamo, da lahko prek identifikacije elementov interakcije in njihovih funkcij sklepamo o komunikacijskem namenu tvorca besedila in posledicah njegovega jezikovnega udejstvovanja.

2 Preučevanje elementov interakcije v govorjenih in pisnih besedilih

Pomen raziskovanja značilnosti jezika, ki ga na spletu vsakodnevno uporabljajo blogerji, komentatorji ter uporabniki forumov in družbenih omrežij, prvi sistematično izpostavi Crystal (2001) z ugotovitvijo, da se prvine t. i. spletnega jezika (v izvirniku *netspeak*) širijo na druga področja jezikovne rabe in niso več omejene na računalniško posredovano komunikacijo. Pri tem imajo pomembno vlogo specifični jezikovni vzorci (Wray, 2005), ki nastopajo kot orodje za manipulacijo z naslovniki, vzpostavljanje lastne identitete in zagotavljanje družbene identitete. Herring (2006) kot ključne značilnosti jezikovnih specifik CMC izpostavlja na ravneh strukture (krajšanje), pomena (tematika), interakcije (stopnja strinjanja) in družbene vloge (ekspresivnost, identifikacija).

Tudi v govoru se pojavljajo specifični jezikovni vzorci, ki so posledica sočasnega načrtovanja in tvorjenja besedil in neposredne interakcije z drugimi udeleženci (Morel in Danon Boileau, 1998). Zato je korpusno jezikoslovje govora osredotočeno na rabo diskurznih označevalcev (Adolphs in Carter, 2013) in elementov interakcije med udeleženci, znotraj katere Tracey in Robles (2013) kot ključne strukture preučujeta elemente menjavanja vlog, diskurzne označevalce, vprašanja in odgovore, elemente govorne skupnosti (identifikacije).

Omenjene raziskave na besedilih CMC in govorjenem diskurzu zaznavajo primerljive jezikovne vzorce, ki odstopajo od standardnega pisnega jezika. V raziskavi Zwitter Vitez in Fišer (2015) smo izvedli vzporedno analizo jezikovnih specifik govorjenega diskurza in besedil računalniško posredovane komunikacije glede na standardna pisna besedila. Ta je pokazala, da so prvine interakcije med udeleženci prisotne tudi v primarno drugače razporejenih kategorijah (npr. pri izrazu *pejd*, pri katerem lahko zaznamo tako odstop od standardnega zapisa kot interakcijo z drugimi udeleženci). Zato se zdi smiselno prvine interakcije v govoru in spletnih besedilih podrobneje preučiti.

3 Metodologija

Cilj analize je bil identificirati točke prekrivnosti in razhajanj nestandardnih in interakcijskih elementov govorjenih in spletnih prvin, kar smo izvedli s primerjavo treh korpusov slovenskega jezika:

- korpus pisnih besedil Kres, ki zajema 100 milijonov pojavnic (Logar Berginc et al., 2012),
- korpus govorjene slovenščine Gos v obsegu milijon pojavnic (Verdonik in Zwitter Vitez, 2011),
- korpus spletnih besedil Janes s 161 milijoni pojavnic (Fišer et al., 2014).

Najprej smo izluščili tipične besedne oblike (pojavnice) korpusa Gos in korpusa Janes glede na korpus Kres. Sezname ključnih besednih oblik smo izdelali z orodjem SketchEngine (Kilgarriff et al., 2004). Na korpusu Gos smo izvedli eno samo analizo, korpus Janes pa smo razdelili na tri podkorpuse (Tviti, Forumi in Komentarji), ker smo zaradi različnih okoliščin nastajanja omenjenih spletnih žanrov pričakovali različne rezultate. Ker nas je zanimala neformalna komunikacija, smo v korpusu Janes analizirali le nestandardna besedila (tista, ki imajo pripisano 2. ali 3. raven jezikovne nestandardnosti) (Ljubešić et al., 2015).

GOS	Forumi	Twitter	Komentarji
eee	*avto*	*btw*	*ane*
mhm	*tud*	*oz.*	*nebi*
eem	*mal*	*cca*	*nevem*
sej	*tko*	*slo*	*ala*
tud	*blo*	*lol*	*kriv*
zdej	*tut*	*cez*	*krivi*
tko	*gor*	*bos*	*obsojen*
aha	*jst*	*nic*	*fajn*
blo	*mam*	*prevec*	*cel*
tak	*gume*	*mogoce*	*neprimerno*

Tabela 1: 10 najbolj ključnih besednih oblik v analiziranih korpusih.

Na prvih 200 besednih oblikah z vsakega seznama smo identificirali elemente, pri katerih je zaznati interakcijo z drugimi udeleženci. Nato smo elemente interakcije analizirali z metodologijo Tracey in Robles (2013), razširjeno s kategorijami, ki sledijo shemi jezikovnih funkcij (Jakobson, 1966):

deiktika	*jst*
vprašalnice	*kaj*
modus	*sigurno*
konativnost	*gremo*
ekspresivnost	*uau, vidim*
fatičnost	*dobro jutro, evo*
performativ	*obljubim*
metajezik	*tvitam*
referenca	*glede, rekel*

Tabela 2: Kategorizacija elementov interakcije.

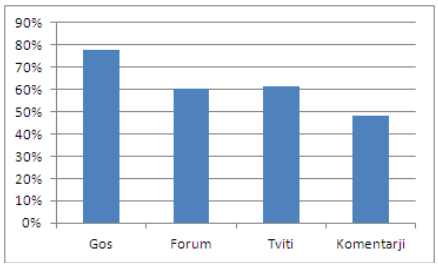
Na koncu smo primerjali rezultate analize govornega korpusa in podkorpusov novih medijev ter poskušali določiti stopnjo in mesta prekrivanja elementov interakcije v govorjenem in spletnem neformalnem diskurzu.

4 Analiza

Elemente interakcije v govoru in uporabniških vsebinah smo analizirali na dveh ravneh. S kvantitativno raziskavo smo želeli proučiti tipe prvin, ki so za interakcijo v proučevanih besedilnih žanrih najznačilnejši, ter preveriti, kakšne podobnosti in razlike veljajo med njimi. Kvantitativno analizo smo nato nadgradili še s kvalitativno, pri kateri smo za vsako od proučevanih kategorij preverili, v kakšni meri so posamezni interakcijski elementi univerzalni oz. žanrsko specifični.

4.1 Kvantitativna analiza

Nabor interaktivnih besednih oblik se razteza čez vse kategorije nestandardnosti, identificirane v raziskavi Zwitter Vitez in Fišer (2015), najpogosteje pa je interakcija poleg primarne kategorije prisotna v kategoriji specifik izgovorjave oz. zapisa (43 % Gos, 25 % Forum, 37 % Tviti). Le pri podkorpusu komentarji je največ elementov interakcije vezanih na tematiko besedila (17 %). To dokazuje, da podrobnejša analiza zgolj kategorije primarno interaktivnih besed ne bi podala celotne slike, temveč je interaktivne elemente v komunikaciji treba opazovati na celotnem besedišču.



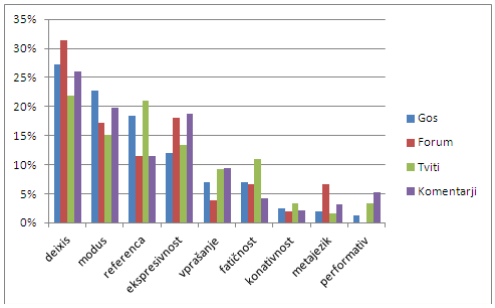
Graf 1: Prisotnost elementov interakcije v korpus Gos in podkorpusih Janes.

Končna slika analize korpusov Gos, Forum, Tviti in Komentarji kaže, da pri treh od štirih analiziranih medijev več kot 60 % analiziranih ključnih besed vsebuje elemente interaktivnosti.

Najvišjo stopnjo interakcije najdemo v govoru, razlike pa se kažejo tudi med različnimi družbenimi mediji. Največ elementov interakcije vsebuje podkorpus Tviti, ki ga zaznamujejo hitro izmenjana kratka sporočila (pogosto prek pametnih telefonov). Pri spletnih forumih so besedila daljša, a v njih kljub temu opazimo izjemno veliko interakcije med uporabniki (svetovanje, izmenjevanje mnenj in izkušenj). Najmanj interaktivni so komentarji na spletne novice, saj je primarni motiv komentatorjev sporočanje osebnega mnenja o članku/dogodkih/osebah v članku, ne toliko interakcija z drugimi.

Prevladujoča kategorija v vseh podkorpusih so deiktični izrazi, največ jih je v forumih. Gre večinoma za osebne zaimke (*jst*, *tale*, *tist*) in časovne in prostorske deiktične izraze (*zdaj*, *tam*, *ven*), ki se nanašajo na zunanji kontekst konverzacije in udeležence v njej. Te izraze verjetno pogojujejo tehnične okoliščine konverzacije na spletu, saj udeleženci ne delujejo v istem fizičnem in

časovnem kontekstu, ki mora biti zato bolj eksplicitno izražen.



Graf 2: Elementi interakcije v korpusu Gos in podkorpusih Janes.

V korpusu Gos so zelo pogosti elementi modalnosti (*dejansko, itak, pač*), ki pogosto "definirajo stopnjo gotovosti, s katero govorec izraža svoje mnenje" (Morel in Danon Boileau, 1998). Največ elementov modalnosti najdemo v komentarjih, kar je skladno z namenom komentiranja na novičarskih portalih. Podobno vlogo imajo ekspresivni izrazi, le da so bolj osredotočeni na avtorja samega (*haha, kul, super*), in jih je na forumih in komentarjih celo več kot v govoru. To je verjetno posledica dejstva, da korpus Gos zajema formalne in neformalne sporočanjske položaje, v Janesu pa pogosto predstavljajo ventil za izražanje mnenj, deljenje izkušenj in sproščanje frustracij uporabnikov.

Posebnost tvitov je izstopanje referencialnih elementov, ki se nanašajo na udeležence v komunikaciji (*poznam, videl, imajo*), ki je najvišja med vsemi analiziranimi podkorpusi. To kaže na visoko stopnjo interaktivnosti tega medija, v katerem udeleženci sebe in svoje sogovornike eksplicitno vključujejo v sporočila. Tviti izstopajo tudi po deležu fatičnih elementov (*aja, btw, čao*), saj gre za priljubljen kanal efemerne komunikacije, medtem ko komentarji niso namenjeni vzpostavljanju in vzdrževanju stika med sogovorci, temveč za izražanje lastnega mnenja.

Po številu vprašalnih izrazov izstopajo tviti in komentarji, kjer jih je celo več kot v korpusu Gos, ki ga zaradi majhnosti nismo razdelili na podkorpuse, tako da smo vanj zajeli tudi manj interaktivne govorne zvrsti, kot so radijske oddaje, predavanje ipd. V kategorijo metajezik smo zajeli izraze, ki se nanašajo na izrekanje samo, kar lahko je pogosto povezano z glagoli izrekanja (*povedal, govorim, reku*), v novih medijih pa tudi z glagoli *tvitnil, napisal* ipd. Najmanj izraziti kategoriji predstavljajo konativni elementi in performativi, ki jih v vsakem korpusu najdemo le po nekaj primerov.

4.2 Kvalitativna analiza

4.2.1 Osrednje kategorije

Kvalitativno analizo pričenjamo s kategorijami, ki zavzemajo največji delež analiziranega ključnega besedišča, obenem pa pri njih na kvantitativni ravni prihaja do največjih razlik med posameznimi proučevanimi korpusi, in sicer elementi modalnosti, referencialni elementi, deiktika in vprašalnice. Podrobnejša analiza elementov modalnosti pokaže, da lahko večino razlik med rezultati pripišemo različnim izgovornim in pisnim variantam besed, ki so precej bogatejše v korpusu Gos, kar nakazuje na večjo regionalno razpršenost zajetega gradiva, kot ga vsebuje korpus Janes.

To lahko ponazorimo z deiktičnimi elementi, od katerih so v korpusu Gos najznačilnejše različne izgovorne variante osebnih in kazalnih zaimkov, kot so *jaz, tisti* (npr. *jst, jest, jez, jaz,*) ter časovnih in prostorskih prislovov, kot so *zdaj/prej/potem* ter *tukaj/tam* (npr. *zdej, zaj, zej, zdele*). Čeprav se jih večina pojavi tudi v vseh podkorpusih Janes, jih tam najdemo le v eni ali dveh variantah. Za spletne žanre pa so specifični svojilni in kazalni zaimki (npr. *moj, tvoj, tale, tole*), ki za razliko od govora v računalniško posredovani komunikaciji eksplicitno izraženi, saj je v tem načinu komunikacije onemogočena gestikulacija, sogovorci pa si tudi ne delijo fizičnega in/ali časovnega konteksta. Podobno je s kategorijo vprašalnic, ki imajo v korpusu Gos precej več variant (npr. *kok, kuk, kolko*) kot v tvitih (npr. *kolk, kok*), še manj pa jih je v forumih in komentarjih (npr. *koliko*). Sicer se vprašanja *kdo/kje/kdaj/koliko* pojavijo v vseh analiziranih žanrih, povsod razen v forumih, ki v primerjavi s preostalimi korpusi močno izstopajo po majhnem naboru identificiranih vprašalnic, najdemo *zakaj*, v tvitih, ki jih od spletnih žanrov vsebujejo največ, pa še *koga* in *kdaj*. Tudi analiza referencialnih elementov, ki večinoma sestoji iz različnih oblik glagolov *biti, imeti, videti, iti* pokaže podobne rezultate, vendar tokrat za razliko od ostalih kategorij več variantnosti zanje najdemo v spletnih žanrih (npr. v tvitih *imaš, maš, mas*), najverjetneje zato, ker v govoru varianta brez začetnega *i* tako močno prevladuje, poleg tega pa je za Twitter tipično izpuščanje diakritik *čšž*.

Sekundarni razlog za razlike smo našli v mediju, preko katerega poteka komunikacija, ki diktira težavnost in posledično ekonomičnost izražanja. V govoru in na forumih, kjer oteževalnih okoliščin praktično ni, je tako elementov modalnosti nekoliko več in so hkrati tudi bolj raznoliki kot v tvitih in komentarjih, kjer uporabnika že tehnične okoliščine (omejena dovoljena dolžina, uporaba mobilnih naprav) in komunikacijske konvencije (komuniciranje v realnem času) spodbujajo k čim krajšim besedilom in posledično omejeni rabi elementov modalnosti. Tako v govoru kot vseh tipih spletnih uporabniških vsebin se pojavljajo tipični prislovi za neformalne govorne okoliščine, kot so *ful, itak, kao, pač*. Za spletne žanre so specifični omilitveni izrazi, kot so *baje, verjetno, zgleda*, pa tudi izrazi, *nebi, nevem, pomoje* (zaradi drugačnih načel transkripcije govora, ki ju obravnava kot niza dveh besed), medtem ko v domeni govora ostajajo regionalno zaznamovani izrazi *glih, provzaprov, zlo*, pa tudi *okej* (zaradi drugačnih načel transkripcije govora, saj je isti izraz v korpusu Janes zapisan kot *ok*).

4.2.2 Obrobne kategorije

V drugi del kvalitativne raziskave smo zajeli kategorije, ki predstavljajo manjši delež analiziranega ključnega besedišča, obenem pa pri njih na kvantitativni ravni nismo zaznali velikih razlik med korpusi. Med te kategorije prištevamo ekspresivne, fatične in konativne izraze.

Čeprav je delež zastopanosti ekspresivnih elementov v analiziranih korpusih praktično identičen, se ti med seboj močno razlikujejo glede na vrsto ekspresivnih elementov, ki se v njih pojavljajo. V korpusu Gos prevladujejo sredstva sočasnega načrtovanja in tvorjenja besedil (npr. *eee, eem, mmm*), ki jih v pisnih žanrih zaradi drugačnega načina tvorjenja in organizacije besedil ne srečamo. Po drugi strani pa so v spletnih žanrih pogosto eksplicitno verbalizirani elementi komunikacije, ki tipično govor spremljajo kot neverbalni elementi komunikacije (npr. *hehe, hahaha, lol*). Zanimive so tudi razlike med posameznimi žanri uporabniško posredovane komunikacije. Medtem ko se ekspresivni element *super* pojavlja tako v korpusu Gos kot v vseh podkorpusih Janes, tviti vsebujejo še številne druge pozitivne ekspresivne elemente (npr. *kul, top, fajn, všeč*), medtem ko v komentarjih najdemo predvsem negativne (npr. *sramota, škoda, žal, brez veze*). Zelo malo prekrivanja med analiziranimi korpusi smo zaznali tudi pri fatičnih elementih. Medtem ko jih v Gosu večina služi za začetek komunikacije (npr. *dobro jutro*), so v spletnih žanrih uporabljeni predvsem za njeno zaključevanje (npd. *lep pozdrav, lep dan, lep vikend*), za tvite pa so značilna še kratka, pomensko izpraznjena sporočila, katerih primarna funkcija je vzdrževanje komunikacijskega kanala (npr. *sej, ane, aja*).

Povsem drugače je s konativnimi elementimi, pri katerih ni praktično nobenih razlik med žanri (npr. *dej, gremo, morš*), z eno samo izjemo, in sicer spodbujanje, naj sogovornik *prijavi* (najverjetneje neprimerno vsebino), kar smo zasledili samo v komentarjih. Razen besedišča, specifičnega za posamezni medij (npr. *odgovor, komentar, forum*), so povsem prekrivni tudi elementi metajezika, kar je zanimivo, saj so tudi v negovorjenih žanrih povsem enakomerno kot v govoru zastopani izrazi, kot so *reku, povem, govorim*, le v forumu poleg naštetih zasledimo še *napisal*. Podobno velja za performative, od katerih se v vseh žanrih pojavi *hvala*, v jeziku vseh analiziranih družbenih omrežij za razliko od govora obstaja potreba po eksplicitnem izražanju strinjanja s sogovorci (*strinjam*), za tvite in komentarje je značilno še voščilo *bravo*, medtem ko smo vljudnostni *prosim* poleg govora identificirali samo v komentarjih.

5 Zaključek in nadaljnje delo

Z raziskavo smo analizirali elemente interakcije v govorjenem diskurzu in besedilih računalniško posredovane komunikacije glede na standardna pisna besedila.

Kvantitativna analiza je pokazala visoko stopnjo ujemanja med kategorijami interakcije v korpusu Gos in pokorpusih Janes, saj se tovrstni elementi raztezajo čez večino analiziranih ključnih besed. Najbolj značilna kategorija interaktivnih elementov so deiktični izrazi, ki so posebej pogosti v forumih, elementi modalnosti, ki jih je največ v komentarjh, ter referencialni in ekspresivni izrazi, ki prevladujejo v tvitih.

Kvalitativna analiza pa je pokazala, da se dejanske oblike interakcije v govorjenih in spletnih besedilih realizirajo z drugačnimi sredstvi in imajo posledično tudi drugačno vlogo v diskurzu. Pri vseh kategorijah se je namreč izkazalo, da je variantnost izgovora v korpusu Gos precej večja kot variantnost zapisa v korpusu Janes, kar je lahko posledica regionalne razpršenosti diskurza v korpusu Gos ali pa nezmožnost identifikacije prek dejanske glasovne podobe nekaterih pogostih izrazov (napr. *zej* za *zdaj*). Drugo razliko pogojuje prostorska oddaljenost udeležencev, kar se kaže skozi prevladujoči obliki svojilnih in kazalnih zaimkov med deiktičnimi izrazi v spletnih žanrih (npr. *tale, tvoj*). Tretja opazna točka razhajanja med govorom in analiziranimi spletnimi žanri je potek načrtovanja in tvorjenja besedil, zaradi česar se razlikujejo predvsem ekspresivni izrazi (*eee* v govoru, *lol* v spletnih žanrih).

V nadaljnjih raziskavah nameravamo preučiti večji nabor ključnih besednih oblik, podrobneje raziskati ostale nestandardne kategorije (izgovoru podoben zapis, neformalni izrazi) in preveriti možnost uporabe metapodatkov v korpusu Gos za potrebe profiliranja avtorjev spletnih besedil.

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.

6 Literatura

Svenja Adolphs in Ron Carter. 2013. Spoken corpus Linguistics. From monomodal to multimodal. Routledge.

David Crystal. 2007. How language works. Penguin Books.

Darja Fišer, Tomaž Erjavec, Ana Zwitter Vitez in Nikola Ljubešić. 2014. JANES se predstavi: metode, orodja in viri za nestandardno pisno spletno slovenščino. V: Proceedings of the 17th International Multiconference Information Society - IS 2014, str. 56–61, Ljubljana.

Susan Herring. 2002. Interactional coherence in CMC. Journal of computer-mediated communication.

Roman Jakobson. 1976. Six leçons sur le son et le sens. Minuit.

Adam Kilgarriff, Pavel Rychly, Pavel Smrz in David Tugwell. 2004. The Sketch Engine. Proceedings EURALEX 2004, str. 105–116, Lorient.

Nataša Logar Berginc in Simon Krek. 2012. New Slovene Corpora within the Communication in Slovene Project, Prace Filologiczne, 63, str. 197–207.

Nikola Ljubešić, Darja Fišer, Tomaž Erjavec, Jaka Čibej, Dafne Marko, Senja Pollak in Iza Škrjanec. 2014. Predicting the level of standardness of text in user generated content, Proceedings of the RANLP15 Conference. Hissar, Bulgaria.

Mary-Annick Morel in Laurent Danon Boileau. 1998. Grammaire de l'intonation. Ophrys.

Karen Tracey in Jessica S. Robles. 2013. Everyday talk, Second Edition: Building and Reflecting Identities. Guilford Press.

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

Ana Zwitter Vitez in Darja Fišer. 2015. From mouth to keyboard: the place of non-canonical written and spoken structures in lexicography. str. 250–267, Proceedings of the eLex 2015 conference, United Kingdom.

Indeks avtorjev