Uredila Darja Fišer

VIRI, ORODJA IN METODE ZA ANALIZO SPLETNE SLOVENŠČINE

Zbirka Prevodoslovje
in uporabno jezikoslovje

Ljubljana 2018

VIRI, ORODJA IN METODE ZA ANALIZO SPLETNE SLOVENŠČINE
ZBIRKA PREVODOSLOVJE IN UPORABNO JEZIKOSLOVJE
ISSN 2335-335X

Urednica: Darja Fišer
Recenzenta: Darinka Verdonik, Reinhild Vandekerckhove
Lektor: Damjan Popič
Tehnični urednik: Jure Preglau

Založila: Znanstvena založba Filozofske fakultete Univerze v Ljubljani
Izdal: Oddelek za prevajalstvo
Za založbo / For the Publisher: Roman Kuhar,
dekan Filozofske fakultete / Dean of the Faculty of Arts

Ljubljana, 2018
Prva izdaja, e-izdaja / First Edition, Digital Edition

Oblikovna zasnova: Kofein, d. o. o.
Prelom: Jure Preglau

Publikacija je brezplačna. / Publication is free of charge.

Knjiga v digitalni obliki (PDF) je dosegljiva na https://e-knjige.ff.uni-lj.si/
The book is available in e-form (PDF) at https://e-knjige.ff.uni-lj.si/

DOI: 10.4312/9789610600701

Knjiga je izšla s podporo Javne agencije za raziskovalno dejavnost Republike Slovenije.

Raziskovalni projekt št. J6-6842 je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna.

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

COBISS.SI-ID=294718208
ISBN 978-961-06-0069-5 (epub)
ISBN 978-961-06-0070-1 (pdf)

Kazalo vsebine

KAZALO

Uvod

Z razmahom svetovnega spleta so skokovito narasle tudi uporabniško generirane vsebine, kot so blogi, forumi in družbena omrežja, ki so v zadnjih dveh desetletjih postale pomemben vir človeškega znanja in mnenj ter katalizator bottom-up komunikacijskih praks, ki prispevajo k demokratizaciji jezika. Posledično narašča tudi potreba po temeljitem, multidisciplinarnem razumevanju tovrstnega sporazumevanja, ki ga pomembno določajo specifične družbene in tehnične okoliščine, v katerih nastaja in zaradi katerih se tudi precej razlikuje od pisnega standarda. Zanj je značilna raba pogovornih in tujejezičnih izrazov, nekanoničnih zapisovalnih praks in skladnje, idiosinkratičnih okrajšav in hiter dotok novega besedišča. Dodatna pomembna značilnost tovrstne visoko participatorne, interaktivne in multimodalne komunikacije so bogati in lahko dostopni (sociodemografski) metapodatki, ki po eni strani odpirajo številne nove priložnosti za raziskovalce, ne samo v korpusnem in računalniškem jezikoslovju, temveč tudi v digitalni humanistiki in družboslovju, po drugi pa raziskovalcem prinašajo nove tehnične, jezikoslovne, pravne in etične izzive.

Medtem ko je bila slovenščina z viri, metodologijo in priročniki za standardni jezik že do sedaj razmeroma dobro opremljena, je do začetka projekta JANES na področju nestandardnega pisnega jezika zevala globoka vrzel, saj nismo imeli ne reprezentativnih korpusov za študije tega segmenta jezika ne orodij za njihovo analizo in obdelavo, prvine nestandardne slovenščine pa so prav tako komajda vključene v jezikovne opise, priročnike in pedagoško prakso. Zato je bil projekt, ki bi zagotovil infrastrukturo in metodologijo za analizo sodobne nestandardne pisne slovenščine, kot se uporablja v novih medijih, v slovenskem okolju že dolgo težko pričakovan.

PROJEKT JANES

V nacionalnem temeljnem raziskovalnem projektu »Viri in metode za raziskovanje nestandardne spletne slovenščine« z neformalnim imenom JANES – Jezikoslovna analiza nestandardne slovenščine (J6–6842), ki ga je od 1. 7. 2014 do 1. 2. 2018 financirala Javna agencija za raziskovalno dejavnost Republike Slovenije, pri njem pa je sodelovalo 12 raziskovalcev in 16 študentov s Filozofske fakultete Univerze v Ljubljani in Instituta »Jožef Stefan«, smo zgradili obsežen korpus internetne slovenščine in na podlagi obsežne jezikoslovne analize razvili metode za izboljšanje avtomatskega procesiranja nestandardne pisne slovenščine. Korpus vsebuje tvite, bloge, forume in komentarje na novice in članke v Wikipediji, torej najpomembnejše zvrsti besedil, ki jih uporabniki ustvarjajo na spletu. Korpus je jezikoslovno označen na več ravneh in prosto dostopen preko zmogljivih konkordančnikov, zato je dobrodošel za teoretično in uporabno jezikoslovje.

Na osnovi korpusa smo izvedli vrsto jezikoslovnih raziskav, in sicer primerjavo internetne slovenščine s pisnim standardom in govorom, študijo žaljivega govora

na družbenih omrežjih, poglobljeni analizi večbesednih zvez in terminologije v internetnih besedilih ter identifikacije pomenskih premikov besed. Izvedene raziskave so s kombinacijo najsodobnejših korpusnih metod in metod s področja računalniškega jezikoslovja omogočile celovit vpogled v segment jezika, ki se hitro spreminja, dobiva vse pomembnejšo vlogo na vseh področjih našega udejstvovanja ter je bil doslej iz različnih razlogov zanemarjan.

Rezultat projekta so tudi ročno označni korpusi, s pomočjo katerih smo razvili metode za izboljšanje avtomatskega procesiranja nestandardne pisne slovenščine, ter spletni slovar tviterščine,[1] ki je namenjen učiteljem, učencem, jezikoslovcem, leksikografom in širši zainteresirani javnosti. Vsi razviti jezikovni viri so ponujeni v odprti dostop pod licenco Creative Commons, razvita orodja za označevanje pa vključena v spletne delotoke, razvili pa smo tudi prototipni sistem za kontinuirano gradnjo spremljevalnega korpusa. Vsi razviti viri in orodja so objavljeni in se bodo vzdrževali v okviru slovenske raziskovalne infrastrukture CLARIN.[2]

ODMEVNOST PROJEKTA JANES

Poleg pričujoče monografije smo v okviru projekta objavili 1 tujo monografijo, 1 tematsko številko znanstvene revije, 6 prispevkov v znanstvenih revijah, 8 poglavij v monografskih publikacijah in 50 prispevkov na mednarodnih in domačih znanstvenih konferencah. Vse publikacije so zbrane na projektni strani: http://nl.ijs.si/janes/publikacije/.

Na temo projekta smo izvedli 9 vabljenih predavanj na mednarodnih konferencah, poletnih šolah (Nemčija, Italija, Hrvaška) in tujih univerzah (Švica, Hrvaška, Srbija, Italija, Češka) ter opravili 2 enomesečna znanstvena obiska v okviru akcije COST EnELL na Češkem in 1 enomesečni raziskovalni obisk v Italiji. Rezultati projekta so bili nagrajeni s 1. mestom na mednarodni deljeni nalogi CLIN2017 pri normalizaciji nestandardnih besedil (Sang et al. 2017). V okviru projekta sta nastali 2 magistrski nalogi (Reher 2017 in Škrjanec 2017).

Oganizirali smo 2 poletna tabora spletne slovenščine za dijake, 1 poletno šolo za študente, 1 seminar in 1 mednarodni seminar empiričnega jezikoslovja za jezikoslovce, 1 domačo in 4 mednarodne znanstvene delavnice ter 1 domačo, 1 mednarodno znanstveno konferenco in 1 okroglo mizo. Vsi dogodki z gradivi so dostopni na projektni spletni strani: http://nl.ijs.si/janes/dogodki/.

Med njimi posebno mesto zaseda serija izobraževanj JANES Ekspres, ki smo jo s finančno podporo ARRS za promocijo slovenske znanosti v tujini izvedli v Lju-

1 http://nl.ijs.si/janes/viri/slovarcek-tviterscine/

2 http://www.clarin.si/

bljani, Zagrebu in Beogradu. Na izobraževanjih smo domačim in tujim kolegom predstavili smernice za ročno označevanje učnih korpusov za nestandardni jezik ter platformo za označevanje WebAnno. Rezultat teh izobraževanj so 3 ročno označeni zlati standardi za označevanje, lematizacijo in normalizacijo nestandardnega jezika za slovenščino, hrvaščino in srbščino. S tem smo metodo, razvito v okviru projekta, preizkusili in prenesli na dva druga jezika, ki smo ju obogatili s pomembnimi novimi jezikovnimi viri.

Projekt je bil širši javnosti predstavljen v 10 radijskih oddajah, 2 televizijskih in 2 časopisnih intervjujih ter v 1 prispevku v reviji. Odzivi v medijih so zbrani na projektni strani: http://nl.ijs.si/janes/publikacije/odzivi-v-medijih/.

POMEN PROJEKTA JANES

Projekt JANES je postavil pomemben mejnik v slovenskem teoretičnem, uporabnem in računalniškem jezikoslovju, saj obsežnih raziskav na področju razvoja in uporabe jezikovnih virov in orodij, prilagojenih za nestandardno pisno slovenščino, do sedaj praktično ni bilo. Rezultati projekta JANES pomembno prispevajo k boljšemu poznavanju značilnosti nestandardne internetne slovenščine, ki z razmahom informacijskih tehnologij postaja vse pomembnejši in pogost način pisne komunikacije. Vrsta raziskav, ki smo jih izvedli v projektu, je kontrastivno, tako kvalitativno kot kvantitativno, osvetlila različne ravni takega jezika, kar je uporabno v nadaljnjih raziskavah in aplikacijah na področju leksikologije in sociolingvistike ter usvajanja in poučevanja jezika. Poleg tega smo v projektu intenzivno uporabljali in razvijali inštrumentarij korpusnega jezikoslovja (korpusnoprimerjalne študije, metode za luščenje kolokacij, pomenskih premikov itd.), kar bo koristilo vsem nadaljnjim korpusnim raziskavam slovenskega jezika.

Projekt je zagotovil tudi razvoj računalniških metod za identifikacijo in luščenje ciljnih besedil s spleta, normalizacijo, oblikoskladenjsko označevanje in lematizacijo nestandardnega jezika ter profiliranje avtorjev in identifikacijo pomenskih premikov, ki so pomembni ne samo za slovenščino, temveč so izrazito aktualne tudi mednarodno, saj se s podobnimi izzivi spopadajo tudi drugi jeziki.

Projekt je omogočil izdelavo kvalitetnih virov nestandardnega jezika, kar je najbolj oprijemljiv rezultat projekta in vključuje obsežen, sodoben in jezikoslovno označen korpus, serijo ročno označenih učnih množic oz. zlatih standardov za jezikoslovne in jezikovnotehnološke raziskave in razvoj ter leksikalno bazo oz. spletni slovarček nestandardne slovenščine. Vsi ti izdelani viri prvič omogočajo celovit vpogled v značilnosti in razvoj spletnega jezika ter odpirajo možnosti za številne nove raziskave s področja jezikoslovja in jezikovnih tehnologij za

slovenščino, in to ne samo projektnim partnerjem, temveč vsem slovenskim in tujim raziskovalcem.

V projektu smo veliko pozornosti posvetili implementaciji in promociji dobrih znanstvenih praks, ki jih v slovenskem prostoru, posebej na področju jezikoslovja, še močno primanjkuje. Projekt je za razliko od introspektivnega jezikoslovja ali jezikoslovnih analiz na napaberkovanih primerih uporabe promoviral empirično podprto korpusno jezikoslovje, jeziko(slo)vne podatke strukturiral po mednarodnih priporočilih in ne v nedokumentiranih ad-hoc formatih, najbolj pomembno pa je, da so viri, izdelani v okviru projekta, dostopni pod licenco Creative Commons, kar omogoča odprto diseminacijo izdelanih podatkov in s tem njihovo maksimalno izkoriščenost, saj mdr. omogoča preverljivost rezultatov, preprečuje dvojno financiranje raziskav in spodbuja gospodarski napredek. S tem smo bistveno pripomogli k preseganju prevladujočega stanja v Sloveniji, kot ga opisujejo Štebe in dr. (2013): »Kljub nekaterim zametkom [omogočanja dostopa do raziskovalnih podatkov] je to področje kritično podhranjeno zaradi prevladujoče kulture zapiranja in monopoliziranja podatkov.«

Projekt je bil tudi izrazito interdisciplinaren, saj je združeval znanstvenike s humanistične in naravoslovne institucije s področja jezikoslovja in računalništva. S tem je projekt prispeval k povezovanju dveh tradicionalno ločenih strok, ki pa sta obe potrebni za razvoj sodobnega empirično osnovanega jezikoslovja in jezikovnih tehnologij. Glede na to, da je metodologija izgradnje virov in orodij jezikovno neodvisna, so pristopi uporabni, in tudi že bili uporabljeni, za sorodne jezike (hrvaščina, srbščina), ki tovrstnih virov in orodij še nimajo, kar rezultatom daje pomembno večjezično razsežnost.

Razviti viri, orodja in metode bodo omogočili prenos znanj na vsa področja, ki uporabljajo spletne vsebine, ki jih ustvarjajo uporabniki. Zaradi odprtosti in zapisa leksikalne baze je ta tako avtorskopravno kot tehnično primerna za vključitev v druge leksikalne vire, kot npr. v načrtovani novi slovar sodobnega slovenskega jezika, s čimer bo integrirana v širši slovarski projekt, od katerega bodo imeli korist vsi govorci slovenskega jezika. Predlagani projekt je ponudil tako kontrastivne raziskave kot konkretne in dostopne vire nestandardne pisne slovenščine, ki jih je mogoče uporabiti pri pouku oz. načrtovanju pouka slovenščine na osnovni, srednješolski in univerzitetni ravni, pa tudi pri učenju slovenščine kot tujega jezika.

Glede na to, da razviti viri in orodja niso omejeni le na nekomercionalno rabo, bodo neposredno uporabni tudi za slovenska podjetja, ki se ukvarjajo z informacijsko-komunikacijskimi tehnologijami, npr. v povezavi s semantičnim spletom, poizvedovanjem po informacijah, rudarjenjem po besedilih ali povzemanjem besedil, in ki bodo vedno pogosteje v svoje produkte želela vključevati

tudi obdelavo slovenskega jezika, pri čemer se bodo morala spopadati tudi ali celo predvsem z nestandardnim jezikom, saj količina takih besedil bliskovito narašča.

Ob neposredni koristi za podjetja, ki bodo lahko bolje procesirala slovenski jezik, bodo sledile tudi koristi za družbo, saj bo govorcem slovenščine omogočen dostop do produktov, ki bolje podpirajo slovenski jezik. S tem bomo pomagali zmanjšati e-izključenost govorcev, ki so pogosto priklenjeni na tujejezične aplikacije, slovenščini pa omogočili boljšo funkcionalnost in razvoj v digitalni dobi, kar je tudi izpostavljeno v Resoluciji o nacionalnem programu za jezikovno politiko 2014–2018.

VSEBINA MONOGRAFIJE

Monografija Viri, orodja in metode za analizo spletne slovenščine v štirinajstih poglavjih združuje vse pomembnejše raziskave, ki so se zvrstile v projektu JANES, tem pa se s prispevki pridružujejo tudi nekateri domači in tuji kolegi, ki se pri svojem delu posvečajo tej raziskovalni tematiki, a pri projektu formalno niso sodelovali. Poglavja smo organizirali v dva večja sklopa: v prvem delu monografije predstavimo izdelavo virov in orodij, v drugem pa nanizamo metodološko heterogene raziskave posameznih vidikov spletne slovenščine, ki so bile na razvitih virih opravljene na različnih ravneh jezikovnega opisa.

1. del: Viri in orodja za analizo spletne slovenščine

V prvem poglavju **Tomaž Erjavec**, **Nikola Ljubešić** in **Darja Fišer** predstavijo korpus slovenskih spletnih uporabniških vsebin Janes, ki vsebuje tvite, spletne forume, novice in uporabniške komentarje nanje, uporabniške in pogovorne strani na Wikipediji ter blogovske zapise in komentarje nanje. Opišejo postopek zajema in obdelave besedil za vsakega od vključenih virov, predstavijo avtomatske in ročne postopke za obogatitev korpusa z dragocenimi metapodatki ter oblikovanje in objavo korpusa, prav tako pa podajo tudi kvantitativno analizo zgrajenega korpusa, ki nudi vpogled v naravo zajetega gradiva.

V drugem poglavju **Jaka Čibej**, **Špela Arhar**, **Tomaž Erjavec** in **Darja Fišer** predstavijo družino ročno označenih korpusov, ki so bili izdelani za potrebe učenja jezikovnotehnoloških orodij za izboljšanje stavčne segmentacije, tokenizacije, normalizacije, oblikoskladenjskega označevanja in lematizacije, prav tako pa tudi za natančnejše proučevanje jezikovnih pojavov v slovenski računalniško posredovani komunikaciji, kot so skladenjske značilnosti, pojave krajšanja, raba vejice, preklapljanje med jeziki in regionalnih prvin.

V tretjem poglavju **Nikola Ljubešić**, **Tomaž Erjavec** in **Darja Fišer** predstavijo razvoj orodij za avtomatsko procesiranje nestandardne slovenščine, s katerimi merijo stopnjo standardnosti besedil, izboljšujejo stavčno segmentacijo, izvajajo normalizacijo nestandardno zapisanih besed in oblikoskladenjsko označevanje ter v besedilih razpoznavajo imenske entitete. Z vrednotenjem rezultatov, dobljenih z razvitimi orodji, pokažejo znatno povečanje kvalitete procesiranja nestandardnega jezika, prav tako pa tudi nakažejo možnosti prenosa razvitih orodij na druge jezike.

V četrtem poglavju **Matej Martinc**, **Senja Pollak** in **Ana Zwitter Vitez** predstavijo vgradnjo razvitih orodij za procesiranje spletnih vsebin v okolje ClowdFlows, ki omogoča lažji in hitrejši zajem, obdelavo in analizo lastnih korpusov tudi brez programerskega znanja. Delotok preizkusijo na leksikalni, besednovrstni in skladenjski analizi pozitivnih in negativnih komentarjev festivala Evrovizija.

2. del: Raziskave spletne slovenščine

V petem poglavju **Darja Fišer**, **Maja Miličević** in **Nikola Ljubešić** obravnavajo značilnosti nestandardnega zapisa besed v slovenskih tvitih z opazovanjem pretvorb iz standardnega v nestandardni zapis besed treh različnih vrst transformacij: izpust, dodajanje in zamenjavo črk. Največ transformacij identificirajo med polnopomenskimi besedami, najpogostejše pa so tiste, do katerih prihaja pri slovničnih besedah. Najpogostejši so izpusti, predvsem samoglasnikov, do česar največkrat prihaja na koncu besed, s čimer se neformalna komunikacija v tvitih približuje govoru.

V šestem poglavju **Damjan Popič** in **Darja Fišer** proučita rabe vejice v slovenskih tvitih, kjer ju zanima, v kolikšni meri je stava v tovrstnih vsebinah rabljena v skladu z jezikovnim standardom in na katerih mestih imajo uporabniki s stavo vejice največ težav. Za opis rabe vejice razvijeta tipologijo ter smernice za označevanje, ki ju preizkusita z pomočjo dveh označevalcev. Ugotovita, da tudi v nestandaradni računalniško posredovani komunikaciji raba standardne vejice prevladuje. Pri veliki večini nestandardne rabe vejice gre za izpuste, ki ne izkazujejo posebnosti medija, identificirata pa tudi namerne izpuste vejice, ki so rezultat tehničnih in družbenih okoliščin tovrstnega komuniciranja.

V sedmem poglavju **Jaka Čibej** predstavi analizo regionalnih jezikovnih različic v spletni slovenščini, za kar v skladu z razvito tipologijo ročno označi tvite uporabnikov iz različnih slovenskih regij. Višjo zastopanost nestandardnih prvih zazna pri uporabnikih iz gorenjske in dolenjske regije, te pa so najpogosteje in najbolj sistematično izkazane v omejenem naboru zaprtih besednih vrst, predvsem v

obliki izpustov samoglasnikov. Nestandardnega besedišča je bilo v označenem vzorcu za tretjino, nestandardnih oblikoslovnih prvin pa zelo malo.

V osmem poglavju **Darja Fišer** in **Nikola Ljubešić** opišeta pristop za polavtomatsko prepoznavanje pomenskih premikov, ki temelji na primerjavi semantičnih profilov besed v korpusu tvitov in v referenčnem korpusu. Za vrednotenje rezultatov razvijeta tipologijo pomenskih premikov, pri čemer kandidate preverita s pomočjo ročne, korpusno podprte leksikografske analize. Na tak način identificirata največ pomenskih premikov, ki se v tvitih pojavijo zaradi dnevnih dogodkov in neformalnih sporočanjskih okoliščin.

V devetem poglavju **Špela Arhar Holdt** analizira skladenjske značilnosti računalniško posredovane slovenščine na ravni besednega reda, ki jo izvede s pomočjo skladenjsko ročno označenega vzorca z odvisnostnim sistemom JOS, prilagojenim specifikam računalniško posredovane slovenščine, kot so novomedijski komunikacijski elementi (emojiji, sklici na uporabniška imena, heštegi), tujejezične prvine, jezikovni fragmenti in nestandardna raba ločil. S kategorizacijo označenih besednorednih značilnosti osvetli doslej še neraziskan segment skladnje računalniško posredovane komunikacije in nakaže možnosti nadaljnjih raziskav besednega reda v slovenščini, s primerjavo ujemanja oznak različnih označevalcev pa izpostavi tudi zelo različno zaznavo (ne)zaznamovanosti in (ne) standaradnosti besednega reda pri govorcih.

V desetem poglavju **Ana Zwitter Vitez** in **Darja Fišer** s triangulacijo korpusov Janes, Gos in Kres preverita prisotnost prvin govorjenega jezika v računalniško posredovani komunikaciji na ravni ključnih besednih oblik in ključnih besed. Ugotovita, da so na besednovrstni ravni besedila računalniško posredovane komunikacije bistveno bliže pisnim besedilom kot govorjenemu diskurzu. Na ravni besedišča so posebej zanimivi za govor specifični elementi interakcije z drugimi udeleženci, ki so najpogosteje prisotni v tvitih in komentarjih.

V enajstem poglavju **Mija Michelizza** obravnava rabo ključnikov v slovenskih tvitih, ki so primarno namenjeni kategoriziranju, lahko pa opravljajo tudi različne komunikacijske vloge. Ugotavlja, da čeprav ključniki predstavljajo novejši jezikovni element, ki v računalniško posredovani komunikaciji izstopa, ti večinoma ostajajo znotraj svoje primarne kategorizacijske vloge, tisti, ki opravljajo komunikacijske vloge, pa so pogosteje skladenjsko vpeti v besedilo. Med identificiranimi fenomeni izstopajo ključniki, ki izkazujejo povezanost objavljanja tvitov s spremljanjem drugih medijev.

V dvanajstem poglavju **Špela Reher** in **Darja Fišer** proučita preklapljanje med jeziki v slovenskih tvitih. Analiza temelji na ročno označenem vzorcu tvitov, za kar tudi razvijeta lastno označevalno shemo. Ugotovita, da preklapljanje ni redek pojav, da med jeziki preklopov izrazito prevladuje angleščina, da do preklopov

pogosteje prihaja znotraj stavkov, da so preklopi v enaki meri eno- in večbesedni in da vključujejo tudi slovnične besedne vrste. Kodno preklapljanje opravlja številne diskurzivne funkcije, kot so referenčna, ekspresivna in poudarjalna, glede na semantično polje pa so preklopi pogosto povezani s popularno kulturo, zlasti TV-oddajami, športom, hrano in Twitterjem.

V trinajstem poglavju **Helena Dobrovoljc** obravnava spremembe pri pozdravljanju, poslavljanju in naslavljanju v pisemskem sporazumevanju v digitalnem okolju. Osvetli spontane in naučene spremembe pisemskega diskurza, ki prodirajo v vse tipe elektronskega pisemskega sporazumevanja in prikazujejo, kako so se v različnih obdobjih spreminjala družbena razmerja in piščev odnos do pisemskega sporazumevanja.

V štirinajstem poglavju **Iza Škrjanec**, **Nada Lavrač** in **Senja Pollak** predstavijo avtomatsko razpoznavo spola avtorjev in avtoric slovenskih blogovskih zapisov, za kar preizkusijo model z ročno zgrajenimi pravili in model, zgrajen z metodami strojnega učenja. Kot najuspešnejši se izkaže model strojnega učenja, ki je naučen na unigramih pojavnic s pomočjo metode podpornih vektorjev. Analiza najbolj informativnih značilk tega modela pokaže, da se besedila blogerk in blogerjev razlikujejo po uporabljenih slovničnih in slogovnih sredstvih ter glede na tematiko.

ZAHVALA

Raziskave, opisane v monografiji, so bile opravljene v okviru nacionalnega temeljnega projekta »Viri, orodja in metode za raziskovanje nestandardne spletne slovenščine« (J6-6842, 2014–2017), ki ga financira ARRS.

Urejanje te monografije je bil velik, a slasten zalogaj. Za vse napake in nekonsistentnosti, ki so se pritihotapile vse do končne verzije, sem kriva urednica, za kar se opravičujem vsem prizadetim. Monografije pa seveda ne bi bilo brez izjemno predanega dela vseh sodelujočih avtorjev in avtoric, za kar se jim iskreno zahvaljujem: Špela Arhar Holdt, Darja Fišer, Nikola Ljubešić, Maja Miličević Petrović, Jaka Čibej, Helena Dobrovoljc, Tomaž Erjavec, Nada Lavrač, Matej Martinc, Mija Michelizza, Senja Pollak, Damjan Popič, Špela Reher, Iza Škrjanec, Ana Zwitter Vitez.

Posebna zahvala gre tudi recenzentom, ki so prispevke skrbno in natančno pregledali ter s konstruktivnimi pripombami pomembno prispevali k višji znanstveni kvaliteti monografije: Maja Bitenc, Polona Gantar, Vojko Gorjanc, Monika Kalin Golob, Petra Kralj Novak, Simon Krek, Tina Lengar Verovnik,

Nataša Logar, Dunja Mladenić, Marko Robnik Šikonja, Tadeja Rozman, Marko Stabej, Darinka Verdonik, Reinhild Vandekerckhove, Jana Zidar Forte. Za prevode rokopisov, ki so zaradi mednarodne avtorske zasedbe nastali v angleščini, se zahvaljujem Dafne Marko in Lei Anžur. Za neumorno lektorsko in korektorsko prečesavanje monografije sem neizmerno hvaležna Damjanu Popiču, za profesionalno oblikovanje in neusahljiv vir dobre volje pa Juretu Preglauu.

Darja Fišer

Ljubljana, 30. marec 2018

Literatura

Tjong Kim Sang, Erik, Marcel Bollmann, Remko Boschker, Francisco Casacuberta, Feike Dietz, Stefanie Dipper, Miguel Domingo, Rob van der Goot, Marjo van Koppen, Nikola Ljubešić, Robert Östling, Florian Petran, Eva Pettersson, Yves Scherrer, Marijn Schraagen, Leen Sevens, Jörg Tiedemann, Tom Vanallemeersch in Kalliopi Zervanou, 2017: The CLIN27 Shared Task: Translating Historical Text to Contemporary Language for Improving Automatic Linguistic Annotation. *Computational Linguistics in the Netherlands Journal* 7/1. 53–64.

Škrjanec, Iza, 2017: *Gender-based analysis of Slovene user-generated content*. Master thesis. Jožef Stefan Postgraduate School.

Reher, Špela, 2017: *Slovenščina na prepihu: kodno preklapljanje v objavah slovenskih uporabnikov Twitterja*. Magistrsko delo. Filozofska fakulteta Univerze v Ljubljani.

Resolucija o Nacionalnem programu za jezikovno politiko 2014–2018. http://www.mk.gov.si/fileadmin/mk.gov.si/pageuploads/Ministrstvo/Zakonodaja/2013/Resolucija_-_sprejeto_besedilo__15.7.2013_.pdf.

Štebe, Janez, Sonja Bezjak in Sanja Lužar, 2013: *Odprti podatki: načrt za vzpostavitev sistema odprtega dostopa do raziskovalnih podatkov v Sloveniji*. Ljubljana: Fakulteta za družbene vede.

Korpus slovenskih spletnih uporabniških vsebin Janes

Tomaž Erjavec, Nikola Ljubešić, Darja Fišer

Izvleček

V poglavju predstavimo korpus spletne slovenščine Janes, ki vsebuje tvite, spletne forume, novice in uporabniške komentarje nanje, uporabniške in pogovorne strani na Wikipediji ter blogovske zapise in komentarje nanje. Najprej opišemo postopek zajema besedil za vsakega od vključenih virov in podamo kvantitativno analizo zgrajenega korpusa. Sledi predstavitev avtomatskih in ročnih postopkov za obogatitev korpusa s koristnimi metapodatki, kot so tip in spol avtorja ter sentiment in stopnja tehnične in jezikovne standardnosti posameznega besedila. Poglavje nato poda zapis korpusa in postopek izdelave ter dostopnost njegove javne različice.

Ključne besede: gradnja korpusa, računalniško posredovana komunikacija, uporabniške spletne vsebine, spletna slovenščina, nestandardna slovenščina

1 UVOD

Slovenščina je razmeroma dobro podprta s korpusi, tako referenčnimi kot specializiranimi,[1] vendar ti ne vsebujejo besedil, ki jih na spletu ustvarjajo uporabniki družbenih omrežij. Edina delna izjema je slWaC (Erjavec et al. 2015), ki vsebuje spletna besedila z domene *.si*, vendar pa v korpusu ni metapodatkov, ki bi razločevali besedila poklicnih piscev, ki so potencialno tudi lektorirana in uredniško pregledana, od tistih, ki so jih ustvarili uporabniki spletnih portalov. Zaradi množične razširjenosti spletnih uporabniških vsebin (Statistični urad RS 2015) in posledičnim naraščanjem njihovega pomena za jezikoslovje, tehnologije pa tudi za družbo nasploh in ker številne tuje (Crystal 2011, Baron 2008, Beißwenger 2013) ter prve domače jezikoslovne raziskave (Dobrovoljc 2012, Erjavec in Fišer 2013, Michelizza 2015) kažejo, da se jezik v njih v marsičem razlikuje od pisnega standarda, smo za omogočanje celovitega in podrobnega proučevanja slovenske računalniško posredovane komunikacije zgradili obsežen, heterogen, jezikoslovno označen in z bogatim naborom metapodatkov opremljen korpus spletnih uporabniških vsebin, imenovan Janes (Jezikoslovna analiza nestandardne slovenščine).

Korpus je bil izdelan v več različicah, pri čemer je bila predzadnja korpus Janes 0.4, ki smo jo že opisali v prispevku Fišer et al. (2016b). V pričujočem poglavju opišemo zadnjo različico korpusa, Janes 1.0, za katero smo dopolnili podkorpusa tvitov in komentarjev na Wikipediji, predvsem pa smo jo v celoti na novo jezikoslovno označili z uporabo najnovejših orodij oz. modelov, ki so bili naučeni na končnih ročno označenih podatkih. S ciljem podpreti odprto znanost smo korpus – ob poprejšnji anonimizaciji – naredili tudi odprto in javno dostopen, tako prek spletnega konkordančnika kot tudi za prevzem v repozitoriju raziskovalne infrastrukture CLARIN.SI.

Poglavje ima naslednjo strukturo. V drugem razdelku predstavimo sorodne raziskave. V tretjem razdelku opišemo zvrstnost korpusa, načela vključevanja virov in postopek zbiranja besedil ter korpus kvantificiramo. V četrtem razdelku predstavimo metapodatke, s katerimi so opremljena besedila v korpusu in ki omogočajo širok nabor natančnejših in primerjalnih jezikoslovnih analiz, podamo pa tudi analize korpusa po posameznih metapodatkih. Peti razdelek opiše zapis korpusa, šesti pa postopke za izdelavo javne različice korpusa in njegovo dostopnost, čemur sledijo sklepne ugotovitve in načrti za nadaljnji razvoj korpusa.

1 Glavni referenčni korpus je Gigafida s pridruženimi korpusi KRES, ccGigafida in ccKRES (Logar et al. 2012), izmed velikega števila ostalih pa omenimo korpus govorjene slovenščine Gos (Verdonik in Zwitter Vitez 2011) in korpus starejše slovenščine IMP (Erjavec 2015).

2 PREGLED SORODNIH RAZISKAV

Glede na to, da so raziskave računalniško posredovane komunikacije (RPK) v korpusnem in računalniškem jezikoslovju pa tudi v družboslovju izrazito empirično naravnane, je presenetljivo, da je raziskovalcem dostopnih razmeroma malo korpusov RPK (Beißwenger in Storrer 2008). Med največjimi, ki jih je možno tudi prenesti na svoj računalnik, so finski Suomi24 (Lagus et al. 2016), ki vsebuje 2,4 milijarde pojavnic s spletnih forumov, nemški DEREKO-Wikipedia (Margaretha in Lüngen 2014), ki vsebuje 580 milijonov pojavnic iz člankov in uporabniških pogovornih strani na Wikipediji, ter francoski CoMeRe (Chanier et al. 2014) z 80 milijoni pojavnic iz elektronskih pisem, forumov, klepetalnic, tvitov in Wikipedije.

Za jezikoslovne analize se tipično uporabljajo precej manjši, a skrbneje označeni korpusi. Nemški Dortmund Chat Corpus (Beißwenger et al. 2015), ki je na voljo za prenos prek raziskovalne infrastrukture CLARIN, tako vsebuje le milijon pojavnic iz spletnih klepetalnic, a so besedila ročno anonimizirana, značilni elementi računalniško posredovane komunikacije v njih pa ročno označeni. Sms2science.ch (Dürscheid in Stark 2011) je korpus, za katerega so prostovoljci prispevali 650 tisoč pojavnic iz SMS-sporočil, napisanih v nemščini, francoščini, švicarski nemščini, italijanščini in retoromanščini, in je ročno normaliziran ter dostopen prek spletnega konkordančnika. Zelo podoben je korpus DiDi (Frey et al. 2015) s 570.000 pojavnicami, ki so jih v nemščini, italijanščini in južni tirolščini prispevali uporabniki Facebooka iz Južne Tirolske v Italiji.

Poleg korpusov so bile razvite posebne učne množice RPK, namenjene razvoju računalniških orodij, kot so analiza sentimenta (Barbieri et al. 2016), prepoznavanje in povezovanje imenskih entitet (Derczynski et al. 2015, Rei et al. 2016, Derczynski et al. 2016) ter razdvoumljanje večpomenskih besed (Jonansson et al. 2016).

Kot že omenjeno, za slovenščino še ni bil izdelan noben velik in javno dostopen specializirani korpus RKP, z izjemo korpusa tvitov Tweet-sl, ki zajema tvite iz obdobja 2007–2011. Vendar je ta korpus, vsaj v primerjavi s korpusom Janes-Tweet, opisanim v tem prispevku, razmeroma majhen (6.300.000 pojavnic), ni opremljen z metapodatki in je dostopen samo prek konkordančnika,[2] tako da ni voljo za prevzem. Poleg tega sta bili opravljeni dve raziskavi (Bučar et al. 2015, Kadunc in Robnik Šikonja 2016), ki sta se osredotočili na označevanje in modeliranje sentimenta v RKP, pri obeh pa so bili izdelani korpusi

2 Korpus Tweet-sl je dostopen za pregledovanje prek konkordančnika na naslovu https://www.clarin.si/noske/run.cgi/corp_info?corpname=tweet_sl.

dani v odprti dostop v okviru repozitorija CLARIN.SI (Bučar 2017, Kadunc in Robnik Šikonja 2017).

3 GRADNJA IN ZVRSTNOST KORPUSA

V korpus Janes 1.0 je vključenih pet zvrsti javno objavljenih uporabniških spletnih vsebin, in sicer tviti, forumi, novice in komentarji nanje, uporabniške in pogovorne strani na Wikipediji ter blogi. Teh pet zvrsti besedil je bilo izbranih iz več razlogov. Tvite smo vključili, ker so v zadnjem desetletju v svetovnem merilu postali izjemno množična oblika RPK in ji veliko pozornosti posvečajo raziskovalci iz različnih disciplin. Vsebine na Wikipediji imajo dragoceno prednost, da ni problemov z njihovo redistribucijo, saj so dostopne pod izrazito liberalno licenco Creative Commons. Ostale tri zvrsti (forumi, komentarji na novice in blogi) pa so zanimive z različnih vidikov, od proučevanja specializirane komunikacije na določeno temo posameznih spletnih skupnosti do opazovanja učinkov samozaložništva, pluralizacije mnenj in demokratizacije jezika. Čeprav se te zvrsti že pojavljajo v korpusu slWaC, besedila v njem nimajo pripisanih dragocenih sociodemografskih metapodatkov in niso strukturirana v pogovorne niti. Za zagotavljanje čim večje pokritosti bi bilo sicer smiselno vključiti tudi druge družbene platforme, predvsem Facebook, ki je v Sloveniji najbolj razširjeno družbeno omrežje (Statistični urad RS 2015), vendar na njem prevladuje zasebna komunikacija, za katero ponudnik izrecno prepoveduje zbiranje in distribucijo vsebin.

Zajem tvitov in uporabniških ter pogovornih strani na Wikipediji je bil celovit, v smislu, da smo v korpus vključili vse uporabnike in njihove objave s teh platform, ki smo jih identificirali. Zaradi časovnih in finančnih omejitev pa smo za zajem forumskih sporočil, komentarjev na novice in blogov izbrali zgolj manjši nabor virov, ki so v slovenskem spletnem prostoru najbolj priljubljeni, tj. da ponujajo največ jezikovne produkcije in/ali so tematsko najbolj zanimivi. To smo ocenili na podlagi števila registriranih uporabnikov, števila in dinamike objavljenih sporočil ter nabora aktivnih tem. Izbor in zajem posamezni virov sta podrobneje opisana v nadaljevanju razdelka. Čeprav se zavedamo, da s tem še zdaleč nismo zajeli vseh tem, s katerimi se spletne uporabniške vsebine ukvarjajo, in besedišča, ki je v njih uporabljeno, predvidevamo, da smo kljub vsemu zbrali zadovoljiv vzorec jezikovne rabe, ki je za ta način komunikacije med govorci slovenščine značilna. V nadaljevanju razdelka opišemo vire in metode, ki smo jih uporabili za zajem posameznih zvrsti besedil, zajetih v korpusu.

3.1 Zajem besedil

3.1.1 Tviti

Tvite smo zajeli z namenskim orodjem TweetCat[3] (Ljubešić et al. 2014), ki je bilo izdelano prav za gradnjo korpusov tvitov manjših jezikov. Orodje uporablja Twitter Search API,[4] da najde uporabnike, ki tvitajo v ciljnem jeziku (v primeru korpusa Janes je to slovenščina). V začetni fazi išče tvite, ki vsebujejo semenske besede izbranega jezika. Te morajo biti visoko frekventne in specifične za ciljni jezik korpusa ter se ne smejo prekrivati z besedami v sorodnih jezikih. Seznam semenskih besed, ki smo jih uporabili za zajem slovenskih tvitov, je sledeč: *ampak, če, jutri, kaj, kdaj, kje, končno, mogoče, očitno, oziroma, preveč, ravnokar, še, spet, sploh, tudi, vendar, všeč, zdaj, že*.

Ko orodje identificira uporabnike, ki potencialno tvitajo v ciljnem jeziku, izvede pravo identifikacijo jezika uporabnika na njegovi časovnici, saj je točnost določanja jezika močno odvisna od količine besedila. Avtorji pretežno ciljnega jezika so dodani v indeks uporabnikov, ki jim orodje ves čas sledi in shranjuje njihove tvite. V množico potencialno zanimivih uporabnikov so zajeti tudi vsi uporabniki, ki jim že identificirani tviteraši sledijo, s čimer se število zajetih uporabnikov, posledično pa tudi količina zajetih tvitov ves čas povečujeta.

Pred dokončno vključitvijo podatkov, zbranih z orodjem TweetCat, v korpus, smo izvedli dodaten korak filtriranja uporabnikov, kjer s Pythonovim modulom *langid.py* identificiramo jezik še vsakemu zajetemu tvitu posameznega tviteraša in odstranimo tiste uporabnike, pri katerih večinski jezik ni slovenščina. To zaporedje filtrov je potrebno, da bi res zajeli čim več slovenskih in čim manj tujejezičnih tviterašev ob zavedanju, da je identifikacija jezika težak problem, toliko bolj za besedila na Twitterju, ki so zelo kratka, pogosto niso napisana v standardnem jeziku in lahko vsebujejo veliko tujejezičnih prvin, kar potrjujejo tudi naše raziskave: kot bo obravnavano v nadaljevanju, ocenjujemo, da je prek 40 % zbranih tvitov pisanih v nestandardnem jeziku (razdelek 4.6) ter da jih je skoraj 10 % napisanih v angleščini (razdelek 4.5). Zato so vsa filtriranja opravljena na uporabnikih in ne na tvitih: ti so za uporabnike, ki pretežno tvitajo v slovenščini, vsi vključeni v korpus, ne glede na to, v katerem jeziku so napisani.

3.1.2 Forumi

V korpus smo vključili zdravstvene posvetovalnice s foruma *med.over.net* ter specializirana foruma s področja avtomobilizma *avtomobilizem.com* in znanosti

3 Orodje je dostopno na https://github.com/clarinsi/tweetcat.

4 https://dev.twitter.com/rest/public/search

kvarkadabra.net, s čimer smo želeli zajeti najaktivnejše forume, pokriti raznovrsten nabor tem in zaobjeti raznolike segmente jezikovne rabe v slovenskih forumih. To smo ocenili na podlagi števila registriranih uporabnikov posameznega foruma ter števila in dinamike objavljenih sporočil. Izbor je bil opravljen z analizo sedanjega stanja za 96 slovenskih forumov s seznama Lebar et al. (2012). Ker se spletna mesta po sestavi med seboj razlikujejo, smo morali za vsak vir posebej napisati ekstraktor besedila,[5] kar je bilo ozko grlo pri nadaljnjem širjenju virov besedil. Iz zajetega materiala smo na ta način izluščili le tiste podatke, ki smo jih želeli vključiti v korpus, in se tako izognili velikemu deležu šumnih prvin, kot so oglasna sporočila, nerelevantne povezave ipd. Ekstraktor ohrani izvorno strukturo vira, tako da so pri forumih zajeti prispevki organizirani v posamezne podforume in teme.

3.1.3 Komentarji na novice

Z novičarskih portalov smo zajeli osrednji nacionalni javni medij *rtvslo.si* ter dva ožje usmerjena politična tednika, levi politični opciji naklonjeni *mladina.si*[6] in desno usmerjeni *reporter.si*. Za vključitev vira v korpus je bila ključna politika novičarskih portalov ob začetku zbiranja, saj številni portali dostop do novic zaračunavajo (npr. *finance.si*), po določenem času komentarje avtomatsko izbrišejo (npr. *siol.net*) ali pa imajo komentiranje člankov zaklenjeno (npr. *dnevnik.si*), s čimer je zajem komentarjev tehnično onemogočen. Zajem je tudi potekal s pomočjo namenskih ekstraktorjev, napisanih za vsak vir posebej, podobno kot zajem forumov.

Ker je analiza komentarjev na novice neločljivo povezana z novico, na katero se komentarji nanašajo, smo kontekstualno celovito analizo komentarjev omogočili tako, da smo pri zajemu komentarjev zajeli tudi novice, čeprav le-te ne sodijo med uporabniško generirane vsebine in so zato v korpusu od njih jasno ločene.

3.1.4 Blogi

Za zajem blogov in komentarjev nanje smo se želeli izogniti težavnemu identificiranju posameznih slovenskih blogov na najpopularnejših tujejezičnih blogerskih portalih (npr. *blogger.com*) in izbrali dva slovenska, ki sta med najpopularnejšimi med laičnimi uporabniki za objavo amaterskih blogov. Tudi pri izboru blogerskih

5 Za luščenje besedil iz zajetih spletnih strani smo uporabili Pythonovo knjižnico Beautiful Soup: https://www.crummy.com/software/BeautifulSoup/.

6 V času zajema je tednik Mladina še omogočal komentiranje spletnih novic, vendar ga je kasneje onemogočil, tako da lahko bralci novic na njihovem portalu v času pisanja prispevka komentarje posredujejo le v obliki pisem bralcev.

portalov so pomembno vlogo odigrale tehnične okoliščine, kjer smo dali prednost tistim portalom in blogom, ki so imeli poenoteno strukturo, saj nam je to omogočilo hkratni zajem večje količine blogov različnih avtorjev in komentarjem nanje. Navedenim kriterijem sta ustrezala blogerska portala *publishwall.si* in *rtvslo.si*, žal pa ne sicer zelo popularna slovenska blogerska portala *blog.siol.net* in *ednevnik.si*.

Tudi tu je zajem potekal z namenskimi ekstraktorji, napisanimi za vsak vir posebej, tako kot pri zajemu forumov in komentarjev na spletne novice.

3.1.5 Uporabniške in pogovorne strani na Wikipediji

Zajem pogovornih strani z Wikipedije smo opravili z lastnim orodjem,[7] ki obdela izvoz Wikipedije.[8] Edina jezikovno odvisna podatka, ki ju orodje potrebuje, sta niz, ki določa uporabnika, in koda jezika (»*uporabnik*« in »*sl*« za slovenščino). Strani, ki komentirajo posamezne Wikipedijine strani (*pagetalk*), smo za omogočanje natančnejših analiz v korpusu eksplicitno ločili od komentarjev na uporabniških straneh posameznih avtorjev slovenske Wikipedije (*usertalk*).

3.2 Postprocesiranje

Zajete podatke vseh petih podkorpusov smo dodatno očistili, predvsem glede kodnih sistemov. V tej fazi smo za vsak podkorpus posebej popravili najpogostejše napake v kodiranju (predvsem kar se tiče šumnikov), saj so se vrste napak med viri zelo razlikovale. Pri sistematičnih napakah smo pretvorili znake ali nize v ustrezen znak Unikod, pri ostalih identificiranih napakah pa smo izbrisali bodisi znake, ki po standardu Unikod niso dovoljeni, bodisi celotno besedilo. V tej fazi smo poskrbeli tudi, da podkorpus ne vsebuje praznih besedil in da se zapiše kot veljaven dokument XML.

3.3 Velikost korpusa

Korpus Janes 1.0 vsebuje skoraj 13 milijonov besedil, v katerih je preko 268 milijonov pojavnic oz. 226 milijonov besed. Zgrajeni korpus je zelo heterogen,

7 Dostopno na https://github.com/nljubesi/wikitalk-extractor.

8 Dostopen na https://dumps.wikimedia.org.

tako glede na količino, dolžino in starost vključenih besedil kot tudi glede na avtorstvo, kar prikažemo s kvantitativno analizo korpusa v nadaljevanju razdelka.

Tabela 1: Velikost podkorpusov Janes po vrsti besedila in posameznih virih.

(Pod)korpus in vir	Št. besedil	Št. besed	Št. pojavnic	Št. besed/ besedilo
Tweet	11.336.646	135.478.891	160.404.265	12,0
Forum	772.953	39.769.122	47.066.575	51,5
avtomobilizem	569.594	21.927.000	25.629.275	38,5
medovernet	122.613	11.618.053	13.799.211	94,8
kvarkadabra	80.746	6.224.069	7.638.089	77,1
Blog	404.281	28.816.954	34.534.431	71,3
rtvslo.post	23.515	8.082.628	9.621.808	343,7
rtvslo.comment	324.586	11.616.062	14.070.220	35,8
publishwall.post	18.515	7.295.274	8.634.274	394,0
publishwall.comment	37.665	1.822.990	2.208.129	48,4
News	308.130	18.153.521	21.442.211	58,9
rtvslo.article	5.074	2.699.423	3.164.041	532,0
rtvslo.comment	267.909	10.346.527	12.239.673	38,6
mladina.article	2.924	2.626.867	3.090.377	898,4
mladina.comment	26.011	1.890.301	2.253.521	72,7
reporter.article	913	302.083	349.719	330,9
reporter.comment	5.299	288.320	344.880	54,4
Wiki	78.765	4.041.123	5.008.067	51,3
pagetalk	25.981	1.245.428	1.545.321	47,9
usertalk	52.784	2.795.695	3.462.746	53,0
Σ	12.900.775	226.259.611	268.455.549	17,5

Kot prikazuje Tabela 1, je v korpusu Janes največji podkorpus tvitov s preko 160 milijoni pojavnic, s čimer predstavlja skoraj dve tretjini celotnega korpusa. Sledijo mu podkorpusi forumskih sporočil, blogov in komentarjev na novice, najmanj pa je komentarjev z Wikipedije. Tabela poda tudi razdelitev po virih znotraj posameznih besedilnih zvrsti, pri čemer pri blogih ločujemo tudi izvirne zapise (*post*) in komentarje nanje (*comment*), pri spletnih novicah pa novice (torej *news.post)* in komentarje nanje. Kot lahko vidimo, je med forumi z dobrimi 25 milijoni pojavnic največji *avtomobilizem*, *medovernet* (od katerega smo zajeli večinoma le zdravstvene posvetovalnice, ostalih podforumov pa ne) je skoraj polovico manjši, *kvarkadabra* pa je manjši še za polovico. Pri blogih je zanimivo, da so kljub temu, da smo z obeh platform zajeli približno enako količino blogovskih zapisov (9,6 v primerjavi z 8,6 milijoni pojavnic), blogi s platforme *rtvslo* pospremljeni s šestkrat več komentarji kot blogi na platformi *publishwall.* Pri komentarjih na novice so

razlike še večje, saj tisti z *rtvslo* vsebujejo prek 12 milijonov pojavnic, kar je petkrat več od števila zajetih komentarjev s portala *mladina*, medtem ko nam je s portala *reporter* uspelo zajeti zgolj dobrih tristo tisoč pojavnic komentarjev.

V Tabeli 1 je podana tudi povprečna dolžina besedil v besedah. Besedila v korpusu so tipično zelo kratka, saj v povprečju vsebujejo manj kot 18 besed, kar je značilno za zajete besedilne zvrsti. Če izvzamemo novice, so po pričakovanju najdaljši blogovski zapisi s skoraj 400 besedami na besedilo na portalu *publishwall*, najkrajši pa tviti z 12 besedami, dolžina katerih je zaradi odločitve ponudnika platforme omejena na največ 140 znakov.[9] Zanimivo je, da je dolžina ostalih besedil precej bolj primerljiva, od malo pod 39 besed za forum *avtomobilizem* do skoraj 95 za *medovernet*, kar razkrije, da so med posameznimi viri precejšnje razlike, ki so celo večje kot med posameznimi besedilnimi zvrstmi.

4 KORPUSNI METAPODATKI

Pomembna odlika korpusa Janes je bogastvo metapodatkov o posameznih besedilih ali skupinah besedil, kar nam omogoča bistveno bogatejše jezikoslovne analize, pa tudi uporabo korpusa za različne sociolingvistične in družboslovne raziskave. Nekateri metapodatki so bili zajeti neposredno, predvsem URL izvornega besedila, pri tvitih pa preko Twitter API-ja identifikator tvita, uporabniško ime avtorja, datum in čas pošiljanja, število posredovanj (*retweets*) in všečkov (*favourites*).

Osnovne metapodatke za ostale vire smo izluščili iz posameznih besedil v procesu čiščenja, pri čemer je potrebno izpostaviti, da uporabljene hevristike niso vedno popolne in zato vsa besedila nimajo vseh pripadajočih metapodatkov, včasih pa pri njihovi ekstrakciji pride tudi do napak. Za vse zvrsti besedil smo tako pridobili uporabniško ime avtorja,[10] naslov in datum objave besedila, za forume pa tudi naslov posameznega podforuma in teme.

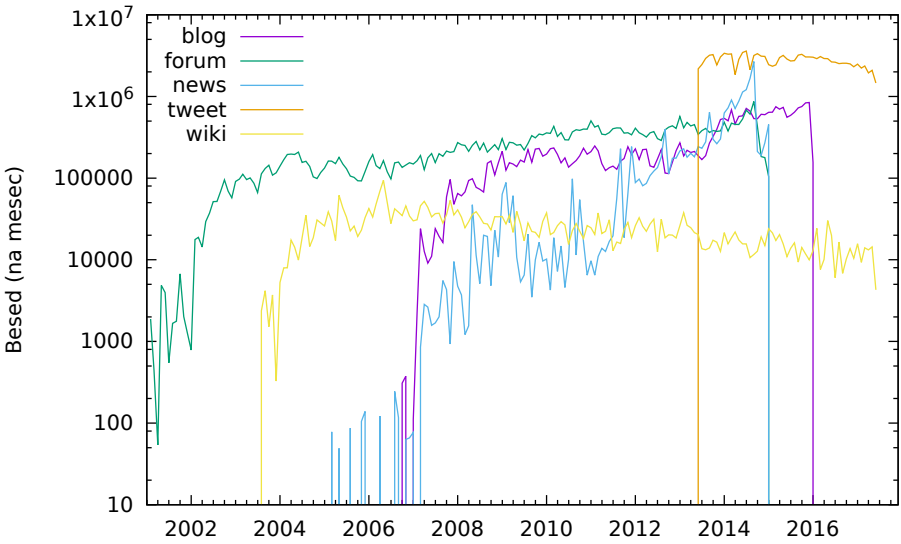
Poleg metapodatkov, ki jih je bilo možno zajeti iz besedila, smo celoten korpus oz. posamezne podkorpuse dodatno obogatili z metapodatki, ki so bili dodani bodisi avtomatsko bodisi ročno. V nadaljevanju razdelka predstavimo statistike za najpomembnejše metapodatke, podrobneje pa razložimo tudi bolj zanimive postopke dodajanja metapodatkov.

9 Čeprav ta podatek drži za vse tvite, vključene v korpus Janes v1.0, je ponudnik platforme Twitter 7. 11. 2017 omejitev s 140 razširil na 280 znakov na posamezni tvit: https://www.theguardian.com/technology/2017/nov/08/twitter-to-roll-out-280-character-tweets-to-everyone. Vplivov te spremembe na jezik tvitov še nismo analizirali, vendar večjih sprememb ne pričakujemo, saj je po podatkih podjetja Twitter dosedanjo zgornjo mejo 140 znakov dosegalo le 9 % vseh tvitov, objavljenih v angleščini, novo zgornjo mejo 280 v preizkusnem obdobju pa le 1 % vseh angleških tvitov.

10 Izjema so novice, ki velikokrat niso podpisane ali imajo v najboljšem primeru samo okrajšavo imena avtorja, zato pri *news.post* imena avtorja ne navajamo.

4.1 Starost besedil

Za večino zvrsti besedil smo zajem izvedli samo enkrat, in sicer februarja 2015 za forumske objave, novice in komentarje nanje, januarja 2016 za bloge in komentarje nanje, julija 2017 pa za komentarje na Wikipediji.



Slika 1: Starost besedil v podkorpusih.

Za razliko od teh spletnih vsebin, ki na spletu ostanejo razmeroma dolgo (delna izjema so komentarji, ki jih nekateri ponudniki platform po določenem času brišejo), vrača uporabljeni Twitter API samo zadnjih 500 tvitov posameznega uporabnika, zato je pomembno, da se tviti zbirajo sproti. TweetCat je obratoval skoraj neprekinjeno od začetka zbiranja junija 2013 pa do konca zbiranja julija 2017, pri čemer smo v Janes 1.0 vključili vse zajete tvite. Ob začetku zbiranja smo pridobili tudi manjše število starejših tvitov, ki pa jih v korpus nismo vključili, saj so na voljo samo pri uporabnikih, ki tvitajo zelo malo.

Kot prikazuje Slika 1, kjer je ordinata logaritemska, so bila besedila, vključena v korpus, objavljena v obdobju 2001–2017. Najstarejši vir so forumi, ki so očitno dovolj stabilni, da je z njih mogoče pridobiti objave vse od februarja 2001, stabilni pa so tudi komentarji na Wikipediji (od avgusta 2003) in blogi (od oktobra 2006). Pri komentarjih na Wikipediji je zanimiv uvid, da njihovo število strmo narašča do konca 2006, nato pa začne počasi upadati, tako da jih je ob koncu zbiranja več kot desetkrat manj na mesec kot v obdobju največjega navdušenja

nad Wikipedijo. Najstarejše novice oz. komentarji nanje so sicer iz leta 2005, vendar je teh zelo malo, medtem ko jih je velika večina iz 2014, kar je najverjetneje posledica tehničnih rešitev novičarskih portalov. Kot rečeno je v povprečju najmlajši vir besedil družbeno omrežje Twitter, pri čemer občasna nihanja niso posledica začasne neuporabe Twitterja, temveč kažejo na obdobja, ko zaradi težav s strežnikom zbiranje tvitov ni delovalo.

Idealen korpus uporabniških spletnih vsebin, kjer se načini komuniciranja in obravnavane teme lahko hitro spreminjajo, bi vseboval enakomerno časovno razporejena besedila po vseh zajetih zvrsteh besedil, kar pa za Janes 1.0, kot je razvidno iz napisanega, ne drži. Kot omenjeno je razlog v različni dinamiki zajema posameznih zvrsti in v različni obstojnosti besedilnih zvrsti na spletu. Zato je pri uporabi korpusa potrebna previdnost, saj je med najstarejšimi in najmlajšimi besedili v korpusu kar 15 let razlike. Čeprav bi lahko vzorčili korpus tako, da bi dobili bolj uravnoteženo diahrono razporeditev, smo se odločili, da raje zadržimo v korpusu celotne zajeme posameznih zvrsti, saj je s tem korpus bistveno večji, raziskovalci, ki se zanimajo za diahrono komponento, pa še vedno lahko izdelajo podkorpus določenega časovnega obdobja, saj so vsa besedila opremljena z metapodatkom o času nastanka.

4.2 Avtorstvo besedil

Besedila v korpusu je napisalo več kot 96.000 avtorjev (uporabnikov), kjer kot enega avtorja štejemo eno uporabniško ime znotraj enega podkorpusa. Število avtorjev je tako zgolj ocena, saj lahko ista oseba uporablja različna uporabniška imena znotraj enega vira ali enako ime v različnih virih, zgodi pa se celo, da ima več oseb enako uporabniško ime v istem viru.

Kot kaže Tabela 2, je posamezni avtor v povprečju napisal skoraj 2.300 besed oz. 130 besedil, pri čemer se tudi tu številke močno razlikujejo glede na podkorpus in vir. Izstopajo predvsem avtorji blogov, ki jih je malo, a objavljajo dolga besedila, ter uporabniki omrežja Twitter, ki objavljajo veliko sicer zelo kratkih besedil. Velika nihanja v številu avtorjev in številu besed oz. besedil na komentatorja opazimo pri forumih, kjer posamezni uporabnik na forumu *avtomobilizem* objavi kar 18-krat več besedil kot uporabnik foruma *medovernet*, ki v korpus prispeva tudi najmanj besed, je pa zato teh avtorjev skoraj 50.000, bistveno več kot pri forumih *avtomobilizem* (13.000) ali *kvarkadabra* (2.200). Tako posamezni avtor na forumu *kvarkadabra* objavi bistveno več besed kot ostali forumski uporabniki. Komentatorji spletnih novic so, ne glede na spletni portal, sestavili okoli 21 besedil, se pa zelo razlikuje število komentatorjev – daleč največ jih je na *rtvslo*, skoraj 13.000, na portalu *mladina* nekaj manj kot 1.300, na portalu *reporter* pa samo 240.

Tabela 2: Avtorstvo besedil v korpusu Janes.

(Pod)korpus in vir	Št. uporabnikov	Št. besed na uporabnika	Št. besedil na uporabnika
Tweet	10.239	13.231,7	1.107,2
Forum	64.489	616,7	12,0
avtomobilizem	12.793	1.714,0	44,5
medovernet	49.484	234,8	2,5
kvarkadabra	2.212	2.813,8	36,5
Blog	7.036	4.095,6	57,5
rtvslo.post	243	33.261,8	96,8
rtvslo.comment	3.138	3.701,7	103,4
publishwall.post	615	11.862,2	30,1
publishwall.comment	3.040	599,7	12,4
News	14.430	1.258,0	21,4
rtvslo.comment	12.921	800,8	20,7
mladina.comment	1.273	1.484,9	20,4
reporter.comment	236	1.221,7	22,5
Wiki	2.496	1.619,0	31,6
pagetalk	940	1.324,9	27,6
usertalk	1.556	1.796,7	33,9
Σ	98.693	2.292,6	130,7

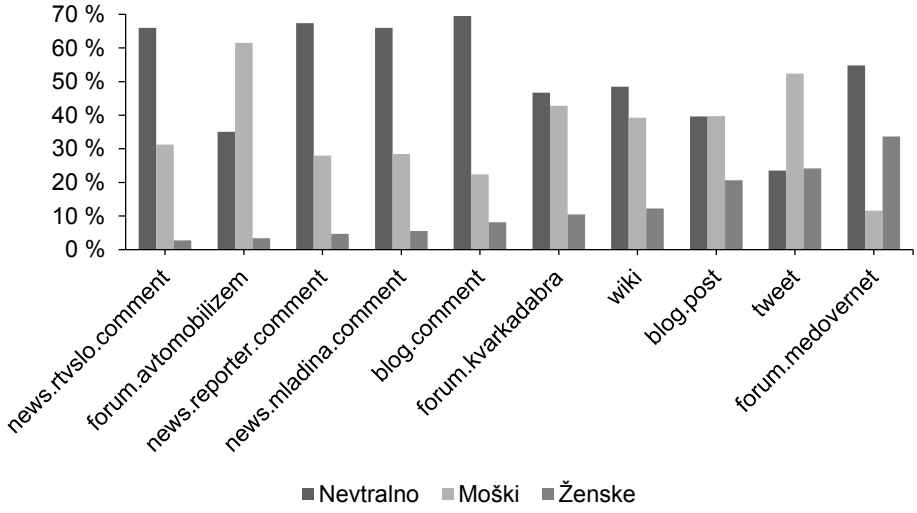
4.3 Spol avtorja

Eden najpomembnejših sociodemografskih podatkov v sociolingvističnih in drugih raziskavah je spol avtorja (Murphy 2010, Baker 2010), ki je v korpusu Janes 1.0 pripisan vsem avtorjem. Spol smo, glede na uporabniško ime, profil uporabnika in vsebino, za avtorje tvitov in blogovskih zapisov določili ročno.

Za vse ostale podkorpuse, vključno s komentarji na bloge, smo spol določili avtomatsko. V slovenščini je spol v glagolskih oblikah v pretekliku in prihodnjiku namreč eksplicitno izražen, kar omogoča njegovo določanje na podlagi prevladujoče oblike v besedilih posameznega avtorja. Za določanje spola avtorjev smo uporabili oblikoskladenjsko označeni korpus, v katerem smo iskali povedi, ki vsebujejo eno od prvoosebnih edninskih oblik pomožnega glagola (*sem, nisem, bom*) in deležnik na *-l* (npr. *mislil* ali *mislila*). V teh stavkih je vsak tak deležnik prispeval 1 točko k indikatorju ustreznega spola. Za vsa besedila nekega avtorja smo potem primerjali število odkritih ženskih in moških indikatorjev: če je bilo razmerje enih do drugih večje od 0,7 in je vsaj 1 % besedil vseboval take indikatorje, smo avtorju

pripisali prevladujoči spol, sicer smo mu pripisali nevtralnega. Ta hevristika je seveda približna, saj bi za natančnejšo opredelitev spola potrebovali skladenjsko razčlenjen korpus, ker lahko samo tako določimo celoten povedek v pretekliku ali prihodnjiku, pa še tu ostaja problem z navedki iz objav drugih uporabnikov.

Natančnost metode smo evalvirali s pomočjo ročno pregledanega seznama oznak za spol za avtorje tvitov. Evalvacija je pokazala, da smo z avtomatskim pristopom pravilni spol ugotovili pri 76 % avtorjev, vendar je bilo napak, kjer je bil moškim pripisan ženski spol in obratno, samo 5 %. Z drugimi besedami, metoda je konservativna in avtorju raje pripiše nevtralni spol, kot da bi se motila pri pripisovanju dejanskega spola.



Slika 2: Spol avtorjev besedil v posameznih podkorpusih.

Slika 2[11] poda razporeditev spolov po podkorpusih in posameznih virih, urejena pa je po naraščajočem deležu ženskih avtorjev. Kot omenjeno je bil za razliko od ostalih podkorpusov spol avtorjev tvitov in blogovskih zapisov pripisan ročno, kar je opazno v tem, da imajo ti podkorpusi manj nevtralnega spola kot ostali (z izjemo foruma *avtomobilizem*), saj avtomatska metoda preferira ta spol na račun moškega in ženskega.

Moških je v vseh virih več kot žensk, razen na forumu *medovernet*, na katerem sodeluje trikrat več žensk kot moških. Poleg že omenjenega foruma *avtomobilizem*, kjer je moških 60 %, jih največ naštejemo še na družbenem omrežju

11 V slikah smo zaradi boljše preglednosti združili nekatere podkategorije iz Tabel 1 in 2, saj med njimi ni bilo večjih razlik po opazovanih kriterijih.

Twitter (53 %) in na forumu *kvarkadabra* (40 %). Najmanj žensk sodeluje v komentarjih na spletne novice in bloge ter na forumu *avtomobilizem* (3 %), največ pa na že omenjenem forumu *medovernet* (34 %), na Twitterju (24 %) in na blogovskih portalih.

4.4 Tip avtorja

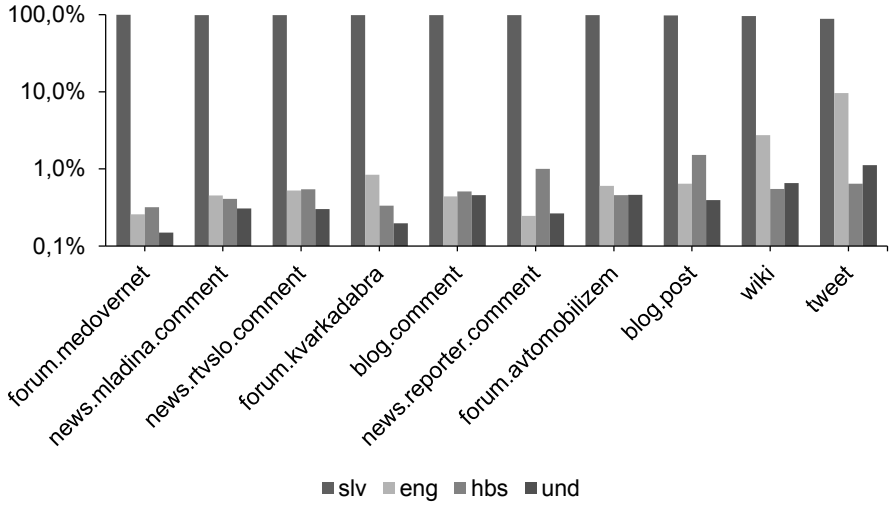
Glede na to, da namen sporočanja močno opredeljuje izbiro jezikovnih sredstev, smo nekatere podkorpuse opremili tudi s podatkom o tipu avtorja, pri čemer ločujemo med osebnimi računi posameznikov, ki uporabniške spletne vsebine objavljajo v svojem imenu kot obliko preživljanja prostega časa, in uradnimi računi medijskih hiš, institucij in podjetij, v imenu katerih spletne vsebine objavljajo za to šolani in plačani predstavniki. Tip avtorja smo označili ročno, pri čemer smo preučili tako profil uporabniškega računa kot zgodovino objav. Ker je število avtorjev za ročni pregled v celotnem korpusu previsoko in ker tip avtorstva v vseh zvrsteh uporabniških spletnih vsebin, ki so zajete v korpusu, niti ni relevanten, smo enako kot za spol tudi tip avtorja pripisali le avtorjem v podkorpusih tvitov in blogov, kjer so poleg individualnih uporabnikov zelo aktivne tudi medijske hiše, javne ustanove in zasebna podjetja. Čeprav smo tudi na forumu *medovernet* poleg individualnih uporabnikov identificirali zdravnike in terapevte, ki uporabnikom odgovarjajo na vprašanja, tipa uporabnikov na forumih nismo določali, ker je to zgolj posebnost podforuma zdravstvene posvetovalnice, ne pa značilnost vseh forumov, vključenih v korpus.

Analiza je pokazala, da 75 % uporabnikov, zajetih v podkorpusu tvitov, tvita v osebnem imenu, medtem ko je korporativnih računov oz. računov javnih ustanov 25 %, pri blogovskih zapisih je samo 51 % osebnih uporabnikov, ostalih 49 % pa je korporativnih. Zanimiva je tudi primerjava tipa uporabnika z njegovim spolom, saj bi pričakovali, da so objave ustanov po spolu avtorja vedno nevtralne. To sicer večinoma drži, ne pa vedno, saj je za 20 % institucionalnih uporabniških računov tvitov spol mogoče določiti: ta je v 15 % moški, v 5 % pa ženski, skoraj identična razmerja (21 % z razmerjem 16 % proti 4 %) pa najdemo tudi pri blogovskih zapisih.

4.5 Jezik besedil

Kljub temu da smo besedila za korpus zbirali iz slovenskih virov oz. pri tvitih slovenskih uporabnikov, je za vse spletne korpuse značilno, da se med besedili najdejo tudi tujejezična. Razlogi za to so raznovrstni, od tega, da v tujem jeziku

pišejo slovenski uporabniki, do tega, da na slovenskih spletnih platformah v svojem jeziku pišejo tuji uporabniki. Da lahko takšna besedila ustrezno izločimo ali se nanje osredotočimo, smo jezik vseh besedil v korpusu Janes avtomatsko označili s programom *langpy*,[12] ki je izšolan za prepoznavanje več sto jezikov, poleg dvočrkovne kode jezika ISO 639-1 pa vrne tudi oceno verjetnosti identificiranega jezika. Rezultati označevanja so uporabni samo pogojno, saj modeli niso najboljši, poleg tega pa so besedila v korpusu Janes velikokrat kratka, napisana nestandardno (npr. brez šumevcev) in vsebujejo mešanico jezikov. Neposredno označevanje korpusa z *langpyem* zato vrne veliko število jezikov (92), od katerih je večina uporabljena malokrat in so tipično tudi napačno identificirani. Rezultate označevanja s programom *langpy* smo zato hevristično popravili tako, da smo vsakemu besedilu pripisali eno od štirih kod ISO 639-2: *slv* (slovenščina), *eng* (angleščina), *hbs* (hrvaščina, srbščina ali bosanščina) ali *und* (nedoločeno).



Slika 3: Zastopanost jezikov po podkorpusih.

Kot vidimo na Sliki 3, kjer je ordinata logaritemska, stolpci pa urejeni po padajočem deležu slovenskih besedil, je velika večina besedil identificiranih kot slovenskih in praktično vse zvrsti oz. viri izkazujejo zanemarljiv tujejezični delež (< 3 %), z izjemo komentarjev na Wikipediji in tvitov. Pri podkorpusu *wiki* je 2,6 % besedil identificiranih kot angleških, medtem ko je pri podkorpusu *tweet* takih besedil kar 9,6 % in 1,1 % nedoločenih, kar je verjetno posledica dejstva, da uporabniki v komentarjih na Wikipediji citirajo angleške članke, na Twitterju pa tvitajo tudi v tujih jezikih, mdr. kadar so tviti namenjeni (tudi) tujejezičnim sledilcem.

12 Dostopno kot del distribucije Pythona.

4.6 Standardnost besedila

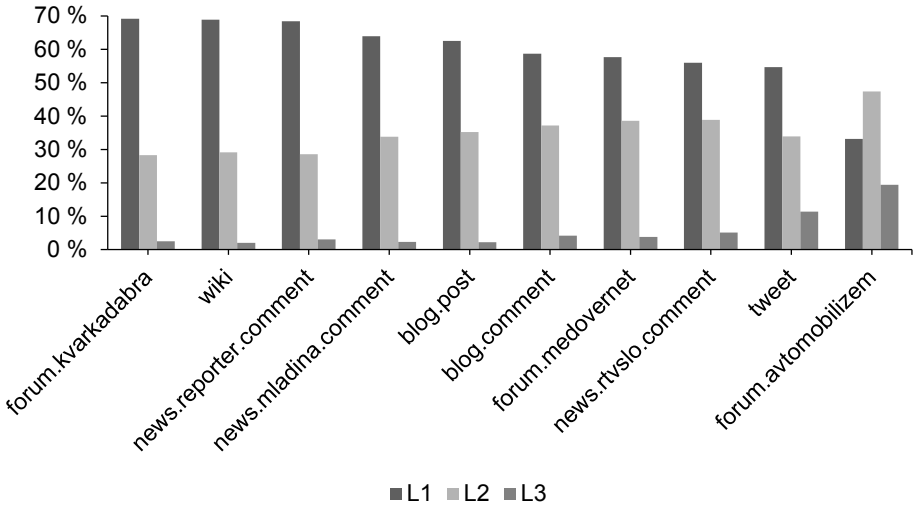
Ker so že prve analize pokazale, da zgrajeni korpus vsebuje številna besedila podjetij (novice, oglasi) in javnih ustanov (obvestila), ki tako po komunikacijskem namenu kot jezikovni podobi v ničemer ne odstopajo od klasičnih besedil na njihovih spletnih straneh, smo se odločili razviti postopek, ki vsakemu besedilu pripiše stopnjo (ne)standardnosti, kar uporabniku korpusa omogoča, da izbere samo besedila, ki ustrezajo tisti stopnji standardnosti, ki ga za konkretno raziskavo zanima. Razvita avtomatska metoda je podrobneje opisana v Ljubešić et al. (2018), na tem mestu pa želimo pripomniti zgolj to, da ločimo tehnično (T) in jezikovno (L) nestandardnost, katerima so pripisane vrednosti od 1 (povsem standardno) do 3 (zelo nestandardno). Tako npr. T1L2 pomeni tehnično povsem standardno, jezikovno pa delno nestandarno besedilo. Z izdelanim orodjem smo določili obe stopnji standardnosti vsem besedilom v korpusu.

Slika 4 podaja podatke o razmerju stopenj jezikovne standardnosti besedil po posameznih podkorpusih in nekaterih bolj zanimivih virih, pri čemer so stolpci urejeni padajoče glede na L1. Gledano v celoti so besedila v korpusu bolj standardna, kot bi morda pričakovali, saj je povsem standardnih več kot polovica besedil v vseh virih, razen v forumu *avtomobilizem*. Poleg njega, kjer je zelo nestandardnih 20 % besedil, po nestandardnosti izstopajo še tviti, ki vsebujejo 12 % besedil stopnje L3, medtem ko je v vseh ostalih virih zelo nestandardnega gradiva zanemarljivo malo, še posebej na forumu *kvarkadabra* in na Wikipediji, kjer je takšnih besedil le okoli 2 %.

4.7 Sentiment besedila

Označevanje sentimenta na področju uporabniško ustvarjenih vsebin postaja vse bolj priljubljeno (Liu 2015). Z analizo sentimenta besedila lahko namreč ugotovimo, ali je javnost neki temi (npr. predsedniškemu kandidatu, predlaganemu zakonu, izdelku) naklonjena ali ne, spremljamo pa lahko tudi trende v sentimentu na določeno temo. Najbolj popularna kategorizacija sentimenta besedila razvršča v negativna, pozitivna in nevtralna, pri čemer se kot nevtralna kategorizira tudi besedila, katerih je sentiment mešan.

Za določanje sentimenta besedilom v celotnem korpusu Janes smo uporabili metodo podpornih vektorjev, naučen pa je bil na večji ročno označeni zbirki raznovrstnih slovenskih tvitov (Smailović et al. 2014), ki žal niso dostopni za neposredno uporabo v našem korpusu.



Slika 4: Jezikoslovna (L) standardnost podkorpusov Janes.

Natančnost smo evalvirali na vzorcu 555 besedil (Fišer et al. 2016a). Vsakemu besedilu v vzorcu je bil pripisan avtomatsko določen sentiment, polega tega pa so ga besedilu ročno pripisali tudi trije anotatorji. Oznake anotatorjev smo primerjali med seboj, avtomatske oznake pa z večinsko oznako anotatorjev. Za izračun ujemanja smo uporabili koeficient alfa po Krippendorffu (2012), pri katerem rezultat 1 pomeni popolno, 0 pa naključno ujemanje. Za naloge, kot je bila naša, velja, da je ujemanje sprejemljivo, kadar je koeficient alfa vsaj 0,4. (Mozetič et al. 2016). Rezultati so pokazali, da je določanje sentimenta precej subjektivna naloga in težak problem za računalnike. Rezultati ročnega ujemanja so pod 0,6, kar je sicer sprejemljivo, a daleč od popolnega ujemanja. Avtomatsko pripisovanje sentimenta je bilo pričakovano slabše od ujemanja med označevalci. Čeprav je bil skupni rezultat nad pragom sprejemljivosti 0,4, ta za tri od petih tipov besedil ni bil dosežen. Tu je potrebno dodati, da je bila evalvacija avtomatskega pripisovanja sentimenta precej stroga, saj smo ga primerjali z večinskimi odgovori označevalcev tudi, kadar se anotatorji med seboj niso strinjali. S tem smo sistem kaznovali tudi, kadar se je morda ujemal z enim od označevalcev.

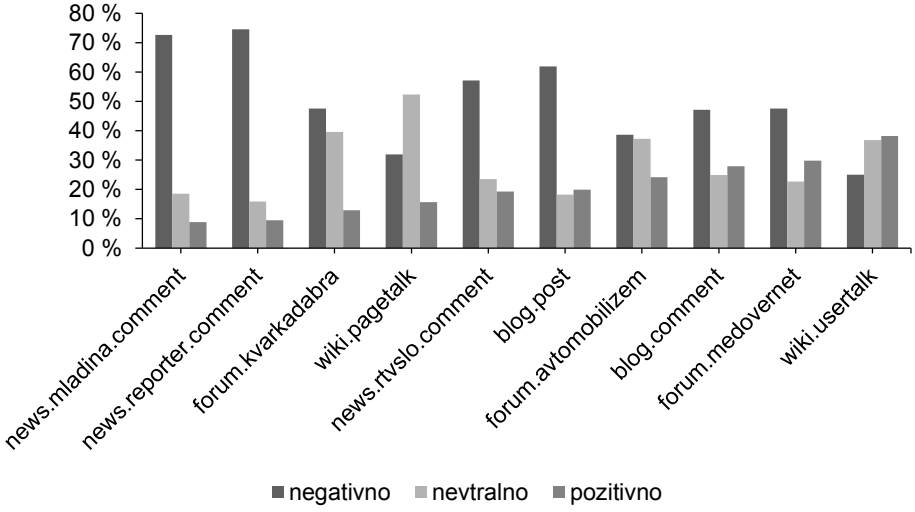
Ob zavedanju, da avtomatska kategorizacija ni zelo zanesljiva, Slika 5 vizualizira razporeditev sentimenta v posameznih virih v korpusu, ki so urejeni naraščajoče glede na pozitivni sentiment. V večini virov (komentarji na novice, forumi in blogi) prevladuje negativni sentiment, najizraziteje na portalih *reporter* in *mladina*, kjer je negativnih kar tri četrt komentarjev. Nevtralni sentiment prevladuje na pogovornih straneh Wikipedije in v tvitih, ki vsebujeta približno polovico nevtralnih vsebin, kar prav tako ustreza glavnemu namenu komuniciranja v teh medijih. Pozitivni

sentiment prevladuje edinole na uporabniških straneh Wikipedije, ki avtorjem predstavlja kanal za pohvale, voščila in druge skupnostno-povezovalne dejavnosti.

5 ZAPIS KORPUSA

Korpus Janes je zapisan v jeziku XML, ki omogoča strukturiranje korpusa, zapis metapodatkov in jezikoslovnih oznak ter strojno preverljivost pravilnosti zapisa. Do različice 0.4 je bil vsak podkorpus kodiran po lastni shemi XML, ki je čim bolje izražala strukturo podkorpusa in njegovih metapodatkov. Za Janes 1.0 smo podkorpuse zapisali v enotnem formatu Iniciative za kodiranje besedil TEI (TEI 2016).

Vsak od petih podkorpusov je zapisan kot svoj dokument TEI, ki je sestavljen iz kolofona TEI in telesa korpusa. Kolofon vsebuje metapodatke o korpusu, kot so naslov, avtorje, dostopnost, opis virov, uporabljene taksonomije in standardizirane vrednosti ter število in opis uporabljenih elementov XML v besedilih podkorpusa.



Slika 5: Sentiment podkorpusov.

5.1 Strukture podkorpusov

Telo podkorpusa vsebuje besedila, ki pa so, odvisno od njegove zvrsti, organizirana v več hierarhičnih razdelkov, tj. TEI-elementov *div*, ki so kvalificirani z atributom *@type*. Vsak razdelek se začne s strukturo lastnosti (*fs*), ki vsebuje metapodatke o

razdelku, kodirane kot lastnosti (*f*). Slika 6 ilustrira to strukturo na primeru začetka podkorpusa Janes-Forum, kjer imamo gnezdenje treh nivojev razdelkov, vsak s svojimi metapodatki, na spodnjem nivoju je najprej naslov objave, nato pa se začne prvi odstavek in besedilo, kar obravnavamo v naslednjem razdelku.

```
<text xml:id="janes.forum.text" xml:lang="slv">
  <body>
    <div type="platform" xml:id="janes.forum.medovernet">
      <fs>
        <f name="platform">medovernet</f>
      </fs>
      <div type="thread" xml:id="janes.forum.medovernet.416.9676700">
        <fs>
          <f name="path">Zdravstvene posvetovalnice &gt; Dermatologija &gt;
            Kožna znamenja, luskavica in druge težave s kožo</f>
          <f name="url">http://med.over.net/forum5/read.php?416,9676700</f>
        </fs>
        <div type="post" xml:id="janes.forum.medovernet.9676700">
          <fs>
            <f name="time">2014-05-30T10:22:00</f>
            <f name="url">http://med.over.net/forum5/read.php?416,9676700,9676700#m
            sg-9676700</f>
            <f name="user">katica1</f><f name="lang">slv</f>
            <f name="std_tech">T1</f><f name="std_tech_n">1.1</f>
            <f name="std_ling">L2</f><f name="std_ling_n">1.6</f>
            <f name="sex">neutral</f><f name="source">private</f> <f name="sentiment">neutral</f>
          </fs>
        <head>Znamenje na nosu odstranitev</head>
        <p xml:id="janes.forum.medovernet.9676700.1">
        …
```

Slika 6: Primer TEI-strukture podkorpusa.

Kot rečeno se strukture posameznih podkorpusov medsebojno razlikujejo, saj so odvisne od lastnosti platforme. Imena podkorpusov in njihova notranja struktura je sledeča:

- **Janes-Tweet**: *div[@type='tweet']* (posamezen tvit)
- **Janes-Forum**: *div[@type='platform']* (vir); *div[@type='thread']* (nit); *div[@type='post']* (objava)
- **Janes-News**: *div[@type='platform']* (vir); *div[@type='text']* (besedilo objave); *div[@type='article' | @type='comment']* (novica, ki ji sledijo komentarji)
- **Janes-Blog**: *div[@type='platform']* (vir); *div[@type='blog']* (besedilo objave); *div[@type='post' | @type='comment']* (blogovski zapis, ki mu sledijo komentarji)

- **Janes-Wiki**: *div[@type='platform']* (vir); *div[@type='page']* (spletna stran, na katero se nanašajo komentarji); *div[@type='topic']* (tema pogovora); *div[@type='comment']* (posamezen komentar)

```
<p>
  <s>
    <name type="per">
      <w lemma="@73cesar" ana="#Xa">@73cesar</w>
    </name><c> </c>
    <choice>
      <orig><w>Dej</w></orig>
      <reg><w lemma="dati" ana="#Vmem2s">daj</w></reg>
    </choice><c> </c>
    <w lemma="ne" ana="#Q">ne</w><c> </c>
    <w lemma="rtjati" ana="#Vmpm2s">RTjaj</w><c> </c>
    <w lemma="z" ana="#Si">z</w><c> </c>
    <name type="per">
      <w lemma="@Delo_Ozadja" ana="#Xa">@Delo_Ozadja</w>
    </name>
    <pc ana="#Z">,</pc><c> </c>
    <w lemma="fejker" ana="#Cs">fejker</w>
  </s>
</p>
```

Slika 7: Jezikoslovno označen tvit »@73cesar Dej ne RTjaj z @Delo_Ozadja, fejker«.

Omenimo še, da s predstavljeno shemo opisa strukture in metapodatkov v TEI (uporaba *div/fs*) odstopamo od predloga Beißwenger et al. (2012), ki so osnovna priporočila TEI nadgradili z vrsto posebnih elementov, namenjenih prav opisu računalniško posredovane komunikacije. Pri tem predlogu je namreč problematična velika parametrizacija TEI, ki jo je težko vzdrževati oz. poskrbeti za skladnost z drugimi pretvorbami TEI, npr. za navpični format, ki ga potrebuje konkordančnik.

5.2 Oznake v besedilu

Znotraj razdelkov najnižjega nivoja so odstavki (element *p*), ki vsebujejo jezikoslovno označeno besedilo (Slika 7). Postopek jezikoslovnega označevanja je zajemal naslednje korake:

1. **tokenizacija** in **stavčna segmentacija**: elementi *w* (beseda), *pc* (ločilo), *c* (presledek) in element *s* (poved)
2. **normalizacija** (kjer je potrebna): elementi *choice* (izbira med izvorno / normalizirano obliko), *orig* (izvorna oblika), *reg* (normalizirana oblika)
3. **oblikoskladenjsko označevanje** in **lematizacija**: atributi *w/@ana* oz. *pc/@ana* (kazalec na definicijo oblikoskladenjske oznake), *w/@lemma* (osnovna oblika besede)
4. **določanje imenskih entitet**: element *name*, pri čemer atribut *@type* poda vrsto imena, vrednosti so: *per* (oseba, npr. »@ZigaTurk«), *deriv-per* (ime, izpeljano iz osebe, npr. »Sizifovo«), *loc* (lokacija, npr. »Slovenija«), *org* (organizacija, npr. »TV Pink«), *misc* (drugo, npr. »Brothers Empire«).

Orodja, s katerimi je bil korpus označen, in ocena njihove točnosti so podrobno opisani v Ljubešić et al. (2018).

6 JAVNA RAZLIČICA KORPUSA

Evropska listina za raziskovalce – Kodeks ravnanja pri zaposlovanju raziskovalcev (Evropska komisija 2006: 13) v zvezi s širjenjem in izkoriščanjem rezultatov navajata, da morajo vsi raziskovalci zagotoviti, da bodo rezultate raziskav širili v druga raziskovalna okolja, rezultati pa naj bodo tržno izkoriščeni in/ali dostopni javnosti, kadarkoli se za to pojavi priložnost.

Tudi načrt projekta JANES je predvideval distribucijo zgrajenih korpusov, saj so korpusi podlaga za sodobno slovaropisje, empirično jezikoslovje in razvoj jezikovnih tehnologij. Vendar se pri njihovi distribuciji pojavljajo problemi in omejitve, in sicer pogoji uporabe spletnih portalov, varovanje osebnih podatkov, vključno s pravico do pozabe, in v manjši meri tudi avtorske pravice nad izvornimi besedili. Te ovire smo podrobno obdelali v Erjavec et al. (2016), kjer smo tudi predlagali načine, da te omejitve lahko presežemo in ki smo se jih v veliki meri držali tudi pri zagotavljanju odprtega in prostega dostopa korpusa Janes in njegovih podkorpusov.

Dostop do korpusov smo zagotovili na dva načina. Za zagotavljanje dostopnosti jeziko(slo)vnih podatkov za humanistične in družboslovne raziskave, s tem pa spodbujanje večkratne uporabe jezikovnih podatkov, je bil v Sloveniji in za slovenščino ustanovljen konzorcij CLARIN.SI (Erjavec et al. 2014).[13] Infrastruktura CLARIN.SI vzdržuje repozitorij, ki omogoča hranjenje jezikovnih virov in je certificiran repozitorij s strani DSA (Data Seal of Approval) in evropskega CLARIN-a. Podkorpuse Janes

13 http://www.clarin.si

smo vnesli v repozitorij CLARIN.SI in na ta način omogočili njihovo trajno in stabilno hrambo ter enostaven prenos in najdljivost. Dodatno smo dostop do korpusov omogočili tudi skozi lastno instalacijo spletnega konkordančnika noSketch Engine (Rychlý 2007), s čimer smo jih naredili uporabne tudi za jezikoslovce.

Tabela 3 podaja podatke, vezane na dostopnost korpusa Janes in njegovih podkorpusov ter njihovih virov. Za **Janes-Tweet** nimamo dovoljenja lastnika platforme za nadaljnje razširjanje podatkov, kar Twitter s standardno licenco celo izrecno prepoveduje. Zato smo omogočili prevzem (pod licenco CC BY-NC) prek CLARIN.SI (Ljubešić et al., 2017a) tako, kot je stalna praksa pri večini raziskovalcev, ki želijo redistribuirati tvite, namreč, da v korpusu ni besedila tvitov, temveč samo njihove identifikacijske številke, del distribucije pa je program, ki prek Twitter API-ja omogoči ponovni zajem vsebovanih tvitov. Prednost tega pristopa je, da z njim ne kršimo pogojev uporabe Twitterja, slabost pa, da ponovno ustvarjeni korpus ne vsebuje nujno vseh izvornih tvitov, če so bili ti zbrisani s strani avtorjev ali pa je bil zbrisan uporabniški račun, in s tem tudi vsi njegovi tviti, s čimer je oteženo reproduciranje in primerljivost eksperimentalnih rezultatov. Dodaten zaplet pri programu povzroča dejstvo, da naš korpus vsebuje tudi normalizirane oblike pojavnic in njihove leme – če bi bili ti podatki neposredno dostopni, bi s tem že dobili dober približek izvornega tvita, kar ni dovoljeno. Zato korpus vsebuje v normaliziranih oblikah in lemah zgolj razlike glede na izvorne pojavnice, in šele s pomočjo ponovno zajetega tvita generira tudi normalizirane oblike in leme. Zato pri tem korpusu ni potrebe po anonimizaciji uporabniških imen oz. lastnih imen ter imen organizacij. Korpus je tudi dostopen v sklopu konkordančnika, kjer pa so odstranjena imena uporabnikov, URL-ji in lastna imena.

Za podkorpus **Janes-Forum** smo pridobili dovoljenja lastnikov portalov za vse tri vire za nadaljnje razširjanje njihovih podatkov pod pogojem, da so iz korpusa odstranjena uporabniška imena, osebna imena v besedilih kot tudi imena organizacij, kar smo tudi storili in s tem zavarovali zasebnost avtorjev in omenjenih oseb. Korpus je v tako anonimizirani obliki dostopen v repozitoriju CLARIN.SI pod licenco CC BY (Erjavec et al. 2017b), ravno tako pa je javno dostopen (tudi v anonimizirani obliki) prek konkordančnika.

Za podkorpus **Janes-Blog** smo dobili od RTV Slovenija ustno zagotovilo, da smemo redistribuirati njihove vsebine, žal pa nam tega dovoljenja ni uspelo dobiti od lastnikov portala *publishwall*. Menimo, da javni interes v tem primeru prevlada nad pomanjkanjem dovoljenja, zato smo tudi ta korpus anonimizirali (vendar ne imen organizacij, saj RTV Slovenija ni postavil tega pogoja) in ga tudi ponudili v prevzem v CLARIN.SI pod licenco CC BY (Erjavec et al. 2017a).

Za **Janes-News** smo pridobili pisna dovoljenja Mladine in Reporterja ter, kot rečeno, ustno dovoljenje RTV Slovenija, vendar v zadnjem primeru samo za

komentarje na novice, ne pa za novice. Zaradi uniformnosti podkorpusa, kot tudi zaradi dejstva, da novice niso uporabniško generirane vsebine in so bile v korpus vključene samo zaradi kontekstualizacije komentarjev, smo novice odstranili tudi iz ostalih dveh virov, poleg tega pa smo korpus anonimizirali in takega ponudili v prevzem v CLARIN.SI pod licenco CC BY (Erjavec et al. 2017c) ter prek konkordančnika.

Podkorpus **Janes-Wiki** je najmanj problematičen, saj je prenesen z Wikipedije, ki ima licenco CC BY-SA, zato tega korpusa ni bilo potrebno anonimizirati, v repozitoriju CLARIN.SI je dostopen pod enako licenco (Ljubešić et al., 2017b), dostop pa je omogočen še prek konkordančnika.

Celoten korpus **Janes** ni na voljo za prevzem, saj bi moral upoštevati vse omejitve posameznih podkorpusov in zato uporabniki laže prevzamejo posamezne korpuse in jih, po želji, sestavijo. Zato pa je v celoti dostopen prek konkordančnika, vendar v maksimalno anonimizirani obliki.

Tabela 3: Dostopnost in anonimizacija javne različice podkorpusov Janes.

(Pod)korpus in vir	Dov.	Prevzem + licenca	Anonim. uporab.	Anonim. os. im.	Anonim. organ.	Konkordančnik
Tweet	NE	CC BY-NC (+ API)	NE	NE	NE	DA
Forum		CC BY	DA	DA	DA	DA
avtomobilizem	DA	DA	DA	DA	DA	DA
medovernet	DA	DA	DA	DA	DA	DA
kvarkadabra	DA	DA	DA	DA	DA	DA
Blog		CC BY	DA	DA	NE	DA
rtvslo.post	DA	DA	DA	DA	NE	DA
rtvslo.comment	DA	DA	DA	DA	NE	DA
publishwall.post	NE	DA	DA	DA	NE	DA
publishwall.comment	NE	DA	DA	DA	NE	DA
News		CC BY	DA	DA	NE	DA
rtvslo.article	NE	NE	-	-	-	NE
rtvslo.comment	DA	DA	DA	DA	NE	DA
mladina.article	DA	NE	-	-	-	NE
mladina.comment	DA	DA	DA	DA	NE	DA
reporter.article	DA	NE	-	-	-	NE
reporter.comment	DA	DA	DA	DA	NE	DA
Wiki		CC BY-SA	NE	NE	NE	DA
pagetalk	DA	DA	NE	NE	NE	DA
usertalk	DA	DA	NE	NE	NE	DA
Janes	NE	NE	DA	DA	DA	DA

7 SKLEP

V poglavju smo predstavili gradnjo, opremljanje z metapodatki, zapis in distribucijo prvega velikega korpusa slovenskih spletnih uporabniških vsebin Janes v1.0 ter podali statistike po korpusnih (meta)podatkih. V primerjavi s tipičnimi spletnimi korpusi se predstavljeni razlikuje po tem, da smo vložili veliko napora v ohranitev strukture izvornih virov in zajemu čim več (sociodemografskih) metapodatkov, ki omogočajo številne sociolingvistične, družboslovne in jezikovnotehnološke raziskave. Posebej smo se posvetili tudi vidiku nestandardnosti jezika v korpusih, kjer smo pred oblikoskladenjskim označevanjem in lematizacijo besedila tokenizirali s posebej za nestandardni jezik prilagojenim tokenizatorjem, zapis besed standardizirali, besedilom v korpusih pa smo dodali tudi oznako za stopnjo standardnosti na tehnični in jezikovni ravni.

Korpus oz. njegovi podkorpusi so dostopni za prevzem pod licencami Creative Commons v repozitoriju raziskovalne infrastrukture CLARIN.SI, ravno tako pa so na voljo tudi prek konkordančnika. Zavedamo se, da z izdelavo velikih javno in odprto dostopnih korpusov trčimo ob številne zakonske omejitve, povezane z avtorskimi pravicami in varovanjem osebnih podatkov. V okviru projekta smo se trudili omiliti, če že ne odpraviti, nesmiselne zadržke do čim večje dostopnosti podatkov o slovenskem jeziku družbenih omrežij.

Korpus Janes je že bil uporabljen v številnih raziskavah s področja korpusnega in računalniškega jezikoslovja. Poleg raziskav, predstavljenih v tej monografiji, je bil korpus uporabljen tudi v prispevkih v 33 mednarodnih in 34 domačih znanstvenih revijah, poglavjih v monografijah, konferenčnih zbornikih ter strokovnih publikacijah.14 Na korpusu temeljita 2 magistrski nalogi, korpus pa je bil tudi povod za 3 poletne šole za srednješolce in študente.

Pri nadaljnjem razvoju korpusa načrtujemo izboljšati kvaliteto jezikoslovnega označevanja (glej Ljubešić et al. (2018)), prav tako pa tudi velikost in raznovrstnost korpusa, pri čemer se nameravamo osredotočiti predvsem na katere od doslej še nepokritih družbenih platform, kot je na primer Facebook, z zavedanjem, da bo tu (še) težje zagotoviti možnost javne redistribucije korpusa. Predvsem pa bi si želeli, da bi korpus Janes 1.0 za proučevanje in poučevanje uporabljal čim širši krog slovenistov in drugih jezikoslovcev pa tudi družboslovcev (novinarjev, politologov, sociologov), saj je bil to tudi naš glavni cilj pri prizadevanjih za zagotovitev javno in odprto dostopnega korpusa.

Zahvala

Avtorji se za dovoljenje za objavo besedil v korpusu zahvaljujejo uredništvom Avtomobilizem.net, Kvarkadabra, MedOverNet, Mladina, Reporter in RTV Slovenija. Prav tako se zahvaljujemo Jaki Čibeju, Teji Goli, Dafne Marko, Eneji Osrajnik, Senji Pollak in Izi Škrjanec za ročno pripisovanje metapodatkov v korpusu.

Literatura

Baker, Paul, 2010: *Sociolinguistics and Corpus Linguistics*. Edinburg: Edinburgh University Press.

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

Beißwenger, Michael, 2013: Raumorientierung in der Netzkommunikation. Korpusgestützte Untersuchungen zur lokalen Deixis in Chats. Frank-Job, Barbara, Alexander Mehler in Tilmann Sutter (ur.): *Die Dynamik sozialer und sprachlicher Netzwerke: Konzepte, Methoden und empirische Untersuchungen an Beispielen des WWW*. 207–258. Springer.

Beißwenger, Michael, Eric Ehrhardt, Andrea Horbach, Harald Lüngen, Diana Steffen in Angelika Storrer, 2015: Adding value to CMC corpora: CLARINification and part-of-speech annotation of the Dortmund Chat Corpus. *Proceedings of the 2nd Workshop on Natural Language Processing for Computer-Mediated Communication/Social Media (NLP4CMC2015)*. 12–16.

Beißwenger, Michael, Maria Ermakova, Alexander Geyken, Lothar Lemnitzer in Angelika Storrer, 2012: A TEI Schema for the Representation of Computer-mediated Communication. *Journal of the Text Encoding Initiative* 3/2012.

Beißwenger, Michael in Angelika Storrer, 2008: Corpora of Computer-Mediated Communication. Lüdeling, Anke in Merja Kytö (ur.): *Corpus Linguistics. An International Handbook*. Berlin: de Gruyter. 292–308.

Bučar, Jože, Janez Povh in Martin Žnidaršič, 2015: Sentiment classification of the Slovenian news texts. *Proceedings of the 9th International Conference on Computer Recognition Systems CORES 2015* (Advances in intelligent systems and computing, Vol. 403). Cham: Springer. 777–787. doi: 10.1007/978-3-319-26227-7_73

Bučar, Jože, 2017: *Automatically sentiment annotated Slovenian news corpus AutoSentiNews 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1109

Chanier, Thierry, Céline Poudat, Benoit Sagot, Georges Antoniadis, Ciara R. Wigham, Linda Hriba, Julien Longhi in Djamé Seddah, 2014: The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. *JLCL-Journal for Language Technology and Computational Linguistics* 29/2. 1–30.

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

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

Dürscheid, Christa in Elisabeth Stark, 2011: SMS4science: An international corpus-based texting project and the specific challenges for multilingual Switzerland. Thurlow, Crispin in Kristine Mroczek (ur.): *Digital Discourse. Language in the New Media*. Oxford: Oxford University Press. 299–320.

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

Erjavec, Tomaž, Jaka Čibej in Darja Fišer, 2016b: Omogočanje dostopa do korpusov slovenskih spletnih besedil v luči pravnih omejitev. *Slovenščina 2.0* 4/2. 189 –219.

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

Erjavec, Tomaž, Jan Jona Javoršek in Simon Krek, 2014: Raziskovalna infrastruktura CLARIN.SI. *Zbornik Devete konference Jezikovne tehnologije*. Ljubljana: Institut »Jožef Stefan«. 19–24.

Erjavec, Tomaž, Nikola Ljubešić in Nataša Logar, 2015: The slWaC corpus of the Slovene Web. *Informatica* 39/1. 35.

Erjavec, Tomaž, Nikola Ljubešić in Darja Fišer, 2017a: *Blog post and comment corpus Janes-Blog 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1138

Erjavec, Tomaž, Nikola Ljubešić in Darja Fišer, 2017b: *Forum corpus Janes-Forum 1.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1139

Erjavec, Tomaž, Nikola Ljubešić in Darja Fišer, 2017c: *News comment corpus Janes-News 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1140

Evropska komisija, 2006: *Evropska listina za raziskovalce. Kodeks ravnanja pri zaposlovanju raziskovalcev*. http://ec.europa.eu/euraxess/pdf/brochure_rights/kina21620b7c_si.pdf

Fišer, Darja in Tomaž Erjavec, 2016a: Analysis of sentiment labelling of Slovene user generated content. *Proceedings of the 4th conference on CMC and Social Media Corpora for the Humanities*, 27.-28.9. 2016. Ljubljana: Filozofska fakulteta.

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

Fišer, Darja, Jasmina Smailović, Tomaž Erjavec, Igor Mozetič in Miha Grčar 2016b: Sentiment Annotation of the Janes Corpus of Slovene User-Generated Content. *Proceedings of the 10th Language Technologies and Digital Humanities Conference*, 29.9.-1.10. 2016. Ljubljana: Filozofska fakulteta.

Frey, Jennifer-Carmen, Aivars Glaznieks in Egon Stemle, 2015: The DiDi Corpus of South Tyrolean CMC Data. *Proceedings of the 2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media*. GSCL2015 (NLP4CMC2015). 1–6.

Kadunc, Klemenc in Marko Robnik Šikonja, 2016: Analiza mnenj s pomočjo strojnega učenja in slovenskega leksikona sentimenta. Erjavec, Tomaž in Darja Fišer (ur.): *Zbornik konference Jezikovne tehnologije in digitalna humanistika 2016*. http://www.sdjt.si/wp/dogodki/konference/jtdh-2016/zbornik/

Kadunc, Klemenc in Marko Robnik Šikonja, 2017: *Opinion corpus of Slovene web commentaries KKS 1.001*, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1115

Krippendorff, Klaus, 2012: *Content Analysis: An Introduction to its Methodology*. Los Angeles, London, New Delhi, Singapur, Washington: Sage Publications.

Lagus, Krista, Mika Pantzar, Minna Ruckenstein in Marjoriikka Ylisiurua, 2016: *Suomi24 – muodonantoa aineistolle*. Technical report. Helsinki: Unigrafia. http://blogs.helsinki.fi/citizenmindscapes/files/2016/05/257383_HY_VALT_suomi24_muodonantoa_aineistolle.pdf

Lebar, Lea, Andraž Petrovčič in Gregor Petrič, 2012: *Analiza slovenskih spletnih forumov*. Poročilo. http://www.nebojse.si/portal/Dokumenti/Analiza_slovenskih_spletnih_forumov.pdf

Liu, Bing, 2015: *Sentiment analysis. Mining opinions, sentiments, and emotions*. Cambridge: Cambridge University Press.

Ljubešić, Nikola, Darja Fišer in Tomaž Erjavec, 2014: TweetCaT: a tool for building Twitter corpora of smaller languages. *Proceedings of LREC'14 Conference*. Reykjavik, Iceland. 2279–2283.

Ljubešić, Nikola, Tomaž Erjavec in Darja Fišer, 2017a: *Twitter corpus Janes-Tweet 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1142

Ljubešić, Nikola, Tomaž Erjavec in Darja Fišer, 2017b: *Wikipedia talk corpus Janes-Wiki 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1137

Ljubešić, Nikola, Tomaž Erjavec in Darja Fišer, 2018: Orodja za procesiranje nestandardne slovenščine. Fišer, Darja (ur.): *Viri, orodja in metode za analizo spletne slovenščine*. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani. 74–99.

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

Margaretha, Eliza in Harald Lüngen, 2014: Building Linguistic Corpora from Wikipedia Articles and Discussions. *JLCL* 29/2. 59–82.

Michelizza, Mija, 2015: *Spletna besedila in jezik na spletu. Primer blogov in Wikipedije v slovenščini*. Ljubljana: Založba ZRC.

Mozetič, Igor, Miha Grčar, Jasmina Smailović, 2016: Multilingual Twitter sentiment classification: The role of human annotators. *PLoS ONE* 11/5. e0155036.

Murphy, Bróna, 2010: *Corpus and sociolinguistics: Investigating age and gender in female talk* (Vol. 38). Amsterdam, Philadelphia: John Benjamins Publishing.

Rychlý, Pavel, 2007: Manatee/Bonito - A Modular Corpus Manager. *Proceedings of the Workshop on Recent Advances in Slavonic Natural Language Processing*. Brno: Masaryk University. 65–70.

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

Statistični urad Republike Slovenije, 2015: *Uporaba interneta v gospodinjstvih in pri posameznikih v Sloveniji*. http://www.stat.si/StatWeb/prikazi-novico?id=5509&idp=10&headerbar=8

TEI Consortium, 2016: *Guidelines for Electronic Text Encoding and Interchange.* http://www.tei-c.org/P5/

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

Ročno označeni korpusi Janes za učenje jezikovnotehnoloških orodij in jezikoslovne raziskave

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

Izvleček

V tem poglavju najprej predstavimo splošni postopek in delotok izdelave ročno označenih korpusov (od priprave podatkov, izdelovanja smernic za označevanje, dela z označevalno platformo in poteka označevalne kampanje do pretvorbe v končni format ter objave in distribucije), pri čemer se podrobneje posvetimo največjima tako nastalima korpusoma Janes-Norm (približno 185.000 pojavnic) in Janes-Tag (približno 75.000 pojavnic), katerih glavni namen je izboljšava jezikovnotehnoloških orodij za tokenizacijo, stavčno segmentacijo, normalizacijo, lematizacijo in oblikoskladenjsko označevanje. Drugi del poglavja poda pregled vseh ročno označenih korpusov Janes: poleg že omenjenih Janes-Norm in Janes-Tag še Janes-Syn (skladnja v RPK), Janes-Kratko (pojavi krajšanja v RPK), Janes-Vejica (raba vejice v RPK), Janes-Preklop (preklapljanje koda v RPK) in Janes-Geo (raba nestandardnih jezikovnih prvin v RPK v odvisnosti od regionalnega izvora uporabnikov). V njem na kratko predstavimo vsebino in strukturo vsakega korpusa ter opišemo njegov predvideni namen.

Ključne besede: slovenščina, računalniško posredovana komunikacija, lematizacija, normalizacija, oblikoskladenjsko označevanje, odprti podatki, Text Encoding Initiative, CLARIN.SI

1 UVOD

Ročno označene podatkovne množice so v okviru proučevanja računalniško posredovane komunikacije (RPK) in razvoja jezikovnotehnoloških orodij za računalniško obdelavo naravnega jezika ključnega pomena. Z vidika jezikoslovnih raziskav omogočajo sistematičen in natančen kvantitativni in kvalitativni vpogled v proučevane pojave, z vidika jezikovnih tehnologij (kot je to podrobneje opisano v Ljubešić et al. (2018)) pa služijo kot podlaga za učenje in izboljšavo orodij, ki olajšajo nadaljnje analize na večjih korpusih, avtomatizirajo označevalne postopke (npr. prepoznavanje imenskih entitet in pripisovanje metapodatkov besedilom ali uporabnikom) in izboljšajo natančnost jezikoslovnega označevanja besedil (npr. normalizacija, lematizacija in oblikoskladenjsko označevanje).

Čeprav je slovenska RPK (kot to velja za druge jezike) zbir različnih komunikacijskih praks oz. besedilnih žanrov, je mogoče v njej v splošnem opaziti jezikovne značilnosti, ki se razlikujejo od standardne pisne slovenščine, kakršna je zbrana in reprezentirana tudi v referenčnih korpusih za slovenščino. Orodja za jeziko(slo)vno označevanje, ki so bila razvita na osnovi standardnega gradiva, zato pri jeziku RPK izkazujejo slabšo natančnost. Težave se pojavljajo na vseh označevalnih ravneh, od tokenizacije, stavčne segmentacije, oblikoskladenjskega označevanja in lematizacije do višjih označevalnih ravni, kot je npr. skladnja. Dodaten izziv za optimizacijo označevalnih postopkov predstavlja dejstvo, da številne značilnosti nestandardne slovenščine tudi v jezikoslovnem smislu še niso dobro raziskane, ne same na sebi ne v razmerju do standardnega jezika. Rešitev, ki jo ponuja projekt JANES, je zato razvoj ročno označenih korpusov ciljno za namene učenja jezikovnotehnoloških orodij in proučevanja jezikovnih pojavov v slovenski RPK. S tem se slovenščina pridružuje jezikom, pri katerih so bile potrebe po prilagoditvi označevalnih orodij že identificirane in rešitve uspešno uporabljene v praksi. Med sorodnimi raziskavami za tuje jezike velja npr. izpostaviti izboljšanje oblikoskladenjskega označevanja nemških tvitov (Rehbein et al. 2013) ter avtomatske normalizacije nemških družbenomedijskih besedil (Laarmann-Quante in Dipper 2016, Ueberwasser 2013), ročno označeni korpusi pa so bili uporabljeni tudi za izboljšanje razreševanja anaforičnih sklicev v angleških besedilih (Poesio et al. 2017) in prepoznavanja imenskih entitet (Bontcheva et al. 2017, Benikova et al. 2014).

V poglavju predstavljamo ročno označene korpuse, ki so bili izdelani v okviru projekta. Vsa besedila, ki so vključena vanje, so bila vzorčena iz korpusa Janes, ki je podrobneje predstavljen v Erjavec et al. (2018). V 2. razdelku opišemo splošni postopek različnih stopenj izdelave korpusov, posebno pozornost pa posvetimo največjima ročno označenima korpusoma Janes-Norm in Janes-Tag (Erjavec et al. 2016c), ki služita kot ponazoritvena primera. Temu sledita še kratek pregled

poglavitnih značilnosti vseh tako nastalih korpusov (3. razdelek) ter sklep, v katerem nakažemo možnosti za izboljšave in prihodnje delo.

2 Izdelava ročno označenih korpusov

V tem razdelku predstavljamo postopek izdelave ročno označenih korpusov od priprave podatkov in označevanja do končne pretvorbe v format TEI (*Text Encoding Initiative*).[1] Postopek se je od korpusa do korpusa nekoliko razlikoval glede na kompleksnost problema in število vpletenih označevalcev/razsodnikov, a je v vseh primerih sledil naslednjim stopnjam:

1. priprava podatkov,
2. izdelava tipologije in smernic za označevanje,
3. označevanje oz. označevalna kampanja z razsojanjem,
4. končni izvoz in pretvorba podatkov.

V naslednjih podrazdelkih ta postopek podrobneje predstavimo na primeru označevalne kampanje, katere cilj je bila izdelava ročno označenega korpusa za izboljšanje avtomatskega označevanja slovenske RPK na petih ravneh: tokenizacija, stavčna segmentacija in normalizacija (korpus Janes-Norm) ter dodatno lematizacija in oblikoskladnja (korpus Janes-Tag).

2.1 Priprava podatkov

Besedila, ki sestavljajo ročno označene korpuse, predstavljene v tem prispevku, so bila vzorčena iz korpusa Janes. V prvi fazi izdelave korpusov Janes-Norm in Janes-Tag sta bila izdelana dva vzorca:

1. Kons1, ki ga sestavljajo tviti, in
2. Kons2, ki sestoji iz forumskih sporočil in komentarjev na blogovske zapise ter novice.

Kons1 vsebuje 4.000 tvitov, ki so bili vzorčeni naključno, a z upoštevanjem določenih omejitev: nismo upoštevali tvitov, ki so bili daljši od 120 znakov,[2] saj so ti pogosto okrajšani oz. odrezani pri koncu. Prav tako nismo upoštevali tvitov, ki so jih objavili računi nezasebnih uporabnikov (npr. organizacije,

1 http://www.tei-c.org/index.xml

2 Tviti so bili zajeti v času, ko je bila dolžina tvita omejena na 140 znakov.

agencije, podjetja), saj ti največkrat ne izkazujejo tipičnih značilnosti jezika RPK. Obenem smo poskrbeli, da je vzorec vseboval tako relativno standarden jezik (s čimer smo želeli v korpus vključiti standardne, a kljub temu za RPK značilne jezikovne prvine) kot zelo nestandardnega: v vzorec smo zato dodali po 1.000 tvitov z različnimi kategorijami jezikovne (L1–L3) in tehnične (T1–T3) nestandardnosti, avtomatsko pripisane po metodi, opisani v Ljubešić et al. (2018). Pri tem 1 označuje visoko stopnjo standardnosti, 3 pa visoko stopnjo nestandardnosti. V Kons1 smo vključili štiri kategorije, ki so prispevale po 1000 tvitov: prve tri kategorije (T1L3, T3L1 in T3L3) vsebujejo tvite z najvišjo stopnjo nestandardnosti (tehnične, jezikovne ali obeh), zadnja (T1L1) pa tvite, ki ne kažejo znakov nestandardnosti.

Tudi Kons2 vsebuje 4.000 besedil, vzorčen pa je bil po enakih kriterijih kot Kons1. Ker za razliko od tvitov forumska sporočila in komentarji na novice in blogovske zapise niso omejeni z dolžino, smo upoštevali samo besedila dolžine 20-280 znakov, da bi zagotovili medsebojno primerljivost vzorcev Kons1 in Kons2 glede na dolžino.

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

2.2 Označevalne smernice

Na podlagi ročne analize manjšega podkorpusa z 200 naključno vzorčenimi tviti iz vseh štirih kategorij smo pripravili označevalne smernice, ki obravnavajo tehnične in jezikoslovne vidike označevalnega postopka.

Tehnične smernice (Erjavec et al. 2016a) vsebujejo celotno označevalno shemo v WebAnnu (Eckart de Castilho et al. 2014; platforma je podrobneje predstavljena v razdelku 2.3.1) in predstavljajo splošna navodila za delo s platformo (npr. kako združujemo ali ločujemo pojavnice, kako brišemo nerelevantne ali avtomatsko generirane tvite, kako ravnamo s kompleksnejšimi večplastnimi oznakami), jezikoslovne smernice (Čibej et al. 2016c) pa pojasnjujejo merila, ki jih mora označevalec upoštevati med označevanjem. Smernice za označevanje RPK so v glavnem sledile smernicam za označevanje standardnih (Holozan et al. 2008) in historičnih (Erjavec 2015) slovenskih besedil, a z nekaterimi prilagoditvami, saj gre za drugačen medij. Vidike smernic, ki so najbolj značilni za RPK, na kratko povzemamo v nadaljevanju.

2.2.1 Stavčna segmentacija

Stavčna segmentacija je delitev besedila na stavke.[3] Pri segmentiranju tvitov na stavke smo kot glavni kriterij upoštevali končna ločila (pike, klicaje, vprašaje, tri- ali večpičje ter narekovaje). Tviti pa vsebujejo tudi nekatere druge prvine, ki se lahko sopojavljajo s končnimi ločili oziroma, če končnih ločil ni, same delujejo na podoben način. Te prvine so emotikoni in emojiji (;-), =D, 😊, 😈), ključniki (#justsayin), sklici na uporabniška imena (@avtor) in URL-naslovi (http://t.co/fqVqV92mzc). Ob odsotnosti končnih stavčnih ločil lahko te prvine prevzamejo njihovo vlogo in končajo stavek. Ker se lahko stavek konča tudi s serijo več tovrstnih prvin, kot konec stavka obravnavamo zadnjo prvino:

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

Če se prvine pojavljajo skupaj z ločilom, jih obravnavamo kot ločen stavek:

> Življenje Je Cirkus. js sm pa cefur. Luka Stigl js sm se poscal v hlace k sm se vidu. bolano. ¶ :) ... ¶ http://t.co/QyzKRZqZnS ¶

2.2.2 Tokenizacija

Pri avtomatski tokenizaciji (tj. delitvi besed na pojavnice) so se pojavile napake, ki jih je bilo treba popraviti ročno. Med najpogostejšimi napakami so bile okrajšave, emotikoni, obrazila in besede, ki so vsebovale ločila. Pri okrajšavah (npr. slov. za »slovenski«) je tokenizator piko pogosto obravnaval kot končno stavčno ločilo in kot posamezno pojavnico. V takšnih primerih je bilo piko treba združiti z okrajšavo.

Emotikoni so se pogosto pojavljali v serijah in brez vmesnih presledkov, zaradi česar jih je tokenizator obravnaval kot ločila in jih delil. V tovrstnih primerih serije nismo delili na posamezne emotikone, temveč smo jih združili v eno pojavnico:

> :) ¶ :) ¶ : ¶ * ¶ * → :):):**
>
> \ ¶ m ¶ / ¶ (¶ - ¶ _ ¶ - ¶) → \m/(-_-)

3 Izraz »stavek« uporabljamo v širšem pomenu, ki zajema logično strukturno celoto. V nekaterih primerih je to poved, v drugih zgolj ena sama (ne)jezikovna prvina (npr. emotikon ali URL-naslov). V primerih konec stavka (oziroma meje med pojavnicami v primerih, ki prikazujejo popravke tokenizacije) označujemo s simbolom ¶.

Podoben pristop smo ubrali pri ločenih obrazilih in besedah, ki so vsebovale ločila:

TV ¶ - ¶ ja → TV-ja

sms ¶ - ¶ i → sms-i

žen ¶ (¶ sk ¶) ¶ am → žen(sk)am

politik ¶ (¶ e ¶ / ¶ o ¶) → politik(e/o)

2.2.3 Normalizacija

Pojem *normalizacija* v našem kontekstu pomeni pripisovanje oblike, ki je po zapisu čim bolj prilagojena standardni. Pojavnice smo normalizirali samo na nivoju zapisa, ne pa denimo na nivoju besedišča (*farbat – farbati*, ne **barvati*) ali skladnje (*nisem bral knjigo – nisem bral knjigo*, ne **knjige*). Na ravni normalizacije sta se za najbolj problematični izkazali dve kategoriji besed, in sicer:

1. nestandardne besede z več različicami zapisa in brez neposredne standardne ustreznice, in
2. tujejezične jezikovne prvine z različnimi stopnjami prevzetosti.

Besede iz prve kategorije (npr. *orng*, *ornk*, *oreng*, *orenk*; *fovš*, *favš*, *fouš*, *fauš*, *fowš*) se najpogosteje pojavljajo le v govorjenem jeziku in v standardni slovenščini nimajo neposredne ustreznice, zato je določanje normalizirane oblike težavno. Normalizirano obliko smo v teh primerih določili s pomočjo dodatnega merila: označevalec je v podkorpusu tvitov korpusa Janes z regularnimi izrazi poiskal vse prisotne različice zapisa, kot normalizirano obliko pa je izbral najpogostejšo (v zgornjih primerih sta to *ornk* in *fouš*).

Na podoben način je bila problematična normalizacija besed iz druge kategorije, ki je vsebovala tujejezične prvine z različnimi stopnjami prilagoditve slovenskemu zapisu in oblikoslovju (npr. *updateati*, *updajtati*, *updejtati*, *apdejtati*). Normalizacija z izvirnimi/citatnimi oblikami (npr. *po-update-ati*) bi v mnogih primerih v korpus uvedla umetne oblike, ki jih v realni jezikovni rabi ni, zato smo pri označevanju tujejezičnih prvin upoštevali naslednji merili:

a) če je bil zapis docela fonetiziran (npr. *danke schön* → *dankešn*, *appreciate* → *apršiejt*), smo besedo obravnavali na enak način kot slovensko nestandardno besedo z več različicami zapisa (glej primer *ornk* zgoraj);

b) če je beseda izkazovala kakršnekoli tujejezične značilnosti (npr. neslovenske črke ali tujejezični zapis), smo normalizirano obliko določili tako, da

smo iz podkorpusa tvitov Janes izbrali najpogostejši zapis med tistimi, ki so še vedno izkazovali tujejezične značilnosti (npr. *updateati*, *updajtati*, *updejtati* → *updejtati*).

Nekaterih prvin, ki so značilne za tvite in RPK (npr. sklici na uporabniška imena, ključniki, URL-naslovi, emotikoni in emojiji), nismo normalizirali, ne glede na to, ali so bile njihove oblike v skladu s standardno ali ne. Pri normalizaciji prav tako nismo popravljali skladnje (npr. nepravilne rabe sklonov ali napak pri ujemanju – niti naključnih), pogostih leksikalnih napak (*moči* – *morati*) ali napak v slogu ali registru (*rabiti* – *potrebovati*).

2.2.4 Lematizacija

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

Podobno kot pri normalizaciji je tudi pri lematizaciji med večjimi izzivi označevanja določanje meje med tujejezičnim in slovenskim besediščem. V tvitih se tujejezične prvine pojavljajo kot posamezne besede različnih besednih vrst in variant zapisa (*share*, *shareati*, *share-ati*, *šerati*), kot besedne zveze ali daljši segmenti. Zadnje smo označevali kot niz pojavitev v tujem jeziku, pri čemer so leme enake oblikam, oblikoskladenjska oznaka pa je *Nj*. Podobno velja za občnoimenske besedne zveze (*bonus score*, *sugar rush*) in posamezne besede, ki so v besedilu zapisane brez jasno razvidnih prilagoditev slovenskemu zapisu oz. pregibanju (*jailbreak*, *hrvatskog*).

Pri besedah, ki prilagoditev izražajo, smo lemo določili v skladu s slovenskimi oblikoslovnimi načeli (*benchmarki* → *benchmark*, *chatala* → *chatati*). Pri odločanju, ali besedo obravnavati kot tujejezično ali prevzeto, so bili uporabljeni tudi referenčni leksikalni viri, predvsem SSKJ[4] in SNB[5] ter leksikon besednih oblik Sloleks. Vprašanja uvrščanja kratičnih poimenovanj med kratice in okrajšave na eni strani ter občna (*lol*, *drž.*) in lastna imena (*Sds*, *Slo.*) na drugi so bila razrešena že na ravni normalizacije, označevalci pa so v teh primerih pri pripisovanju lem (in oblikoskladenjskih oznak) sledili normaliziranim oblikam.

4 http://www.fran.si/130/sskj-slovar-slovenskega-knjiznega-jezika

5 http://www.fran.si/131/snb-slovar-novejsega-besedja

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

2.2.5 Oblikoskladnja

Tudi na oblikoskladenjski ravni so bile osnovno izhodišče za označevanje smernice korpusa ssj500k po sistemu za oblikoskladenjsko označevanje JOS. Med razlikami gre v prvi vrsti omeniti širitev sistema z naslednjimi novimi oznakami: *Nh* za ključnike; *Nw* za URL- in e-naslove; *Na* za sklice na uporabniška imena; in *Ne* za emotikone in emojije. Z naštetimi oznakami in načelom lematizacije, pri katerem lema sledi izvorni obliki, smo na enostaven in sledljiv način rešili vprašanje označevanja tvitersko specifičnih prvin. Glede na sistem JOS smo uvedli še eno pomembno novo oznako, in sicer za ločila: v sistemu JOS za ločila ni bila predpisana specifična oznaka (v formatu XML/TEI so bila identificirana s posebnim elementom, ne pa tudi z posebno oznako), v projektu JANES pa smo za to uporabili oznako *U*. Vse uporabljene oznake sledijo sistemu MULTEXT-East V5 (v izdelavi), dostopne pa so na http://nl.ijs.si/ME/V5/msd/.

Pri ročnem označevanju nista bili uporabljeni oznaki *Nt* (zatipkana beseda) ter *Np* (tokenizacijska napaka), saj so bile tovrstne težave ročno odpravljene že na ravni normalizacije in tokenizacije.

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

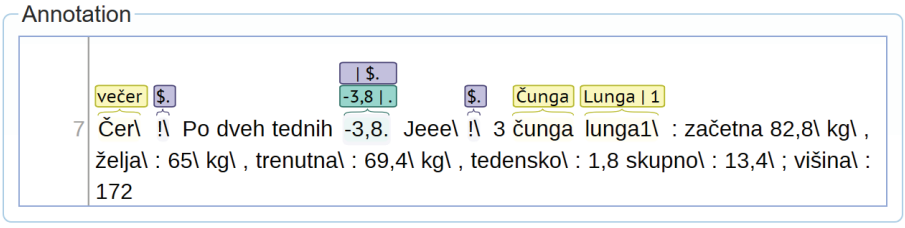
Nazadnje je treba omeniti še označevanje zaprtih besednih vrst, ki je skladno z načeli ssj500k v izhodišču potekalo leksikonsko pogojeno, a z možnostjo dodajanja nestandardnega besedišča. Kategorija, ki je na ta način dobila največ novih elementov, je členek (npr. *eto*, *evo*, *ajde*, *naka*, *kao*,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

2.3 Postopek ročnega označevanja

Ker so bila obstoječa orodja za obdelavo besedil naučena na standardni slovenščini in bi se na nestandardni odrezala bistveno slabše, je bilo za večjo natančnost pri lematizaciji in oblikoskladenjskem označevanju besedil ključno, da v vzorcih Kons1 in Kons2 najprej ročno popravimo napake v tokenizaciji in stavčni segmentaciji ter obenem normaliziramo besede z nestandardnim zapisom. Pri označevanju smo zato v prvi fazi ročno popravili avtomatsko označene nivoje tokenizacije, stavčne segmentacije in normalizacije, zatem pa smo ročno popravljeni podmnožici ponovno uvozili v označevalno orodje kot vzorca Kons1-MSD in Kons2-MSD ter ročno popravili še napake na ravneh lematizacije in oblikoskladenjskih oznak. V podmnožice, ki smo jih izbrali za drugo fazo označevalne kampanje, smo vključili več nestandardnih besedil kot standardnih, s čimer smo v korpusu želeli zagotoviti čim več prvin, ki so značilne za RPK.

2.3.1 Označevalna platforma

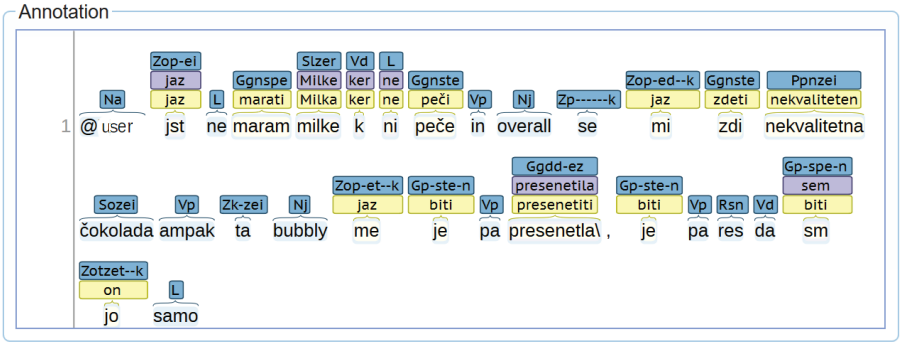
Označevanje je potekalo v spletni označevalni platformi WebAnno (Eckart de Castilho et al. 2014), ki med drugim omogoča večplastno označevanje, vsaki plasti pa lahko pripišemo tudi več kot eno vrednost. Slabost orodja je ta, da ni namenjeno popravljanju tokenizacijskih napak, zato je to precej zapleteno. Iz tega razloga smo poleg tokenizacijske plasti z več vrednostmi v orodje uvedli tudi nabor posebnih simbolov, s pomočjo katerih smo lahko ločevali in združevali pojavnice ter določali meje med stavki. Primer prikazuje Slika 1: tokenizator je v tem primeru niz »-3,8.« napačno obravnaval kot eno pojavnico, označevalec pa jo je ločil na dve in dodal stavčno mejo na drugo pojavnico (pika). S poševnicami smo zaznamovali, da v izvirnem besedilu med pojavnicama ni bilo presledka.[6]



Slika 1: Popravljanje mej med pojavnicami in stavki v WebAnnu (oznake za normalizacijo so označene z rumeno, za tokenizacijo z zeleno in za stavčno segmentacijo z vijolično).

6 Če je bila poševnica prisotna tudi v izvirnem besedilu, je bila v označevalni platformi prikazana kot *\$0*. Na ta način smo ločili prave poševnice od tistih, s katerimi smo zaznamovali pomanjkanje presledkov med pojavnicami.

Slika 2 prikazuje primer z oznakami za oblikoskladnjo (modra), lematizacijo (rumena) in normalizacijo (vijolična):



Slika 2: Označevanje lem in oblikoskladenjskih oznak v WebAnnu.

Platforma omogoča tudi t. i. razsojanje, pri katerem razsodnik sprejme dokončno odločitev v primerih, kjer prihaja do razhajanja med različnimi označevalci; primer je podan v Sliki 3.

Slika 3: Razsojanje v WebAnnu.

2.3.2 Označevalna kampanja

Označevalne kampanje vseh korpusov so se razlikovale v številu označevalcev in obsegu označevanega gradiva, a so potekale na podoben način, in sicer v več fazah. V nadaljevanju predstavljamo pregled in opis označevalne kampanje za korpusa Janes-Norm in Janes-Tag, ki je zajemala tri stopnje:

a) NTS-Kons1 – normalizacijo, tokenizacijo in stavčno segmentacijo vzorca Kons1 (od decembra 2015 do marca 2016);

b) NTS-Kons2 – normalizacijo, tokenizacijo in stavčno segmentacijo vzorca Kons2 (od marca 2016 do maja 2016); in

c) LO-Kons1&2 – lematizacijo in oblikoskladenjsko označevanje vzorcev Kons1 in Kons2 (od marca 2016 do oktobra 2016).

Ob začetku prvega dela označevalne kampanje (NTS-Kons1) smo priredili dvodnevno delavnico, na kateri so se označevalci seznanili z delom v WebAnnu in z označevalnimi smernicami. Na delavnici je sodelovalo 11 študentov jezikoslovnih smeri na magistrski stopnji.

Teoretičnemu uvodu v WebAnno s praktičnim delom in predstavitvi smernic je sledila uvajalna označevalna faza, med katero so udeleženci označili manjše število tvitov. Cilji označevanja so bili naslednji:

a) vsak tvit mora biti pravilno razdeljen na stavke;

b) vsak tvit mora biti pravilno razdeljen na pojavnice; in

c) vse pojavnice morajo imeti pripisano normalizirano obliko; dvoumne pojavnice ohranijo izvirno, nenormalizirano obliko.

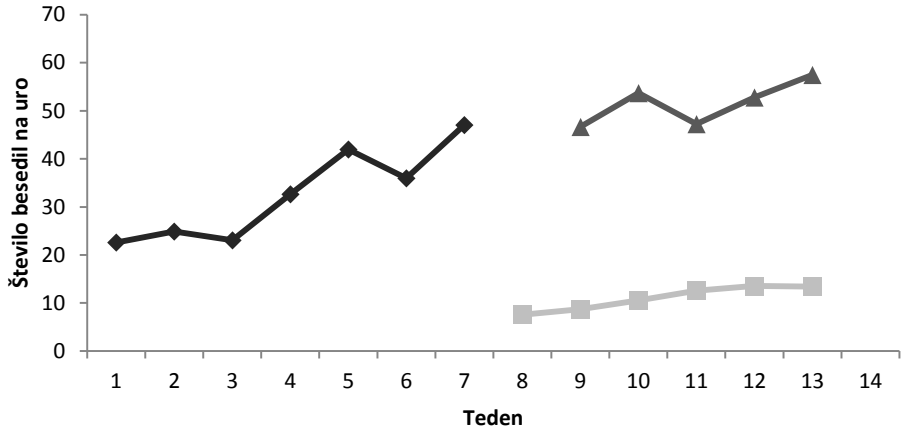
Uvajalni označevalni fazi je sledila diskusija, na kateri smo z označevalci razpravljali o njihovih odločitvah in razhajanjih med njihovimi oznakami, podali pa smo tudi pravilne rešitve in razloge zanje, da bi čim bolj uskladili odločitve označevalcev in izboljšali njihovo ujemanje. V drugem delu kampanje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

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

Delotok označevanja je vključeval skupino označevalcev in dva razsodnika z dobrim poznavanjem smernic za označevanje. Razsodnika, ki sta bila zadolžena tudi za vodenje označevalne kampanje, sta v tedenskih fazah posamezni skupini

označevalcev[7] dodelila določeno število datotek, po koncu vsake faze pa sta oznake ročno preverila in, če je bilo potrebno, označevalcem podala konstruktivno povratno informacijo ter na ta način odstranila najpogostejše oz. najresnejše napake. Če so označevalci med delom naleteli na posebno problematično dilemo, so bile z njo dopolnjene tudi smernice za označevanje. Ustvarjen je bil tudi e-poštni seznam, na katerem so lahko označevalci razsodnikoma zastavljali vprašanja in razreševali problematične ali dvoumne primere, ki niso bili vključeni v smernice.

Med delom smo spremljali učinkovitost označevalcev, tako da smo v vsaki fazi merili razmerje med časom označevanja in številom označenih besedil (glej Sliko 4).



Slika 4: Učinkovitost označevalcev pri označevanju vzorcev Kons1 in Kons2. Črni del predstavlja NTS-Kons1, temno sivi NTS-Kons2 in svetlo sivi LO-Kons1&2.

Z grafa je razvidno, da normalizacija, tokenizacija in stavčna segmentacija potekajo mnogo hitreje od lematizacije in oblikoskladenjskega označevanja. Dobro usposobljeni označevalci lahko v eni uri normalizirajo med 45 in 55 besedil, lematizirajo in oblikoskladenjsko označijo pa le nekaj nad 10 besedil. Na tej točki je treba znova poudariti, da so označevalci preverjali in popravljali napačne avtomatsko pripisane oznake. Označevanje brez predhodnih avtomatskih postopkov bi po vsej verjetnosti vzelo bistveno več časa.

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

2.4 Izvoz in končni zapis označenega korpusa

Posebno pozornost smo namenili formatu podatkov, da bi vse ročno preverjene oznake združili v enovit zapis. Pri zapisu korpusa Janes uporabljamo priporočila za kodiranje besedil TEI, ki so v uporabi tudi pri večini obstoječih slovenskih korpusov. Ker WebAnno formata TEI ne podpira, smo med razpoložljivimi formati izbrali tabelarični format TSV, v katerem je vsaka pojavnica skupaj z identifikatorjem in vsemi oznakami zapisana v svoji vrstici.

Izdelali smo pretvornik med izvornim formatom TEI in formatom TSV, ki omogoča ciklično izvažanje in uvažanje ter združevanje oznak v skupni TEI. Rezultat tako vsebuje vse oznake izvornega TEI, dopolnjene s popravki ročnega označevanja, pretvornik pa je uporaben tudi za druge označevalne kampanje z drugim naborom oznak.

Končne različice ročno označenih korpusov so torej kodirane kot datoteke XML v formatu TEI P5 (TEI Consortium), ki vključujejo kolofon TEI z metapodatki o korpusu ter telo, ki ga sestavljajo anonimni bloki (<ab>), od katerih vsak vsebuje po eno besedilo. Poleg tega vsak dokument vsebuje tudi oblikoskladenjske specifikacije, ki so kodirane kot knjižnica struktur lastnosti TEI. To omogoča, da oblikoskladenjsko oznako razgradimo na posamezne sestavne dele (pare atributov in vrednosti) oz. da jo lokaliziramo v slovenščino.

```
<ab xml:id="janes.blog.publishwall.4264.3" type="blog" subtype="T1L3">
   <s>
      <w lemma="kaj" ana="#Rgp">Kaj</w><c> </c>
      <w lemma="biti" ana="#Va-r3s-y">ni</w><c> </c>
      <w lemma="ta" ana="#Pd-nsn">to</w><c> </c>
      <choice>
         <orig><w>tazadnje</w></orig>
         <reg>
            <w lemma="ta" ana="#Q">ta</w><c> </c>
            <w lemma="zadnji" ana="#Agpnsn">zadnje</w>
         </reg>
      </choice><c> </c>
      <choice>
         <orig><w>AAjevska</w></orig>
         <reg><w lemma="aa-jevski" ana="#Agpfsn">AA-jevska</w></reg>
      </choice><c> </c>
      <w lemma="molitev" ana="#Ncfsn">molitev</w>
      <pc ana="#Z">?</pc>
   </s>
</ab>
```

Slika 5: Izsek iz XML TEI 5.

Kot prikazuje Slika 5, je vsak element <ab> (tj. besedilo) označen s svojo identifikacijsko oznako/kodo iz korpusa Janes, z vrsto vira (tviti, komentarji na novice, forumska sporočila ali blogovski zapisi) in s kategorijo standardnosti (T1L1, T1L3, T3L1 ali T3L3), vsak blok pa vsebuje zaporedne stavke (<s>) iz besedila. Pojavnice so kodirane kot besede (<w>) ali ločila (<pc>), izvirni »jezikovni« presledki pa so ohranjeni z elementom TEI za znak (»character«, <c>). Pojavnice so označene z oblikoskladenjskimi oznakami, ki so kazalci na svoje definicije v knjižnici struktur lastnosti, besede pa so označene tudi z lemami.

Za kodiranje standardne oblike besed z nestandardnim zapisom smo uporabili element TEI <choice> z dvema podrednima elementoma za izvirno (<orig>) in normalizirano obliko (<reg>). Ta pristop ima to prednost, da omogoča večbesedne preslikave in razlikuje med jezikoslovnimi oznakami izvirnika in normalizirane oblike. Trenutno označujemo samo normalizirane oblike.

Kodiranje TEI smo nato prevedli v vertikalni format CQP, ki ga uporablja Sketch Engine (Rychlý 2007), korpus pa namestili na NoSketch Engine (instalacija CLARIN.SI).

2.5 Varovanje osebnih podatkov in avtorskih pravic

Vsi ročno označeni korpusi so po obsegu majhni in ne vsebujejo občutljivih osebnih podatkov, zato jih ne dojemamo kot problematične z vidika varovanja osebnih podatkov ali avtorskih pravic (Erjavec et al. 2016b). Besedil, ki so vključena v korpuse, nismo anonimizirali, v malo verjetnem primeru pritožb pa bomo posamezna problematična besedila odstranili iz javno dostopnih korpusov.

3 ROČNO OZNAČENI KORPUSI

V nadaljevanju predstavljamo vse ročno označene korpuse, ki so bili izdelani v okviru projekta JANES. Delimo jih na dve kategoriji:

a) korpuse za učenje jezikovnotehnoloških orodij in

b) korpuse za jezikoslovne raziskave.

Poleg korpusov Janes-Norm in Janes-Tag, ki smo se jima posvetili že ob opisu postopka izdelave ročno označenih korpusov, v kategorijo korpusov za učenje jezikovnotehnoloških orodij uvrščamo še skladenjsko označeni Janes-Syn, med korpuse za jezikoslovne raziskave pa štejemo Janes-Kratko, Janes-Vejica, Janes-Preklop in Janes-Geo.

Vsi korpusi so prosto dostopni pod licenco CC BY-SA 4.0 na repozitoriju CLARIN.SI. Povezave so navedene v ustreznih podrazdelkih.

3.1 Korpusi za učenje jezikovnotehnoloških orodij

V tem podrazdelku podrobneje predstavimo korpuse za učenje jezikovnotehnoloških orodij Janes-Norm, Janes-Tag in Janes-Syn. Prvi dve sta bili v okviru projekta za te namene že uporabljeni, kar je podrobneje opisano v Ljubešić et al. (2018).

3.1.1 Janes-Norm

Janes-Norm vsebuje besedila iz vzorcev Kons1 in Kons2, njegova glavna vloga pa je učenje in preizkušanje orodij za tokenizacijo, stavčno segmentacijo in normalizacijo slovenske RPK. Tabela 1 prikazuje velikost korpusa oz. njegovih delov glede na različne stopnje standardnosti in besedilne žanre.

Tabela 1: Velikost in sestava korpusa Janes-Norm.

Janes-Norm												
	Besedila	%	Pojavnice	%	Besede	%	Norm.	%	Prave norm.	%	Večbesedne norm.	%
Vse	7.816	100,0	184.755	100,0	142.848	100,0	39.304	27,5	16.604	42,2	815	4,9
T1L1	1.979	25,3	48.437	26,2	37.666	26,4	7.883	20,9	795	10,1	78	9,8
T1L3	1.936	24,8	47.426	25,7	34.861	24,4	12.609	36,2	6.566	52,1	239	3,6
T3L1	1.954	25	41.472	22,4	33.071	23,2	6.458	19,5	1.018	15,8	153	15
T3L3	1.947	24,9	47.420	25,7	37.250	26,1	12.354	33,2	8.225	66,6	345	4,2
Blogi	1.159	14,8	20.981	11,4	16.258	11,4	3.567	21,9	1.621	45,4	87	5,4
Forumi	1.572	20,1	37.647	20,4	30.960	21,7	7.556	24,4	3.796	50,2	214	5,6
Komentarji	1.145	14,6	23.489	12,7	19.083	13,4	4.628	24,3	1.880	40,6	92	4,9
Tviti	3.940	50,4	102.638	55,6	76.547	53,6	23.553	30,8	9.307	39,5	422	4,5

V celoti korpus vsebuje 7.816 besedil, ki so relativno enakomerno razporejena po štirih vključenih kategorijah (ne)standardnosti. Omeniti je treba, da je vrstni red besedil tako v korpusu Janes-Norm kot v korpusu Janes-Tag naključen, saj je korpus tako lažje razdeliti na učno in testno množico ter obenem zajeti vse besedilne tipe. Korpus ne vsebuje vseh 8.000 vzorčenih besedil (4.000 iz Kons1 in 4.000 iz Kons2) oz. 2.000 besedil za vsako od kategorij standardnosti, saj so označevalci imeli možnost, da posamezna besedila označijo kot nerelevantna (npr. če so bila avtomatsko generirana, povsem nerazumljiva, brez kakršnihkoli jezikovnih prvin ali v celoti v tujem jeziku). Teh besedil v končno različico korpusa nismo vključili.

Skupaj besedila vsebujejo skoraj 185.000 pojavnic oz. 144.000 besed (pri čemer kot besedo štejemo vse pojavnice razen ločil, številk in specifičnih elementov RPK, kot so npr. e-naslovi, URL-naslovi, ključniki, sklici na uporabniška imena ter emojiji in emotikoni). Iz Tabele 1 je razvidno, da so deleži med kategorijami standardnosti večinoma ohranjeni tudi pri pojavnicah in besedah. Glede na besedilne tipe približno polovica besedil, pojavnic in besed izvira iz tvitov, po okrog 12 % besedil pa iz blogovskih zapisov in komentarjev na novice.

Stolpec Norm. prikazuje delež normaliziranih pojavnic glede na skupno število besed. Stolpec Prave norm. podaja delež jezikovno kompleksnejših normalizacij glede na vse normalizirane besede. Med jezikovno kompleksnejše normalizacije štejemo vse popravke, ki ne vključujejo zgolj kapitalizacije (*slovenija* → *Slovenija*) ali rediakritizacije (*macka* → *mačka*). Kot je razvidno, je bila normalizirana več kot četrtina besed (27,5 %), 42 % od teh pa je vključevalo kompleksnejše normalizacije. Kot lahko pričakujemo, standardna besedila vsebujejo mnogo manj normaliziranih besed, stopnja jezikovne standardnosti (L) pa korelira s potrebo po normalizaciji. Glede na žanr so v korpusu Janes-Norm na splošno najbolj standardni komentarji na bloge (21,9 %), sledijo pa jim forumska sporočila in komentarji na novice. Največji delež besed, ki jih je bilo treba normalizirati, izkazujejo tviti (30,8 %).

Slika je nekoliko drugačna, če opazujemo samo jezikovno kompleksnejše normalizacije, saj je v tvitih tovrstnih normalizacij le 39,2 %, v komentarjih na novice in bloge 40,6 % in 45,4 %, v forumskih sporočilih pa več kot pol (50,1 %). Iz tega sklepamo, da uporabniki najbolj upoštevajo rabo diakritičnih znamenj na forumih, najmanj pa v tvitih. Vzrok za to je najverjetneje v napravah, s katerih uporabniki objavljajo: forumska sporočila pišejo na računalnikih, tvite pa na telefonih.

Zadnji stolpec podaja število in delež primerov (glede na vse jezikoslovno kompleksnejše normalizacije), pri katerih je normalizacija vključevala ločevanje ali združevanje besed. Kot smo že omenili, je tovrstne normalizacije še posebej težavno modelirati, a rezultati kažejo, da ne gre za pogost pojav, saj zajema le okrog

5 % jezikovnih normalizacij. Povedano drugače, četudi teh primerov sploh ne bi obravnavali, bi končni upad natančnosti ne bil znaten.

Omeniti je treba tudi, da so v Janes-Norm vključene tudi oblikoskladenjske oznake in leme, a so bile za del, ki ni prekriven z Janes-Tag, pripisane le avtomatsko in zato vsebujejo napake.

Janes-Norm je kot podatkovna množica na voljo na repozitoriju CLARIN.SI (Erjavec et al. 2016), iskanje po korpusu pa je omogočeno v konkordančnikih CLARIN.SI, in sicer KonText in NoSketch Engine. Slika 6 prikazuje primer iskanja vseh pojavnic z normalizirano obliko *koliko*.


Query koliko 125 (676.6 per million)

Page 1 of 7 Go Next | Last

janes.forum.medovernet.5809700	moraš uporabit primerno silo Pozdravljeni!	Koliko	dni oziroma mesecev je potrebno jemati
tid.5457101769225175O4	@tomltoml @vinkovasle1 ka pa pol tolk pizdite	kolk	so jih za glavo skrajšali, ..-vaših.. @surfon
tid.5027441568765706 24	kdo ima daljšega v Piranu oz. kdo bo zidal	koliko	in kje. Hot VS Pope let the games begin
tid.5401280977195499 52	komunajzarju", z veseljem ti bom prisluhnil.	Kolko	je blo onih infomatikov v JU, @petrasovdat
janes.blog.publishwall.3702.6	posojajo, pa jim ne bomo nikoli vrnili,...	koliko	smo že dolžni ?? 30 miljard, zato pa moramo
tid.5064982280835235 84	Polom. Fak. Polom. Ja tle sm. Ne. Ne vem	kolk	časa bom tle. Kva ti? Aja. A gre pingvin
janes.forum.avtomobilizem.18.75263.1530748	tut da se avti prodajo tut če jih nevem	kok	spluvamo @chatek hmmm ... florentinca san
janes.forum.medovernet.9184124	fast...tebe bi pa jaz koj mela hahahaha	kok	bi se midva nasmejala Evo, še ena slikca
janes.news.rtvslo.266505.14	sociedad je kr en vaški klub.. a ne? '..	kolk	je že blo? je pa itak "obvezna sestavina
tid.4231098289074995 21	in ste že začeli z antipropagando? jao,	kolk	ste poceni.. @Sraqa Sorry :) We need to
tid.3897948595487617 024	@PrimozP @PEroCaks pojma nimam. Toliko da lahko	kolko	tolko normalno oddajajo. Videl sem razmere
tid.3862089329951498 25	potem si Janko ne bo mogel več zmišljevati,	koliko	so vredni slabi krediti, ki bodo prenešeni
janes.news.mladina.163356.9	prebrati več kot komentar pod člankom in tvit?	Koliko	strani že ima ta popravek zgodovine? Več
janes.news.rtvslo.301991.1317	http://t.co/lxLrPc87pp gapponJa vidimo,	kolko	jih podpira 3500, 3500 - mogoče 5000 jih
tid.5046017830341877 76	in vodi. @m4tija cuj nism :-) ne vem od	kok	jim je uspel... sm pa ze zamenju geslo
tid.3869063368079114 24	sem bil v paru z Mertljem, ne glede nato,	koliko	mislijo, da ne moreva skupaj igrat. " @NK_MARIBOR
janes.news.rtvslo.306503.3	mogoče da je res okusno... sam me zanima	kolk	porcij bi mogu pojest en delavc k pride
janes.news.rtvslo.344929.19	gledam studio na planet tv pa ne morem verjet	kolk	se je dani bavec postaral... sploh ne zgleda

Slika 6: Iskanje pojavnic z normaliziranо obliko *koliko* v korpusu Janes-Norm preko vmesnika NoSketch Engine.

3.1.2 Janes-Tag

Janes-Tag je podmnožica korpusa Janes-Norm in vsebuje vzorca Kons1-MSD in Kons2-MSD, zasnovan pa je bil kot zlati standard za lematizacijo in označevanje oblikoskladenjskih oznak oz. za učenje in preizkušanje slovenskih oblikoskladenjskih označevalnikov in lematizatorjev (več o tem v poglavju Ljubešić et al. (2018)). Različica 2.0 poleg oblikoskladenjskih oznak in lem vsebuje še oznake lastnih imen, zato se jo lahko uporabi tudi kot učno množico za prepoznavalnike imenskih entitet v slovenski RPK.

Tabela 2 prikazuje velikost celotnega korpusa Janes-Tag in njegovih podkategorij glede na stopnjo standardnosti in vključene žanre.

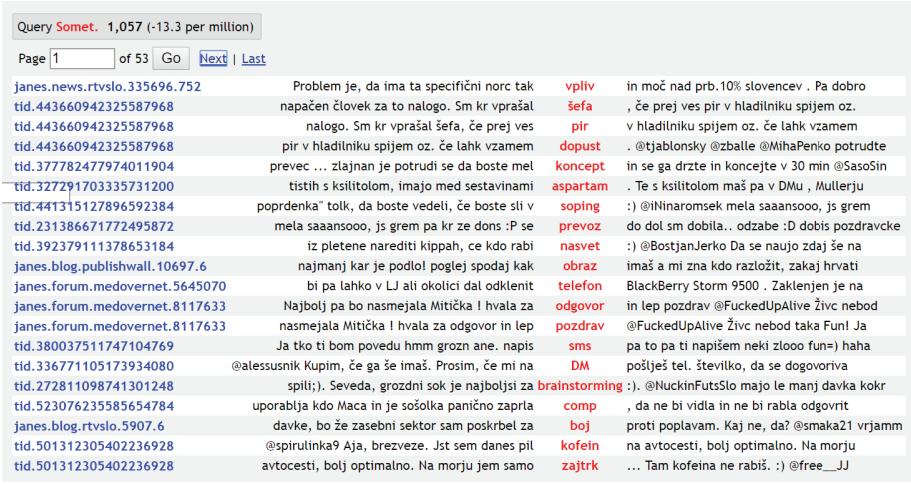
Tabela 2: Velikost in sestava korpusa Janes-Tag.

Janes-Tag														
	Besedila	%	Imena	%	Pojavnice	%	Besede	%	Norm.	%	Prave norm.	%	Večbesedne norm.	%
Vse	2.958	100,0	4.780	100,0	75.276	100,0	56.555	100,0	18.829	33,3	10.103	53,7	379	3,8
T1L1	275	9,3	254	5,3	6.695	8,9	5.400	9,5	954	17,7	77	8,1	11	14,3
T1L3	1.219	41,2	2.040	42,7	32.329	42,9	23.159	40,9	8.762	37,8	4.447	50,8	150	3,4
T3L1	245	8,3	159	3,3	4.559	6,1	3.788	6,7	589	15,5	126	21,4	12	9,5
T3L3	1.219	41,2	2.327	48,7	31.693	42,1	24.208	42,8	8.524	35,2	5.453	64,0	206	3,8
Blogi	269	9,1	180	3,8	5.046	6,7	3.952	7,0	848	21,5	370	43,6	24	6,5
Forumi	403	13,6	211	4,4	9.445	12,5	7.761	13,7	1.894	24,4	935	49,4	46	4,9
Komentarji	303	10,2	339	7,1	6.097	8,1	4.801	8,5	1.250	26	522	41,8	20	3,8
Tviti	1.983	67,0	4.050	84,7	54.688	72,6	40.041	70,8	14.837	37,1	8.276	55,8	289	3,5

Korpus v celoti vsebuje nekaj manj kot 3.000 besedil in dobrih 75.000 pojavnic oz. 56.000 besed, torej je približno pol manjši od korpusa Janes-Norm. Korpus kot učna množica torej ni velik (za primerjavo: zajema približno desetino velikosti ročno označenega učnega korpusa ssj500k (Krek et al. 2015)), a je kljub temu znatno pripomogel k izboljšanju oblikoskladenjskega označevanja in lematizacije slovenske RPK. Poleg tega kot zlati standard omogoča tudi preizkušanje oblikoskladenjskih označevalnikov in lematizatorjev za slovensko RPK.

Zaradi kriterijev vzorčenja Kons1-MSD in Kons2-MSD so deleži besedil, pojavnic in besed glede na različne stopnje standardnosti precej drugačni kot pri korpusu Janes-Norm, saj smo se pri korpusu Janes-Tag osredotočili predvsem na besedila v kategoriji L3, ki zajemajo več kot 80 % celotnega korpusa. Z vidika žanrov večina besedil (67 %) in še večji delež pojavnic (72,8 %) izvira iz tvitov, kar odseva dinamiko označevalne kampanje. Deleži normaliziranih besed v Tabeli 2 so podobni tistim iz korpusa Janes-Norm, razlike pa lahko verjetno pripišemo naključnim dejavnikom vzorčenja.

Janes-Tag je kot podatkovna množica na voljo na repozitoriju CLARIN.SI (Erjavec et al. 2016), iskanje po korpusu pa omogočeno v konkordančnikih CLARIN.SI. Slika 7 npr. prikazuje iskanje vseh samostalnikov v tožilniški obliki v noSketch Engine (v vrstico CQL vpišemo *[tag="Somet."]*).



Query Somet. 1,057 (-13.3 per million)

Page 1 of 53 Go Next | Last

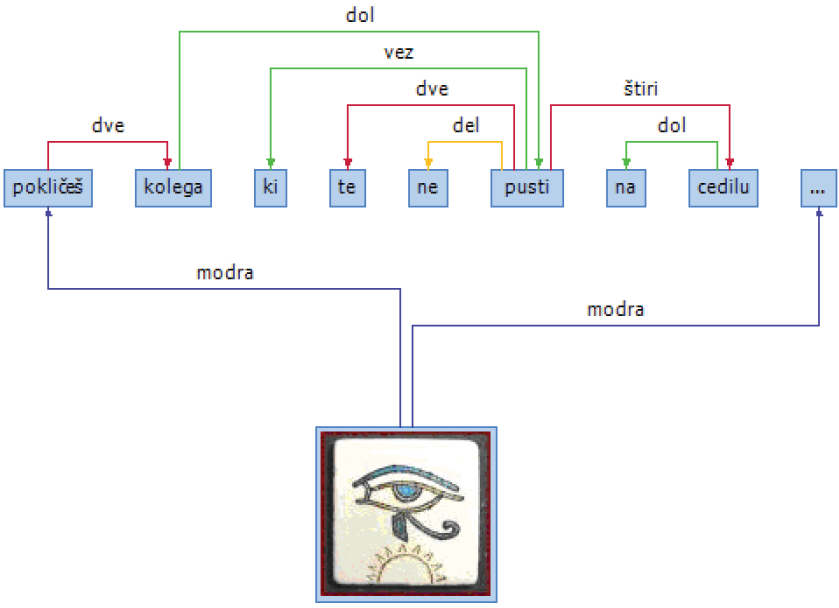
janes.news.rtvslo.335696.752	Problem je, da ima ta specifični norc tak	vpliv	in moč nad prb.10% slovencev . Pa dobro
tid.443660942325587968	napačen človek za to nalogo. Sm kr vprašal	šefa	, če prej ves pir v hladilniku spijem oz.
tid.443660942325587968	nalogo. Sm kr vprašal šefa, če prej ves	pir	v hladilniku spijem oz. če lahk vzamem
tid.443660942325587968	pir v hladilniku spijem oz. če lahk vzamem	dopust	. @tjablonsky @zballe @MihaPenko potrudte
tid.377782477974011904	prevec ... zlajnan je potrudi se da boste mel	koncept	in se ga drzte in koncejte v 30 min @SasoSin
tid.327291703335731200	tistih s ksilitolom, imajo med sestavinami	aspartam	. Te s ksilitolom maš pa v DMu , Mullerju
tid.441315127896592384	poprdenka" tolk, da boste vedeli, če boste sli v	soping	:) @iNinaromsek mela saaansooo, js grem
tid.231386671772495872	mela saaansooo, js grem pa kr ze dons :P se	prevoz	do dol sm dobila.. odzabe :D dobis pozdravcke
tid.392379111378653184	iz pletene narediti kippah, ce kdo rabi	nasvet	:) @BostjanJerko Da se naujo zdaj še na
janes.blog.publishwall.10697.6	najmanj kar je podlo! poglej spodaj kak	obraz	imaš a mi zna kdo razložit, zakaj hrvati
janes.forum.medovernet.5645070	bi pa lahko v LJ ali okolici dal odklenit	telefon	BlackBerry Storm 9500 . Zaklenjen je na
janes.forum.medovernet.8117633	Najbolj pa bo nasmejala Mitička ! hvala za	odgovor	in lep pozdrav @FuckedUpAlive Živc nebod
janes.forum.medovernet.8117633	nasmejala Mitička ! hvala za odgovor in lep	pozdrav	@FuckedUpAlive Živc nebod taka Fun! Ja
tid.380037511747104769	Ja tko ti bom povedu hmm grozn ane. napis	sms	pa to pa ti napišem neki zlooo fun=) haha
tid.336771105173934080	@alessusnik Kupim, če ga še imaš. Prosim, če mi na	DM	pošlješ tel. številko, da se dogoviva
tid.272811098741301248	spili;). Seveda, grozdni sok je najboljsi za	brainstorming	:). @NuckinFutsSlo majo le manj davka kokr
tid.523076235585654784	uporablja kdo Maca in je sošolka panično zaprla	comp	, da ne bi vidla in ne bi rabla odgovrit
janes.blog.rtvslo.5907.6	davke, bo že zasebni sektor sam poskrbel za	boj	proti poplavam. Kaj ne, da? @smaka21 vrjamm
tid.501312305402236928	@spirulinka9 Aja, brezveze. Jst sem danes pil	kofein	na avtocesti, bolj optimalno. Na morju
tid.501312305402236928	avtocesti, bolj optimalno. Na morju jem samo	zajtrk	... Tam kofeina ne rabiš. :) @free__JJ

Slika 7: Iskanje samostalnikov v tožilniški obliki v korpusu Janes-Tag preko vmesnika NoSketch Engine.

3.1.3 Janes-Syn

Kot pilotski vir za učenje skladenjskega razčlenjevanja slovenske RPK je bil pripravljen korpus Janes-Syn, ki je vzorčen iz Janes-Tag in obsega 200 tvitov (475 stavkov). Tviti so izbrani iz besedil zasebnih uporabnikov (ne pa tudi korporativnega tvitanja, ki je glede na preliminarna opažanja jezikovno bolj v skladu s standardom), vključujejo pa primere, ki so daljši od 120 znakov.

Janes-Syn je bil avtomatsko označen s skladenjskim razčlenjevalnikom SSJ (Dobrovoljc et al. 2012) in uvožen v program za pregledovanje drevesnic SSJ (avtor J. Brank, glej Sliko 8). Avtomatsko pripisane skladenjske povezave so bile ročno popravljene skladno z označevalnimi smernicami (Holozan et al. 2008), pri čemer smo sistem označevanja nadgradili z rešitvami za specifike nestandardnega jezika (Arhar Holdt 2016), in sicer z novostmi pri označevanju žanrsko specifičnih elementov, rabi tujejezičnih prvin, obravnavi eliptičnosti in fragmentarnosti jezika ter nestandardni rabi ločil. Več o pripravi Janes-Syn je mogoče prebrati v Arhar Holdt et al. (2016) in v Arhar Holdt (2018).



Slika 8: Primer označenega tvita v Označevalniku SSJ.

Janes-Syn je prosto dostopen za uporabo v nekoliko skrajšani različici (4.000 pojavnic oz. 170 besedil). Kot podatkovna množica je na voljo na repozitoriju CLARIN.SI (Arhar Holdt et al. 2017), iskanje po korpusu pa je omogočeno v konkordančnikih CLARIN.SI. Korpus je v konkordančnik umeščen tako, da je mogoče iskati tudi po skladenjskih oznakah: Slika 9 npr. prikazuje iskanje povezave 'ena', ki načeloma označuje stavčne osebke, v polje CQL vpišemo *[deprel="ena"]*.

Query ena 217 (49,498.2 per million)

Page 1 of 11 Go Next | Last

tid.266122536432062464	videti vsak tweet kandidatov, bi enostavno	sledili	njim?#predsednik12@petrasovdat V priemerjavi
tid.311838678605512704	priemerjavi s svojim predhodnikom nedvomno.	Okoliščine	in pogoji dela pa so mu bili vse prej kot
tid.342718195666399232	mu bili vse prej kot naklonjeni.Čeferin:	sodišče	podlega javnemu mnenju.Ni samo obsodilna
tid.342718195666399232	podlega javnemu mnenju.Ni samo obsodilna	sodba	tista, ki kaže na to, da pravna država
tid.342718195666399232	obsodilna sodba tista, ki kaže na to, da pravna	država	funkcionira.#poglediSlovenijeLažji med
tid.352691378020548608	#Glave s pregledom 3. sezone #GoT.Gobcamo	@anzet	@BokiNachbar @WiC_HmR @matevzluzar http://t.co/LWHogSK9nj
tid.361498211262791681	Blagovici je ustavljen lažno okvarjen romunski	kombi	; ustavljajo nič hudega sluteče naivne voznike
tid.366156416806944768	na SD kartico sem probala že stokrat.Tudi	telefon	to ponudi kot možnost.Rezultat?"Ni predmetov
tid.374883326189764608	@kricac; Bom ugibala - pripadnost?Današnja	mladina	tako zelo hlepi po tem.Mi pa tudi verjetno
tid.374883326189764608	Današnja mladina tako zelo hlepi po tem.	Mi	pa tudi verjetno nismo bili tako zelo drugačni
tid.381732876586217473	Čeprav ti letos ni šlo brez tebe ne bi bili	#junaki	.Rečem ti lahko le SREČNO.Šele zdaj videl
tid.390441098822561792	ti lahko le SREČNO.Šele zdaj videl kakšna	drama	je bila v CONCACAF,ko so ZDA v zadnjih
tid.390441098822561792	videl kakšna drama je bila v CONCACAF,ko so	ZDA	v zadnjih minutah priigrale Mehiki dvoboj
tid.397719565875953665	izkušnjah se da vzgajati brez nasilja.In tudi	sam	nisem bil nikoli tepen, za kar sem staršem
tid.412943107395973120	hvaležen.Ce vprasate mene (in mislim, da se	@anakobal_kobe	strinja), Tini do res odlicnega rezultata
tid.412943107395973120	do res odlicnega rezultata manjka le malo	teze	na spodnji smucki.:)Pa ti @stanka_d, si
tid.429342290239582209	manjka le malo teze na spodnji smucki.:)Pa	ti	@stanka_d, si čisto prestreljena.Od branjenja
tid.429342290239582209	zločincev se ti že pošteno blede.Si tudi	ti	del te "slavne" mreže #UDBA?@RomanLeljak
tid.439337873708679168	#novaplata http://t.co/mytlEbvTUzSuspendirana	tozilka	ne more vec preganjati novinarke.Odgovornost
tid.439337873708679168	vec preganjati novinarke.Odgovornost nosi	tozilec	, ki je prevzel zadevo.In njegov sef.http://t.co/4TRmccuocN

Slika 9: Primer iskanja skladenjskih oznak po Janes-Syn v NoSketchEnginu.

3.2 Korpusi za jezikoslovne raziskave

Z ročno označenimi korpusi za raziskave smo proučili štiri zanimive vidike slovenske računalniško posredovane komunikacije: načine krajšanja besedila, kodno preklapljanje, rabo vejice in rabo regionalnih jezikovnih različic na Twitterju.

3.2.1 Janes-Kratko

Janes-Kratko je korpus tvitov, ki je ročno označen z načini krajšanja po izdelani tipologiji (Goli et al. 2016a), ki pojave krajšanja uvršča na tri ravni: zapis (npr. krajšanje z ločili; *Slovenija - Slo.*), leksika (npr. nadomeščanje s kraticami; bruto domači proizvod – BDP) in skladnja (npr. izpust glagola; *Bi blo treba [∅] tiralico?*). Vsi nivoji so razdeljeni na skupno 32 podkategorij, ki med drugim na nivoju zapisa zajemajo npr. izpuščanje črk, opuščanje presledkov in ločil ter nadomeščanje s krajšimi nizi (npr. s številčnimi homofoni ali s tujejezičnimi črkami), na leksikalnem nivoju pa npr. zapis s kraticami in ustaljenimi okrajšavami s piko.

Glavni namen korpusa je ponuditi kvantitativen pregled nad vrsto in pogostostjo načinov krajšanja v slovenskih tvitih, zajema 777 tvitov (okoli 20.000 pojavnic), ki so bili vzorčeni iz korpusa Janes-Norm. V njem je zabeleženih skupno 3.464 pojavov krajšanja. Od tega je 87 % krajšanja na nivoju zapisa, na ostalih nivojih pa je krajšanja občutno manj: približno 12 % na leksikalnem in le dober 1 % na skladenjskem.

Query LN.* 90 (4,433.7 per million)

Page 1 of 5 Go Next | Last

tid.567399946710966272	drugih zornih kotov + počasni posnetki +	stat	! http://t.co/HheCGiojpc @MatevzNovak @cashkee
tid.617733529640824832	Glede na kratek čas od razpisa do izvedbe	ref.	v Grčiji, se sprašujem, kako uspešno je
tid.617733529640824832	Grčiji, se sprašujem, kako uspešno je bilo	info.	volivcev, ki je ključno za legitimnost.
tid.508689157519310848	moje je dobro, da gate trga. *in prestavi	tevejko	* Resna zadeva. V Mb med starejšimi pustoši
tid.598512578357231616	. *in prestavi tevejko * Resna zadeva. V	Mb	med starejšimi pustoši virusna pljučnica
tid.590962437379198976	npapi. chrome://flags/#enable-npapivpišeš v	nasl.	vrstico in klikneš "enable". @freeeky ok
tid.289085142964793346	#danasnjidan pred 10 leti je vlada razrešila	gen.	dir. Vursa Zorana Kovača, ker je na nov.
tid.289085142964793346	#danasnjidan pred 10 leti je vlada razrešila gen.	dir.	Vursa Zorana Kovača, ker je na nov. konf.
tid.289085142964793346	gen. dir. Vursa Zorana Kovača, ker je na	nov.	konf. kritiziral prenizka proračunska sredstva
tid.289085142964793346	dir. Vursa Zorana Kovača, ker je na nov.	konf.	kritiziral prenizka proračunska sredstva
tid.622036427992367104	segrevanje. Predstavljajte si zdej vse na	minimalcu	. Photo: V dogovoru s Sunčanom Stoneom
tid.605739619754364928	#lokalnevolitve2014 Mandarič, ki ga je Tito izgnal kot	kapital.	izdajalca, prihaja v deželo kjer prirejajo
tid.512893001082101760	letnika '96? Dajte se pripravit tudi na to. /	cc	@VitezDobTeka @BMWSlovenija @JelenaJal
tid.494794705264455680	@IgorGaberc @IgorLuksicSD @strankaSDS V Slo ni	polit.	higiene, kvečjemu politični waterboarding
tid.420282861996875776	sledila na TW, pa je tip vse preveč sral po	TL	, sem ga odsledil. Ku se pa celotna generacija
tid.519363999776133121	again. #matura Kakšno zgražanje zaradi "	nepreprič.	in sploš" odg. AB ; da "smo" izvolili PV
tid.519363999776133121	nepreprič. in sploš" odg. AB ; da "smo" izvolili	PV	, ki v celi kampanji ni dal enega konkr.
tid.519363999776133121	izvolili PV , ki v celi kampanji ni dal enega	konkr.	odg. nikogar ne moti! Hm. Al se on to poskuša
tid.380965979821318144	modre črte, zdej mi interactions ne folgajo,	app	ima krizo identitete in ne prikzuje fotk
tid.394481745154043905	nogavic pa ni #slabo #facepalm Setamo po	Lj	in pride Zoki mim, se ustavi pa da Emanuelu

Slika 10: Iskanje neustaljenih krajšav v korpusu Janes-Kratko preko vmesnika NoSketch Engine.

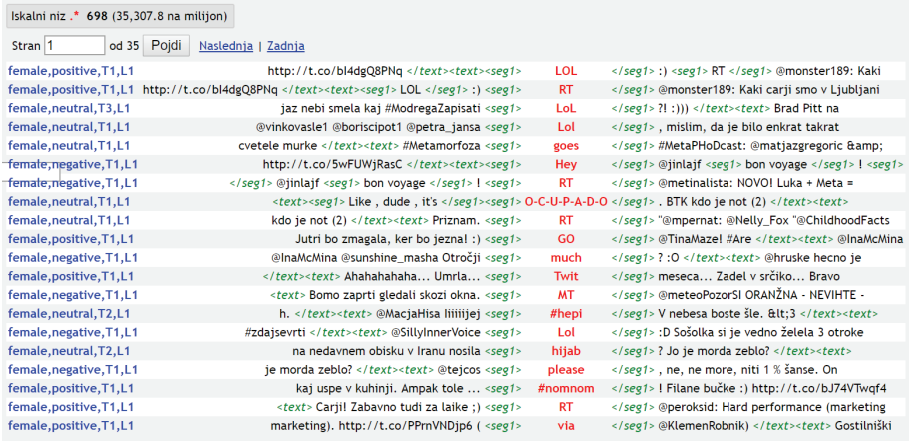
Rezultati raziskave na korpusu so podrobneje predstavljeni v Goli et al. (2016b), korpus pa je na voljo na repozitoriju CLARIN.SI (Goli et al. 2017) in v konkordančnikih CLARIN.SI. Slika 10 prikazuje iskanje vseh neustaljenih krajšav v korpusu (v polje CQL vpišemo *[seg="LN.*"]*).

3.2.2 Janes-Preklop

Korpus Janes-Preklop (glej Reher in Fišer 2018) vsebuje 1.104 tvite (19.769 pojavnic) in je namenjen proučevanju preklapljanja med jeziki v slovenskih tvitih. Preklopi so označeni na več nivojih: jezik preklopa, tip preklopa (medstavčno, zunajstavčno), stopnja prilagojenosti slovenskemu zapisu, stopnja razvidnosti oblikoslovne prilagojenosti (razvidnost iz obrazil in končnic) in vrsta besedne zveze preklopa (samostalniška besedna zveza ipd.).

V korpusu je označenih približno 1.400 preklopov, od tega sta približno dve tretjini znotrajstavčnih, tretjina pa medstavčnih. Jezikov, v katere preklapljajo uporabniki v korpusu, je skupno 9, najpogostejši pa so angleščina (90 % preklopov), hrvaščina/bosanščina/srbščina (4,5 %) in nemščina (3,5 %).

Podrobnejši rezultati raziskave kodnega preklapljanja so predstavljeni v Reher (2017), korpus pa je na voljo na repozitoriju CLARIN.SI (Reher et al. 2017) in v konkordačnikih CLARIN.SI. Slika 11 prikazuje iskanje vseh enobesednih preklopov v korpusu (v polje CQL vpišemo *<seg1>".*"</seg1>*).



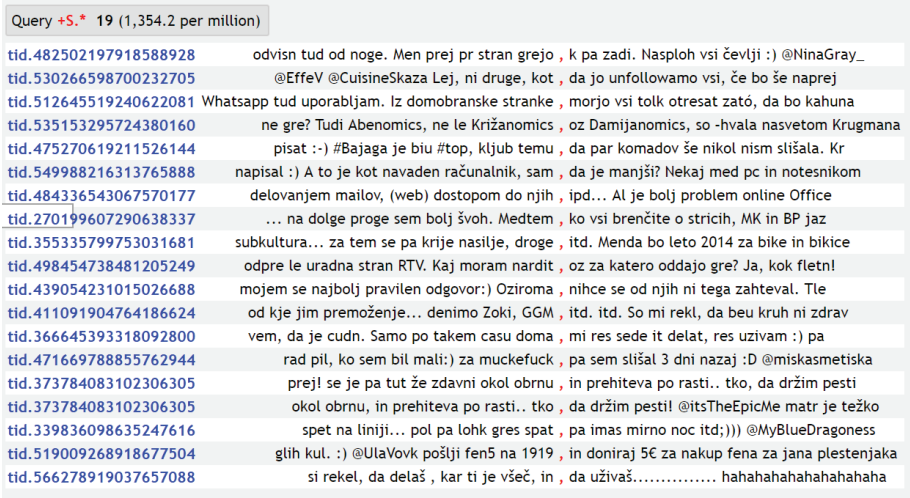
Slika 11: Iskanje enobesednih preklopov po korpusu Janes-Preklop v vmesniku NoSketch Engine.

3.2.3 Janes-Vejica

Janes-Vejica (Popič et al. 2017) je korpus tvitov, v katerih je v skladu z izdelano tipologijo ročno označena nestandardna (ne)raba vejice. Korpus vsebuje naključen vzorec 495 tvitov iz korpusa Janes v0.4, natančneje po 250 iz kategorij z visoko jezikovno nestandardnostjo (T1L3 in T3L3), pri čemer je bilo 5 tvitov izločenih iz prvotnega vzorca, ker so bili nerelevantni za jezikoslovne raziskave. V okviru raziskave na korpusu je bila razvita tudi sistemsko osnovana tipologija za opis nestandardne stave vejice (Popič et al. 2016a).

Glavni namen raziskave na korpusu Janes-Vejica je bil začrtati nadaljnje raziskave stave vejice v slovenščini, zlasti v primerjavi s standardnojezikovnim gradivom, ter določiti, v kolikšni meri raba vejice na Twitterju odstopa od jezikovnega standarda. Rezultati izpostavljajo nekaj novih težišč pri tej problematiki (npr. da je nestandardna raba vejice na Twitterju vezana predvsem na skladenjsko rabo). Izsledke so podrobneje predstavili Popič et al. (2016b).

Korpus Janes-Vejica je prosto in odprto dostopen na repozitoriju CLARIN.SI in njegovih konkordančnikih. Slika 12 prikazuje primer iskanja odvečnih vejic (v polje CQL vpišemo *[seg="\+S.*"]*).


Query +S.* 19 (1,354.2 per million)

tid.482502197918588928	odvisn tud od noge. Men prej pr stran grejo ,	k pa zadi. Nasploh vsi čevlji :) @NinaGray_
tid.530266598700232705	@EffeV @CuisineSkaza Lej, ni druge, kot ,	da jo unfollowamo vsi, če bo še naprej
tid.512645519240622081	Whatsapp tud uporabljam. Iz domobranske stranke ,	morjo vsi tolk otresat zató, da bo kahuna
tid.535153295724380160	ne gre? Tudi Abenomics, ne le Križanomics ,	oz Damijanomics, so -hvala nasvetom Krugmana
tid.475270619211526144	pisat :-) #Bajaga je biu #top, kljub temu ,	da par komadov še nikol nism slišala. Kr
tid.549988216313765888	napisal :) A to je kot navaden računalnik, sam ,	da je manjši? Nekaj med pc in notesnikom
tid.484336543067570177	delovanjem mailov, (web) dostopom do njih ,	ipd... Al je bolj problem online Office
tid.270199607290638337	... na dolge proge sem bolj švoh. Medtem ,	ko vsi brenčite o stricih, MK in BP jaz
tid.355335799753031681	subkultura... za tem se pa krije nasilje, droge ,	itd. Menda bo leto 2014 za bike in bikice
tid.498454738481205249	odpre le uradna stran RTV. Kaj moram nardit ,	oz za katero oddajo gre? Ja, kok fletn!
tid.439054231015026688	mojem se najbolj pravilen odgovor:) Oziroma ,	nihce se od njih ni tega zahteval. Tle
tid.411091904764186624	od kje jim premoženje... denimo Zoki, GGM ,	itd. itd. So mi rekl, da beu kruh ni zdrav
tid.366645393318092800	vem, da je cudn. Samo po takem casu doma ,	mi res sede it delat, res uzivam :) pa
tid.471669788855762944	rad pil, ko sem bil mali:) za muckefuck ,	pa sem slišal 3 dni nazaj :D @miskasmetiska
tid.373784083102306305	prej! se je pa tut že zdavni okol obrnu ,	in prehiteva po rasti.. tko, da držim pesti
tid.373784083102306305	okol obrnu, in prehiteva po rasti.. tko ,	da držim pesti! @itsTheEpicMe matr je težko
tid.339836098635247616	spet na liniji... pol pa lohk gres spat ,	pa imas mirno noc itd;))) @MyBlueDragoness
tid.519009268918677504	glih kul. :) @UlaVovk pošlji fen5 na 1919 ,	in doniraj 5€ za nakup fena za jana plestenjaka
tid.566278919037657088	si rekel, da delaš , kar ti je všeč, in ,	da uživaš..... hahahahahahahahahaha

Slika 12: Iskanje odvečnih vejic v korpusu Janes-Vejica preko vmesnika NoSketch Engine.

3.2.4 Janes-Geo

Janes-Geo (Čibej 2018) je korpus tvitov, ki so bili vzorčeni iz podkorpusa tvitov Janes-Tweet v0.3 glede na metapodatke o regionalni pripadnosti uporabnikov, avtomatsko pripisane s pomočjo geolokacije njihovih tvitov (Čibej in Ljubešić 2015, Čibej 2016). Skupno 321 uporabnikov, ki so vključeni v korpus, je bilo razporejenih v 9 regij: Ljubljano, Maribor in 7 regij, ki predstavljajo glavne narečne skupine (Ramovš 1931): Primorska, Rovtarska, Gorenjska, Dolenjska, Štajerska, Koroška in Panonska. Iz vsake regije je bilo vzorčenih po 500 tvitov (z izjemo Rovtarske, Koroške in Maribora, ki so prispevali 400, 260 in 330 tvitov) s kategorijo nestandardnosti L3 (v nekaterih manjših regijah zaradi pomanjkanja podatkov tudi L2), zato korpus vsebuje skupno nekaj manj kot 4000 tvitov oz. 64.000 pojavnic.

Glavni namen korpusa Janes-Geo je proučevanje medregionalnih jezikovnih razlik v slovenski RPK. Da bi ugotovili, ali se pogostost določenih nestandardnih jezikovnih pojavov razlikuje med uporabniki iz različnih regij, so bili tviti ročno označeni v skladu z za ta namen izdelano tipologijo nestandardnih jezikovnih prvin v slovenski RPK (z nekaterimi izjemami, kot sta raba ločil in skladnja). Oznake delimo na 6 glavnih kategorij: izpusti, transformacije, nestandardno besedje, nestandardno oblikoslovje, variantne različice pogostih besed in drugo.

Podrobnejši izsledki raziskave so predstavljeni v Čibej (2018), korpus pa je na voljo na repozitoriju CLARIN.SI (Čibej et al. 2018).

4 SKLEP

V poglavju smo predstavili postopek izdelave in ročnega označevanja korpusov v okviru projekta JANES s posebnim poudarkom na izdelavi korpusov Janes-Norm in Janes-Tag, ki služita kot zlata standarda za učenje in preizkušanje orodij za tokenizacijo, stavčno segmentacijo in normalizacijo na eni strani ter lematizacijo in oblikoskladenjsko označevanje RPK na drugi. Za odvisnostno označevanje skladenjske ravni jezika je bil pripravljen korpus Janes-Syn. Korpusa Janes-Norm in Janes-Tag sta bila že uporabljena za učenje normalizatorjev in oblikoskladenjskih označevalnikov, prilagojenih za nestandardni jezik (Ljubešić et al. 2018) kot tudi za označevanje celotnega korpusa Janes 1.0 (Erjavec et al. 2018).

Na drugi strani predstavljajo ročno označeni korpusi Janes-Kratko, Janes-Preklop, Janes-Vejica in Janes-Geo podlago za empirične analize jezikovnih prvin, ki vstopajo v središče raziskovalnega interesa s pojavom RPK: (ne)standardna stava vejice, raba tujejezičnih elementov in preklapljanje med jeziki, načini in

pogostost krajšanja jezikovnih elementov v besedilih ter regionalno specifične jezikovne prilagoditve pisne komunikacije.

Poleg ročno označenih korpusov velja kot pomemben projektni doprinos omeniti dobro dokumentirane in prosto dostopne označevalne tipologije ter smernice, ki bodo v korist nadaljnjim raziskavam na področju računalniško posredovane komunikacije. Predlagane rešitve je mogoče prilagoditi tudi za označevanje sorodnih jezikov, kar se je že izkazalo kot uspešno v okviru projekta ReLDI,[8] ki je organiziral označevalno kampanjo za normalizacijo, lematizacijo in oblikoskladenjsko označevanje besedil v okviru razvoja orodij za obdelavo hrvaške in srbske RPK (Miličević et al. 2016), učni množici, izdelani po vzoru Janes-Tag, pa tudi objavil v repozitoriju CLARIN.SI (Ljubešić et al. 2017a, Ljubešić et al. 2017b).

Prosta in odprta dostopnost rezultatov je bila med glavnimi vodili projekta in vsi korpusi so na voljo v repozitoriju CLARIN.SI in v konkordančniku NoSketch Engine. Pri prenosu v slednjega so bile dodatno upoštevane specifike vključenega gradiva, zaradi česar je mogoče s kombinacijo iskanja po korpusnospecifičnih oznakah in njihovega prikaza v konkordančnem nizu dobiti kvalitetnejši vpogled v jezikovne podatke. Nekaj možnosti za iskanje je opisanih na spletni strani projekta JANES, kjer so javno dostopne tudi vse navedene označevalne smernice, vključene pa so tudi v vnose v repozitoriju CLARIN.SI. Vse gradivo je za razliko od dosedanjih praks pri distribuciji korpusov RPK (npr. Frey et al. 2015, Chiari in Canzonetti 2014) na voljo pod zelo liberalno licenco Creative Commons Priznanje Avtorstva, s čimer so viri na voljo za nadaljnje raziskave in za razvoj komercialnih produktov tudi izven okvirov projekta JANES, licenca pa omogoča tudi morebitne izboljšave označevalnih tipologij, smernic in korpusov ter njihovo redistribucijo. Pomembna naloga za nadaljnje delo pa vsekakor ostaja nadaljnja nadgradnja označevalnih orodij, preizkus njihove točnosti na različnih vrstah jezikovnega gradiva ter nadaljnji koraki v smeri njihove optimizacije.

Zahvala

Avtorji prispevka se najlepše zahvaljujejo Kaji Dobrovoljc, Simonu Kreku in Katji Zupan za konstruktivne pripombe pri izdelavi smernic za označevanje korpusov Janes-Norm in Janes-Tag. Posebna zahvala gre vsem označevalcem, ki so sodelovali v označevalnih kampanjah korpusov: Teji Goli, Melaniji Kožar, Vesni Koželj, Poloni Logar, Klari Lubej, Dafne Marko, Barbari Omahen, Eneji Osrajnik, Predragu Petroviću, Poloni Polc, Aleksandri Rajković, Špeli Reher, Izi Škrjanec in Katji Zupan.

8 Regional Linguistic Data Initiative: https://reldi.spur.uzh.ch/.

Literatura

Arhar Holdt, Špela, 2016: Smernice za označevanje z odvisnostnim sistemom JOS: nestandardna slovenščina, v1.0: specifikacije projekta Jezikoslovna analiza nestandardne slovenščine. http://nl.ijs.si/janes/wp-content/uploads/2014/09/Janes-skladnja-v1.0.pdf

Arhar Holdt, Špela, Darja Fišer, Tomaž Erjavec in Simon Krek, 2016: Syntactic annotation of Slovene CMC: first steps. Fišer, Darja in Michael Beißwenger (ur.). *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities, 27-28 September 2016, Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia*. 1st ed. Ljubljana: Znanstvena založba Filozofske fakultete. 3–6.

Arhar Holdt, Špela, Tomaž Erjavec in Darja Fišer, 2017: CMC training corpus Janes-Syn 1.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1086

Arhar Holdt, Špela, 2018: Korpusni pristop k skladnji računalniško posredovane slovenščine. Fišer, Darja (ur.): *Viri, orodja in metode za analizo spletne slovenščine*. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani. 228–253.

Benikova, Darina, Chris Biemann in Marc Reznicek, 2014: NoSta-D Named Entity Annotation for German: Guidelines and Dataset. LREC 2014.

Kalina Bontcheva, Leon Derczynski in Ian Roberts, 2017: Crowdsourcing Named Entity Recognition and Entity Linking Corpora. Ide, Nancy in James Pustejovsky (ur.): *Handbook of Linguistic Annotation*. Dordrecht: Springer. 875–892.

Chiari, Isabella in Alessio Canzonetti, 2014: Le forme della comunicazione mediata dal computer: generi, tipi e standard di annotazione. Garavelli, Enrico in Elina Suomela-Härmä (ur.): *Dal manoscritto al web: canali e modalità di trasmissione dell'italiano. Tecniche, materiali e usi nella storia della lingua*. Firence: Franco Cesati Editore. 595–606.

Čibej, Jaka in Nikola Ljubešić, 2015: "S kje pa si?" – Metapodatki o regionalni pripadnosti uporabnikov družbenega omrežja Twitter. Fišer, Darja (ur.): *Zbornik konference Slovenščina na spletu in v novih medijih*. Ljubljana: Znanstvena založba Filozofske fakultete. 10–14.

Čibej, Jaka, Darja Fišer in Tomaž Erjavec, 2016a: Normalisation, Tokenisation and Sentence Segmentation of Slovene Tweets. *Proceedings of the Workshop on Normalisation and Analysis of Social Media Texts (NormSoMe)*. Portorož: ELRA. 5–10.

Čibej, Jaka, Špela Arhar Holdt, Tomaž Erjavec in Darja Fišer, 2016b: Razvoj učne množice za izboljšano označevanje spletnih besedil. Erjavec, Tomaž in Darja Fišer (ur.): *Proceedings of the Conference on Language Technologies and Digital Humanities*. Ljubljana: Znanstvena založba Filozofske fakultete. 40–46.

Čibej, Jaka, Špela Arhar Holdt, Tomaž Erjavec, Darja Fišer in Katja Zupan, 2016c: *Smernice za označevanje računalniško posredovane komunikacije: tokenizacija, stavčna segmentacija, normalizacija, lematizacija in oblikoskladenjsko označevanje (v1.0)*. http://nl.ijs.si/janes/viri/

Čibej, Jaka, Tomaž Erjavec in Darja Fišer, 2018: *Tweet corpus of Slovene regional language variants Janes-Geo v1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1174

Čibej, Jaka, 2016: Framework for an Analysis of Slovene Regional Language Variants on Twitter. *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities*. Ljubljana: Znanstvena založba Filozofske fakultete. 17–21.

Čibej, Jaka, 2018: Regionalne jezikovne različice v slovenski računalniško posredovani komunikaciji: korpusni pristop z ročno označenim korpusom Janes-Geo. Fišer, Darja (ur.): *Viri, orodja in metode za analizo spletne slovenščine*. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani. 160–197.

Dobrovoljc, Kaja, Simon Krek in Jan Rupnik, 2012: Skladenjski razčlenjevalnik za slovenščino. Erjavec, Tomaž in Jerneja Žganec Gros (ur.): *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. 42–47.

Eckart de Castilho, Richard, Chris Biemann, Iryna Gurevych in Sied Muhie Yimam. 2014. WebAnno: a flexible, web-based annotation tool for CLARIN. Proceedings of the CLARIN Annual Conference (CAC) 2014. Soesterberg, Netherlands. https://www.clarin.eu/sites/default/files/cac2014_submission_6_0.pdf

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

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

Erjavec, Tomaž, Cyprian Laskowski, Jaka Čibej, Darja Fišer in Kaja Dobrovoljc, 2016a: *Navodila za označevanje računalniško posredovane komunikacije v WebAnno (v1.0)*. http://nl.ijs.si/janes/viri/

Erjavec, Tomaž, Jaka Čibej in Darja Fišer, 2016b: Omogočanje dostopa do korpusov slovenskih spletnih besedil v luči pravnih omejitev. *Slovenščina 2.0* 4/2. 189–219.

Erjavec, Tomaž, Darja Fišer, Jaka Čibej, Špela Arhar Holdt, Nikola Ljubešić in Katja Zupan, 2017: *CMC training corpus Janes-Tag 2.0*. Slovenian Language Resource Repository CLARIN.SI. http://hdl.handle.net/11356/1123

Erjavec, Tomaž, Jaka Čibej, Špela Arhar Holdt, Nikola Ljubešić in Darja Fišer, 2016c: Gold-Standard Datasets for Annotation of Slovene Computer-Mediated Communication. *Proceedings of the Tenth Workshop on Recent Advances in Slavonic Natural Languages Processing*, *RASLAN 2016*. Brno, Češka.

Erjavec, Tomaž, Darja Fišer, Jaka Čibej in Špela Arhar Holdt, 2016d*: CMC training corpus Janes-Norm 1.2*. Slovenian Language Resource Repository CLARIN.SI. http://hdl.handle.net/11356/1084

Erjavec, Tomaž, Nikola Ljubešić in Darja Fišer, 2017: Adapting a state-of-the-art tagger for South Slavic languages to non-standard text. *Proceedings of the EACL workshop*. The 6th Workshop on Balto-Slavic Natural Language Processing, April 4, 2017 Valencia, Spain. Stroudsburg: The Association for Computational Linguistics. 60–68. http://bsnlp-2017.cs.helsinki.fi/bsnlp2017-book.pdf

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

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

Frey, Jennifer-Carmen, Aivars Glaznieks in Egon Stemle, 2015: The DiDi Corpus of South Tyrolean CMC Data. *Proceedings of the 2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media*. GSCL2015 (NLP4CMC2015). 1–6.

Goli, Teja, Eneja Osrajnik in Darja Fišer, 2016a: *Strategije krajšanja tvitov: tipologija oznak, v1.0*. http://nl.ijs.si/janes/viri/#Janes-Kratko

Goli, Teja, Eneja Osrajnik in Darja Fišer, 2016b: Analiza krajšanja slovenskih sporočil na družbenem omrežju Twitter. Erjavec, Tomaž in Darja Fišer (ur.): *Proceedings of the Conference on Language Technologies and Digital Humanities*. Ljubljana: Znanstvena založba Filozofske fakultete. 77–82.

Goli, Teja, Eneja Osrajnik in Darja Fišer, 2017: *CMC shortening corpus Janes--Kratko 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1087

Holozan, Peter, Simon Krek, Matej Pivec, Simon Rigač, Simon Rozman in Aleš Velušček, 2008: *Specifikacije za učni korpus*. Kamnik: Projekt »Sporazumevanje v slovenskem jeziku« ESS in MŠŠ. http://www.slovenscina.eu/Vsebine/Sl/Kazalniki/K2.aspx

Krek, Simon, Kaja Dobrovoljc, Tomaž Erjavec, Sara Može, Nina Ledinek in Nanika Holz, 2015: *Training corpus ssj500k 1.4*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1052

Laarmann-Quante, Ronja in Stefanie Dipper, 2016: An Annotation Scheme for the Comparison of Different Genres of Social Media with a Focus on Normalization. *Normalisation and Analysis of Social Media Texts (NormSoMe) Workshop*. 23–30. https://www.linguistics.ruhr-uni-bochum.de/~dipper/pub/normsome16_webVersion.pdf

Ljubešić, Nikola, Katja Zupan, Darja Fišer in Tomaž Erjavec, 2016: Normalising Slovene data: historical texts vs. user-generated content. Dipper, Stefanie, Friedrich Neubarth in Heike Zinsmeister (ur.): *Proceedings of the 13th Conference*

on Natural Language Processing (KONVENS), September 19-21, 2016, Bochum, Germany. 146–155. https://www.linguistics.rub.de/konvens16/pub/19_konvensproc.pdf

Ljubešić, Nikola, Daša Farkaš, Filip Klubička, Tomaž Erjavec, Maja Miličević, Mateja Filko, Denis Kranjčić in Barbara Dujmić, 2017: *Croatian Twitter training corpus ReLDI-NormTag-hr 1.1*. Slovenian Language Resource Repository CLARIN.SI. http://hdl.handle.net/11356/1121

Ljubešić, Nikola, Daša Farkaš, Filip Klubička, Tomaž Erjavec, Maja Miličević in Teodora Vuković, 2017: *Serbian Twitter training corpus ReLDI-NormTag-sr 1.1*. Slovenian Language Resource Repository CLARIN.SI. http://hdl.handle.net/11356/1120

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

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

Ljubešić, Nikola, Tomaž Erjavec in Darja Fišer, 2018: Orodja za procesiranje nestandardne slovenščine. Fišer, Darja (ur.): *Viri, orodja in metode za analizo spletne slovenščine*. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani. 74–99.

Maja Miličević in Nikola Ljubešić, 2016: Tviterasi, tviteraši or twitteraši? Producing and analysing a normalised dataset of Croatian and Serbian tweets. *Slovenščina 2.0* 4/2. 156–188. http://dx.doi.org/10.4312/slo2.0.2016.2.156-188

Poesio, Massimo, Jon Chamberlain in Udo Kruschwitz, 2017: Phrase Detectives. Ide, Nancy in James Pustejovsky (ur): *Handbook of Linguistic Annotation*. Dordrecht: Springer. 1149–1176.

Popič, Damjan, Darja Fišer, Katja Zupan in Polona Logar, 2016b: Raba vejice v uporabniških spletnih vsebinah. *Proceedings of the Conference on Language Technologies & Digital Humanities, September 29th – October 1st, 2016 Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia, 2016*. 106–110.

Popič, Damjan, Katja Zupan in Darja Fišer, 2016a: *Smernice za označevanje nestandardne rabe vejice v uporabniških spletnih vsebinah*. http://nl.ijs.si/janes/viri/#Janes-Vejica

Popič, Damjan, Katja Zupan, Polona Logar, Teja Kavčič, Tomaž Erjavec in Darja Fišer, 2017: *Tweet comma corpus Janes-Vejica 1.0*. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1088

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

Rehbein, Ines, Emiel Visser in Nadine Lestmann, 2013: Discussing best practices for the annotation of Twitter microtext. *Proceedings of the Third Workshop on Annotation of Corpora for Research in the Humanities* (ACRH-3). 73.

Reher, Špela, 2017: *Slovenščina na prepihu: kodno preklapljanje v objavah slovenskih uporabnikov Twitterja. Kvalitativna in kvantitativna analiza tvitov iz korpusa nestandardne slovenščine Janes.* Magistrsko delo. Ljubljana: Filozofska fakulteta.

Reher, Špela, Tomaž Erjavec in Darja Fišer, 2017: *Tweet code-switching corpus Janes-Preklop 1.0*, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1154

Reher, Špela in Darja Fišer, 2018: Kodno preklapljanje v objavah slovenskih uporabnikov Twitterja. Fišer, Darja (ur.): *Viri, orodja in metode za analizo spletne slovenščine*. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani. 294–323.

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

TEI Consortium (ur.): *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/

Ueberwasser, Simone, 2013: Non-standard data in Swiss text messages with a special focus on dialectal forms. Zampieri, Marcos in Sascha Diwersy (ur.): *Non-standard Data Sources in Corpus-based Research*. Aachen: Shaker Verlag. 7–24. https://ueberwasser.eu/UeFiles/uni/Tagungen/2012Koeln/ueberwasser.pdf

Orodja za procesiranje nestandardne slovenščine

Nikola Ljubešić, Tomaž Erjavec, Darja Fišer

Izvleček

Poglavje je posvečeno težavam, povezanim z avtomatskim procesiranjem nestandardnega jezika, in orodjem, ki smo jih razvili za reševanje teh težav. V poglavju obravnavamo merjenje standardnosti besedil, stavčno segmentacijo, normalizacijo, rediakritizacijo, oblikoskladenjsko označevanje in razpoznavanje imenskih entitet. Pokažemo, da se število napak, ki jih povzročijo orodja, naučena na standardnem jeziku, pri uporabi za nestandardni jezik sicer močno poveča, vendar predhodna normalizacija nestandardnih besedil ali prilagoditev orodij zanje bistveno povečata kvaliteto procesiranja nestandardnega jezika. Za ta namen potrebujemo dovolj ročno označenih nestandardnih besedil, ki jih nato uporabimo za učenje ali posodabljanje modelov za nadzorovano strojno učenje.

Ključne besede: jezikovne tehnologije, nestandardni jezik, normalizacija besedil, oblikoskladenjsko označevanje, razpoznavanje imenskih entitet

1 UVOD

Z vidika procesiranja podatkov sodi jezik med zahtevnejše naloge tudi zaradi znatne večpomenskosti. Ta naloga postane še težja, ko besedila odstopajo od pravopisnih in slovničnih norm, kar je pri jeziku na spletu zelo pogost pojav. V slovenščini najpogostejša odstopanja od jezikovne norme predstavljajo opuščanje strešic, nestandardno črkovanje in pogosta uporaba pogovornih izrazov (Fišer et al. 2015).

Tovrstni pojavi močno vplivajo na avtomatsko procesiranje besedil. Gimpel et al. (2011) poročajo, da so pri učenju in testiranju modela na podatkih iz korpusa Wall Street Journal pri pripisovanju oblikoskladenjskih oznak dosegli 97-% točnost, ko so isti model uporabili na besedilih s Twitterja, pa je točnost znašala le 85 %, kar predstavlja petkratno povečanje števila napak. Ljubešić et al. (2017) so izvedli eksperiment z oblikoskladenjskim označevanjem slovenščine, ki ima v primerjavi z angleščino veliko kompleksnejšo označevalsko shemo. Standardni označevalnik je pri testni množici s standardnimi besedili dosegel 94-% točnost, pri nestandardnih besedilih pa le 69-%, kar ponovno predstavlja petkratno povečanje števila napak.

V pričujočem poglavju predstavimo različne pristope za zmanjševanje težav pri procesiranju nestandardnih besedil. V naslednjem razdelku opišemo postopek identifikacije nestandardnih besedil, v tretjem razdelku pa prilagoditve orodja za segmentacijo, tj. ločevanje niza znakov na stavke in pojavnice. V naslednjih štirih razdelkih predstavimo dva pristopa, ki se pogosto uporabljata za nadaljnje procesiranje nestandardnih besedil (Eisenstein, 2013): (1) normalizacija besedil v standardno obliko in uporaba standardnih orodij ter (2) prilagoditev orodij za nestandardne vhodne podatke. Za prvi pristop opišemo postopek normalizacije, ki temelji na statističnem strojnem prevajanju na nivoju znakov, in orodje za rediakritizacijo besedil, za drugi pristop pa prilagoditve orodij za oblikoskladenjsko označevanje in razpoznavanje imenskih entitet. Čeprav se v poglavju posvečamo predvsem slovenščini, smo hkrati razvili tudi orodja za hrvaščino in srbščino, zato navajamo rezultate za vse tri jezike.

Večina orodij, ki jih predstavimo v poglavju, temelji na paradigmi nadzorovanega strojnega učenja, kar pomeni, da moramo podatke za obravnavani problem najprej ročno označiti. Če želimo denimo napovedati standardnost besedila, označimo vzorec besedil glede na njihovo standardnost. Drugi primer je pripisovanje oblikoskladenjskih oznak besedam v vzorcu. Pripisane oznake (npr. stopnja standardnosti ali oblikoskladenjske oznake) imenujemo *odvisne spremenljivke*, specifične spremenljivke, ki jih izluščijo razvita orodja, pa *neodvisne spremenljivke* ali značilke. Orodja nato modelirajo odvisnost med odvisnimi in neodvisnimi spremenljivkami. Ko je proces modeliranja, imenovan tudi *učni proces*, zaključen,

lahko orodja *napovejo* vrednosti odvisnih spremenljivk za nova besedila, in sicer tako, da iz teh besedil izluščijo neodvisne spremenljivke in uporabijo izdelani napovedni model.

2 NAPOVEDOVANJE STANDARDNOSTI

Splošna domneva je, da jezik na spletu odstopa od norme (Crystal 2011), vseeno pa ta pojav navadno ni kvantitativno izmerjen. Da bi to izboljšali, smo razvili posebno orodje – kolikor nam je znano, prvo te vrste –, ki napove stopnjo standardnosti določenega besedila (Ljubešić et al. 2015). Orodje je uporabno z dveh vidikov: (1) izboljša lahko postopek procesiranja korpusov, saj pomaga pri odločitvi, ali za izbrano besedilo uporabimo model za standardni ali nestandardni jezik, in (2) omogoča, da podatek o standardnosti besedil uporabimo v korpusnojezikoslovnih analizah. Drugi vidik je bil tudi naš glavni cilj pri gradnji orodja, torej omogočanje korpusnim jezikoslovcem, da se pri raziskovanju osredotočijo bodisi na standardni bodisi na nestandardni spletni jezik.

2.1 Določanje standardnosti besedila

Pojem standardnosti besedila razumemo kot avtorjevo upoštevanje jezikovnih norm, opisanih v pravopisnih, slovničnih in slogovnih priročnikih. Avtomatsko določanje standardnosti besedila ni lahka naloga. Medtem ko lahko večino pojavov obravnavamo kot eno dimenzijo besedila (kot na primer pri označevanju sentimenta), se izkaže, da standardnost zajema zelo raznolike značilnosti. Nekateri avtorji na primer uporabljajo standardno črkovanje, a opuščajo veliko začetnico. Drugi napravijo veliko tipkarskih napak, tretji pa se držijo standardne uporabe ločil, a uporabljajo pogovorno ali narečno besedišče ter oblikoskladnjo.

Da bi zagotovili pravšnje razmerje med ustreznostjo in kompleksnostjo oznak, smo se odločili za uporabo dveh dimenzij standardnosti: tehnično in jezikovno. Stopnja tehnične standardnosti besedila (okrajšana s »T«) pokriva uporabo velike začetnice, uporabo ločil, prisotnost tipkarskih napak ali ponovljenih znakov v besedi (npr. *napreeej*). Stopnja jezikovne standardnosti (okrajšana z »L«) upošteva črkovanje, besedišče, oblikoslovne lastnosti in besedni red. Za obe dimenziji smo uporabili tri vrednosti: 1 (povsem standardno), 2 (nekoliko nestandardno) in 3 (zelo nestandardno).

Sistem dveh dimenzij s tremi razredi je zasnovan tako, da označevalcem omogoča enostavno pripisovanje stopnje standardnosti, je dovolj informativen za orodja za procesiranje naravnega jezika za izbiro različnih metod normalizacije in hkrati

dovolj relevanten za jezikoslovce, da lahko filtrirajo besedila, ko raziskujejo nestandardni jezik. Primera z ekstremnimi vrednostmi na obeh dimenzijah sta podana v Tabeli 1.

Tabela 1: Stopnja standardnosti dveh besedil.

T1L3	Tehnično povsem standardno, jezikovno zelo nestandardno
Izvirnik	*Ma men se zdi tole s poimenovanji oz s poslovenjenjem imen mest čist mem.*
Standardizirano	*Meni se zdi to s poimenovanji oz. s poslovenjenjem imen mest čisto mimo.*

T3L1	Tehnično zelo nestandardno, jezikovno povsem standardno
Izvirnik	*se pravi,da predvidevaš razveljavitev*
Standardizirano	*Se pravi, da predvidevaš razveljavitev?*

2.2 Izdelava podatkovne množice

Podatke za eksperimente smo zajeli iz vmesne različice korpusa Janes, vsebujejo pa tri vrste besedil: tvite, objave na forumih in komentarje pod spletnimi novicami. Podatkovna množica za eksperimente vsebuje posamezna besedila, ki smo jih vzorčili iz korpusa po postopku, opisanem v nadaljevanju. Besedila so bila nato ročno označena.

Da bi zagotovili uravnoteženo množico podatkov, smo zbrali enak delež (eno tretjino) besedil za vsako vrsto, prav tako pa smo pri objavah na forumih in komentarjih pod novicami vključili enak delež besedil iz vseh šestih virov. Da bi iz korpusa, v katerem prevladuje standardni jezik, pridobili uravnoteženo podatkovno množico glede jezikovne (ne)standardnosti, smo na podlagi postopka normalizacije (Ljubešić et al. 2014) približno ocenili stopnjo (ne)standardnosti besedil. Za vsako besedilo smo izračunali razmerje med številom pojavnic, ki so bile v procesu avtomatske normalizacije spremenjene, in skupno dolžino besed. Besedila z razmerjem 0,1 ali manj smo obravnavali kot standardna, ostala pa kot nestandardna. To merilo smo določili ročno na podlagi pregledovanja podatkov. Podatkovno množico smo nato izdelali tako, da je vsebovala enako število domnevno standardnih in nestandardnih besedil. Poudariti je treba, da smo pri vzorčenju uporabili precej grobo metodo za zagotavljanje uravnotežene podatkovne množice in na ta način zaobšli razmeroma nizek odstotek nestandardnih besedil v celotnem korpusu. Za vzorčenje bi lahko uporabili tudi druga groba merila standardnosti, npr. delež besed zunaj besedišča (angl. *out-of vocabulaty ratio*) glede na leksikon standardnih besednih oblik.

2.3 Ročno označevanje in dobljena podatkovna množica

Označevalcem, študentom 2. stopnje jezikoslovnih smeri, smo predstavili smernice in kriterije za označevanje dveh dimenzij (ne)standardnosti. Vsakemu besedilu so nato pripisali stopnjo standardnosti za obe dimenziji, pri tem pa uporabili vrednost 1 (povsem standardno), 2 (nekoliko nestandardno) ali 3 (zelo nestandardno). Za tristopenjsko lestvico smo se odločili, saj naloga ni (in bi težko bila) zelo natančno definirana.

Po učni fazi, med katero so vsi označevalci označili manjšo množico besedil in prediskutirali rezultate, je označevanje potekalo v dveh kampanjah. V prvem delu je besedila označeval po en označevalec, to pa smo uporabili kot razvojno množico za eksperimente. V drugem delu sta besedila označila po dva označevalca, vzorec pa smo v nadaljevanju uporabili kot testno množico. Končne vrednosti odvisnih spremenljivk za eksperimente smo izračunali kot povprečje vrednosti, ki sta jih pripisala dva označevalca.

2.4 Uporabljene značilke

Za opis tehničnih in jezikovnih značilnosti besedil smo definirali 29 neodvisnih spremenljivk oz. značilk. Za napovedovanje tehnične in jezikovne standardnosti smo uporabili isti nabor značilk. Značilke lahko razdelimo v dve glavni kategoriji.

Značilke na nivoju znakov vključujejo napačno rabo ločil in presledkov, ponavljanje znakov, razmerje med abecednimi in neabecednimi znaki, razmerje med samoglasniki in soglasniki ipd.

Značilke na nivoju pojavnic opisujejo lastnosti besed. Med njimi ločujemo značilke, vezane na niz besed, in značilke, vezane na leksikon, ki temeljijo na zunanjih virih podatkov. Za značilke, vezane na niz besed, smo izračunali delež besed, zapisanih z veliko začetnico ali samimi velikimi črkami, ponovitve besed, delež besed, sestavljenih iz samih soglasnikov, in delež zelo kratkih besed. Za značilke, vezane na leksikon, smo uporabili oblikoslovni leksikon Sloleks (Krek in Erjavec 2009), ki vsebuje slovenske besede z vsemi pregibnimi oblikami. Značilke, ki temeljijo na leksikonu Sloleks, vključujejo delež besed zunaj besedišča, delež besed zunaj besedišča z manjkajočim samoglasnikom, delež kratkih besed zunaj besedišča ipd. Del značilk temelji na korpusu Kres, uravnoteženem korpusu standardne slovenščine (Logar Berginc et al. 2012). Te značilke med drugim vključujejo delež besed zunaj besedišča glede na leksikon pregibnih oblik, ki se v korpusu Kres pojavijo vsaj desetkrat.

2.5 Eksperimenti in rezultati

Za modeliranje odvisnosti med vsako izmed dveh odvisnih spremenljivk (tehnična in jezikovna standardnost) in 29 neodvisnimi spremenljivkami smo testirali številne regresijske modele, s katerimi lahko na podlagi neodvisnih spremenljivk napovemo zvezno vrednost, tj. vrednost med 1 in 3. Na koncu smo se odločili za regresijo z metodo podpornih vektorjev (angl. *SVM regressor*), ki kot jedro uporablja funkcijo RBF (angl. *Radial Basis Function*). Točnost regresorjev smo evalvirali z izračunom povprečne absolutne napake, tj. povprečja razlik med napovedano vrednostjo in vrednostjo, ki so jo pripisali označevalci. Glede na to, da smo računali napako, pomeni nižji rezultat večjo točnost modela. Pri tehnični dimenziji je povprečna absolutna napaka pri najboljšem rezultatu znašala 0,377, pri jezikovni dimenziji pa 0,424, kar nakazuje, da je napovedovanje tehnične standardnosti lažja naloga, vsaj z uporabo našega nabora značilk.

V nadaljevanju smo želeli preveriti, kolikšno izboljšanje modela dosežemo pri uporabi 29 značilk v primerjavi z eno samo, predvsem deležem besed zunaj besedišča, kar je pogost pristop, opisan v literaturi. V ta namen smo zgradili model, ki uporablja samo to značilko. Za tehnično standardnost je povprečna absolutna napaka znašala 0,594, za jezikovno standardnost pa 0,597. Ti rezultati kažejo, da uporaba dodatnih značilk občutno izboljša natančnost modela, predvsem za napovedovanje tehnične standardnosti. Dejstvo, da je ta model boljši pri določanju jezikovne standardnosti, ne preseneča, saj je edina uporabljena značilka (delež besed zunaj besedišča) vezana na jezikovne prvine.

Na koncu smo preverili, kako se naš model razlikuje od osnovnega modela, ki vsakemu besedilu naključno pripiše vrednost med 1 in 3. Za tehnično standardnost znaša povprečna absolutna napaka 0,713, za jezikovno standardnost pa 0,749.

Glede na rezultate lahko sklenemo, da (1) ti še zdaleč niso popolni, vendar (2) so boljši od tistih, ki temeljijo samo na eni značilki (delež besed zunaj besedišča), in (3) veliko boljši od naključnih.

Poleg gradnje modela za napovedovanje standardnosti za slovenščino smo organizirali tudi dodatne označevalske maratone, v okviru katerih smo zbrali podatke za hrvaščino in srbščino, označene na podlagi istih smernic. Fišer et al. (2015) so na podlagi korpusa slovenskih, hrvaških in srbskih tvitov analizirali razlike med napovedano standardnostjo besedil in ugotovili, (1) da so v vseh treh jezikih tviti v veliki večini napisani v standardnem jeziku, (2) da je število nestandardnih besedil najvišje v slovenščini, sledijo jim hrvaška besedila, najmanj pa je srbskih, kar je verjetno posledica različnih stopenj narečne raznolikosti v teh treh jezikih, in da (3) so za slovenske nestandardne tvite značilne predvsem nestandardne pravopisne prvine, medtem ko je nestandardno besedišče pogostejše v hrvaških tvitih, najpogostejše pa v srbskih.

3 SEGMENTACIJA

Delitev besedil na pojavnice (besede in ločila) in stavke na splošno velja za enostavno nalogo. To drži predvsem za jezike, kjer so besede ločene s presledkom, in standardni jezik, kjer so načela za ločevanje besed in stavkov točno določena. Če so ta načela kršena, kar je zelo pogost pojav pri spletnih uporabniških vsebinah, pa postane naloga veliko težja.

3.1 Metoda segmentacije

Za tokenizacijo in stavčno segmentacijo smo razvili orodje v programskem jeziku Python, ki trenutno pokriva slovenščino, hrvaščino in srbščino. Za razliko od ostalih orodij, opisanih v poglavju, ki temeljijo na nadzorovanem strojnem učenju, temelji ta model, kakor večina orodij za segmentacijo, na ročno določenih pravilih, ki so implementirana kot regularni izrazi. Regularni izrazi so definirani na podlagi leksikonov za specifičen jezik, kot so npr. seznami okrajšav. Posebnost tega tokenizatorja je v tem, da ima dva načina: za procesiranje standardnega in nestandardnega jezika.

Način za nestandardni jezik se od standardnega razlikuje v dveh vidikih: (1) definirana pravila so bolj ohlapna kot tista za standardni jezik in (2) dodana so specifična pravila, ki opisujejo pojave, značilne za spletno komunikacijo. Primer takšnega pravila je, da lahko pika konča poved, tudi če se naslednja beseda ne začenja z veliko začetnico ali od pike celo ni ločena s presledkom. Pri tem pa vseeno drži, da se s pojavnico, ki se sicer konča s piko, a je na seznamu okrajšav, ki ne končujejo povedi, npr. *prof.*, poved ne konča. Prav tako je eden izmed dodanih regularnih izrazov za nestandardni način namenjen prepoznavi emotikonov, npr. *:-]*, *:-PPPP*, ^_^ itd.

3.2 Podatkovne množice

Za vse tri jezike smo izvedli evalvacijo nestandardnega načina orodja, pri tem pa smo kot zlati standard uporabili tri podatkovne množice (podrobneje opisane v Čibej et al. 2018), in sicer Janes-Tag 2.0 (Erjavec et al. 2016), ReLDI-NormTagNER-sr 2.0 (Ljubešić et al. 2017a) in ReLDI-NormTagNER-sr 2.0 (Ljubešić et al. 2017a). Poleg drugih nivojev označevanja sta pri teh podatkovnih množicah ročno popravljeni tudi stavčna segmentacija in tokenizacija. Vse tri množice vsebujejo besedila s stopnjami standardnosti T1L1, T1L3, T3L1 in T3L3.

3.3 Eksperimenti in rezultati

Rezultati testiranja orodja na treh podatkovnih množicah so podani v Tabeli 2. Pri računanju natančnosti stavčne segmentacije smo uporabili strogo merilo, in sicer penaliziranje tako v primeru, ko sistem ni ločil povedi, pa bi moral, kot tudi v primeru, ko je izvedel segmentacijo, pa je ne bi smel. Nasprotno smo orodje pri tokenizaciji penalizirali samo za vsako izvirno pojavnico, ki je bila tokenizirana napačno.

Tabela 2: Evalvacija segmentacije.

Jezik	Povedi	Napačnih	Natančnost	Pojavnic	Napačnih	Natančnost
sl	19.009	2.497	86,86 %	184.896	2.067	98,88 %
hr	7.942	1.328	83,28 %	89.208	562	99,37 %
sr	6.902	733	89,38 %	91.853	546	99,41 %

Kot je razvidno iz tabele, znaša natančnost stavčne segmentacije za slovenščino skoraj 87 %, za srbščino je nekoliko višja, za hrvaščino pa nižja. Pri pregledovanju rezultatov za slovenščino se izkaže, da je večina napak posledica povedi, ki se končajo z emotikonom namesto s končnim ločilom, ta pojav pa je izven obsega orodja. Orodje se moti tudi pri nekaterih neprepoznanih okrajšavah, saj končno piko razume kot signal za konec povedi.

Pri tokenizaciji so rezultati najslabši za slovenščino (natančnost znaša tik pod 99 %), najboljši pa ponovno za srbščino z natančnostjo 99,4 %. Pri pregledovanju rezultatov za slovenščino se izkaže, da se največ napak pojavi zaradi vezajev, npr. enota »*nm-lj*«, ki bi morala biti ločena na tri pojavnice. Do napak pride tudi pri napačno tokeniziranih emotikonih in spletnih naslovih – čeprav orodje vsebuje regularne izraze za ti dve vrsti pojavnic, ne pokriva vseh oblik, ki se pojavijo v besedilih. Kot je bilo omenjeno, določene napake povzročijo tudi okrajšave, ki niso zajete v leksikonu orodja.

4 NORMALIZACIJA

Kot smo omenili v uvodu, se za prilagajanje jezikovnih tehnologij za procesiranje nestandardnih besedil uporabljata dva glavna pristopa: (1) normalizacija besedil in uporaba orodij za standardni jezik ter (2) prilagoditev orodij. Prvi pristop je bolj ekonomičen, saj celotnemu postopku zgolj dodamo eno komponento, medtem ko drugi pristop zahteva, da prilagodimo vsak korak procesiranja besedil.

Normalizacija besedil ima še dodatno prednost, ki je pomembna predvsem za (korpusne) jezikoslovce: normalizirani korpus omogoča iskanje besed, ne da bi upoštevali ali sploh poznali vse različice zapisa.

4.1 Metoda normalizacije

Pri projektu JANES smo za normalizacijo besednih oblik uporabili pristop statističnega stojnega prevajanja na nivoju znakov. Pri tem ne gre za prevajanje besed in zvez iz izvornega v ciljni jezik, temveč za pretvorbo znakov in nizov znakov v nestandardnih različicah besed v znake in nize znakov v standardnih različicah.

Model za pretvorbo izvornih nizov znakov v ciljne nize se imenuje *prevodni model*, zgrajen (naučen) pa je na zbirki paralelnih podatkov, tj. obstoječih prevodov ali normalizacij. V paradigmi statističnega strojnega prevajanja obstaja tudi drug zelo uporaben vir informacij, in sicer verjetnostni model različnih nizov v ciljnem jeziku, ki ga imenujemo *jezikovni model*. Glede na to, da je pri normalizaciji ciljni jezik pravzaprav standardni jezik, za gradnjo takšnega jezikovnega modela ni težko zbrati velike količine podatkov. Za slovenščino lahko uporabimo korpus Kres, korpus Gigafida (Logar, 2012) ali slovenski spletni korpus slWaC (Erjavec et al. 2015). Medtem ko se prevodni model uporablja za generiranje hipotez za normalizacijo, se oba modela, skupaj z vrsto dodatnih manjših modelov, uporabljata za identifikacijo najbolj verjetne normalizirane oblike.

Primer iz nestandardne slovenščine, ki se ga model v tej paradigmi nauči zelo hitro, je nestandardna končnica pri deležnikih na *-l* določenih glagolov (*naredil* vs. *naredu*). Če imamo v učni množici primere takšnih transformacij (besede, ki se končajo na »du«, pretvorjene v besede, ki se končajo na »dil«), lahko prevodni model zlahka generira hipotezo *pobegnil* za nestandardno obliko *pobegnu* in ji pripiše precej visoko verjetnost, čeprav te pojavnice v učni množici še ni videl. Jezikovni model lahko pripisano verjetnost še poveča, saj je verjetnost za pojavitev niza znakov *pobegnil* v standardni slovenščini zelo velika in veliko večja kot verjetnost za niz *pobegnu* ali katero drugo možno hipotezo.

Prednost pristopov nadzorovanega strojnega učenja je v tem, da lahko isti učni algoritem uporabimo za reševanje podobnih problemov, če le imamo na voljo učne podatke. Med eksperimenti smo se poleg normalizacije spletnih uporabniških vsebin ukvarjali tudi s prevajanjem zgodovinskih besedil v sodobni jezik. Drugi problemi, ki bi jih lahko reševali na podoben način, vključujejo prevajanje med narečji, odpravljanje pravopisnih in slovničnih napak pa tudi popravljanje napak, ki jih napravijo osebe z različnimi jezikovnimi motnjami.

V nadaljevanju predstavimo rezultate eksperimentov, ki so podrobno opisani v Ljubešić et al. (2016). V članku smo identificirali optimalni način za uporabo strojnega prevajanja na nivoju znakov tako pri normalizaciji spletnih uporabniških vsebin kot pri prevajanju zgodovinskih besedil v sodobni jezik. Orodje, ki smo ga uporabili za slovenščino, smo uporabili tudi pri eksperimentih za prevajanje narečne švicarske nemščine v metašvicarsko nemščino, ki se precej približa standardni, opisanih v Scherrer in Ljubešić (2016).

4.2 Podatkovne množice

Eksperimente smo izvedli na podatkih iz zgodnje različice korpusa Janes-Norm (glej Čibej et al. 2018), ki je vseboval 1.000 tvitov, označenih kot povsem standardnih (L1), in 1.000 tvitov, označenih kot zelo nestandardnih (L3). Tako je podatkovna množica za kategorijo spletnih uporabniških vsebin vključevala za normalizacijo težjo množico podatkov L3, kot tudi lažjo množico podatkov L1.

Dodatno smo izvedli eksperimente z zgodovinskimi besedili, kjer smo uporabili ročno označeni korpus starejše slovenščine goo300k (Erjavec, 2015), ki vključuje transkripcije 1.100 strani besedil (približno 300.000 pojavnic), vzorčenih iz 88 knjig in enega časopisa, izdanih med letoma 1584 in 1899. Vsaki pojavnici v korpusu je pripisana normalizirana (sodobna) besedna oblika. Za potrebe eksperimenta korpus ločimo na dva dela: težjo in lažjo množico podatkov, pri čemer težja vsebuje besedila, napisana v bohoričici, ki se precej razlikuje od sodobnega jezika, lažja pa gajici in je bližje sodobnemu jeziku.

4.3 Eksperimenti in rezultati

Z eksperimenti smo skušali odgovoriti na dve glavni raziskovalni vprašanji: (1) ali obstaja en sam model statističnega strojnega prevajanja na nivoju znakov, ki je najbolj učinkovit za normalizacijo ne glede na to, ali normaliziramo spletne uporabniške vsebine ali zgodovinska besedila, in (2) ali lahko izboljšamo tradicionalno normalizacijo po pojavnicah tako, da uporabimo prevajanje celotnih segmentov in s tem upoštevamo kontekst.

Drugo vprašanje se nanaša na dejstvo, da večina sodobnih pristopov temelji na normalizaciji na nivoju pojavnic, kar pomeni, da ne upoštevajo konteksta, v katerem se besede pojavljajo. Tovrstni pristopi ne sledijo jezikovni intuiciji o pomembnosti konteksta za ustrezno normalizacijo.

Za evalvacijo modela smo uporabili prilagojeno Levenshteinovo razdaljo, tj. odstotek znakov, ki bi jih morali zamenjati, da bi bila normalizacija identična referenčni normalizaciji. Kot referenco smo uporabili mero *Leave-As-Is* (LAI), tj. proces, ki vhodnih podatkov ne spremeni. Razlog za uporabo te reference je v tem, da lahko za določeno podatkovno množico tako izmerimo kompleksnost problema in dodani odstotek primerov, ki smo jih uspeli razrešiti z določenim pristopom. Rezultati eksperimentov so prikazani v Tabeli 3.

Tabela 3: Vrednosti relativne Levenshteinove razdalje za normalizacijo težjih in lažjih primerov pri spletnih uporabniških vsebinah in pri starejši slovenščini.

	LAI	pojavnica	segment	+JM pojavnica	+JM segment
L3	5,15	2,19	2,12	1,76	1,58
L1	0,75	0,41	0,43	0,34	0,38
Bohorič	17,63	1,55	1,92	1,51	1,33
Gaj	3,13	1,01	1,15	0,91	0,93

V prvem stolpcu so prikazane podatkovne množice, v drugem stolpcu pa mera LAI, ki kaže, kako daleč je določena podatkovna množica od standardizirane različice. Pri spletnih uporabniških vsebinah mora biti za manj standardno množico (L3) popravljenih 5 % znakov, za bolj standardno množico (L1) pa manj kot 1 %. Pri zgodovinskih besedilih mora biti za množico z bohoričico spremenjenih 18 % znakov, pri množici z gajico pa 3 % znakov.

Učenje normalizacije na nivoju pojavnic (stolpec *pojavnica*) je v primerjavi z referenco (LAI) zelo uspešno. Pri najmanj standardnih podatkih (bohoričica) je kljub relativno majhni učni množici napačnih zgolj 9 % transformacij, 91 % primerov pa je že razrešenih (relativna Levenshteinova razdalja se s 17,63 zniža na 1,55). Pri preostalih treh podatkovnih množicah število napak prav tako pomembno upade, čeprav v manjši meri. Zanimivo je, da je prevajanje celotnih segmentov (stolpec *segment*) v večini primerov manj uspešno kot normalizacija posameznih pojavnic. Razlog za to bi lahko bil v tem, da sta učna množica in jezikovni model precej majhna.

Zadnja dva stolpca prikazujeta rezultate, kjer smo uporabili jezikovne modele, naučene na velikih zbirkah večinoma standardnih, sodobnih podatkov. Prva ugotovitev je, da uporaba dodatnih jezikovnih modelov izboljša rezultate pri obeh nalogah. Še zanimivejša je ugotovitev, da je pri težjih množicah podatkov (L3 in bohoričica) normalizacija na nivoju segmentov bolj uspešna kot normalizacija na nivoju pojavnic. Odstotek primerov, ki so bili razrešeni, se giblje med 49 % in 92 %.

Rezultate za slovenščino smo primerjali z rezultati za narečno švicarsko nemščino, opisanimi v Scherrer in Ljubešić (2016). Pri uporabi normalizacije celotnih povedi namesto posameznih pojavnic se je pri tej nalogi število napak zmanjšalo za 20 %. Pri podatkovni množici z bohoričico se je število napak pri istem scenariju zmanjšalo za 12 %, pri podatkovni množici L3 pa za 10 %. Nasprotno se je za podatkovno množico z gajico in podatkovno množico L1 pri prevajanju celotnih povedi število napak povečalo. Pregledali smo vse podatkovne množice in predlagali metriko, izračunano na pojavnicah v izvornem jeziku, ki bi lahko pokazala, ali bi normalizacija na nivoju segmentov izboljšala rezultate: izračunali smo število pojavnic, ki imajo več možnih normalizacij, izbrana pa mora biti manj pogosta. Na ta način smo izmerili odstotek pojavnic, pri katerih je za pravilno normalizacijo nujen kontekst, saj bi bila pri normalizaciji na nivoju pojavnic izbrana napačna, najpogostejša normalizacija. Pri švicarski množici je takšnih pojavnic 7 %, medtem ko je pri množici z bohoričico 5,5 %, pri množici z gajico 3,5 %, pri množici L3 6,1 % in pri množici L1 1,6 % pojavnic, ki za pravilno normalizacijo zahtevajo kontekst. Definirali smo hevristiko, ki predvideva uporabo normalizacije celotnih povedi v primeru, da več kot 4 % pojavnic ne moremo pravilno normalizirati brez konteksta.

Najbolj uspešen sistem za vse opisane eksperimente je bil objavljen kot orodje, imenovano *csmtiser*. S tem orodjem smo sodelovali tudi na tekmovanju CLIN2017 v prevajanju starejše nizozemščine v sodobno nizozemščino, kjer smo med 8 evropskimi univerzami pri normalizaciji dosegli najboljši rezultat (Tjong Kim Sang et al. 2017).

5 REDIAKRITIZACIJA

Pri računalniško posredovani komunikaciji uporabniki, ki pišejo v latinici, znake s strešicami iz ergonomskih razlogov pogosto zamenjajo z ustreznicami iz nabora ASCII, predvsem pri tipkanju na tablicah ali pametnih telefonih. Branje takšnih besedil ljudem praviloma ne povzroča težav, računalniško procesiranje pa je zelo zahtevno, saj je veliko besed brez šumnikov neznanih ali dvoumnih. Iz tega razloga je rediakritizacija besedil zelo aktivno raziskovalno področje. Razvili smo orodje za avtomatsko rediakritizacijo (Ljubešić et al. 2016), ki smo ga naučili in testirali na slovenskih, hrvaških in srbskih besedilih (srbska besedila so bila napisana v latinici). Tega orodja za končno normalizacijo korpusa Janes nismo uporabili (za simultano rediakritizacijo in standardizacijo smo uporabili orodje *csmtizer*, glej razdelek 4), vseeno pa je v nadaljevanju podrobneje opisano, saj je uporabno za kakovostno rediakritizacijo, računalniško manj kompleksno in zato tudi hitrejše, prav tako pa ga je lažje namestiti kot celotno normalizacijsko orodje.

5.1 Podatkovne množice

Za učenje modela smo za vse tri jezike uporabili tri vrste besedil: besedila z Wikipedije, spletna besedila in nestandardna besedila s Twitterja. Ker smo želeli, da orodje pokriva tako besedila, napisana v standardnem jeziku, kot tista, ki so pogosto nestandardna, smo za standardni nabor kot testno množico uporabili besedila z Wikipedije, za nestandardni pa besedila s Twitterja. Korpuse za Wikipedijo smo zgradili s pomočjo splošne skripte za zajem besedil z Wikipedije. Obdržali smo samo povedi, ki vsebujejo 100 znakov ali več, in na ta način odstranili večino preostalih šumnih podatkov. Podatkovna množica za slovenščino je vsebovala približno 20 milijonov, za hrvaščino 28 milijonov in za srbščino skoraj 34 milijonov besed.

Za gradnjo spletnih korpusov smo uporabili korpuse WaC za tri obravnavane jezike (Ljubešić in Klubička 2014; Erjavec in Ljubešić 2014). Ker smo želeli pridobiti dobro učno množico, besedila na spletu pa so lahko napisana tudi brez strešic, smo vključili zgolj tista besedila, pri katerih vsaj 20 % pojavnic vsebuje diakritična znamenja. Čeprav gre za precej strogo merilo, na podlagi katerega smo izključili tudi besedila s pravilno rabo strešic, je spletnih besedil toliko, da so bile dobljene podatkovne množice vseeno zelo velike. Podatkovna množica s spletnimi besedili vsebuje za slovenščino več kot 130 milijonov, za hrvaščino skoraj 270 milijonov in za srbščino 103 milijone besed. Uporabljeno merilo z 20 % šumnikov je bilo definirano na podlagi ročnega pregledovanja podatkov in z glavnim ciljem zagotoviti čim večjo natančnost modela.

Za gradnjo korpusov s Twitterja smo uporabili veliko zbirko tvitov, ki smo jih zbirali od sredine leta 2013 z orodjem TweetCat (Ljubešić et al. 2014). Ker sta hrvaščina in srbščina zelo podobna jezika, smo za razlikovanje med hrvaškimi in srbskimi uporabniki uporabili namensko razvito orodje (Ljubešić in Kranjčić 2015). Ker smo želeli, da podatkovne množice s Twitterja vsebujejo predvsem nestandardna besedila, smo vključili samo tiste tvite, ki so bili avtomatsko označeni kot nekoliko ali zelo jezikovno nestandardni (L2 in L3). Na koncu smo izključili še vse tvite, pri katerih manj kot 10 % pojavnic vsebuje diakritična znamenja. V tem primeru smo uporabili manj strogo merilo kot pri spletnih besedilih, saj je število podatkov s Twitterja veliko nižje. Podatkovna množica za slovenščino vsebuje le 6,7 milijona, za hrvaščino 1,9 milijona in za srbščino 13,7 milijona besed. Merilo (10 %) je bilo tako kot prej definirano na podlagi ročnega pregledovanja podatkov, glavna cilja pa sta bila visoka natančnost in priklic.

Podatki z Wikipedije so bili ločeni na stavke in pojavnice, pri podatkih s Twitterja pa smo kot osnovno enoto uporabili celoten tvit. Prav tako smo za vse pojavnice uporabili zapis z malimi črkami in na ta način zmanjšali razpršenost podatkov. Pri

začetnih eksperimentih se je namreč izkazalo, da ohranitev velikih in malih črk za obravnavane jezike nima informativne vrednosti. Pretvorba pojavnic v prvotni zapis z velikimi in malimi črkami po rediakritizaciji ne predstavlja težav, saj se v večini primerov število črk v pojavnici ne spremeni.

Podatke smo razdelili na učne, razvojne in testne množice in iz učnih množic odstranili vse duplikate. Da bi sistem prilagodili za standardne in nestandardne podatke, smo iz zbirk besedil z Wikipedije in Twitterja izločili razvojne množice, ki so vsebovale po 10.000 besedil za vsako vrsto besedila in jezik. Za testiranje modela na standardnih podatkih smo uporabili dodatnih 10.000 besedil iz podatkovnih množic za Wikipedijo. Glede na to, da so med filtriranimi podatki lahko napake oz. izpusti strešic, smo za testiranje na nestandardnih podatkih za vsak jezik izdelali testno množico po 2.000 tvitov, ki so jih pregledali jezikoslovci. Te tvite smo zajeli iz prvotne množice tvitov, vendar se pri njih nismo držali omejitve števila pojavnic, ki vsebujejo diakritična znamenja, saj smo želeli zagotoviti, da bo testna množica reprezentativna za nestandardni jezik na splošno. Iz razvojnih in testnih množic prav tako nismo odstranili duplikatov.

5.2 Eksperimenti in rezultati

Rediakritizacijo smo definirali kot prevajalski problem na nivoju pojavnic, kjer je vsaka pojavnica »prevedena« v različico s pravilno postavljenimi strešicami na šumnikih. Za učenje sistema smo iz izvornih besedil preprosto odstranili strešice in tako ustvarili paralelno podatkovno množico, poravnano na nivoju pojavnic. Za reševanje tega prevajalskega problema smo uporabili dva pristopa:

- *leksikonski pristop* (leksikon) – uporaba najpogostejšega prevoda iz učne množice
- *korpusni pristop* (TM+LM) – kombiniranje informacij o tem, kako verjeten je prevod (*translation model* – TM) in kako verjetna je pojavitev prevoda v določenem kontekstu (*language model* – LM), z uporabo log-linearnega modela, ocenjenega na podlagi razvojne množice

Pri obeh pristopih smo za ocenjevanje verjetnosti prevoda uporabili oceno največje verjetnosti (angl. *maximum likelihood estimate*) za obliko s strešicami glede na obliko brez strešic. Za ocenjevanje kontekstualne verjetnosti smo uporabili dobro poznano orodje za modeliranje jezika KenLM (Heafield, 2011) s privzetimi parametri.

Za ocenjevanje dveh parametrov log-linearnega modela smo izvedli izčrpno iskanje med vsemi kombinacijami v intervalu [0,0, 1,0] s korakom 0,1. Kot ciljno

funkcijo smo uporabili točnost pojavnice. Za glajenje podatkov v preiskovalnem prostoru smo za vsako kombinacijo parametrov povprečili rezultate za en korak večjih in manjših vrednosti parametra. Tako smo za kombinacijo parametrov (0,2, 0,3) vzeli povprečje meritev naslednjih kombinacij: (0,1, 0,3), (0,3, 0,3), (0,2, 0,3), (0,2, 0,2) in (0,2, 0,4).

Tabela 4 prikazuje rezultate najuspešnejšega pristopa, in sicer metode TM+LM, naučene na vseh podatkih, tj. skupku podatkov z Wikipedije, Twitterja in spleta. Podatki označujejo točnost na besedo, ne glede na to, ali je ta beseda kandidat za rediakritizacijo ali ne. Kot je razvidno iz preglednice, so rezultati precej dobri; v vseh primerih je stopnja napake manjša kot 1 %. Rezultati prav tako kažejo, da naše korpusne metode delujejo veliko bolje kot edino prosto dostopno orodje za rediakritizacijo, imenovano *charlifter*.[1]

Na splošno se testna množica z Wikipedije izkaže za najenostavnejšo in za slovenščino doseže najboljše rezultate, medtem ko za hrvaščino najboljše rezultate dosežejo podatki s Twitterja. V zadnji vrstici je prikazano, kolikšno zmanjšanje napake zagotovi naša metoda v primerjavi z enostavno metodo z leksikonom, kjer za vsako besedo z morebitnimi diakritičnimi znamenji poiščemo besedo v Sloleksu in ji v skladu s tem dodamo strešice. Pri naši metodi stopnja napake znatno upade, in sicer zagotovi od skoraj četrtine do več kot polovice manj napak.

Tabela 4: Rezultati uporabe orodja za rediakritizacijo z najboljšim modelom za različne podatkovne množice in jezike. Zadnja vrstica prikazuje zmanjšanje števila napak v primerjavi z enostavno metodo z leksikonom.

	Wiki-sl	Wiki-hr	Wiki-sr	Tweet-sl	Tweet-hr	Tweet-sr
brez intervencije	0,8615	0,8614	0,8844	0,8715	0,8397	0,8730
točnost *charlifter*	0,9790	0,9674	0,9706	0,9508	0,9436	0,9330
točnost TM+LM	0,9962	0,9957	0,9947	0,9912	0,9938	0,9917
Δ metode z leksikonom	32,81%	30,26%	43,28%	25,30%	22,43%	51,11%

Da bi dobili vpogled v naravo napak, ki se pojavljajo pri najuspešnejšem modelu, smo opravili tudi analizo napak za podatkovni množici z besedili z Wikipedije in Twitterja v slovenščini. V vsaki podatkovni množici smo ročno pregledali 100 napak in jih razvrstili v 9 kategorij, ki so opisane v Tabeli 5.

1 https://sourceforge.net/projects/lingala/files/charlifter/

Tabela 5: Rezultati ročne analize napak za slovenščino s primeri.

Tip napake	Wikipedija	Twitter	Primeri
lastno ime	30	6	*petar *šegvić*, mesto *kiš**
redka beseda	28	6	*osemkotno *užleblјenje*, *šamaševa* tablica*
dvoumna beseda	21	37	*šoja / soja, teza / teža*
tuja beseda	8	3	*DE *das* antlitz der erde, kaj potegnit za your *case**
tipkarska napaka	6	6	*naj *počakojo* nasprotnika, operacijski *ojačvealniki**
težava s tokenizacijo	4	31	*zaostajali ali *bilispuščeni* → bili spuščeni, Wiki: en - *sipad* - zid - ana *laraški* → en-sipad-zin-ana laraški*
pravilna različica	3	3	*inštitucija / institucija, špirala / spirala*
ponovljene črke	0	5	*kolk sem *žiivčna*, *sončči* *špeeegliiii**
napaka testne množice	0	3	*član los *angaleske* skupine → član losangeleške skupine, pa *se* hipster si → pa še hipster si*
	100	100	

Rezultati analize kažejo, da se razlogi za napake v obeh podatkovnih množicah precej razlikujejo. Pri Wikipediji največjo težavo povzročajo lastna imena, ki jih v učni množici ni bilo (30 %, npr. *japonski umetnik Hirošige, hrvaški pevec Vinko Coce, sumersko mesto Ešnuna, Jangončani – prebivalci burmanskega mesta Jangon*), in redke besede, značilne za specifično domeno, pogosto izpeljane iz tujih besed ali lastnih imen (28 %, npr. *senponški škof, pižanski koncil, komodoški varan, Jastrebova stela*). Pri tvitih sta glavna razloga za napačno rediakritizacijo dvoumnost besed – besede, ki obstajajo s strešicami in brez njih (37 %, npr. *se/še, reci/reči, car/čar, nas/naš, poklice/pokliče*), in izpuščanje presledkov (31 %, npr. *splohnisemnatekocem, #sanjskikrozek*) bodisi za varčevanje s prostorom in časom bodisi kot pogost fenomen pri ključnikih. Najhujše napake se pojavijo pri dvoumnih besedah (21 % pri standardnih besedilih in 37 % pri nestandardnih besedilih), zato bi morali v prihodnje največ pozornosti nameniti odpravljanju teh napak.

6 OBLIKOSKLADENJSKO OZNAČEVANJE

Prednosti normalizacije nestandardnih podatkov in uporabe standardnega pristopa procesiranja besedil so bile opisane, v naslednjih dveh poglavjih pa podamo

še argumente za prilagoditev orodij za nestandardni jezik. Najpomembnejši argument za neposredno procesiranje nestandardnega jezika je ta, da s predhodno normalizacijo izgubimo določene informacije in zato kasnejša označevanja ne morejo delovati tako dobro kot z uporabo modelov, naučenih na nestandardnih podatkih. Vsakršno avtomatsko procesiranje prav tako pripelje do napak, ki se prenesejo v nadaljnje faze procesiranja, hkrati pa lahko negativno vplivajo na procesiranje sosednjih ali kako drugače povezanih pojavnic. Prilagajanje orodij za nestandardni jezik prav tako ni nujno zelo potratno, saj že majhna učna množica domensko specifičnih/nestandardnih besedil zadošča, da se sistem nauči vsaj zelo pogostih fenomenov.

V tem razdelku predstavimo pristop prilagajanja najsodobnejšega označevalnika slovenskih (Ljubešić in Erjavec, 2016), hrvaških in srbskih besedil (Ljubešić et al. 2016a) za nestandardni jezik s primerom za slovenščino (Ljubešić et al. 2017). Vse metode temeljijo na učenju pogojnih naključnih polj, tj. metodi za učenje zaporednega označevanja.

Izvedli smo dve vrsti prilagoditev: (1) z vključitvijo nestandardnih učnih podatkov (nadzorovana prilagoditev) in (2) z vključitvijo dodatnih informacij, naučenih iz velikih zbirk surovih nestandardnih podatkov, v obliki Brownovih gruč (nenadzorovana prilagoditev).

Brownovo gručenje (Brown et al. 1992) je tehnika hierarhičnega gručenja besed, tj. procesa razvrščanja besed v skupine glede na podobnost konteksta, v katerem se pojavljajo. Ta nenadzorovana metoda (ne zahteva ročno označenih podatkov) je pri procesiranju naravnega jezika zelo priljubljena, saj predstavlja zelo poceni način prilagajanja orodij za različne domene.

V Tabeli 6 so prikazani rezultati Brownovega gručenja za slovenski spletni korpus. Prva gruča vsebuje niz nedoločnikov, med katerimi so nekateri zapisani v nestandardni okrajšani obliki brez končnega *-i*. Z informacijo, da je določena beseda v tej gruči, označevalniku omogočimo, da se nauči odvisnosti med oznako za glagolsko nedoločnost in identifikatorjem te gruče. Na ta način lahko označevalnik na podlagi konteksta in informacije o gruči za fenomen, ki ga še ni videl, npr. okrajšano različico nedoločnika, napove pravilno oznako. Druga gruča vsebuje različne oblike črkovanja za osebni zaimek, tretja in četrta pa standardne in nestandardne oblike prislovov. Kot lahko vidimo, vsebuje četrta gruča tudi oblike, ki niso prislovi, kar pomeni, da s to metodo pridobimo zgolj približne rezultate. Vseeno pa te informacije kljub šumnim podatkom pripomorejo k izboljšanju orodij za procesiranje naravnega jezika.

Tabela 6: Primeri Brownovih gruč, izračunanih iz slovenskega spletnega korpusa.

narediti storiti naučiti napisati početi izgubiti naredit napraviti poskusiti zaslužiti pojesti obleči tvegati plačat postoriti shujšati žrtvovati narest prebrat popiti zamenjat nardit rešit skuhati počet spremenit zapraviti popravit potrpeti privarčevati poizkusiti pojest spiti menjat naučit ukreniti požreti prodat izmisliti zmenit nastavit dodat pripravit uredit
jaz jst jest js jz
marsikaj karkoli kej kj karkol kàj
itak tud kr tut kao skoz ziher tle lahk skor tm zdele prov valda tuki skos dons zihr lohk una nonstop edin dans prou loh itaq napisu non-stop valjda kle počas ponavad već kmal nardil tamo clo nešto prec čak tukej opet lohka ratala verjetn

6.1 Podatkovne množice

Za nadzorovano prilagoditev orodij smo uporabili podatkovno množico Janes-Tag (glej poglavje Čibej et al. 2018). Za namene eksperimentov smo podatke razdelili na deleže 80 : 10 : 10, ki so služili kot učna, razvojna in testna množica.

Za računanje Brownovih gruč smo uporabili (1) 1,2 milijarde pojavnic iz spletnega korpusa za slovenščino slWaC v2.0 (Erjavec et al. 2015), (2) korpus Janes v.04 in (3) skupek obeh korpusov. Za vsak vir smo besede, ki se pojavijo vsaj 50-krat, združili v 2.000 gruč.

Na koncu smo preverili tudi, kako se rezultati spremenijo, če v nabor značilk vključimo normalizirane oblike. Za normalizacijo smo uporabili delež podatkovne množice Janes-Norm (opisana v poglavju Čibej et al. 2018), ki se ne prekriva z množico Janes-Tag, saj smo želeli zagotoviti, da učenje normalizatorja ne bo potekalo na istih podatkih, ki jih mora kasneje normalizirati.

6.2 Značilke

Osnovne značilke, ki smo jih uporabili za označevalnik, so naslednje: pojavnice, zapisane z malimi črkami, na položajih {-3, -2 ... 3}; pripone obravnavane pojavnice (pojavnica na položaju 0, ki jo sistem trenutno označuje) dolžine {1, 2, 3, 4}; hipoteze za oznako, ki jih pridobimo iz oblikoslovnega leksikona, za pojavnice na položajih {-2, -1 ... 2} in predstavitve obravnavanih pojavnic, ki zaznamujejo, ali pojavnica vsebuje številke, velike črke, male črke ali druge simbole (npr. pojavnica *Gr8t* je zaznamovana z »uldl«, saj je sestavljena iz velike

črke (*upper*), male črke (*lower*), številke (*digit*) in male črke), in ali se pojavi na začetku povedi.

Za dodajanje informacij o Brownovem gručenju smo v skupino značilk vključili različne informacije o gručah, npr. celotni identifikator gruč in binarne poti različnih dolžin v binarnem drevesu, in na ta način zagotovili informacije o gručenje z različnimi nivoji podrobnosti.

Na koncu smo v nabor značilk dodali tudi podatke o normalizaciji, in sicer hipoteze za ustrezno oblikoskladenjsko oznako, izračunane iz oblikoslovnega leksikona na podlagi normalizirane oblike pojavnice.

6.3 Eksperimenti in rezultati

S prvim eksperimentom smo merili, kako se stopnja napake poveča, če s standardnim modelom označimo nestandardne podatke. Za standardno testno množico je bila točnost (odstotek pravilno označenih pojavnic) 94-%, pri nestandardnih podatkih pa je model dosegel le 69-% točnost. Nadzorovana prilagoditev označevalnika z učenjem na podatkovni množici Janes-Tag je točnost zvišala na 84 %. Pri kombiniranju standardne učne množice in množice Janes-Tag se je točnost še povišala, in sicer na 86 %.

Z nadaljnjimi eksperimenti smo preverjali, kako dodajanje Brownovih gruč v nabor značilk vpliva na točnost modela. Za te eksperimente smo uporabili sistem, naučen zgolj na podatkih iz množice Janes-Tag. Z dodajanjem informacij o gručenju se je prvotna točnost modela (84 %) zvišala na 86 %. Dodajanje značilk, ki smo jih pridobili iz normaliziranih oblik besed, je rezultate izboljšalo samo za pol odstotka, v primeru, da bi bile na voljo popolne normalizacije besed, pa bi točnost znašala 88 %.

Z združevanjem standardne in nestandardne učne množice in uporabo dodatnih značilk z informacijami o Brownovem gručenju in avtomatski normalizaciji smo dosegli najboljši rezultat, in sicer 88-% točnost. Medtem ko je ta rezultat v primerjavi s prvotno 69-% točnostjo veliko boljši, je še vseeno precej daleč od točnosti za standardna besedila, ki znaša 94 %.

7 RAZPOZNAVANJE IMENSKIH ENTITET

Zadnje orodje, ki ga opišemo v tem poglavju, je tudi zadnje, ki smo ga razvili v okviru projekta JANES. Označevanje imenskih entitet je v splošnem zelo

uporabno, pri projektu pa je bil glavni namen za razvoj tega orodja avtomatska anonimizacija besedil, potrebna za objavo zgrajenih korpusov.

7.1 Metoda

Orodje je zelo podobno oblikoskladenjskemu označevalniku, saj vključuje iste osnovne značilke (1) brez hipotez na podlagi oblikoslovnega leksikona, (2) z značilkami glede Brownovega gručenja in (3) z dvema dodatnima značilkama, ki opisujeta napovedano besedno vrsto in celotno oblikoskladenjsko oznako.

7.2 Podatkovne množice

Za učenje orodja smo uporabili delež nove različice korpusa ssj500k (Krek et al. 2015), ki vsebuje oznake imenskih entitet, in podatkovne množice Janes-Tag (opisana v Čibej et al. 2018), ki je bil označen na enak način kot nova različica ssj500k. Imenske entitete v obeh korpusih so klasificirane v osebna imena (*oseba*), svojilne pridevnike, izpeljane iz osebnega imena (*izpeljano iz osebe*), krajevna imena (*lokacija*), imena organizacij (*organizacija*) in druga imena (*drugo*).

7.3 Eksperimenti in rezultati

Orodje je bilo izdelano na podlagi številnih predhodnih izkušenj tako v označevanju zaporedij (Ljubešić in Erjavec, 2016; Ljubešić et al. 2016a) kot tudi v razpoznavanju imenskih entitet (Ljubešić et al. 2013), tako da med izdelavo orodja nismo izvedli obsežnejših eksperimentov. Evalvacijo smo opravili naknadno, pri tem pa 80 % podatkov uporabili za učenje in 20 % za testiranje. Glede na to, da podatkovna množica vsebuje standardna in nestandardna besedila, v nadaljevanju poročamo o rezultatih, ki smo jih dosegli pri uporabi (1) samo standardnih podatkov, (2) samo nestandardnih podatkov in (3) kombinacije standardnih in nestandardnih podatkov. Dodatno smo izvedli tudi eksperiment (4), pri katerem smo model učili na standardnih, testirali pa na nestandardnih podatkih. Rezultati evalvacije so podani v Tabeli 7. Za vsako kategorijo so podani rezultati za natančnost, priklic in F1 – harmonično povprečje natančnosti in priklica.

Rezultati prvih treh eksperimentov kažejo, da je razpoznavanje imenskih entitet najuspešnejše pri kategoriji *oseba*. Kot je bilo pričakovano, je na drugem mestu kategorija *lokacija*, sledita pa ji kategorija *organizacija* in na koncu kategorija

drugo. Rezultati za novo kategorijo *izpeljano iz osebe*, ki do sedaj za slovenščino še ni bila vključena, so slabši, kot bi pričakovali, saj ima večina izpeljank obliko svojilnih pridevnikov s točno določeno končnico. Ta kategorija se v nestandardni testni množici pojavi zgolj enkrat in ima zato pri teh pogojih slabe rezultate. Pri primerjavi standardnih in nestandardnih besedil opazimo, da boljši rezultat pri nestandardnih besedilih dosega kategorija *oseba*, in sicer zaradi omemb uporabniških imen, ki se začenjajo z @ in se jih zato model zlahka nauči. Pri kategoriji *lokacija* so rezultati primerljivi, kategorija *organizacija* pa ima pri nestandardnih besedilih veliko slabši rezultat.

Rezultati četrtega eksperimenta, učenja modela na standardnih podatkih in testiranja na nestandardnih, izkazujejo znaten upad F1 pri vseh kategorijah, predvsem pri *organizaciji* in *osebi*, kar je podobno kot pri oblikoskladenjskem označevanju. Vzrok za drastični upad pri kategoriji *organizacija* so različni načini označevanja teh entitet v obeh vrstah besedil, upad pri kategoriji *oseba* pa je po vsej verjetnosti posledica pogostih omemb uporabniških imen, ki jih v standardni učni množici ne srečamo.

Tabela 7: Rezultati evalvacije (natančnost, priklic, F1) razpoznavanja imenskih entitet.

	Standardno			Nestandardno			Oboje			Standardno na nestandardnem		
oseba	0,87	0,95	0,91	0,98	1,00	0,99	0,88	0,96	0,92	0,89	0,20	0,34
izpelj.	0,44	0,56	0,49	0,00	0,00	0,00	0,44	0,52	0,48	0,00	0,00	0,00
lokac.	0,85	0,74	0,79	0,79	0,92	0,85	0,85	0,75	0,80	0,57	0,57	0,62
organ.	0,69	0,48	0,57	0,50	0,33	0,40	0,69	0,48	0,56	0,00	0,00	0,00
drugo	0,39	0,24	0,30	0,75	0,21	0,33	0,41	0,24	0,30	0,60	0,21	0,22

Pri standardnih podatkih smo prekosili do sedaj najboljše rezultate za standardno slovenščino (Štajner et al. 2013), kjer F1 za osebna imena znaša 0,84, za zemljepisna imena 0,76 in imena organizacij 0,56, prav tako pa smo kot prvi predstavili rezultate za nestandardno slovenščino.

8 SKLEP

V poglavju smo opisali skupino orodij za procesiranje nestandardnih besedil v slovenščini, ki smo jih razvili v okviru projekta JANES. Pokazali smo, da lahko dosežemo velika izboljšanja pri jezikovnem procesiranju nestandardnih besedil. Največje izboljšave pogosto dosežemo že z majhno množico domenskih podatkov, npr. nestandardnih podatkov, označenih glede na poljubni fenomen.

Vsa predstavljena orodja (in druga) so prosto dostopna v GitHub repozitoriju CLARIN.SI,[2] kar omogoča, da drugi raziskovalci ne le uporabljajo, ampak tudi prispevajo k orodjem, s tem da poročajo o težavah ali iz obstoječih modelov zgradijo lastne, izboljšane različice orodij. Izbrana orodja za označevanje smo vključili tudi v spletno okolje za gradnjo delotokov ClowdFlows (opisano v poglavju Martinc et al. 2018), prav tako pa jih nameravamo vključiti v podobno okolje WebLicht (Ljubešić et al. 2017), ki ga je razvil CLARIN-DE (Hinrichs et al. 2010).

Najsodobnejše jezikovne tehnologije temeljijo na paradigmi nadzorovanega strojnega učenja in zdi se, da se to v prihodnosti ne bo spremenilo. Strojno učenje in s tem povezana področja (npr. procesiranje naravnega jezika) se namreč nagibajo h globokemu učenju, tj. uporabi nevronskih mrež. Pri procesiranju zveznih signalov (npr. zvok, slika, video) uporaba globokega učenja znatno zmanjša število napak v primerjavi s prejšnjimi pristopi, prav tako globoko učenje izboljša jezikovno procesiranje besedil, čeprav v precej manjši meri. *Bilty* (Plank et al. 2016), najsodobnejši označevalnik, ki temelji na globokem učenju, dosega primerljive rezultate kot označevalnik za slovenščino, ki smo ga opisali v tem poglavju. Največja razlika med tradicionalnimi pristopi, opisanimi v tem poglavju, in nevronskimi mrežami pa je v tem, da nevronske mreže ne zahtevajo oblikovanja značilk, temveč relevantne dele signala za določeno nalogo identificirajo same. Tako bo priprava orodij za procesiranje jezika v bližnji prihodnosti vključevala (1) uporabo sodobnega orodja in učenje na podatkovni množici ali (2) razvoj posebnega orodja, ki bo primarno vključeval definiranje arhitekture nevronske mreže, in učenje nevronske mreže na dani podatkovni množici. Glede na to, da se obseg dela razvijalcev orodij manjša, potreba po velikih količinah kakovostnih ročno označenih podatkov pa vztrajno narašča (nevronske mreže so uspešne predvsem pri zelo velikih količinah podatkov), zagovarjamo vse večjo pomembnost strokovnjakov s področij jezikovnih tehnologij (jezikoslovje, obdelava podatkov) in sorodnih disciplin digitalne humanistike in družbenih znanosti, na primer družbenih tehnologij (prediktorji sociodemografskih spremenljivk za določene govorce ipd.). Izdelava kakovostno označenih podatkovnih množic, kot smo jih uporabili pri gradnji orodij, opisanih v tem poglavju, predstavlja kompleksno in drago nalogo, ki pa bo v prihodnosti še pomembnejša.

Zahvala

Ker je avtorska zasedba tega poglavja mednarodna, smo rokopis pripravili v angleščini. V slovenščino ga je prevedla Dafne Marko, ki se ji za natančen in tekoč prevod ter skrbno upravljanje s terminologijo iskreno zahvaljujemo.

2 https://www.github.com/clarinsi/

Literatura

Brown, Peter F., Vincent J. Della Pietra,. Peter V. deSouza, Jenifer C. Lai, and Robert L. Mercer, 1992: Class-based n-gram models of natural language. *Computational Linguistics* 18/4. 467–479.

Crystal, David, 2011: *Internet linguistics. A student guide*. New York: Routledge.

Čibej, Jaka, Špela Arhar Holdt, Tomaž Erjavec in Darja Fišer, 2018: Ročno označeni korpusi Janes za učenje jezikovnotehnoloških orodij in jezikoslovne raziskave. Fišer, Darja (ur.): *Viri, orodja in metode za analizo spletne slovenščine*. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani. 44–73.

Eisenstein, Jacob, 2013: What to do about bad language on the Internet. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*. 359–369. http://www.cc.gatech.edu/~jeisenst/papers/naacl2013-badlanguage.pdf

Erjavec, Tomaž, Nikola Ljubešić in Nataša Logar, 2015: The slWaC corpus of the Slovene Web. *Informatica* 39/1. 35.

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

Erjavec, Tomaž, Nikola Ljubešić in Darja Fišer, 2017: Adapting a state-of-the--art tagger for South Slavic languages to non-standard text. *Proceedings of the EACL workshop*. The 6th Workshop on Balto-Slavic Natural Language Processing, April 4, 2017 Valencia, Spain. Stroudsburg: The Association for Computational Linguistics. 60–68. http://bsnlp-2017.cs.helsinki.fi/bsnlp2017-book.pdf

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

Fišer, Darja, Tomaž Erjavec, Nikola Ljubešić, and Maja Miličević (2015): Comparing the nonstandard language of Slovene, Croatian and Serbian tweets. Smolej, Mojca (ur.): *Obdobja 34. Slovnica in slovar - aktualni jezikovni opis (1. del)*. Ljubljana: Znanstvena založba filozofske fakultete. 225–231.

Gimpel, Kevin, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan in Noah A. Smith, 2011: Part-of-speech tagging for Twitter: Annotation, features, and experiments. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers*. Volume 2. pages Association for Computational Linguistics. 42–47.

Heafield, Kenneth, 2011: KenLM: Faster and smaller language model queries. *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Hinrichs, Erhard, Marie Hinrichs, Thomas Zastrow, 2010: WebLicht: Web-Based LRT Services for German. *Proceedings of the Systems Demonstrations at the 48th Annual Meeting of the Association for Computational Linguistics (ACL-2010)*. Uppsala. 25–29.

Krek, Simon in Erjavec, Tomaž, 2009: Standardised Encoding of Morphological lexica for Slavic languages. *MONDILEX Second Open Workshop*. Kyiv, Ukraine. 24–29

Krek, Simon, Kaja Dobrovoljc, Tomaž Erjavec, Sara Može, Nina Ledinek in Nanika Holz, 2015: *Training corpus ssj500k 1.4*. Slovenian language resource repository CLARIN.SI.

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

Ljubešić, Nikola, Marija Stupar, Tereza Jurić, in Željko Agić, 2013: Combining available datasets for building named entity recognition models of Croatian and Slovene. *Slovenščina 2.0* 1/2. 35–57.

Ljubešić, Nikola, Tomaž Erjavec in Darja Fišer, 2014: Standardizing Tweets with Character-Level Machine Translation. Gelbukh, Alexander (ur.): *CICLing, Lecture notes in computer science*. Berlin, Heidelberg: Springer. 164–175.

Ljubešić, Nikola, Tomaž Erjavec in Darja Fišer, 2018: Orodja za procesiranje nestandardne slovenščine. Fišer, Darja (ur.): *Viri, orodja in metode za analizo spletne slovenščine*. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani. 74–99.

Ljubešić, Nikola, Darja Fišer, Tomaž Erjavec, Jaka Čibej, Dafne Marko, Senja Pollak in Iza Škrjanec, 2015: Predicting the Level of Text Standardness in User-generated Content. *Proceedings of Recent Advances in Natural Language Processing*. 371–378.

Ljubešić, Nikola, Tomaž Erjavec in Darja Fišer, 2016: Corpus-based diacritic restoration for south slavic languages. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA). 3612–3616.

Ljubešić, Nikola in Tomaž Erjavec, 2016: Corpus vs. lexicon supervision in morphosyntactic tagging: the case of Slovene. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA). 1527–1531.

Ljubešić, Nikola, Filip Klubička, Željko Agić, Ivo-Pavao Jazbec, 2016a: New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).*

Ljubešić, Nikola, Tomaž Erjavec, Darja Fišer, Erhard Hinrichs, Marie Hinrichs, Cyprian Laskowski, Filip Petkovski in Wei Qui, 2017: Multilingual Text Annotation of Slovenian, Croatian and Serbian with WebLicht. *Proceedings of the CLARIN Annual Conference*. 18–20 September, Budapest, Hungary.

Martinc, Matej, Senja Pollak in Ana Zwitter Vitez, 2018: Delotoki za nadaljnje analize nestandardne slovenščine. Fišer, Darja (ur.): *Viri, orodja in metode za analizo spletne slovenščine*. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani. 16–43.

Plank, Barbara, Anders Sogaard in Yoav Goldberg, 2016: Multilingual Part-of--Speech Tagging with Bidirectional Long Short-Term Memory Models and Auxiliary Loss. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin, Germany, August 7-12, 2016. 412–418.

Štajner, Tadej, Tomaž Erjavec in Simon Krek, 2013: Razpoznavanje imenskih entitet v slovenskem besedilu. *Slovenščina 2.0* 1/2. 58–81.

Tjong Kim Sang, Erik, Marcel Bollmann, Remko Boschker, Francisco Casacuberta, Feike Dietz, Stefanie Dipper, Miguel Domingo, Rob van der Goot, Marjo van Koppen, Nikola Ljubešić, Robert Östling, Florian Petran, Eva Pettersson, Yves Scherrer, Marijn Schraagen, Leen Sevens, Jörg Tiedemann, Tom Vanallemeersch in Kalliopi Zervanou, 2017: The CLIN27 Shared Task: Translating Historical Text to Contemporary Language for Improving Automatic Linguistic Annotation. *Computational Linguistics in the Netherlands Journal* 7/1. 53–64.

Scherrer, Yves in Nikola Ljubešić, 2016: Automatic normalisation of the Swiss German ArchiMob corpus using character-level machine translation. *Proceedings of the 13th Conference on Natural Language Processing (KONVENS).* September 19-21, 2016, Bochum, Germany. 248–255. https://www.linguistics.rub.de/konvens16/pub/32_konvensproc.pdf

Delotoki za nadaljnje analize nestandardne slovenščine

Matej Martinc, Senja Pollak, Ana Zwitter Vitez

Izvleček

V zadnjih letih se je razmahnil razvoj platform, ki poenostavljajo znanstveno raziskovanje, povezujejo različna področja in omogočajo lažjo dostopnost metod in rezultatov širšemu krogu uporabnikov. V raziskavi, ki jo predstavljamo v tem poglavju, smo v platformo za vizualno programiranje ClowdFlows implementirali množico orodij (gradnikov) za obdelavo naravnega jezika, ki bodo jezikoslovcem in ostalim zainteresiranim uporabnikom omogočila lažjo in hitrejšo analizo besedil. Novi gradniki omogočajo obdelavo naravnega jezika, upravljanje korpusa ter končni prikaz statistik in analiz. Implementacijo novih gradnikov predstavimo na primeru dveh delotokov. Prvi je namenjen izdelavi novega korpusa iz tvitov, zbranih s pomočjo orodja *TweetCaT*, drugi pa analizi komentarjev Evrovizije. Razvita spletna orodja omogočajo gradnjo in obdelavo novih korpusov, razširjajo možnosti kvantitativnih analiz ter poenostavljajo zapletene postopke za obdelavo naravnega jezika.

Ključne besede: obdelava naravnega jezika, orodja za korpusno analizo, vizualno programiranje, ClowdFlows, nestandardna slovenščina

1 UVOD

Interdisciplinarnost je eden ključnih pogojev za izboljšanje kvalitete in kvantitete znanstvene produkcije. V zadnjem času je bilo veliko truda vloženega v razvoj platform za interdisciplinarno sodelovanje, glavni namen tovrstnih platform pa je poenostavitev in pospešitev znanstvenega raziskovanja in s tem lažja dostopnost metod posameznih področij širšemu krogu uporabnikov. V članku predstavimo implementacijo orodij za obdelavo naravnega jezika v platformo za vizualno programiranje z imenom *ClowdFlows* (Kranjc et al. 2012). *ClowdFlows* je spletna platforma za rudarjenje podatkov, namenjena poenostavitvi razvoja kompleksnih metod in procesov rudarjenja podatkov. Platforma omogoča uporabo metod tudi raziskovalcem s področja humanistike in družboslovja, ki bi jim bilo zaradi pomanjkljivega tehničnega znanja takšno raziskovanje sicer onemogočeno.

Zaradi navedenih razlogov je platforma *ClowdFlows* zasnovana za preprosto uporabo, dosegljiva preko spletnega brskalnika in ne potrebuje predhodne namestitve. Paradigma vizualnega programiranja poenostavi uporabo in izdelavo zapletenih procesov na upravljanje gradnikov (angl. *widgets*) s preprosto operacijo *primi-odloži* (angl. *drag-and-drop*). Gradniki so vizualno predstavljeni deli programov, ki imajo definirane vhodne in izhodne oblike podatkov, parametre pa lahko uporabnik ročno izbere. Te gradnike se na delovni površini sestavlja v delotoke (angl. *workflows*), ki jih lahko opišemo kot vizualne predstavitve znanstvenih procesov. Prednost platforme *ClowdFlows* je tudi, da omogoča preprosto deljenje in ponovljivost rezultatov ter ponovno uporabo delotokov (z objavo spletne strani delotoka). Platforma *ClowdFlows* je kolaborativne narave, saj lahko razvijalci v različnih programskih jezikih prispevajo svoje programe v obliki gradnikov, uporabniki pa lahko obstoječe gradnike inovativno povezujejo v nove delotoke.

Eno izmed področij z velikim številom uporabnikov statističnih in računalniških metod je gotovo jezikoslovje. Jezikoslovje je v zadnjih desetletjih doseglo izjemen napredek s pomočjo korpusnih metod raziskovanja avtentične jezikovne rabe, s katerimi lahko analiziramo kolokacije, besedotvorne posebnosti, specifično rabo ločil ipd. v eni ali več besedilnih zvrsteh. Ključno pomanjkljivost korpusnih analiz predstavlja dejstvo, da so v veliki meri odvisne od vključenosti jezikovnega gradiva v razpoložljiva orodja (Anthony 2013), kar jezikoslovcem pogosto onemogoča analizo najbolj relevantnega gradiva ali pa hitro odzivanje in preučevanje aktualnih družbenih procesov. Korpus Janes (Erjavec et al. 2018) tako ne omogoča vpogleda v besedila, povezana z družbenimi tematikami, ki so krojile diskurz na slovenskem spletu od leta 2017 dalje, kakršne so na primer migrantska problematika, spremembe družinske zakonodaje ali

reševanje morskih mejnih vprašanj. Po drugi strani orodja za gradnjo novih korpusov, kot so *SketchEngine* (Kilgarriff et al. 2004), *Wordsmith Tools* (Scott 1998) ali *AntConc* (Anthony 2014), zahtevajo sposobnost avtonomne priprave korpusa v zahtevanem formatu, vključno z izpeljavo postopkov lematizacije in jezikoslovnega označevanja, kar marsikateremu jezikoslovcu predstavlja nepremostljivo oviro. Delotoki lahko ustvarijo sinergijo med analitičnimi metodami jezikoslovja in računalništva ter skozi analizo različnih aktualnih tematik omogočijo boljše razumevanje družbenega dogajanja.

Za obdelavo slovenščine je že na voljo nekaj gradnikov in delotokov, ki omogočajo jezikovno označevanje besedil ter luščenje terminologije in definicij (Pollak et al. 2012a, 2012b). V poglavju predstavimo dvajset novih gradnikov, ki omogočajo sestavo številnih novih delotokov. Posebno pozornost namenimo gradnikom, ki omogočajo pridobivanje in obdelavo besedil z družbenih medijev in s tem korpusno podprto jezikovno analizo aktualnih družbenih tematik.

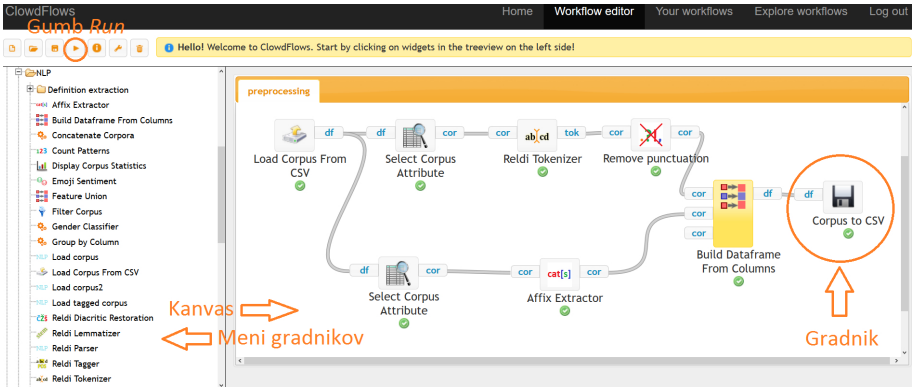
2 PLATFORMA CLOWDFLOWS

Kot je bilo na kratko omenjeno že v uvodu, je spletna platforma ClowdFlows namenjena gradnji, izvedbi in objavljanju interaktivnih delotokov za rudarjenje podatkov. Najpomembnejša značilnost platforme je grafični vmesnik, ki omogoča vizualno programiranje (Slika 1). Ta zajema izdelavo zapletenih delotokov s pomočjo tehnike *primi-odloži*. Uporabnik s klikom na gradnik v meniju na levi izbere želeni gradnik, ki se nato pojavi na delovni površini. Če kliknemo na izhod enega gradnika in nato na vhod drugega, se med gradnikoma vzpostavi povezava. Na ta način lahko hitro in učinkovito izdelujemo zapletene delotoke. Delotoke poženemo s klikom na ikono za pogon zgoraj levo ali s pritiskom na gumb *Run* v meniju, ki se pojavi ob desnem kliku na poljuben gradnik. V meniju je tudi gumb *Run only this*, ki namesto celega delotoka požene le izbrani gradnik. Če uporabnik želi pridobiti več informacij o posameznem gradniku, je v meniju tudi gumb *Help*, ki odpre okno s podrobno dokumentacijo o gradniku in opisom njegove uporabe.

Vizualno programiranje se je v zadnjem času uveljavilo kot učinkovit in priljubljen način za enostavno izvedbo zapletenih procesov, zato se je v preteklih letih pojavilo kar nekaj platform, ki so po funkcionalnosti podobne platformi ClowdFlows. Ena bolj znanih je platforma *RapidMiner* (Mierswa et al. 2006), znana tudi kot *YALE* (*Yet another Learning Environment*), ki jo avtorji opisujejo kot aplikacijo za strojno učenje in odkrivanje znanja v podatkovnih bazah. Dokaj znana je tudi platforma *KNIME* (*The Konstanz Information Miner*) (Berthold et al. 2009), ki jo lahko opišemo kot modularno okolje za izdelavo in interaktivno

izvajanje podatkovnih delotokov. Manj znane so platforme *ORANGE* (Demšar et al. 2004), *MyGrid*,[1] *Language Grid*[2] in *ARGO*.[3]

Platforma ClowdFlows se od zgoraj omenjenih platform razlikuje po posebnosti, da ne potrebuje namestitve na računalnik, saj jo lahko zaženemo v spletnem brskalniku. To poenostavi uporabo platforme in omogoča dostop do zgrajenih delotokov in gradnjo novih delotokov brez namestitve katerih koli programov. Še ena pomembna lastnost platforme *ClowdFlows* je njena odprtokodnost (njena izvorna koda je objavljena na platformi *GitHub*), kar pomeni, da lahko kdor koli v platformo dodaja nove gradnike. Dokumentacija o platformi in navodila za dodajanje novih gradnikov so dosegljivi na spletu.[4] Gradniki so vizualno predstavljeni deli programov, ki imajo definirane vhodne in izhodne oblike podatkov, parametre pa lahko uporabnik ročno izbere. Gradniki so lahko implementirani kot operacije spletnih servisov (angl. *web services*) ali pa lokalno v okolju *ClowdFlows*.



Slika 1: Prikaz grafičnega uporabniškega vmesnika platforme ClowdFlows. Na levi je meni gradnikov, na desni delovna površina za grajenje delotokov, levo zgoraj je gumb za zagon delotoka.

Platforma poleg gradnje novih delotokov omogoča tudi objavo delotokov, kar močno olajša ponovljivost rezultatov. Ponovljivost rezultatov je v znanosti temeljna zahteva, hkrati pa objava delotokov omogoča drugim uporabnikom, da že objavljeni delotok uporabijo na lastnih podatkih ali ga uporabijo kot osnutek za izdelavo lastnega delotoka s podobnimi značilnostmi. Na platformi *ClowdFlows* je trenutno objavljenih že več kot sto delotokov, ki so koristni tudi za učenje uporabe platforme.

1 http://www.mygrid.org.uk

2 http://langrid.org/

3 http://argo.nactem.ac.uk

4 http://clowdflows.readthedocs.io/en/latest/

ClowdFlows vsebuje množico gradnikov za podatkovno rudarjenje, pomanjkljiva pa je bila množica gradnikov za obdelavo naravnega jezika, ki bi jezikoslovcem omogočala kvantitativno analizo besedil in korpusov. V preteklosti je bila razvita posebna veja platforme *ClowdFlows*, poimenovana *TextFlows* (Perovšek et al. 2016), ki je specializirana predvsem za obdelavo naravnega jezika. Platforma *TextFlows* je trenutno namenjena predvsem obdelavi angleških besedil, prav tako pa ima predpisan izhodno/vhodni format gradnikov ADC (*Annotated Document Corpus*). Značilnost ADC-formata je izdelava novega nivoja anotacij za vsako operacijo obdelave naravnega jezika in shranjevanje teh nivojev skupaj z izvornim tekstom. Tak način zapisa omogoča, da so vse anotacije dosegljive na izhodu vsakega gradnika, kar izboljša povezljivost posameznih gradnikov. Po drugi strani je tak način zapisa spominsko potraten, tako da že relativno majhen korpus z okoli 100.000 pojavnicami in z nekaj nivoji anotacij, zapisan kot ADC-objekt, preseže mejo 1 GB, kar je zgornja meja za zapis objekta v podatkovno bazo PostGres[5], ki jo TextFlows uporablja za hranjenje podatkov. Zaradi omejitve procesiranja večjih korpusov in zaradi ciljnega jezika (slovenščine) smo se odločili, da nove gradnike za obdelavo naravnega jezika namesto v platformo *TextFlows*, ki bi bila tematsko bolj primerna, implementiramo v platformo *ClowdFlows*. Dodatni razlog za implementacijo gradnikov v platformo *ClowdFlows* je načrt avtorjev platforme *TextFlows*, da jo postopoma reintegrirajo v platformo *ClowdFlows* 3.0, novo in izboljšano različico.

V okviru raziskave smo tako v platformo *ClowdFlows* implementirali dvajset novih gradnikov za obdelavo naravnega jezika (ki jih predstavimo v tretjem razdelku) in kot primera uporabe implementirali dva funkcionalno različna delotoka za obdelavo naravnega jezika (predstavljena v četrtem razdelku).

3 GRADNIKI ZA KORPUSNO ANALIZO IN OBDELAVO SLOVENŠČINE

Glavni cilj novih gradnikov je obdelava slovenskega jezika, vendar je mnogo na novo implementiranih gradnikov splošno namenskih in omogoča podporo za več jezikov. Jezikoslovci po navadi za osnovno enoto obdelave uporabijo korpus, ki je sestavljen iz množice besedilnih dokumentov, zato smo ta princip upoštevali tudi pri naši implementaciji.

Obstaja množica različnih formatov za predstavitev korpusov, od zelo kompleksnih do precej preprostih. Pri načrtovanju implementacije smo se odločili, da mora biti izbrana predstavitev korpusa uporabniku prijazna in razumljiva, da

5 https://wiki.postgresql.org/wiki/FAQ#What_is_the_maximum_size_for_a_row.2C_a_table.2C_and_a_database.3F

mora omogočati preprosto razširitev in vključitev novih anotacij, hkrati pa mora upoštevati omejitve platforme, ki zaradi svoje spletne zasnove ne omogoča obdelave zelo velikih datotek. Zato smo se na koncu odločili za predstavitev korpusa v tabelarični obliki (angl. *dataframe*) (Slika 2), kjer vrstice predstavljajo posamezne dokumente, posamezni stolpci pa vsebujejo različne vrste anotacij (npr. besedilo, leme, oblikoskladenjske oznake ...).

Example	Sentiment	Subsentiment
Danes imajo naši turisti še zadnji dan turistovanja na davkoplačevalske stroške.	negative	cynicism
Jutri pa domov in davkoplačevalcem plačat račune.	negative	criticism
Potem bomo pa poslušali staro lajno...bila je lepa izkušnja, odziv publike je bil odličen, SLO smo dobro promovirali, dali smo vs	negative	cynicism
upam, da je šlo včeraj pri tinkari vse ok in da nas rtv spet ne zavaja kot nas je lani z naslovom...	positive	support
BO ... tokrat sem optimističen!	positive	positive exp
Tinkara je res profi in vsaka ji čast !!!	positive	support
Dejmo naši	positive	support
SREČNO !	positive	support
Držim pesti!	positive	support
Verjamem v finale in si finale tudi zaslužimo.	positive	positive exp
Pričakujem 12 točk iz Makedonije!	positive	positive exp

Slika 2: Primer korpusa v tabelarični obliki.

Tabelarična oblika je po svoji zasnovi podobna nekaterim že uveljavljenim formatom za predstavitev korpusa, je preprosta, razumljiva in računsko ter spominsko nezahtevna za obdelavo. Pri implementaciji takšne predstavitve smo si pomagali s programsko knjižnico *Pandas* (McKinney 2011), ki poleg drugih orodij vsebuje tudi razred *Dataframe*, ki predstavlja implementacijo tabelaričnega formata v programskem jeziku *Python*.

Na splošno lahko implementirane gradnike razdelimo v naslednje kategorije:

- gradniki za vnos korpusa,
- gradniki za obdelavo naravnega jezika,
- gradniki za upravljanje s korpusom,
- gradniki za izračun in prikaz statistik.

V naslednjih sekcijah, ki razvrščajo nove gradnike glede na funkcionalnost, podrobneje opišemo posamezne gradnike.

3.1 Gradniki za vnos korpusa

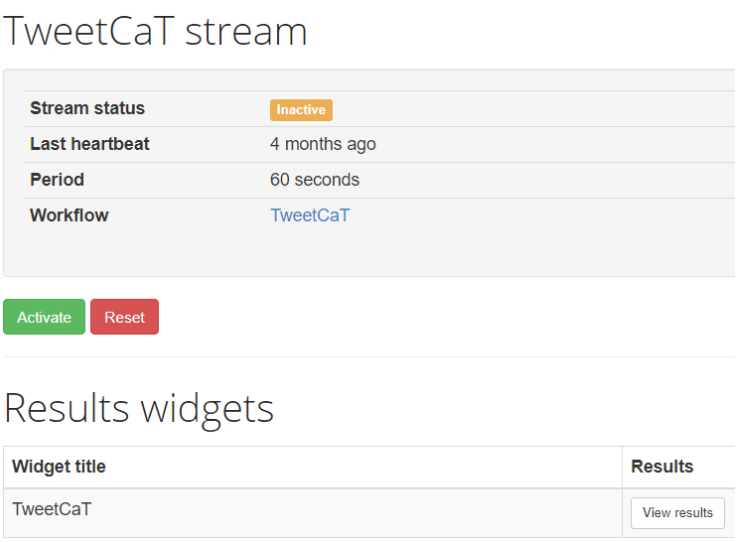
V platformo *ClowdFlows* smo implementirali dva gradnika za vnos korpusa:

- *Load Corpus From CSV:* Ta gradnik kot vhod sprejme datoteko s korpusom v formatu *comma-separated values* (CSV) in korpus pretvori v tabelarično obliko, ki je primerna za nadaljnjo obdelavo. Ta vrsta vnosa je primerna predvsem za jezikoslovce, ki večino svojih analiz izvajajo v programu Microsoft Excel, saj ta program omogoča transformacijo datotek iz formata Excel Workbook v format CSV. Gradnik uporabniku omogoča, da sam definira razločevalni znak (privzeto je ta v formatu CSV vejica), prva vrstica v dokumentu pa mora vsebovati imena stolpcev.
- *TweetCaT* (Ljubešić et al. 2014): Orodje za gradnjo korpusa z zajemom tvitov za jezike z relativno malo govorci, med katere sodi tudi slovenščina. Orodje omogoča zajem tvitov v realnem času (angl. *streaming*) s pomočjo vmesnika *Twitter API* in ima dva načina delovanja. V geografskem načinu (angl. *geographical search*) orodje zajema le tvite znotraj območja, ki je omejeno s štirimi geografskimi koordinatami. Uporabnik koordinate definira kot parametre gradnika v naslednjem zaporedju: minimalna geografska širina, minimalna geografska dolžina, maksimalna geografska širina, maksimalna geografska dolžina. V jezikovnem načinu (angl. *language search*) orodje zajema tvite na podlagi semenskih besed (angl. *seed words*) za posamezen jezik. Gre za besede, ki so specifične ali pogoste v določenem jeziku in jih lahko definira uporabnik ali pa uporabi privzeti seznam besed (za slovenščino so privzete besede definirane v besedilnem okencu parametra *seed words*, za hrvaščino, srbščino in bosanščino pa je množica privzetih besed podana v dokumentaciji gradnika). Z orodjem *TweetCaT* lahko uporabnik zgradi velike baze tvitov na hiter in učinkovit način, saj orodje nabere vse razpoložljive tvite posameznega avtorja, ki je objavil tvit z določeno besedo. Podrobnosti orodja *TweetCaT* so opisane v Ljubešić et al. (2018).

  Ker orodje za svoje delovanje potrebuje vmesnik *Twitter API*, lahko uporabnik pri uporabi gradnika izbira med uporabo že privzetega vmesnika *ClowdFlows Twitter API*, kar stori z izborom parametra *Use ClowdFlows authentication,* ali pa ustvari lastni *Twitter API*. Zaradi množične uporabe privzetega vmesnika *ClowdFlows Twitter API* je njegova uporaba nezanesljiva in močno upočasni zajem tvitov, zato je priporočeno, da uporabnik ustvari lastni *Twitter API* po naslednjem postopku: uporabnik najprej ustvari račun na *Twitterju*, nato pa na spletni strani https://dev.twitter.com/ sledi navodilom za izdelavo API-ja.[6] Po končanem postopku izdelave API-ja uporabnik pridobi uporabniški ključ (angl. *consumer key*), uporabniško geslo (angl. *consumer secret*), žeton za dostop (angl. *access token*) in geslo za žeton za dostop (angl. *access token secret*), ki jih vpiše v besedilna polja kot parametre gradnika.

6 Podrobna navodila za uporabo vmesnika Twitter API so objavljena tudi na spletnem naslovu http://docs.inboundnow.com/guide/create-twitter-application/.

Gradnik *TweetCaT* sodi v posebno kategorijo gradnikov za pretočne podatke (angl. *streaming widgets*), za katere je značilen sprotni zajem in jih v platformi *ClowdFlow*s upravljamo s pomočjo posebej za te gradnike narejene nadzorne plošče. Vsak delotok, ki vsebuje gradnike za zajem, je predstavljen kot tok (angl. *stream*). Če v glavnem meniju *ClowdFlows* kliknemo na gumb *Your workflows,* pridemo do seznama delotokov. Vsak delotok, ki vsebuje gradnike za zajem podatkov, je v seznamu predstavljen z ikono, ki prikazuje trenutno stanje toka delotoka. Tok je lahko aktiviran, kar pomeni, da zajem podatkov trenutno poteka, ali pa deaktiviran, kar pomeni, da se novi podatki trenutno ne nabirajo. Aktivacijo, deaktivacijo in ponastavljanje (izbris do sedaj zbranih podatkov) toka je možno nadzirati s pomočjo nadzorne plošče za vsak posamezen tok (Slika 3), ki je dosegljiva s klikom na ikono za nastavitve poleg ikone za prikaz stanja toka v seznamu delotokov. Nadzorna plošča omogoča tudi prikaz do sedaj zbranih podatkov.



Slika 3: Nadzorna plošča za upravljanje s tokom (angl. stream).

Tipičen scenarij uporabe gradnika *TweetCaT* bi bil naslednji: uporabnik izbere gradnik, ki se prikaže na delovni površini, določi želeni način delovanja (npr. jezikovni način) ter vpiše potrebne vhodne parametre. Tok je potrebno najprej aktivirati po navedenem postopku, saj bi sicer ob kliku na gumb *Run* gradnik javil napako, ki uporabnika opozori, da ni bil zajet še noben tvit. Po aktivaciji toka bo v nekaj minutah nabranih nekaj tisoč tvitov, že en sam tvit pa omogoča zagon gradnika. Pri vsakem

zagonu gradnika se kot izhod vrnejo vsi trenutno nabrani tviti. Ko imamo zajetih dovolj tvitov, lahko tok deaktiviramo, kar pomeni, da zaključimo z zajemom tvitov in na ta način poskrbimo, da se izhod gradnika ne spreminja (dopolnjuje z novimi tviti) več z vsakim novim zagonom.

Naslednji scenarij uporabe gradnika *TweetCaT* bi bila uporaba že obstoječega delotoka, ki vsebuje gradnik in že zajete podatke. Kot je že bilo omenjeno, platforma *ClowdFlows* omogoča objavljanje zgrajenih delotokov, ki tako postanejo dostopni ostalim uporabnikom. Če uporabnik želi uporabiti objavljen delotok, se pravzaprav ustvari kopija tega delotoka z vsemi podatki, ki jo nato uporabnik lahko poljubno spreminja, ne da bi s tem vplival na originalni delotok.

3.2 Gradniki za obdelavo naravnega jezika

Pri implementaciji orodij za obdelavo naravnega jezika smo se osredotočili na orodja za obdelavo slovenščine, hrvaščine in srbščine, implementirali pa smo tudi nekaj jezikovno neodvisnih orodij:

- *Reldi Tokenizer* (Ljubešić in Erjavec 2016): Orodje za tokenizacijo (prepoznavanje pojavnic, tj. besed in ločil) slovenskih, hrvaških in srbskih korpusov. Želeni jezik uporabnik nastavlja kot parameter gradnika, privzeto je nastavljen slovenski jezik. Orodje ima dva načina delovanja, in sicer omogoča tokenizacijo standardnih besedil (npr. znanstveni članki, literarna dela) in nestandardnih besedil (npr. blogi, tviti in komentarji). Kot vhod orodje sprejme stolpec korpusa v tabelarični obliki in ponavadi je to stolpec z besedilom. Orodje privzeto vrne seznam dvojic pojavnic in njihovih pozicij v tekstu, s čimer je izhodni zapis kompatibilen z vhodi gradnikov *Reldi Tagger*, *Reldi Lemmatizer* in *Reldi Diacritic Restoration*. Če uporabnik izbere parameter za sploščen izhod (angl. *Flatten output*), gradnik vrne sploščeno (enodimenzionalno) zaporedje s presledkom združenih pojavnic za vsak tekst v korpusu, ki izhod naredi kompatibilen z nekaterimi gradniki za nadaljnjo obdelavo in izračun statistik, kot je na primer gradnik *Display Corpus Statistics*. Primer tokeniziranega stavka, zapisanega v sploščenem formatu, je »Ta suhi škafec pušča , ker nima dna .«.
- *Reldi Tagger* (Ljubešić in Erjavec 2016): Oblikoskladenjski označevalnik za slovenščino, hrvaščino in srbščino. Posamezen jezik se izbere kot parameter, privzeto pa je nastavljen slovenski jezik. Kot vhod gradnik dobi dvojice pojavnic in njihovih pozicij v tekstu (izhod, ki ga vrne zgoraj opisani gradnik *Reldi Tokenizer*) in za vsako besedilo vrne zaporedje s

presledkom združenih oblikoskladenjskih oznak (angl. *MSD* oz. *morphosyntactic descriptor*) posameznih pojavnic v besedilu.[7]

- *Reldi Lemmatizer* (Ljubešić in Erjavec 2016): Orodje, ki posamezni besedi pripiše njeno osnovno obliko oz. lemo. Lematizator *Reldi* podpira slovenščino, hrvaščino in srbščino, posamezen jezik pa tako kot pri prej opisanem gradniku uporabnik izbere kot parameter. Tudi pri tem gradniku je privzeto nastavljen slovenski jezik. Ta gradnik sprejme isti vhod kot zgoraj opisani *Reldi Tagger* in za vsak tekst vrne s presledkom združene leme besed. Prav tako velja omeniti, da sta gradnika *Reldi Lemmatizer* in *Reldi Tagger* dve povezani operaciji enega in istega orodja, ki sta bili umetno ločeni in implementirani kot dva ločena gradnika za potrebe *ClowdFlows* platforme.
- *Reldi Diacritic Restoration* (Ljubešić et al. 2016): Orodje za obnovo diakritičnih znakov na izpuščenih mestih v nestandardnih zapisih (npr. sola > šola, pac > pač). Podprti jeziki so nestandardna slovenščina, hrvaščina in srbščina (privzeti jezik je slovenščina). Kot vhod ta gradnik dobi dvojice pojavnic in njihovih pozicij (izhod, ki ga vrne gradnik *Reldi Tokenizer*) in privzeto vrne seznam dvojic pojavnic z obnovljenimi diakritiki in njihovih pozicij za vsak vhodni tekst. Tako kot *Reldi Tokenizer* tudi ta gradnik omogoča sploščeni izhod s presledkom združenih pojavnic z obnovljenimi diakritičnimi znaki, ki izhod naredi kompatibilen z nekaterimi gradniki za nadaljnjo jezikovno obdelavo in izračun statistik.
- *Emoji Sentiment* (Novak et al. 2015): Gradnik za izračun sentimenta tvita s pomočjo emojijev kot vhod dobi stolpec korpusa v tabelarični obliki in za vsak dokument vrne sentiment (pozitiven, negativen ali nevtralen). Sentiment dokumenta je izračunan kot vsota sentimentov posameznih emojijev, izračun sentimenta posameznega emojija pa je odvisen od njegove pojavitve v pozitivnih, nevtralnih in negativnih dokumentih učnega korpusa.

Poleg dodajanja že obstoječih orodij za izdelavo korpusnih anotacij smo sami implementirali še naslednje gradnike:

- *Remove Punctuation*: Ta gradnik kot vhod dobi stolpec korpusa v tabelarični obliki (ponavadi stolpec s samim besedilom) in iz njega odstrani najpogostejša ločila:
  - #@!«$%&()*+,-./:;<=>?[\]^_`{|}~'.

  Kot izhod za vsako besedilo v korpusu vrne besedilo brez ločil. Odstranitev ločil se pogosto uporablja pri klasifikacijskih nalogah na nivoju teksta, kot

7 Množica oblikoskladenjskih oznak je bila definirana v okviru projekta Jezikoslovno označevanje slovenščine (http://nl.ijs.si/jos/msd/html-en/index.html).

je npr. klasifikacija spola avtorja teksta, saj ločila predvidoma ne vplivajo na točnost klasifikacijskega modela, hkrati pa njihova odstranitev zmanjša število značilk in s tem kompleksnost klasifikacijskega modela.

- *Remove Stopwords*: Gradnik kot vhod dobi stolpec korpusa v tabelarični obliki, ki vsebuje besedila in iz njih odstrani pomensko prazne besede (angl. *stopwords*). Za vsako besedilo korpusa gradnik vrne besedilo brez pomensko praznih besed. Trenutno gradnik vsebuje sezname slovenskih, angleških, španskih in portugalskih pomensko praznih besed,[8] kot privzeti jezik je nastavljena slovenščina.
- *Affix Extractor*: Gradnik kot vhod dobi stolpec korpusa v tabelarični obliki in privzeto vrne s presledkom ločene predpone besed dolžine tri (prve tri črke vsake besede). Uporabnik lahko kot parameter nastavlja dolžino vrnjenih zaporedij črk, hkrati pa lahko izbira med predponami, končnicami ter tako imenovanimi ločilniškimi končnicami (angl. *beg-puncts*) (Sapkota et al. 2015), kjer gre za zaporedja, ki se pričnejo z ločilom, ostali znaki v zaporedju pa so črke, številke ali presledki (npr., ločilniški končnici dolžine 3 vhodnega zaporedja »Pazi, drek! Ups« bi bili », d« in »! U«). Ločilniške končnice se uporabljajo kot značilke v klasifikacijskih modelih za profiliranje avtorjev teksta (npr. v modelih za napovedovanje spola, starosti, dialekta ...). V primeru ločilniških končnic, ki lahko vsebujejo tudi presledke, ne dobimo s presledkom ločenih predpon, temveč zaporedja, ločena z nizom ### (ta niz je bil izbran zaradi nizke verjetnosti pojavitve v originalnem tekstu). Posebnost gradnika *Affix Extractor* je, da ga lahko smiselno uporabimo tudi na stolpcih korpusa v tabelarični obliki, ki vsebujejo korpusne anotacije. Tako bi lahko na primer gradniku kot vhod dali s presledkom združene oblikoskladenjske oznake, ki jih vrne gradnik *Reldi Tagger,* in iz njih izluščili le predpono dolžine ena. Na ta način bi dobili le besedno vrsto pojavnice namesto celotne oznake, saj prva črka oznake določa besedno vrsto.
- *Count Patterns*: Gradnik za štetje vzorcev v korpusu kot vhod dobi stolpec korpusa v tabelarični obliki in za vsak dokument v korpusu vrne število vzorcev. Uporabnik lahko sam definira besede ali besedne zveze, ki naj jih gradnik prešteje (v obliki seznama z vejicami ločenih besed ali besednih zvez), lahko pa izbere, da naj gradnik prešteje vnaprej definirane vzorce, kot so ponavljajoči se znaki (angl. *character flood*), kot je na primer »jaaaa«, ali emojiji. Privzeta nastavitev je, da kot izhod gradnika dobimo seznam frekvenc vzorca za vsak posamezen dokument, če pa obkljukamo parameter *Count for entire corpus*, dobimo frekvenco pojavitev za celoten korpus. Gradnik privzeto vrne relativno frekvenco

8 Seznami praznih besed za angleščino, španščino in portugalščino so iz knjižnice NLTK, za slovenščino pa je seznam praznih besed izdelala Iza Škrjanec na podlagi korpusa Kres in zajema predloge, veznike, členke in zaimke.

pojavitev (število pojavitev vzorca se normalizira s številom vseh znakov v dokumentu), če pa obkljukamo parameter *Raw frequency*, gradnik vrne absolutno frekvenco oziroma število pojavitev vzorca.

- *Tweet Cleaner*: Ta gradnik kot vhod dobi stolpec korpusa v tabelarični obliki in za vsak dokument vrne dokument brez povezav (URL-jev), omemb (angl. *mentions*) in oznak (angl. *hashtags*). Uporabnik lahko izbira med odstranitvijo teh entitet in njihovo nadomestitvijo z žetoni *HTTPURL*, *TWEETMENTION* in *HASHTAG*. Gradnik je uporaben predvsem v primeru, če želi uporabnik procesirati in izluščiti informacije iz korpusa tvitov, saj so omembe in oznake zelo specifične za to vrsto besedil. Gradnik se lahko uporabi tudi pri obdelavi drugih besedil v vseh jezikih, če bi želeli na primer odstraniti povezave do spletnih strani iz poljubnega besedila.
- *Gender Classifier* (Martinc et al. 2017): Gradnik kot vhod dobi korpus v tabelarični obliki in vrne korpus v tabelarični obliki z dodatnim stolpcem, ki vsebuje avtomatsko pripisano oznako za spol avtorja posameznega teksta v korpusu. Uporabnik mora kot vhodni parameter podati ime stolpca v korpusu, ki vsebuje tekste. Klasifikator je bil naučen na tvitih in omogoča klasifikacijo slovenskih, angleških, španskih, portugalskih in arabskih tvitov (privzet jezik je slovenski). Za točno klasifikacijo potrebujemo besedilo, ki je daljše od enega tvita, idealno je to dokument, sestavljen iz dvesto ali več tvitov posameznega avtorja.

3.3 Gradniki za upravljanje s korpusom

Korpus besedil z vsemi pripadajočimi anotacijami je v platformi *ClowdFlows* zapisan v tabelarični obliki. Ta oblika zapisa omogoča različne vrste pretvorb, izborov in filtriranja korpusa. Za te vrste operacij so bili implementirani naslednji gradniki:

- *Select Corpus Attribute*: Gradnik kot vhod dobi korpus v tabelarični obliki in vrne stolpec korpusa z imenom, ki ga definira uporabnik. Stolpec je vrnjen v obliki pythonskega seznama, kjer posamezen element v seznamu predstavlja eno vrstico stolpca (tekst ali anotacije posameznega dokumenta v korpusu). Na ta način je mogoče iz korpusa v tabelarični obliki izločiti le informacije, ki jih potrebujemo za nadaljnjo obdelavo, kar zmanjša časovno in spominsko kompleksnost nadaljnjega procesiranja.
- *Concatenate Corpora*: Gradnik kot vhod dobi množico korpusov v tabelarični obliki in vrne vse vhodne korpuse, združene v en sam korpus v tabelarični obliki. Omejitev gradnika je, da morajo imeti vsi vhodni korpusi enako obliko (enako število in imena stolpcev).

- *Group by Column*: Gradnik kot vhod dobi korpus v tabelarični obliki in vrne korpus, ki ima vrednosti vseh atributov združene glede na vrednost atributa, ki ga definira uporabnik. Za primer lahko vzamemo korpus 30 tisoč tvitov, napisanih s strani 100 različnih avtorjev. Uporabnik določi stolpec z imeni avtorjev kot stolpec, ki naj vodi združevanje korpusa, gradnik pa vrne 100 dokumentov, pri čimer vsak dokument vsebuje s presledkom združene tvite posameznega avtorja.
- *Build Dataframe From Corpus*: Gradnik kot vhod dobi posamezne stolpce korpusa enake dolžine in vrne korpus v tabelarični obliki, sestavljen iz omenjenih stolpcev. Uporabnik lahko definira imena stolpcev izhodnega korpusa (v obliki seznama z vejico ločenih imen stolpcev v istem vrstnem redu, kot je vrstni red vhodnih stolpcev), če tega ne stori, so stolpci privzeto poimenovani kot *column_1*, *column_2*, *column_3* ...
- *Corpus to CSV*: Gradnik kot vhod dobi korpus v tabelarični obliki in kot izhod vrne datoteko CSV, v kateri je zapisan korpus v formatu CSV.
- *Filter Corpus*: Gradnik, ki omogoča filtriranje dokumentov v korpusu s pomočjo različnih poizvedb. Gradnik kot vhod dobi korpus v tabelarični obliki in vrne filtriran korpus v tabelarični obliki. Uporabnik lahko korpus filtrira s pomočjo petih različnih fraz za filtriranje:
  - *Je enako* – fraza ima obliko *ime_stolpca == poizvedba*, vrne pa se korpus v tabelarični obliki, ki vsebuje le dokumente, pri katerih vrednost posameznega stolpca ustreza iskani poizvedbi.
  - *Ni enako* – fraza ima obliko *ime_stolpca != poizvedba* in vrne le dokumente, pri katerih vrednost v stolpcu ni enaka iskani poizvedbi.
  - *Je večje* – fraza ima obliko *ime_stolpca > število* in vrne le dokumente, pri katerih je vrednost v stolpcu večja od števila. Ta poizvedba deluje le pri stolpcih z numeričnimi vrednostmi.
  - *Je manjše* – fraza ima obliko *ime_stolpca < število* in deluje na enak način kot *Je večje*, le da vrne obraten rezultat.
  - *Je v* – fraza ima obliko *poizvedba in ime_stolpca* in vrne le dokumente, ki vsebujejo iskano poizvedbo.

3.4 Gradniki za izračun in prikaz statistik

V tej kategoriji je bil implementiran le gradnik *Display Corpus Statistics*, ki kot vhod dobi stolpec korpusa. Ta gradnik omogoča izračun več različnih statistik:

- frekvenco posameznih pojavnic v korpusu (absolutne in relativne frekvence),

- izpis *hapax legomena*, ki so besede, ki se v korpusu pojavijo le enkrat, oz. *dis legomena*, ki se pojavijo točno dvakrat,
- izpis bigramskih in trigramskih kolokacij, ki temeljijo na meri povezanosti besed PMI (*pointwise mutual information*) (Church in Hanks 1990).

Pri izračunu statistik lahko uporabnik definira, za kakšne pojavnice želi izračun želene statistike. Izbira lahko med besednimi unigrami, bigrami, trigrami, tetragrami, pentagrami ali sekstagrami. Gradnik pri vsakem zagonu vrne tudi izračunano raznolikost besedišča (angl. type-token ratio), število dokumentov, število pojavnic in povprečno število pojavnic na dokument. Glede na vhodni gradnik lahko frekvence izpišemo tako za leme kot npr. za pojavnice ali oblikoskladenjske oznake.

4 PRIMERI DELOTOKOV

V tem razdelku bomo predstavili dva primera uporabe delotokov za obdelavo naravnega jezika, ki jih je mogoče implementirati v platformi *ClowdFlows* s pomočjo novih gradnikov. Oba primera sta javno objavljena in prikazujeta primer uporabe novih gradnikov na dveh različnih nalogah, s katerima se jezikoslovci pogosto srečujejo - izgradnja novega korpusa in analiza obstoječega korpusa. V obeh primerih gre za procesiranje računalniško posredovane komunikacije.

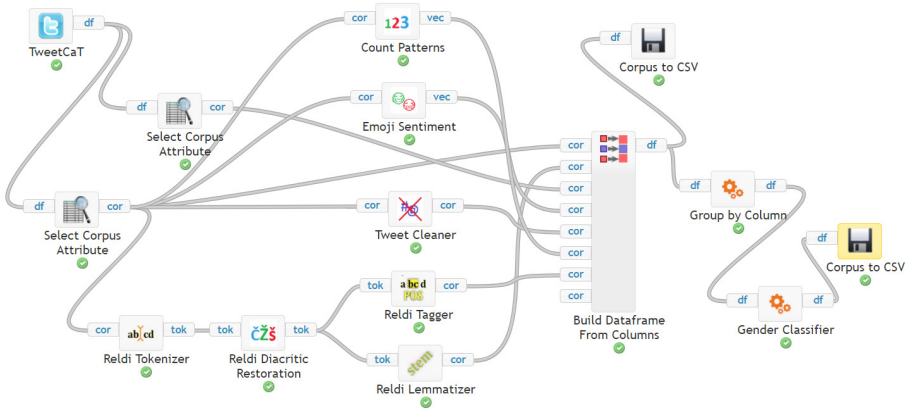
4.1 Nabor korpusa z orodjem TweetCaT

V tem razdelku je prikazano, kako lahko s pomočjo ClowdFlows platforme na preprost način zgradimo nov korpus slovenskih tvitov. Slika 4 prikazuje delotok, prav tako je delotok javno dostopen na spletu.[9] S pomočjo gradnika *TweetCaT* smo najprej nabrali okoli 23.000 slovenskih tvitov. Vsi parametri gradnika so bili privzeti, dodani so bili le podatki, potrebni za uporabo lastnega Twitter API-ja. Z gradnikom *TweetCaT* v naslednji fazi povežemo dva gradnika *Select Corpus Attribute*. Prvi iz korpusa izloči stolpec z imeni avtorjev tvitov, ki jih želimo imeti v izhodnem korpusu, drugi pa izloči stolpec z besedili. Besedila v nadaljevanju obdelamo na več načinov. Najprej s pomočjo gradnika *Count Patterns* izračunamo frekvenco emojijev za vsak posamezen tvit. S pomočjo gradnika *Emoji Sentiment* dobimo sentiment posameznega tvita, ki je izračunan na podlagi emojijev v tvitu. V izhodnem korpusu želimo imeti tudi prečiščene tvite brez oznak, omemb in oznak URL, zato uporabimo gradnik *Tweet Cleaner*.

9 http://clowdflows.org/workflow/10619/

Da bi dobili leme in oblikoskladenjske oznake za vse tvite, izpeljemo naslednje zaporedje operacij. Besedila najprej tokeniziramo s pomočjo gradnika *Reldi Tokenizer*, pri čemer pazimo, da izberemo parameter *Non-standard text*. Izhod gradnika nato povežemo z gradnikom *Reldi Diacritic Restoration*, ki normalizira besedila z obnovo manjkajočih diakritičnih oznak.

Ta gradnik nato povežemo z gradnikoma *Reldi lemmatizer* in *Reldi Tagger*, ki določita leme in oblikoskladenjske oznake. V končni fazi vse stolpce, ki jih želimo imeti v izhodnem korpusu (imena avtorjev, neprečiščena besedila, prečiščena besedila, frekvenco emojijev, sentiment na podlagi emojijev, leme in oblikoskladenjske oznake) združimo v nov korpus s pomočjo gradnika *Build Dataframe From Columns.* Korpus nato zapišemo v datoteko CSV s pomočjo gradnika *Corpus to CSV*.



Slika 4: Delotok TweetCaT.

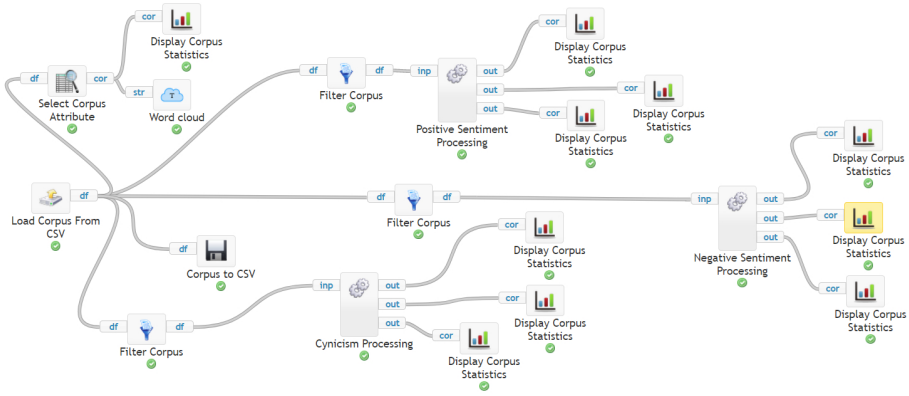
Ker želimo pridobiti tudi informacije o spolu avtorjev tvitov, dodamo še tri dodatne korake obdelave. Za začetek združimo vsa besedila posameznega avtorja v en dokument s pomočjo gradnika *Group by Column*, ki mu kot parameter damo ime stolpca, ki vsebuje imena avtorjev. Ta gradnik nato povežemo z gradnikom *Gender Classifier*, ki korpusu doda stolpec z oznakami spola. Na koncu tudi ta korpus zapišemo v datoteko s pomočjo gradnika *Corpus to CSV*.

Končni rezultat delotoka sta dva na novo izdelana korpusa v obliki CSV. Prvi vsebuje okoli 23.000 dokumentov in 7 stolpcev z različnimi informacijami o tvitih. Drugi vsebuje 10 dokumentov, en dokument na avtorja, sestavljen iz vseh dokumentov posameznega avtorja, in 8 stolpcev, saj je dodan stolpec z oznakami spola avtorjev.

Obstoječi delotok je mogoče uporabiti na dva načina. Če želi uporabnik delotok uporabiti na novih podatkih, bo moral najprej aktivirati tok po postopku, opisanem pri gradniku *TweetCaT*. To bo sprožilo nov zajem tvitov, uporabnik pa bo na ta način pridobil nov korpus tvitov z želenimi parametri, na katerih bodo preostali gradniki v delotoku izračunali enake statistike, kot so bile izračunane na originalnih podatkih. Če pa želi uporabnik preveriti dobljene rezultate, lahko ponovno zažene vse gradnike objavljenega delotoka. Seveda pa se lahko v obeh primerih delotok tudi poljubno prilagaja in spreminja glede na lastne potrebe.

4.2 Analiza ročno izbranega in označenega korpusa: primer Evrovizije

V tem razdelku prikažemo, kako lahko v platformi ClowdFlows obdelujemo obstoječe korpuse z različnimi anotacijami. Za primer vzamemo analizo korpusa Eurosong, ki je bila objavljena v prispevku Zwitter Vitez in Fišer (2016), tokrat pa ji dodajamo možnost avtomatskega označevanja, lematizacije in izračuna osnovnih statistik. Korpus Eurosong zajema komentarje na portalu rtvslo.si, ki so jih uporabniki objavili kot odziv na novico, da se je slovenska predstavnica na tekmovanju za Pesem Evrovizije uvrstila v finale. Komentarji izražajo ali odrekajo podporo slovenski predstavnici in omenjenemu tekmovanju nasploh, zato je bila naravnanost komentarjev najprej ročno označena s kategorijama pozitivnega in negativnega sentimenta, nato pa smo ročno dodali tudi oznake o podsentimentu (cinizem, kritika, optimizem, pesimizem). Cilj analize je identificirati jezikovne razlike med različno naravnanimi komentarji. Slika 5 prikazuje delotok, ki je dostopen na naslednji povezavi: http://clowdflows.org/workflow/10980/.

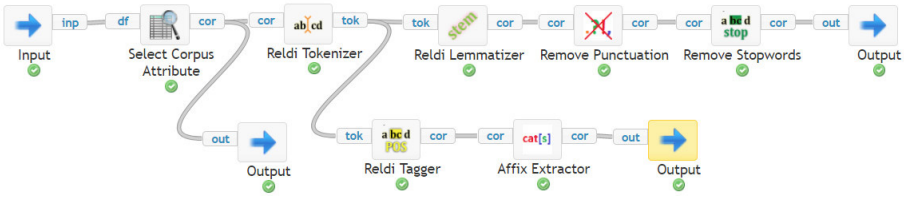


Slika 5: Delotok Eurosong.

Za izhodišče vzamemo korpus komentarjev Evrovizije v obliki CSV in ga naložimo s pomočjo gradnika *Load Corpus From CSV*. Korpus vsebuje ročne anotacije sentimenta in podsentimenta, ni pa še avtomatsko označen in lematiziran. Za začetek si želimo ogledati splošne statistike o besedilih v korpusu, zato s pomočjo gradnika *Select Corpus Attribute* izberemo le stolpec z besedili in ga povežemo z gradnikoma *Display Corpus Statistics* in *Word cloud,* saj nas zanima nekaj splošnih statistik o korpusu (število dokumentov, število pojavnic, povprečno število pojavnic na dokument), ki so prikazane v Tabeli 1, in pa najpogostejše polnopomenske pojavnice v korpusu (Slika 6). Vidimo, da se med zanimivimi polnopomenskimi besedami pojavljajo pojavnice *finale*, *Tinkara*, *uvrstitev, upam* ipd.

V naslednjem koraku želimo ugotoviti, kakšne so jezikovne razlike med pozitivnimi in negativnimi komentarji. Zato korpus razdelimo na dva podkorpusa glede na sentiment komentarjev. To storimo s pomočjo dveh gradnikov *Filter Corpus*, ki ju povežemo z gradnikom *Load Corpus From CSV*. S prvim gradnikom *Filter Corpus* želimo dobiti le podkorpus dokumentov s pozitivnim sentimentom, zato nastavimo parameter *query* na *Sentiment == positive*. Z drugim gradnikom *Filter Corpus* pa želimo dobiti podkorpus dokumentov z negativnim sentimentom, zato nastavimo parameter *query* na *Sentiment == negative*. Posebej nas zanimajo tudi cinični komentarji, zato tudi te izločimo iz korpusa s pomočjo gradnika *Filter Corpus,* ki ima parameter *query* nastavljen na *Subsentiment == cynicism.* Vse tri dobljene podkorpuse v nadaljevanju obdelujemo ločeno v treh skoraj identičnih podprocesih (*Positive Sentiment Processing, Negative Sentiment Processing, Cynicism Processing*), ki jih prikazuje Slika 7. Edina razlika med tremi procesi je, da podprocesa *Positive Sentiment Processing* in *Negative Sentiment Processing* ne vsebujeta gradnika *Remove Punctuation* (kar bo razloženo v nadaljevanju).

Slika 6: Najpogostejše polnopomenske pojavnice v korpusu Eurosong.



Slika 7: Podproces za obdelavo podkorpusov korpusa komentarjev Evrovizije.

V podprocesih iz korpusa izločimo le besedila s pomočjo gradnika *Select Corpus Attribute* ter ta gradnik povežemo z izhodom podprocesa (*Output*). Na ta izhod nato pri vseh podprocesih navežemo gradnik *Display Corpus Statistics*, da dobimo splošne statistike o vseh podkorpusih, ki jih prikazuje Tabela 1. Iz tabele je razvidno, da so dokumenti s pozitivnim sentimentom v povprečju krajši kot tisti z negativnim sentimentom, vendar hkrati opozarjamo pred posplošitvijo teh rezultatov, saj korpus zaradi svoje majhnosti statistično gledano ne dovoljuje zanesljivih zaključkov in je v tej raziskavi uporabljen predvsem za ponazoritev možne uporabe.

Tabela 1: Splošne statistike o vseh, pozitivnih, negativnih in ciničnih komentarjih v korpusu Eurosong.

Korpus	Število dokumentov	Število pojavnic	Povp. število pojavnic na dokument
Vsi dokumenti	70	894	12,77
Dokumenti s pozitivnim sentimentom	33	341	10,33
Dokumenti z negativnim sentimentom	37	553	14,95
Cinični dokumenti	20	284	14,20

Podobne ugotovitve so podane tudi v prispevku Zwitter Vitez in Fišer (2016), kjer je bilo z ročno analizo skladenjske strukture pozitivnih in negativnih komentarjev ugotovljeno, da imajo pozitivni komentarji pogosteje zgradbo enostavčne povedi (*Imam dober občutek*), medtem ko so negativni komentarji pogosteje formulirani v obliki večstavčne povedi (*Če zaupaš našim medijem, so še skoraj vsako leto bile kritike glede naših pesmi pozitivne, ampak rezultata pa nobenega in isto bo letos)*. Možno razlago za zaznano razliko na ravni dolžine in skladnje komentarjev vidimo v dejstvu, da pri izražanju negativnega mnenja uporabnik čuti dolžnost, da svojo kritiko utemelji, pri izražanju podpore pa se s tem ne ukvarja.

Nato se v podprocesih lotimo leksikalne analize besedil, ki jih tokeniziramo s pomočjo gradnika *Reldi Tokenizer* in lematiziramo z uporabo gradnika *Reldi Lemmatizer*, obenem pa odstranimo funkcijske besede z gradnikom *Remove Stopwords*. Seznam lem povežemo z izhodom (*Output*) podprocesa, ki ga v glavnem procesu povežemo z gradnikom *Display Corpus Statistics* (Slika 5).

N-gram	Raw frequency	Frequency	N-gram	Raw frequency	Frequency
,	43	0.1150	,	29	0.1189
.	26	0.0695	.	15	0.0615
...	9	0.0241	!	12	0.0492
finale	7	0.0187	tinkara	9	0.0369
iti	6	0.0160	finale	7	0.0287
spet	5	0.0134	dober	5	0.0205
res	5	0.0134	sreča	4	0.0164
pesem	4	0.0107	upati	4	0.0164
naprej	4	0.0107	:)	3	0.0123
nastop	4	0.0107	pesem	3	0.0123
zadnji	4	0.0107	pripravljen	3	0.0123
leto	3	0.0080	verjeti	3	0.0123
en	3	0.0080	srečno	3	0.0123
izpasti	3	0.0080	zaželeti	3	0.0123
upati	3	0.0080	iti	2	0.0082
....	3	0.0080	pozitiven	2	0.0082
danes	3	0.0080			

Slika 8: Najpogostejše leme v negativnih (levo) in pozitivnih (desno) komentarjih korpusa Eurosong.

Seznam najpogostejših lem v pozitivnih in negativnih komentarjih (Slika 8) pokaže zanimive razlike med obema podkorpusoma. Čeprav se v obeh podkorpusih pojavlja nekaj enakih lem, npr. *iti* in *finale*, opazimo zanimive razlike: v pozitivnih komentarjih se pojavi ime izvajalke *Tinkara*, v negativnih pa ne. V pozitivnih komentarjih prevladujejo leme, kot so *sreča, dober, verjeti, držati pest*, ki bi se lahko pojavili tudi pri izražanju podpore na kakšnem drugem področju (npr. pri športu). Med negativnimi komentarji pa so najpogostejše leme, bolj tipično vezane na tekmovanje za Pesem Evrovizije (*pesem, nastop, izpasti, kuhna*).

Pri analizi pozitivnih in negativnih komentarjev namenoma nismo uporabili gradnika *Remove punctuation*, ker smo želeli preveriti rabo ločil v pozitivnih in negativnih komentarjih. Pike in vejice so v obeh podkorpusih zelo pogoste, takoj za njima pa se v podkorpusu pozitivnih komentarjev pojavita klicaj in dvopičje z oklepajem v funkciji emotikona (*Dajmo klobasica!!!:*), medtem ko je pri negativnih precej bolj prisotno ločilo tri pike (*tokrat bomo res nesrečno izpadli...66.667% imamo možnosti da gremo naprej...hočem rečt, če danes ne zguramo naprej, pol nam res ni več pomoči*). Na podlagi zaznane razlike v rabi ločil

sklepamo, da v pozitivnih komentarjih uporabniki izražajo več čustev kot v negativnih, ko poskušajo svoje mnenje karseda racionalno utemeljiti.

Na koncu izvedemo v vseh podprocesih še analizo na besednovrstni ravni (Slika 9), ki v raziskavi Zwitter Vitez in Fišer (2016) ni bila mogoča. To storimo s pomočjo gradnika *Reldi Tagger,* ki vrne oznake MSD. Zanima nas le prvi znak v oznaki MSD, ki označuje besedno vrsto, zato ta gradnik povežemo z gradnikom *Affix Extractor* s parametroma *Affix type* in *Affix length,* nastavljenima na *prefix* in 1. Ta gradnik nato navežemo na izhod podprocesa, na katerega nato v glavnem procesu (Slika 5) navežemo gradnik *Display Corpus Statistics.*

N-gram	Raw frequency	Frequency	N-gram	Raw frequency	Frequency
N	125	0.1900	N	80	0.1975
V	122	0.1854	V	74	0.1827
R	78	0.1185	Z	60	0.1481
Z	77	0.1170	C	48	0.1185
C	64	0.0973	P	42	0.1037
P	51	0.0775	R	31	0.0765
Q	46	0.0699	A	23	0.0568
S	44	0.0669	S	22	0.0543
A	29	0.0441	Q	15	0.0370
Y	14	0.0213	M	8	0.0198
M	7	0.0106	Y	1	0.0025
X	1	0.0015	X	1	0.0025

Slika 9: Zastopanost besednih vrst (negativni komentarji levo, pozitivni desno).

Poleg samostalnikov in glagolov, ki so izrazito pogosti v obeh podkorpusih, pri pozitivnih komentarjih izstopajo ločila, kar smo pripisali večji čustveni nabitosti sporočil (*Tinkara je res profi in vsaka ji čast !!!*), pri negativnih komentarjih pa izstopajo prislovi (*važno, odvisno, nesrečno*), ki poudarijo naravnanost avtorja besedila do dogajanja v sporočilu.

Poleg osnovne naravnanosti avtorjev smo v korpusu ročno označili tudi podsentiment in posebej analizirali ciničnе pripombe. Te pripombe obdelujemo na enak način kot pozitivne in negativne komentarje, z izjemo že prej omenjenega dodatnega koraka izločitve ločil s pomočjo gradnika *Remove Punctuation*. Seznam najpogostejših lem (Slika 10) v ciničnih komentarjih omogoča nadaljnjo analizo.

Za razliko od specifik negativnih komentarjev, v katerih izstopata npr. lemi *izpasti* in *zadnji*, je v ciničnih komentarjih opaziti nekoliko drugačno leksikalno zastopanost. Na vrhu seznama se znajdejo leme, ki na prvi pogled delujejo nevtralno ali celo pozitivno, kvalitativna analiza pa pokaže, da imajo ciničen ali ironičen prizvok. Tako vlogo imajo na primer leme *finale* (*Finale ja... V soboto med publiko :)*), *dopust* (*važn da smo mišo peljali na dopust*), in *molkov*, ki je lema, pripisana pojavnici Molkova (*Sorry admin me bo brisal a resnica zgleda boli ali je to direktiva RTV-ja, da ne bi kdo*

rekel kaj čez Molkovo?). Ostale primerjave med ciničnimi in neciničnimi komentarji (npr. dolžina komentarjev, oznake MSD) bi bile smiselne pri večjem korpusu.

N-gram	Raw frequency	Frequency
finale	5	0.0360
pesem	3	0.0216
leto	3	0.0216
en	3	0.0216
dati	3	0.0216
dopust	2	0.0144
res	2	0.0144
publika	2	0.0144
kea	2	0.0144
zadnji	2	0.0144
molkov	2	0.0144
potem	2	0.0144
it	2	0.0144
lahko	2	0.0144
skuhati	2	0.0144
dobro	2	0.0144
kuhinja	2	0.0144
mrbit	2	0.0144

Slika 10: Najpogostejše leme v ciničnih komentarjih korpusa Eurosong.

Uporaba delotoka na korpusu Eurosong je pokazala, kako je mogoče preveriti hipoteze, ki so nastale s kvalitativnimi raziskavami (npr. glede leksikalnih značilnosti besedil), in analizirati besedila na novih ravneh (npr. na ravni oznak MSD), ki pred vključitvijo v orodja niso bile mogoče. Delotok prikazuje, kako lahko na hiter in enostaven način izvedemo osnovno statistično in primerjalno analizo besedil na leksikalni, besednovrstni in znakovni ravni tudi na poljubnih novih označenih korpusih.

5 SKLEP

V poglavju smo predstavili implementacijo orodij za obdelavo naravnega jezika v obstoječo platformo za vizualno programiranje *ClowdFlows*, ki omogoča lažjo in hitrejšo analizo besedil. Na kratko smo opisali platformo *ClowdFlows* in njen uporabniški vmesnik za vizualno programiranje ter podrobneje opisali na novo implementirana orodja, ki so bila v okviru projekta JANES razvita za gradnjo korpusa tvitov in obdelavo nestandardne slovenščine, pa tudi vrsto orodij za obdelavo kakršnih koli besedilnih korpusov. Kot primer uporabe orodij smo predstavili delotok za izdelavo in označevanje novega korpusa tvitov ter delotok za analizo komentarjev Evrovizije.

Nova orodja širijo nabor možnosti kvantitativnih analiz, ki jih je mogoče izvajati na obširnejših korpusih besedilih brez znanja programiranja in brez zapletenih pretvorb v formate, primerne za nadaljnje procesiranje v obstoječih orodjih za obdelavo besedil. Pri implementaciji orodij smo poseben poudarek namenili obdelavi slovenskih besedil, saj se ravno pri jezikih z manj govorci najbolj jasno kaže pomanjkanje orodij za analizo.

Primeri delotokov kažejo, da platforma *ClowdFlows* omogoča veliko svobode pri korpusnih analizah in dokaj enostavno implementacijo zapletenih znanstvenih postopkov. Kolaborativna narava in odprtokodnost platforme pa omogočata, da tudi drugi razvijalci dodajo nove gradnike in tako omogočijo nove možnosti analize.

Oba primera delotoka ponujata rešitev, kako omogočiti jezikoslovno analizo besedil, ki nastajajo kot odziv na aktualna družbena dogajanja. Obstoječi zaključeni korpusi, vključno s korpusom Janes, so namreč zelo aktualni z vidika preučevanja jezika računalniško posredovane komunikacije, vendar z vidika družbenih tematik, ki jih pokrivajo, hitro zastarajo. Korpus lahko pripravimo ročno in ga vnesemo, kot smo prikazali na primeru korpusa *Eurosong* v razdelku 4.2, ali pa korpus s poljubnega družbeno aktualnega področja sestavimo z orodjem *TweetCaT*.

Načrt nadaljnjega dela je večplasten. V prvi fazi želimo razširiti nabor orodij za obdelavo naravnega jezika v smislu funkcionalnosti in podpore drugih jezikov. Trenutno platforma *ClowdFlows* nima orodij za lematizacijo in oblikoskladenjsko označevanje angleščine, ki pa so podprta v veji *TextFlows*, kar nameravamo v prihodnje združiti v okviru platforme *ClowdFlows 3.0*. Manjkajo tudi orodja za izdelavo odvisnostne drevesnice (angl. *dependency trees*) in orodja za prepoznavanje imenskih entitet (angl. *named entity recognition*). Druga faza, ki je dolgoročnejša in že poteka, je optimizacija same platforme v smislu hitrejšega procesiranja in manjše porabe pomnilnika. Prav tako se bo izboljšala modularnost strukture platforme, kar bo omogočilo večjo prilagodljivost in lažjo implementacijo novih gradnikov.

Zahvala

Raziskava, opisana v prispevku, je bila delno opravljena v okviru projekta »ClowdFlows spletno tržišče za podatkovno in tekstovno analitiko« (RIA CF-Web, 2017–2018, H2020).

Literatura

Anthony, Laurence, 2013: A critical look at software tools in corpus linguistics. *Linguistic Research* 30/2.141–161.

Anthony, Laurence, 2014: AntConc (različica 3.4.3). Tokio: Waseda University.

Berthold, Michael R., Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Kilian Thiel in Bernd Wiswedel, 2009: KNIME-the Konstanz information miner: version 2.0 and beyond. *AcM SIGKDD explorations Newsletter* 11/1. 26–31.

Church, Kenneth Ward in Patrick Hanks, 1990: Word association norms, mutual information, and lexicography. *Computational linguistics* 16/1. 22–29.

Demšar, Janez, Blaž Zupan, Gregor Leban in Tomaž Curk, 2004: Orange: From experimental machine learning to interactive data mining. *Knowledge discovery in databases: PKDD 2004*. 537–539.

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

Kilgarriff, Adam, Pavel Rychly, Pavel Smrz in David Tugwell, 2004: Itri-04-08 the sketch engine. *Information Technology* 105. 116.

Kranjc, Janez, Vid Podpečan in Nada Lavrač, 2012: ClowdFlows: A Cloud Based Scientific Workflow Platform. *Proceedings of ECML/PKDD (2)*. 816–819.

Ljubešić, Nikola, Darja Fišer in Tomaž Erjavec, 2014: TweetCaT: a tool for building Twitter corpora of smaller languages. *Proceedings of LREC*. 2279–2283.

Ljubešić, Nikola, Tomaž Erjavec in Darja Fišer, 2016: Corpus-based diacritic restoration for south slavic languages. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016*). European Language Resources Association (ELRA). 3612–3616.

Ljubešić, Nikola in Tomaž Erjavec, 2016: Corpus vs. lexicon supervision in morphosyntactic tagging: the case of Slovene. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).* European Language Resources Association (ELRA). 1527–1531.

Ljubešić, Nikola, Tomaž Erjavec in Darja Fišer, 2018: Orodja za procesiranje nestandardne slovenščine. Fišer, Darja (ur.): *Viri, orodja in metode za analizo spletne slovenščine*. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani. 74–99.

Martinc, Matej, Iza Škrjanec, Katja Zupan in Senja Pollak, 2017: PAN 2017: Author Profiling-Gender and Language Variety Prediction. *Working Notes Papers of the CLEF 2017 Evaluation Labs. CEUR Workshop Proceedings*.

McKinney, Wes, 2011: Pandas: a foundational Python library for data analysis and statistics. *Python for High Performance and Scientific Computing*. 1–9.

Mierswa, Ingo, Michael Wurst, Ralf Klinkenberg, Martin Scholz in Timm Euler, 2006: Yale: Rapid prototyping for complex data mining tasks. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. 935–940.

Kralj Novak, Petra, Jasmina Smailović, Borut Sluban in Igor Mozetič, 2015: Sentiment of emojis. *PloS one* 10/12.

Perovšek, Matic, Janez Kranjc, Tomaž Erjavec, Bojan Cestnik in Nada Lavrač, 2016: TextFlows: A visual programming platform for text mining and natural language processing. *Science of Computer Programming* 121. 128–152.

Pollak, Senja, Anže Vavpetič, Janez Kranjc, Nada Lavrač in Špela Vintar, 2012a: NLP workflow for online definition extraction from English and Slovene text corpora. *KONVENS*. 53–60.

Pollak, Senja, Nejc Trdin, Anže Vavpetič in Tomaž Erjavec, 2012b: NLP web services for Slovene and English: morphosyntactic tagging, lemmatisation and definition extraction. *Informatica* 36/4. 441–449.

Sapkota, Upendra, Steven Bethard, Manuel Montes-y-Gómez in Thamar Solorio, 2015: Not All Character N-grams Are Created Equal: A Study in Authorship Attribution. *HLT-NAACL*. 93–102.

Scott, Mike, 1998: WordSmith Tools Version 3. Oxford: Oxford University Press.

Zwitter Vitez, Ana in Darja Fišer, 2016: Linguistic Analysis of Emotions in Online News Comments-an Example of the Eurovision Song Contest. *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities*. Ljubljana, Slovenia, 27–28 September 2016. 74–76.

Zapisovalne prakse v spletni slovenščini

Darja Fišer, Maja Miličević Petrović, Nikola Ljubešić

Izvleček

V poglavju obravnavamo značilnosti nestandardnega zapisa besed v slovenskih objavah na družbenem omrežju Twitter. Analiza temelji na ročno normaliziranih, lematiziranih in oblikoskladenjsko označenih vzorcih tvitov v slovenščini. Podrobneje proučimo distribucijo pretvorb iz standardnega v nestandardni zapis besed glede na besedne vrste in leme ter distribucijo treh različnih vrst transformacij: izpust, dodajanje in zamenjavo črk. Rezultati kažejo, da je največ transformacij med polnopomenskimi besedami, vendar so transformacije slovničnih besed najpogostejše. Najpogostejši tip transformacij so izpusti, predvsem samoglasnikov, do česar največkrat prihaja na koncu besed, s čimer se neformalna komunikacija v tvitih približuje govoru.

Ključne besede: nestandardni zapis besed, računalniško posredovana komunikacija, Twitter, slovenščina

1 UVOD

Nekonvencionalno, nekanonično oz. nestandardno zapisovanje besed je ena od najopaznejših značilnosti računalniško posredovane komunikacije (RPK), prav tako pa povzroča tudi največ težav pri avtomatski obdelavi takšnih besedil, zato ni presenetljivo, da temu fenomenu RPK tako jezikoslovci kot jezikovni tehnologi posvečajo veliko pozornosti. Zgodnje raziskave so bile večinoma opravljene na SMS-sporočilih, predvsem v angleščini (Shortis 2007, Tagg et al. 2012) in francoščini (Anis 2007), danes pa so za številne jezike v ospredju družbena omrežja Facebook (Chariatte 2014), Twitter (Van Halteren in Oostdijk 2012, Sidarenka et al. 2013) in WhatsApp (Ueberwasser 2013).

Metodološko in jezikovnogradivno heterogenim raziskavam so skupne ugotovitve, da nestandardno zapisovanje besed v RPK, če izvzamemo tipkarske napake, večinoma ni kaotično in kataklizmično, temveč v veliki meri namerno, predvsem pa »funkcionalno, načelno in osmišljeno« (Tagg et al. 2012: 367). Funkcionalno, ker se pojavi v interakciji znotraj določene družbene skupine kot odziv na konkretne komunikacijske potrebe in tipično nima negativnega vpliva na razumevanje sporočila med pripadniki te skupine, načelno, ker navadno odseva splošne vzorce ortografske variacije v tem jeziku, in osmišljeno, ker prispeva k udejanjanju družbenih identitet uporabnikov. Thurlow in Brown (2003) navajata tri osnovne značilnosti nestandardno zapisanih besed: kratkost in hitrost (krajšanje, nadomeščanje črk s številčnimi homofoni, opuščanje kapitalizacije, ločil, presledkov, nadomeščanje črk s števil), nadomeščanje paralingvističnih sredstev (raba velikih črk in ponavljanje ločil za nadomeščanje prozodičnih in čustvenih komponent) in fonološka aproksimacija (približevanje zapisa (neformalnemu) govoru).

Zapisovalnim praksam v slovenskih tvitih je bilo posvečenih že kar nekaj študij. Analiza strategij krajšanja v tvitih (Goli et al. 2016) je pokazala močno tendenco po krajšanju, ki je izrazito pogosta v nestandardnih sporočilih in se kaže predvsem kot redukcija na nivoju zapisa besed (*mam* za *imam, anglesk* za *angleško, lah* za *lahko*). Analiza novih skovank v različnih tipih slovenskih uporabniških vsebin iz korpusa Janes, ki vsebujejo homofone s črkami in številkami (Marko 2016), je pokazala, da so ti izrazito značilni za Twitter, da se pojavljajo enako pogosto v tujih in slovenskih besedah in da se isti simbol lahko uporablja grafično (*g33k* za *geek*) ali fonetično (*u3nek* za *utrinek*). Analiza platform, ki omogočajo interaktivno in takojšnjo komunikacijo, kot je na primer Twitter, pa je pokazala, da se v njih brišejo meje med govornim in pisnim diskurzom (Zwitter Vitez in Fišer 2015), kar je med drugim razvidno iz pogoste rabe fonetičnega zapisa besed, pogovornih izrazov, deiktike in nestandardne leksike.

Vsi ti rezultati jasno kažejo, da so nekanonične zapisovalne prakse pri neformalnem komuniciranju slovenskih uporabnikov na družbenih omrežjih zelo razširjene, manjkajo pa raziskave, ki bi jih empirično preverile ter sistematično opisale, kako se razlikujejo od standarda. To je cilj pričujočega prispevka, v katerem se posvečamo značilnostim nestandardnega zapisa besed v slovenskih objavah na družbenem omrežju Twitter. Predstavljena analiza sodi v okvir večjezične raziskave (Miličević et al. 2017), kjer smo primerjali zapisovalne prakse v slovenskih, hrvaških in srbskih tvitih, pri čemer je v pričujočem prispevku poudarek na interpretaciji rezultatov za slovenščino.

Analiza temelji na ročno normaliziranih,[1] lematiziranih in oblikoskladenjsko označenih vzorcih tvitov v slovenščini, ki so bili izdelani za razvoj orodij za avtomatično normalizacijo in označevanje besedil računalniško posredovane komunikacije (Čibej et al. 2018). V nadaljevanju najprej predstavimo vzorec, ki je bil uporabljen za analize, nato pa opišemo postopek in rezultate analize. V prvem delu se osredotočimo na analizo odmikov od standardnega zapisa glede na besedno vrsto in lemo, v drugem pa se posvetimo pregledu vrst odmikov.

2 OPIS VZORCA

V prispevku uporabljamo vzorec, ki smo ga izluščili iz korpusa Janes-Norm in vsebuje slovenske uporabniško generirane vsebine, ki so bile ročno normalizirane, lematizirane in označene (Čibej et al. 2018). Glede na to, da poglavje obravnava nestandardni zapis besed v tvitih, smo v vzorec za analizo zajeli zgolj jezikovno nestandardne tvite (Erjavec et al. 2018), ki predstavljajo 1983 tvitov oz. 54.688 pojavnic.

Primer tvita z nestandardnimi prvinami je prikazan na Sliki 1. Te prvine vključujejo fenomene, značilne za računalniško posredovano komunikacijo na splošno, kot so fonetični zapisi tujih besed (npr. *lajk* za *like*), izpust strešic (*razrednicarka* za *razredničarka*) ali okrajšave (npr. *yt* za *YouTube*), fenomene, značilne za Twitter, kot so ključniki, omembe imen z znakom @ ali emotikoni/emojiji, in fenomene, ki se pogosto uporabljajo v neformalnih komunikacijskih situacijah, kot je uporaba pogovornih in narečnih nestandardnih oblik (npr. *tko* za *tako* ipd.).

Smernice za označevanje so bile oblikovane v okviru projekta JANES, tviti pa so bili označeni na petih ravneh: pojavnica (popravljanje mej med besedami), stavek (popravljanje stavčne segmentacije), normalizacija (standardizacija nestandardnih jezikovnih prvin), lematizacija (pripisovanje osnovne oblike vsaki

1 Normalizacija je postopek pripisovanja standardne ustreznice nestandardni pojavnici v korpusu (npr. *js* – *jaz*, glej Čibej et al. 2018).

besedi v tekočem besedilu, npr. *objavili – objaviti*) in oblikoskladenjski opis (pripisovanje oblikoskladenjskih oznak vsaki besedi v tekočem besedilu glede na standard MULTEXT-East v5.0,[2] npr. *demona* > *Sometd* za *samostalnik*, *občno ime*, *moški spol*, *ednina*, *tožilnik*, *živost*) (Čibej et al. 2018).

@user99 vrjamm [Verjamem] ja :) nm [Nam] pa rece [reče] razrednicarka [razredničarka], da je naj do 6ihne [6-ih ne] budimo, in tko [tako] npr [npr.] smo bli [bili] ze [že] enkrt [enkrat] ob 4 zjutri [zjutraj] pred Louvrom :D

Slika 1: Primer nestandardnega tvita s pripisanimi normaliziranimi oblikami (Tvit [standardna oblika besede]).

Od vseh ravni označevanja, s katerimi je bil korpus Janes-Norm označen, je za pričujočo raziskavo najpomembnejša raven z jezikovno normalizacijo. Normalizacija je bila omejena na nivo besede, kar pomeni, da besedni red, skladnja, uporaba ločil, elipse, uporabniška imena, ključniki, emotikoni/emojiji in leksikalne izbire (npr. pogovorni izraz *mobi* za *mobitel*) niso bile normalizirane. Je pa normalizacija vključevala standardiziranje tako nestandardnih različic črkovanja (npr. *jst* > *jaz*) kot tudi napak v črkovanju in tipkanju (npr. *popoldme* > *popoldne*) ter rediakritizacijo (npr. *vceraj* > *včeraj*). Pri normalizaciji so se označevalci držali načela minimalne intervencije. Z drugimi besedami – osredotočili smo se na nestandardne oblike, ki jih lahko razumemo kot odstopanje od standardnega črkovanja, pri tem pa nismo vplivali na slog, slovnico in fenomene, značilne za Twitter. Po drugi strani pa smo za razliko od nekaterih sorodnih raziskav variantnosti zapisovanja besed (npr. van Halteren and Oostdijk 2012) normalizirali nestandardno morfologijo (npr. *hodu* > *hodil*), saj sta cilja naših raziskav dvojna: zagotavljanje gradiva za razvoj orodij za avtomatsko procesiranje nestandardnega jezika in za raziskovanje pojavov v nestandardni spletni slovenščini.

Pri razreševanju nejasnih in dvoumnih primerov (npr. *k* > *ki* v *stvar k je postala »slavna«*, ampak *k* > *kot* v *ameriške jopice izgledajo, k da so jih babice spletle*) so označevalci upoštevali kontekst, če pa primera niso mogli razrešiti z uporabo danega konteksta, beseda ni bila normalizirana. Nestandardna pojavnica je bila v večini primerov normalizirana v eno standardno pojavnico, v redkih primerih pa je ena nestandardna pojavnica morala biti razdeljena v več standardnih pojavnic (1:n, *nevem – ne vem*) in obratno (n:1, *vse eno – vseeno*). Delež pojavnic z 1:n v vzorcu znaša 0,47 %, z n:1 pa 0,06 %.

Pri določanju smernic za normalizacijo se je bilo potrebno tudi natančno opredeliti do oblik, ki jih obravnavamo kot nestandardne, kar nujno vključuje tudi vprašanje jezikovne norme in referenčnih jezikovnih virov, ki jih pri

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

označevanju upoštevamo. Vloga jezikovne norme in njen odnos do jezikovne rabe (preskriptiven : deskriptiven) sta v slovenistiki neusahljiv vir diskusij in polemik (glej Verovnik 2004 in Smolej 2015). V okviru predstavljene raziskave smo zavzeli deskriptivno stališče in nestandardnih oblik (z izjemo tipkarskih napak) ne obravnavamo kot napake, temveč kot variante, ki so v okoliščinah, v katerih so uporabljene, pretežno funkcionalne, načelne in osmišljene. Zato bi bilo zmotno našo normalizacijo interpretirati kot »popravljanje«, temveč nam služi kot pripomoček za lažjo avtomatsko obdelavo in analizo slovenščine, kot se uporablja v računalniško posredovani komunikaciji, kjer je standard razumljen kot skupni imenovalec, ne pa kot nabor neizpodbitnih pravil.

V smernicah se opiramo na splošno sprejete referenčne vire za standardno slovenščino, hkrati pa si prizadevamo upoštevati tudi realno jezikovno rabo. Zato smo anotatorje prosili, da se pri obravnavi povsem jasnih in neproblematičnih primerov, kot so manjkajoče strešice in očitne tipkarske napake, zanašajo na lastno intuicijo, v vseh ostalih primerih pa uporabijo referenčne priročnike v naslednjem vrstnem redu: (1) spletni portal Fran,[3] na katerem sta dostopna Slovar slovenskega knjižnega jezika in Slovenski pravopis, (2) oblikoslovni leksikon Sloleks,[4] (3) konkordančnik Gigafida[5] in (4) korpus Janes v0.4.[6] Uporaba korpusov je bila potrebna za pojavnice, ki niso zajete v referenčnih virih standardne slovenščine, še posebej pa za tiste, ki se pojavljajo v več različicah (npr. *fouš – fauš – favš*). V teh primerih smo anotatorje prosili, da jih normalizirajo v najpogostejšo obliko (npr. *fouš* v zgornjem primeru).

3 ANALIZA PODATKOV

V tem razdelku predstavimo rezultate analiz, ki so bile opravljene na normaliziranih slovenskih tvitih. Glede na to, da so smernice za normalizacijo temeljile na opisnih kategorijah, ki jih je avtomatsko težko identificirati (npr. fonetična transkripcija ali napačen zapis), smo se v tem prispevku omejili na analizo po avtomatsko določljivih kriterijih. S tem namenom smo se osredotočili na transformacije, tj. modifikacije, ki so se kazale z uporabo nestandardnega jezika v nasprotju s standardnim. Gre torej za nasprotni proces od ročne normalizacije, predstavljene v tretjem razdelku, kjer nestandardnim oblikam pripisujemo standardne različice (npr. *reko* > *rekel*). V naši analizi ta fenomen obravnavamo kot transformacijo standardne oblike *rekel* v nestandardno obliko *reko* z zamenjavo znakov.

3 http://www.fran.si

4 http://www.slovenscina.eu/sloleks

5 http://www.gigafida.net

6 Glej Erjavec et al. (2018).

Analizo smo izvedli za štiri ravni: (1) zapis izvornih pojavnic v primerjavi z (2) normaliziranimi,[7] (3) oblikoskladenjske oznake, pripisane normaliziranim pojavnicam in (4) leme, pripisane normaliziranim pojavnicam. Porazdelitev transformacij opazujemo glede na besedne vrste, prav tako pa izluščimo najpogosteje transformirane leme in pregibne oblike. Ko opazujemo pregibne oblike normaliziranih in izvornih pojavnic, klasificiramo razlike glede na Levenshteinove vrste transformacij (izpust, dodajanje, zamenjava; Levenshtein 1966), prav tako pa smo pozorni na položaj specifične transformacije znotraj besed.

3.1 Skupna pogostost transformacij

Skupni delež transformiranih pojavnic znaša 17,39 % (9.555 pojavnic). Pri nekaterih transformacijah gre zgolj za izpust strešic (č, ć, š, ž, đ > c, c, s, z, dj), ki so posledica tehničnih in ne jezikovnih razlogov (tipkanje na mednarodnih tipkovnicah je hitrejše brez uporabe strešic). Če te izločimo, ostane 15,56 % (8.552) transformiranih pojavnic. Rezultati so v skladu s predhodnimi raziskavami, ki kažejo, da je v slovenščini močnejša tendenca uporabe nestandardnih oblik kot izpuščanja diakritikov (Fišer et al. 2015, Miličević in Ljubešić 2016).

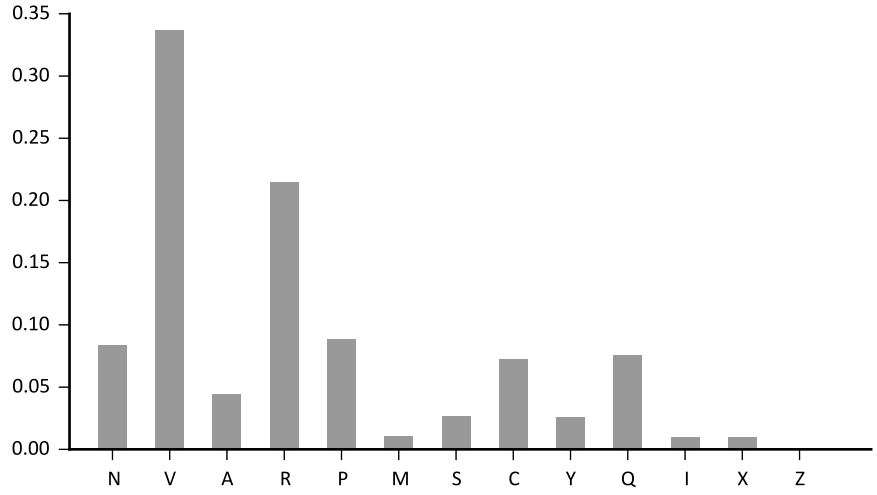
3.2 Analiza glede na besedne vrste

V tej analizi opazujemo porazdelitev transformacij glede na besedno vrsto (tj. koliko transformacij pripada določeni besedni vrsti). Prav tako izračunamo delež oblik, ki so bile transformirane za vsako besedno vrsto (tj. koliko besed izmed vseh, ki pripadajo določeni besedni vrsti, je bilo transformiranih). Obe analizi sta omejeni na pojavnice, kjer transformacija ni zajemala izpusta strešic.

3.2.1 Pogostost transformacij glede na besedno vrsto

Relativne frekvence transformacij glede na besedno vrsto so prikazane na Sliki 2. S slike je razvidno, da so najpogosteje transformirani glagoli, sledijo jim prislovi, zaimki in samostalniki.

7 Zaradi tehničnih omejitev platforme, na kateri je označevanje potekalo, je ena izvorna pojavnica lahko normalizirana v največ štiri pojavnice (npr. 1 > 2: *nevem* > *ne vem*, 1 > 3: *anede* > *a ne da*, 1 > 4: *norostinividetikonca* > *norosti ni videti konca*), prav tako je več izvirnih pojavnic lahko normaliziranih v eno samo pojavnico.



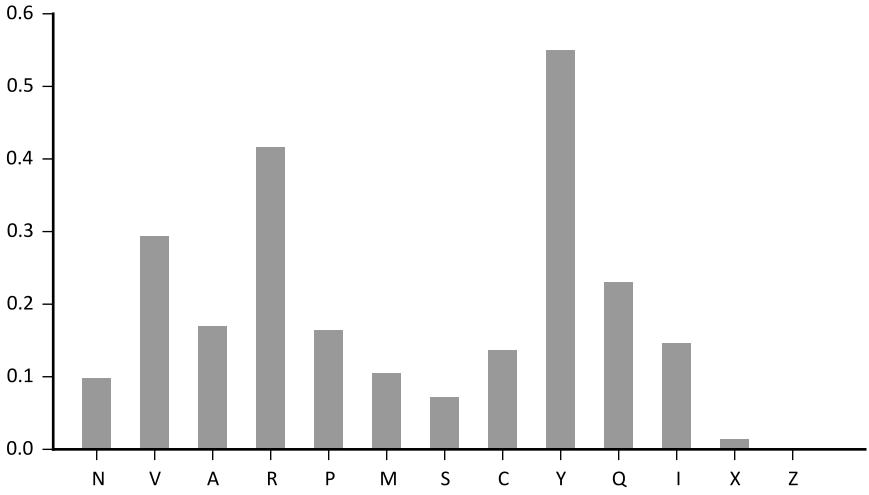
Slika 2: Distribucija transformiranih oblik glede na besedno vrsto[8] in relativno frekvenco.

Transformacije glagolov se v večini primerov nanašajo na pomožni glagol *biti*, predvsem pri obliki za prvo osebo ednine *sem* (pogosto zapisano kot *sm*) in pri obliki za tretjo obliko ednine v pretekliku *bilo* (skrajšano v *blo*). Pogoste so tudi transformacije drugih glagolov s krajšanjem nedoločnika, npr. *gledat* za *gledati*. Prislovi so večinoma okrajšani (npr. *tako* je pogosto skrajšan v *tko*), pojavljajo pa se tudi druge vrste transformacij. Zanimiv primer je *zdaj*, ki se v vzorcu kaže v treh različnih transformacijah, in sicer *zdej*, *zdj* in *zj*. Transformacije medmetov se nanašajo predvsem na ponovitev samoglasnikov ali zlogov (npr. *hahahaha*).

3.2.2 Deleži transformiranih oblik znotraj besedne vrste

Deleži oblik, ki so bile transformirane znotraj določene besedne vrste, kažejo, da so slovnične besedne vrste pogosteje transformirane kot polnopomenske. Najvišji delež transformiranih pojavnic najdemo med okrajšavami (pogosto gre za izpust končne pike, npr. *slo* namesto *slo.* za *slovenski*). Členki in vezniki so večinoma skrajšani z izpustom zadnjega samoglasnika, npr. *al* za *ali* in *kak* za *kako*. Najpogostejša transformirana oblika je osebni zaimek za prvo osebo ednine *jaz*, pogosto zapisana kot *jst*, *js*, *jest* ali *jz*.

8 Oznake besednih vrst so: N – samostalnik, V – glagol, A – pridevnik, R – prislov, P – zaimek, M – številnik, S – predlog, C – veznik, Y – okrajšava, Q – členek, I – medmet, X – neuvrščeno, Z – ločilo.



Slika 3: Deleži transformiranih oblik znotraj posameznih besednih vrst.

Med polnopomenskimi besednimi vrstami smo največ transformacij zasledili med prislovi, glagoli in pridevniki, kar sovpada s tendencami transformacij glede na besedne vrste, opisane v razdelku 3.2.1.

Pri prvi primerjavi zajemajo polnopomenske besede večino vseh transformacij, pri drugi pa vodijo funkcijske besede. Z drugimi besedami – čeprav so leksikalne besede pogostejše, jih transformiramo v manjši meri. To je tudi razlog, da leksikalne besede dominirajo na Sliki 2, na Sliki 3 pa ne.

3.3 Analiza glede na leme in pregibne oblike

V tem razdelku predstavimo analizi glede na najpogosteje transformirane leme (3.3.1) in pregibne oblike (3.3.2).

3.3.1 Analiza lem

V Tabeli 1 so predstavljene najpogosteje transformirane leme z deležem transformiranih oblik, ki jih pokriva določena lema (% skupaj, npr. *biti*) in deležem vseh oblik te leme, ki so bile transformirane (% lema, npr. *sm, blo, bla, nism, bit, bli, nebi, biu, sn, sm, neb, ble*). Transformacije z izpustom strešic ponovno niso upoštevane.

Tabela 1: 20 najpogosteje transformiranih lem v slovenskih tvitih.

Lema	% skupaj	% lema
biti#V	8,33 %	17,02 %
jaz#P	3,24 %	33,9 %
tudi#Q	3,13 %	82,21 %
imeti#V	3,09 %	66,5 %
saj#C	1,61 %	79,77 %
potem#R	1,49 %	73,41 %
tako#R	1,39 %	74,38 %
zdaj#R	1,34 %	76,16 %
malo#R	1,3 %	82,22 %
samo#Q	1,29 %	61,45 %
lahko#R	1,2 %	52,82 %
toliko#R	1,09 %	91,18 %
ne#Q	1,06 %	11,15 %
kaj#P	1,05 %	36,29 %
kar#R	1,04 %	70,08 %
ali#C	1,03 %	63,77 %
videti#V	0,83 %	76,34 %
misliti#V	0,81 %	62,73 %
kot#C	0,72 %	32,46 %
danes#R	0,70 %	61,86 %

Najpogosteje transformirana lema je pomožni glagol *biti*, sledijo ji funkcijske besede in medmeti. Med leksikalnimi besedami prednjačijo prislovi, med glagoli pa je največ nedoločnikov, kjer gre za izpust končnega *i*-ja. Samostalniki in pridevniki se na seznam ne uvrščajo.

3.3.2 Analiza pregibnih oblik

V Tabeli 2 podajamo 20 najpogostejših parov standardnih oblik in njihovih transformacij, pri tem pa izpuščamo tiste, pri katerih smo zabeležili le izpust strešic. Specifične transformacije so zapisane v oklepajih, prav tako so podani deleži teh oblik glede na celotno število transformacij.

Tabela 2: 20 najpogosteje transformiranih oblik v slovenskih tvitih.

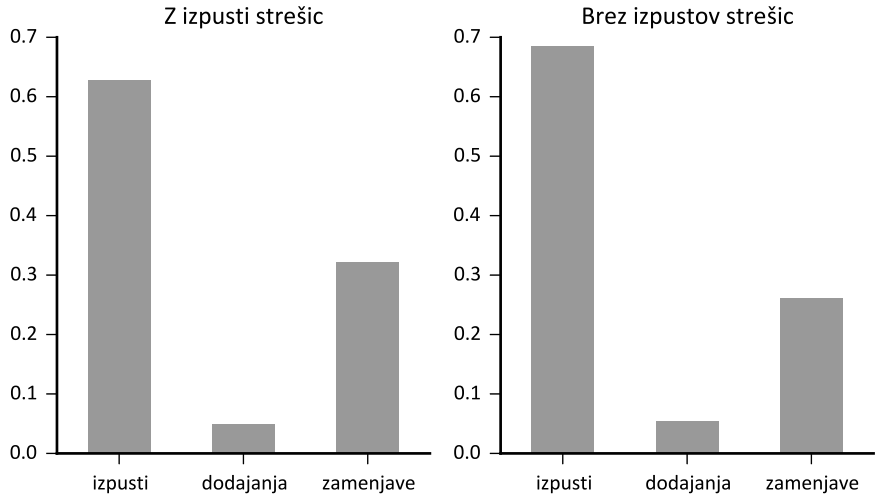
Oblika	% skupaj
sem (sm)	3,37 %
tudi (tud)	2,29 %
samo (sam)	1,93 %
bilo (blo)	1,68 %
potem (pol)	1,39 %
saj (sej)	1,30 %
tako (tko)	1,28 %
jaz (jst)	1,21 %
malo (mal)	1,21 %
kar (kr)	1,10 %
ali (al)	1,07 %
jaz (js)	1,03 %
zdaj (zdej)	0,97 %
tudi (tut)	0,89 %
imam (mam)	0,76 %
pri (pr)	0,70 %
ko (k)	0,70 %
kaj (kej)	0,70 %
nekaj (neki)	0,66 %
toliko (tolk)	0,66 %

Zelo pogosti sta obliki *js* in *jst* za *jaz*, druge transformacije pa zajemajo zamenjavo samoglasnika (tipično *a* > *e*) ali izpust samoglasnika na različnih položajih v besedi. Glede na besedno vrsto je med najpogostejšimi 20 pari največ prislovov.

3.4 Analiza glede na vrsto transformacije

V tem razdelku predstavimo verjetnostno porazdelitev treh vrst Levenshteinovih transformacij (Levenshtein 1996): izpust (npr. *tudi* > *tud*), dodajanje (npr. *super* > *suuuper*) in zamenjava (*zdaj* > *zdej*). Pri tem transformacije opazujemo v smeri od normaliziranih oblik k izvornim oblikam, ki jih najdemo v tvitih. Rezultati so povzeti na Sliki 4. Na levi strani slike so zajete vse transformacije. Najpogostejši so izpusti, sledijo zamenjave, najmanj pa je dodajanj. Na desni strani slike so zajete transformacije brez izpustov strešic, kjer tendence ostajajo podobne.

Najpogostejša vrsta transformacij je opuščanje znakov, sledijo zamenjave, dodajanja pa so v nestandardnem jeziku na Twitterju najredkejši fenomen.



Slika 4: Primerjava distribucij transformacij z upoštevanimi transformacijami zaradi golega izpuščanja strešic (levo) in brez njih (desno).

V naslednjem koraku analiziramo najpogostejše specifične transformacije, kjer ponovno izpuščamo besede, pri katerih gre zgolj za izpuščanje diakritičnih znamenj. V Tabeli 3 je prikazanih najpogostejših 10 transformacij za vsako izmed treh Levenshteinovih vrst, transformacije pa so podkrepljene tudi s pogosto uporabljenimi primeri.

Tabela 3: 10 najpogostejših transformacij za vsako vrsto (s primeri).

Izpust			Dodajanje			Zamenjava		
i	35,04 %	tudi > tud	a	25,8 %	pa > paa	l > u	14,65 %	mogel > mogu
e	17,83 %	sem > sm	h	14,97 %	haha > hahah	a > e	13,32 %	zdaj > zdej
o	13,30 %	lahko > lahk	e	14,17 %	ne > neee	j > i	5,21 %	zjutraj > zjutri
a	11,23 %	tako > tko	j	9,24 %	ne > nej	o > u	4,37 %	ono > uno
j	3,88 %	skoraj > skor	_	4,62 %	odkar > od kar	a > s	4,19 %	jaz > jst
_	3,10 %	ne bi > neb	o	4,14 %	zelo > zelooo	m > l	4,09 %	potem > pol
.	2,79 %	npr. > npr	s	3,98 %	imate > maste	a > o	3,98 %	danes > dons
t	2,73 %	potem > pol	i	3,82 %	vsak > saki	z > s	3,95 %	jaz > js
d	1,77 %	tudi > tut	u	3,82 %	super > suuuper	z > t	3,88 %	jaz > jst
u	1,26 %	tule > tle	m	2,71 %	bi > bim	i > t	3,57 %	tudi > tut

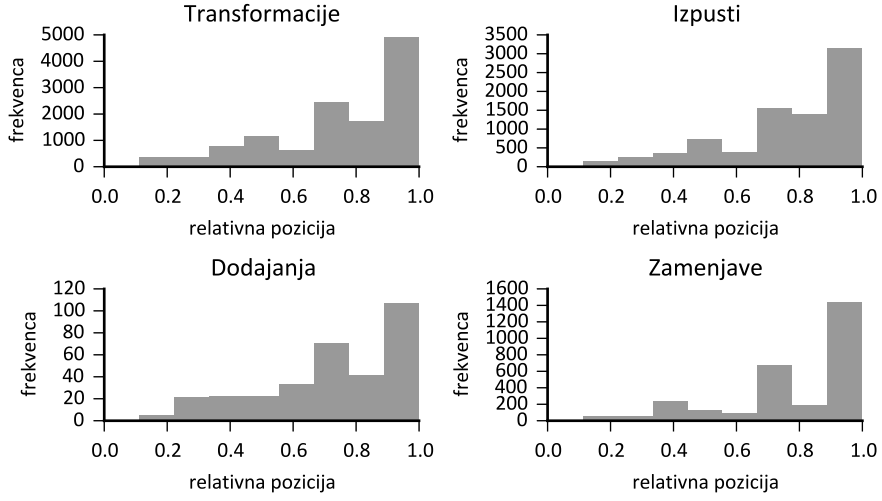
Uporabniki najpogosteje izpuščajo samoglasnike in presledke. Najpogostejši izpust, ki je hkrati tudi najpogostejša transformacija v celotnem korpusu, je samoglasnik *i (tudi > tud)*, ki zavzema več kot tretjino vseh izpustov. Sledijo mu izpusti samoglasnikov *e (sem > sm)*, *o* (*lahko > lahk*) in *a (tako > tko)*, pri čemer je izpuščanje samoglasnika *u* za cel velikostni razred redkejše (*tule > tle*). Samoglasniki so izpuščeni tako na koncu (*tudi > tud*) kot znotraj besede (*tako > tko*). Občutno redkejši so izpusti soglasnikov, med katerimi je največkrat izpuščen *j (skoraj > skor).* Med najpogostejših 10 izpustov se uvršča tudi izpuščanje presledka (*ne bi > neb*) in pike (*npr. > npr*).

Dodajanja so večinoma posledica ekspresivnega ponavljanja zlogov (npr. *hahahahaha*) ali samoglasnikov (npr. *zelooo*) v medmetih in polnopomenskih besedah. Druga najpogostejša kategorija je niz dveh besed, napačno zapisanih kot ena sama ali obratno (npr. *nebo* za *ne bo* ali *od kar* za *odkar*). Sledijo idiosinkratičen zapis domačih besed (npr. *sobitza* za *sobica*), neuveljavljen zapis krajšav (npr. *esemes* za *sms*) ali narečne oblike (npr. *imaste* za *imate*).

Pri zamenjavah je najpogostejša transformacija *l > u* pri deležnikih (*napisal > napisu*, *mogel > mogu*, *mislil > mislu* itd.), sledi pa ji zamenjava samoglasnikov *a > e*, s katero uporabniki zapis besed približujejo izgovoru (*kaj > kej*, *zdaj > zdej* itd.).

3.5. Analiza glede na položaj transformacije

V tem razdelku obravnavamo položaj transformacij (izpusta, dodajanja ali zamenjave) v besedi. Na Sliki 5 je prikazana skupna porazdelitev položaja transformacij, slike 6, 7 in 8 pa prikazujejo relativne položaje izpustov, dodajanj in zamenjav.



Slike 5–8: Distribucije transformacij glede na relativno pozicijo.

S Slike 5 je razvidno, da se transformacije najpogosteje pojavljajo na koncu besede, zelo redko pa na začetku. Podoben trend se pojavlja tudi pri specifičnih vrstah transformacij. Kot je razvidno s Slike 6, so izpusti večinoma vezani na konec besede, predvsem kot posledica izpusta zadnjega samoglasnika (npr. pri funkcijskih besedah in nedoločnikih, kot je prikazano v razdelkih 3.2 in 3.3). Dodajanja (Slika 7) in zamenjave (Slika 8) nakazujejo še močnejšo tendenco pojavljanja na koncu besede (npr. *ne* > *neee*). Podrobnejši pregled dodajanj razkriva, da gre v večini primerov za ponovitev zadnjega samoglasnika. Zamenjave na koncu besede so v veliki meri posledica zamenjave znakov *l* > *u* pri glagolih.

5 SKLEP

V poglavju smo obravnavali zapisovalne prakse v spletni slovenščini. V ta namen smo analizirali vzorec ročno normaliziranih, lematiziranih in oblikoskladenjsko označenih slovenskih tvitov, pri čemer smo se osredotočili na analizo transformacij nestandardnih besednih oblik glede na njihove standardne ustreznice. Analiza transformacij glede na besedne vrste je pokazala, da je teh največ pri polnopomenskih besedah, med katerimi prvo mesto zasedajo prislovi. Analiza znotraj posamičnih besednih vrst pa je pokazala obratno sliko, saj so najpogostejše transformacije slovničnih besed, kar potrjuje tudi ročna analiza najpogosteje transformiranih lem, ki razkriva, da med najpogosteje transformiranimi lemami najdemo največ pomožnih glagolov, medmetov in veznikov.

Z računanjem Levenshteinovih transformacij smo ugotovili, da so daleč najpogostejši tip transformacij izpusti. Glede na to, da smo za analizo uporabili tvite zasebnih uporabnikov, ki vsebujejo nestandardne prvine, je bil tak rezultat pričakovan, ne samo zaradi splošnega načela jezikovne ekonomičnosti, ampak tudi zaradi neformalnega, interaktivnega okolja komunikacije, ki se za povrh pogosto odvija na majhnih prenosnih napravah, ki so opremljene z neergonomičnimi tipkovnicami. Med izpusti prevladuje izpuščanje samoglasnikov, s čimer uporabniki posnemajo govor (glej Zwitter Vitez in Fišer 2018), dodajanja pa so v veliki meri posledica ekspresivnega ponavljanja črk in zlogov, predvsem pri medmetih. Zamenjave so raznorodnejše, vključujejo pa predvsem transformacije v pogovorne oblike in regionalne/narečne variante. Ugotovili smo, da se na začetku besede transformacije pojavljajo le redko, najpogostejše pa so na koncu besede, kar je sicer značilno za nestandardno govorjeno slovenščino (Može 2013), ki se ji neformalno, interaktivno komuniciranje na družbenih omrežjih približuje s fonetiziranim zapisom.

Identificirani fenomeni so primerljivi z raziskavami, opravljenimi na drugih jezikih (glej Miličević et al. 2017 za hrvaščino in srbščino, van Halteren in Oostdijk 2012 za nizozemščino, Sidarenka et al. 2013 za nemščino in Eisenstein 2013 za angleščino),

kjer prav tako prevladujeta težnja po krajšanju besed in prisotnost oblik, ki so značilne za (geografsko in demografsko) različne družbene skupine. Rezultati na nivoju zapisa besed še posebej izrazito kažejo brisanje meja med govorjenim in pisnim jezikom (glej Eisenstein 2013 za angleščino, Zwitter Vitez in Fišer 2015 za slovenščino), kar pomeni, da pojavi, na katere smo naleteli, niso novi, temveč zgolj nekoliko bolj opazni zaradi množičnejše komunikacije in trajnejšega medija v primerjavi z večino neformalnih govornih situacij. Rezultati prav tako nakazujejo povezavo med variantnostjo zapisa z udejanjanjem identitete uporabnikov (glej Tagg 2012), kjer odstopanje od norme z rabo regionalno in demografsko obarvanih različic igra pomembno vlogo. Vprašanje, kako se tovrstne jezikovne prakse vpenjajo v širšo razpravo o demokratizaciji jezika, jezikovne izbire in standardizacije, je kompleksno in bo zagotovo predmet zanimivih prihodnjih raziskav.

Glede na pomanjkanje empiričnih podatkov za računalniško posredovano komunikacijo v slovenščini pričujoča analiza predstavlja dragocen vpogled v naravo odstopanj od jezikovnih norm, prav tako pa služi kot osnova za prihodnje bolj poglobljene raziskave tega jezikoslovnega fenomena, ki bodo osredotočene na preverjanje specifičnih hipotez. Nadaljnje raziskave bi lahko vključevale analizo vpliva sociodemografskih faktorjev na opazovane transformacije, kot so starost, geografsko poreklo, izobrazba uporabnikov ipd. V prihodnosti bi prav tako lahko opravili leksikalno analizo nestandardnih prvin v računalniško posredovani komunikaciji. Tovrstni primeri v uporabljenem označenem korpusu niso zajeti, so pa predhodne raziskave (Fišer et al. 2015) že pokazale, da so ti primeri zelo pomembni za primerjave med jeziki.

Zahvala

Ker je avtorska zasedba tega poglavja mednarodna, smo rokopis pripravili v angleščini. V slovenščino ga je prevedla Dafne Marko, ki se ji za natančen in tekoč prevod ter skrbno upravljanje s terminologijo iskreno zahvaljujemo.

Literatura

Anis, Jacques, 2007: Neography: Unconventional spelling in French SMS. Danet, Brenda in Susan C. Herring (ur.): *The Multilingual Internet: Language, culture, and communication online*. Oxford: Oxford University Press. 87–115.

Chariatte, Nadine, 2014: "Facebook Style": The use of non-standard features in virtual speech conditioned by the medium Facebook. Brumme, Jenny in Sandra Falbe (ur.): *The Spoken Language in a Multimodal Context: Description, Teaching, Translation*. 93–114. Berlin: Frank & Timme.

Čibej, Jaka, Darja Fišer in Tomaž Erjavec, 2016: Normalisation, tokenisation and sentence segmentation of Slovene tweets. *Proceedings of Normalisation and Analysis of Social Media Texts (NormSoMe) 2016*, *LREC 2016*. 5–10. http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-NormSoMe_Proceedings.pdf.

Čibej, Jaka, Špela Arhar Holdt, Tomaž Erjavec in Darja Fišer, 2018: Ročno označeni korpusi Janes za učenje jezikovnotehnoloških orodij in jezikoslovne raziskave. Fišer, Darja (ur.): *Viri, orodja in metode za analizo spletne slovenščine*. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani. 44–73.

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

Eisenstein, Jacob, 2013: What to do about bad language on the Internet. Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics. 359–369. http://www.cc.gatech.edu/~jeisenst/papers/naacl2013-badlanguage.pdf.

Fišer, Darja, Tomaž Erjavec, Nikola Ljubešić, and Maja Miličević (2015): Comparing the nonstandard language of Slovene, Croatian and Serbian tweets. Smolej, Mojca (ur.): *Obdobja 34. Slovnica in slovar - aktualni jezikovni opis (1. del)*. Ljubljana: Znanstvena založba filozofske fakultete. 225–231.

Goli, Teja, Eneja Osrajnik in Darja Fišer, 2016: Analiza krajšanja slovenskih sporočil na družbenem omrežju Twitter. *Proceedings of the Language Technologies and Digital Humanities Conference*. Ljubljana, Slovenia. 77–82.

Levenshtein, Vladimir I, 1966: Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10/8: 707–710.

Marko, Dafne, 2016: The Use of Alphanumeric Symbols in Slovene Tweets. *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities*. Ljubljana, Slovenia. 48–53.

Miličević, Maja, Nikola Ljubešić in Darja Fišer, 2017: Birds of a feather don't quite tweet together: An analysis of spelling variation in Slovene, Croatian and Serbian Twitterese. Fišer, Darja in Michael Beißwenger (ur.): *Investigating Computer-Mediated Communication: Corpus-Based Approaches to Language in the Digital World*. Ljubljana: Znanstvena založba Filozofske fakultete. 14–43.

Miličević, Maja in Nikola Ljubešić, 2016: Tviterasi, tviteraši or twitteraši? Producing and analysing a normalised dataset of Croatian and Serbian tweets. *Slovenščina 2.0* 4/2. 156–188. http://dx.doi.org/10.4312/slo2.0.2016.2.156-188

Može, Sara, 2013: Raba kratkega nedoločnika: korpusni pristop. *Slovenščina 2.0* 1/1: 155–175. http://www.trojina.org/slovenscina2.0/arhiv/2013/1/Slo2.0_2013_1_08.pdf

Shortis, Tim, 2007: Revoicing Txt: Spelling, vernacular orthography and 'unregimented writing'. Posteguillo, Santiago, María José Esteve in M. Lluïsa Gea-Valor (ur.): *The Texture of Internet: Netlinguistics in Progress*. Newcastle: Cambridge Scholars Publishing. 2–23.

Sidarenka, Uladzimir, Tatjana Scheffler in Manfred Stede, 2013: Rule-based normalization of German Twitter messages. *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology*. https://gscl2013.ukp.informatik.tu-darmstadt.de/fileadmin/user_upload/Group_UKP/conferences/ gscl2013/workshops/sidarenka_scheffler_stede.pdf

Smolej, Mojca (ur.), 2015: *Slovnica in slovar – aktualni jezikovni opis*. *Obdobja 34*. Ljubljana: Znanstvena založba Filozofske fakultete.

Tagg, Caroline, Alistair Baron in Paul Rayson, 2012: "i didn't spel that wrong did i. Oops": Analysis and normalisation of SMS spelling variation. *Lingvisticæ Investigationes* 35/2. 367–388.

Thurlow, Crispin in Alex Brown, 2003: Generation Txt? The sociolinguistics of young people's text-messaging. *Discourse Analysis Online* 1/1.

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

Halteren, Hans van in Nelleke Oostdijk, 2012: Towards identifying normal forms for various word form spellings on Twitter. *CLIN Journal* 2. 2–22.

Verovnik Tina, 2004: Norma knjižne slovenščine med kodifikacijo in jezikovno rabo v obdobju 1950–2001. *Družboslovne razprave XX*, 46/47: 241–258.

Zwitter Vitez, Ana in Darja Fišer, 2015: From mouth to keyboard: the place of non-canonical written and spoken structures in lexicography. *Electronic lexicography in the 21st century: linking lexical data in the digital age: proceedings of eLex 2015 Conference*, 11-13 August 2015, Herstmonceux Castle, United Kingdom. Ljubljana: Trojina, zavod za uporabno slovenistiko; Brighton: Lexical Computing. 250–267.

Zwitter Vitez, Ana in Darja Fišer, 2018: Govorne prvine v nestandardni spletni slovenščini. Fišer, Darja (ur.): *Viri, orodja in metode za analizo spletne slovenščine*. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani. 254–273.

(Ne)normativnost računalniško posredovane komunikacije v slovenščini: merilo vejice

Damjan Popič, Darja Fišer

Izvleček

V poglavju predstavimo raziskavo rabe vejice v slovenskih tvitih, in sicer preverimo, v kolikšni meri je stava v tovrstnih vsebinah rabljena v skladu z jezikovnim standardom in na katerih mestih uporabniki s stavo vejice najbolj odstopajo od pravopisa. S tem nadgrajujemo pretekle raziskave, ki so bile usmerjene predvsem v ugotavljanje, v katerih primerih imajo uporabniki največ težav, in poskušamo predstaviti nekoliko bolj celostno sliko rabe vejice v slovenskih uporabniških spletnih vsebinah. Rezultati raziskave kažejo, da je v slovenski računalniško posredovani komunikaciji standardna stava vejice pogostejša kot nestandardna, lahko pa vsaj za določen del gradiva trdimo, da je vejica stavljena nestandardno (tj. izpuščena) namerno, kot del neformalne komunikacije.

Ključne besede: vejica, pravopis, jezikovni standard, računalniško posredovana komunikacija, slovenščina

1 UVOD

Raziskave računalniško posredovane komunikacije, zlasti v razmerju do splošnega jezika,[1] so bile v zadnjem obdobju močno obogatene in razširjene, zlasti s korpusno-jezikoslovno analizo v okviru projekta JANES (prim. npr. Fišer et al. 2017 in Arhar Holdt 2017). V okviru normativnostne obravnave korpusa Janes je bilo še posebej veliko pozornosti posvečene presoji standardnosti računalniško posredovane komunikacije (Arhar Holdt in Dobrovoljc 2016, Fišer et al. 2015, Goli et al. 2015, Osrajnik et al. 2015, Škrjanec et al. 2015), kar je glede na slovensko preskriptivistično tradicijo in direktivno naravo jezikoslovne misli pričakovano, kot je pričakovano tudi to, da je bilo kar nekaj pozornosti posvečene obravnavi vejice (Popič et al. 2016), ki je glede na sodobne empirične raziskave temeljna pravopisna težava v slovenskem jeziku (Kosem et al. 2012, Rozman et al. 2012, Popič 2014, Logar in Popič 2015; Popič in Logar 2015, Popič in Fišer 2015), obenem pa je v slovenski jezikovni tradiciji tudi nosilka prestiža.[2] To pomeni, da vejici pripisujemo status preizkusnega kamna jezikovne omike (Popič in Logar 2015), kultiviranosti in socialne stratifikacije, in čeprav je prestižnost ena od lastnosti knjižnojezikovne norme oz. standardnega jezika (Nebeská 1999), lahko v slovenski računalniško posredovani komunikaciji vidimo, da se je ta vrednotenjski sistem v veliki meri prenesel tudi na splet (glej Popič in Fišer 2017).[3]

Iz tega razloga v pričujočem prispevku analiziramo (ne)standardnost stave vejice v slovenskih uporabniških spletnih vsebinah, in sicer želimo številne (tudi lastne; Popič in Fišer 2015, Popič et al. 2016) raziskave nenormativne stave vejice razširiti z analizo normativno ustrezne stave.[4] Cilj je prikazati nekoliko bolj celostno podobo rabe vejice v tvitih, na način, ki ni obremenjen s predsodki o boljši in slabši slovenščini glede na namen in medij določene jezikovne zvrsti (prim. npr. Jakop 2008). S tem se nanašamo predvsem na prepričanje, da je jezik v računalniško posredovani komunikaciji izrazito nestandarden.

Nedavne raziskave so tovrstne predsodke v precejšnji meri ovrgle. S primerjalno raziskavo korpusov Janes (Erjavec et al. 2018) in Kres (Logar et al. 2012) smo

1 Izraz *splošni jezik* uporabljamo z namenom, da se izognemo diadi med računalniško posredovano komunikacijo in standardnim oz. knjižnim jezikom. Tovrstna delitev je v prvi vrsti nestrokovna, predvsem pa je problematično to, da je lahko tudi diskriminatorna v razmerju do računalniško posredovane komunikacije (predvsem v smislu, da je ena od temeljnih lastnosti knjižnega jezika oz. njegove norme prestižnost). V besedilu se torej z izrazoma *standardna* in *nestandardna slovenščina* nanašamo izključno na dejansko jezikovno skladnost s pravopisnim standardom.

2 Na to kaže tudi nedavno parlamentarno glasovanje o vejici. 21. junija 2017 je državni zbor namreč glasoval o dopolnitvi prve vrstice Župančičeve pesmi Domovina je ena za inskripcijo na Spomeniku vsem žrtvam vojn (in z vojnami povezanim žrtvam na območju Republike Slovenije), in sicer je šlo za glasovanje o izbrisu vejice. Glasovanje, ki se je sicer končalo z rezultatom 54 : 6 v podporo izbrisu, je dober indikator odnosa slovenske družbe do jezika, tudi do vejice, predvsem glede vprašanja jezikovne regulacije – za slovensko okolje je namreč značilno prepričanje, da je mogoče o jezikovnih vprašanjih odločati in tudi razsojati s centralne pozicije moči.

3 To dokazujejo tudi nedavne raziskave odnosa do jezika (in same vejice) na Twitterju (Popič in Fišer 2017), ki kažejo, da se vejica tudi v slovenskem računalniško posredovanem diskurzu uporablja kot orodje prestiža in kot sredstvo dokazovanja, da so argumenti nekoga odvisni od jezikovne pravilnosti ubeseditve tega argumenta.

4 Pričujoče poglavje razširja raziskavo o nestandardni rabi vejice v slovenski računalniško posredovani slovenščini (Popič et al. 2016).

tako v preteklih raziskavah pokazali, da raba vejice na družbenih omrežjih ne peša in da so razlike med rabo vejice v tradicionalnih in novomedijskih besedilih manjše, kot pregovorno velja (Popič in Fišer 2015). Pri tem smo se omejili na ožji nabor tipičnih atraktorjev vejic (Verovnik 2003; Žibert 2006) in na samo pogostnost ter razporejenost vejic, zato smo želeli nadgraditi in pripraviti izhodišče za celovito analizo rabe vejice, tako standardne kot tudi nestandardne, predvsem pa razviti čim bolj univerzalen kategorizacijski sistem za označevanje rabe vejice in ga preizkusiti na širšem vzorcu.

V poglavju tako pokažemo, v čem raba vejice v uporabniških spletnih vsebinah odstopa od pravopisnega standarda in v čem mu sledi, obenem pa poskušamo podati preliminarne ugotovitve, v kolikšni meri je nestandardna stava vejice posledica zavestne odločitve in v kolikšni posledica nepoznavanja jezikovnega standarda. Analiza temelji na označevanju napačno in pravilno stavljene vejice na naključnem vzorcu 500 tvitov iz korpusa Janes v0.4 (Popič et al. 2017). Pri tem smo kot merilo upoštevali še stopnjo jezikovne in tehnične standardnosti besedil (Ljubešić et al. 2015) ter v vzorec zajeli po 250 tvitov z naslednjima oznakama: T1L3 (nestandarden jezik, a standardna raba velikih začetnic, presledkov, ločil) in T3L3 (tako tehnično kot tudi jezikovno povsem nestandardni tviti). Ker nas zanima stava vejice v nestandardni slovenščini, tvitov, zapisanih v standardni slovenščini (L1), v analizo nismo zajeli, kar sicer pomeni, da so rezultati nekoliko manj posplošljivi, a so toliko bolj dragoceni zaradi tega, ker lahko na ta način vidimo dejansko vlogo vejice – v besedilih, ki s standardnostjo zapisa vsaj predvidoma niso obremenjena, lahko sami prisotnosti vejice v besedilu pripišemo določeno vrednost.

Na podlagi označenega korpusa tako prikažemo temeljne težave v zvezi s stavo vejice v spletni slovenščini, obenem pa tudi mesta, na katerih tovrstnih težav uporabniki nimajo. Naša osnovna podmena, ki izhaja iz preteklih raziskav, je, da imajo uporabniki težave z vejico predvsem na mestih, ki so skladenjsko manj predvidljiva in obenem tudi manj formalizirana (z vidika napovedljivosti stave vejice zgolj na podlagi izraženega veznika). Prav tako je utemeljeno pričakovati, da rezultati ne bodo odgovarjali zgolj na vprašanje, kje imajo jezikovni uporabniki težave, temveč tudi, ali in kje se vejica v računalniško posredovani komunikaciji – vsaj v tisti tehnično manj standardni, s katero se v raziskavi ukvarjamo – zavestno izpušča.

2 VEJICA V (SPLETNI) SLOVENŠČINI

V slovenski jeziko(slo)vni tradiciji je vejica malodane razvpita (glej Popič in Fišer 2017), in sicer velja za nekaj, kar je praktično nemogoče popolnoma osvojiti

oz. obvladati. Tovrstno izročilo je morda nekoliko pretirano, a vendarle nedavne empirične raziskave tovrstnim sodbam v veliki meri pritrjujejo (glej Kosem et al. 2012; Popič 2014; Popič in Logar 2015), saj kažejo, da imajo uporabniki z vejico velike težave. Vejica pa je tudi v jezikoslovnih krogih pogosto povzročala težave, saj gre za »ločilo, o stavi katerega so si bili kritiki najbolj in najpogosteje navzkriž« (Dobrovoljc 2004: 188). Verjetno lahko glavne razloge za to iščemo v sami naravi stave vejice v slovenščini, ki je v slovenski preskripciji izrazito sintaktična, to pa velja za celotni proces normiranja jezika od prvega, Levčevega pravopisa (SP 1899) naprej. Ta je kot prvi slovenski pravopis vejico predpisal predvsem na podlagi skladenjsko-logičnih meril, in sicer je Levec kot izpostavne prepoznal naslednje prvine (Levec 1899: 103–108):

- prosti stavki,
- skrčeni stavki,
- priredja,
- podredja,
- skrajšani stavki,
- periode med posameznimi stavki prve in druge polovice (proreka in poreka).

Kot lahko vidimo, je razdelitev izrazito sintaktična in ne omogoča omahovanja na podlagi govora ali pomena. To je sicer značilno za celotno slovensko zgodovino predpisovanja rabe vejice (in drugih elementov jezika nasploh), saj je v zgodovini v resnici prišlo zgolj do enega resnega liberalnega posega v stavo ločil, in sicer s prvim akademijskim pravopisom (SP 1950: 43; poudarila avtorja):

> Modulacijo glasu in presledke, s katerimi v govoru vežemo in ločimo pomenske enote, zaznamujemo v pisavi z ločili. Nekatera teh imajo objektivno logično vrednost in so potrebna za lažje in zanesljivo razumevanje pisanega jezika, zato so določena po pravopisnih pravilih. Včasih pa **pisatelj** z ločili izraža osebno ali čustveno razgibanost; **taka ločila so osebna in zanje ni splošno veljavnih pravil**.

To pomeni, da je bila »stava ločil, zlasti vejice, še v SP 1950 utemeljena predvsem s skladenjskim pomensko-logičnim razmerjem« (Dobrovoljc 2004: 189), v SP 1962 pa je bilo v zametkih »upoštevano tudi stavčnofonetično merilo oz. tonski potek kot posledica oblikovanja osnovnih pomenskih enot« (prav tam), zato je v SP 1962 prišlo do izrazitega posega v stavo ločil – od uvajalnega odstavka v rabo ločil je ostal zgolj naslednji zapis:

> Modulacijo glasu in presledke, s katerimi v govoru ločimo pomenske enote, zaznamujemo v pisavi z ločili. Potrebna so za lažje in zanesljivo razumevanje pisanega jezika, zato so določena s pravopisnimi pravili (SP 1962: 81).

Z vidika obravnave vejice se aktualni slovenski kodifikacijski priročnik (SP 2001) nekoliko razlikuje, saj kot prvi uvaja skladenjsko in neskladenjsko rabo ločil. Poleg vzpostavljenega razmerja pomen – oblika je z novim slovenskim pravopisom (oz. z načrtom zanj) v slovensko tradicijo prišlo tudi upoštevanje okoliščin oz. sotvarja (Dobrovoljc 2004: 189).

Na področju analiz spletnega jezika se obstoječe študije v veliki meri nanašajo na jezik v novih medijih v razmerju do jezikovnega standarda v tradicionalnih, prim. npr. Jakop (2008). Ugotovitev »večine razprav, ki se skoraj praviloma sklicujejo tudi na ugotovitve tujega jezikoslovja, je, da je jezik, uporabljen v novomedijskih besedilih, slogovno poln posebnosti, kot so kratkost sporočil, okrajšanost besed in besednih zvez ter povedi, izpustnost in neformalnost /…/, /v/endar pa je to ugotovitev mogoče podkrepiti le z jezikovno analizo ustreznih besedil« (Dobrovoljc 2008: 297). Za slovensko tradicijo je v odnosu do slovenščine v novih medijih značilno tudi nerazumevanje, da gre za *nove* medije in da obstoječega normativnega modela, ki temelji na knjižnojezikovni normi, pogosto ni mogoče aplicirati na *spletno slovenščino*, to nerazumevanje pa pogosto spremlja tudi splošno negativno vrednotenje jezika v tovrstnih besedilih, prim. npr. analizo spletnih forumov (Jakop 2008: 326):

> Raziskava je pokazala, da je jezikovna podoba slovenskih spletnih forumov z vidika upoštevanja obstoječe pravopisne norme na zelo nizki ravni. Opazna so kršenja pravil o rabi velikih in malih črk, ločil in pisanja besed skupaj ali narazen. Pisanje prevzetih besed na spletnih forumih pa kaže spodbudne težnje jezikovnih uporabnikov po zapisovanju prevzetega besedja v podomačeni obliki.

Pri teh težnjah bi težko domnevali, da gre za prizadevanje piscev za »dobro slovenščino«, temveč za drugačne komunikacijske okoliščine, s katerimi jezikovni standard, temelječ na knjižnojezikovni normi, nima prav dosti skupnega oz. za njimi v marsičem zaostaja. Razlogi za tovrstne odmike od standardne slovenščine so odvisni tudi (ali predvsem) od prenosnika, ne le od jezikovnih ali metajezikovnih dejavnikov, ki se po navadi znajdejo na tnalu kot počelo nestandardnosti slovenščine v novih medijih, prim. npr. Jakop (2008: 326): »Med jezikovnimi dejavniki so v ospredju nepoznavanje oziroma slabo razumevanje nekaterih poglavij iz slovenskega pravopisa, zlasti pravil o rabi ločil in pisanju besed skupaj ali narazen.«

Pomislek, da lahko v novomedijskih besedilih najdemo pravopisne napake, zlahka zvrnemo tudi na kateri koli segment ubesedovanja v slovenščini, težave z obvladovanjem jezikovnega standarda pa so akutne tako pri poklicnih piscih (glej Popič 2014) kot tudi pri šolarjih (Kosem idr. 2012; Popič in Logar 2015), vendarle pa bi le stežka pričakovali, da se bodo uporabniki novih medijev »kultivirali«, predvsem v smislu, kot ga navaja Jakop (2008: 327):

> Morda smo preveč optimistični, vendar verjamemo, da z izobraževanjem mladih glede sporazumevanja na spletu dolgoročno prispevamo k višji kultiviranosti tvorcev besedil v spletnem diskurzu in s tem tudi k bolj kultiviranemu dialogu nasploh. Vse skupaj v duhu pregovora: Kar se Janezek nauči, to Janez zna.

Malo verjetno je, da se bodo uporabniki v spletnem okolju obnašali podobno kot v drugih, formaliziranih okoljih, saj so ta okolja pač neformal(izira)na in jim je zato neformalno izražanje inherentno. Ravno to je eden od ciljev pričujočega prispevka – preveriti, ali kljub težnji uporabnikov, da ločila uporabljajo, vendarle obstaja tudi močna tendenca po manj standardiziranem izražanju. Prav zaradi tega ima jezikoslovna stroka toliko težav že pri poimenovanju jezika v novih medijih, saj ga ni mogoče opisati z obstoječo socialno-funkcijsko stratifikacijo. Računalniško posredovana komunikacija lahko namreč zaseda katero koli mesto v tej stratifikaciji, od izrazito privzdignjene govorice do povsem neartikuliranega, profanega izražanja.

3 OPIS STAVE VEJICE V SLOVENSKI RAČUNALNIŠKO POSREDOVANI KOMUNIKACIJI

V pričujočem poglavju predstavimo procesa označevanja standardne in nestandardne stave vejic v našem korpusu. Delo je potekalo v orodju WebAnno (Eckart de Castilho et al. 2014), za označevanje pa smo uporabili tipologijo, predstavljeno v Tabeli 1. V nadaljevanju najprej predstavimo zasnovo in izpopolnjevanje tipologije za označevanje rabe vejic, zatem pa opišemo še proces označevanja in delotok.

3.1 Tipologija

Za opis rabe vejice smo razvili tipologijo, pri tem pa smo se oprli na aktualni jezikovni predpis (Slovenski pravopis 2001, Pravila), z nekaj (minimalnimi) dopolnitvami in prilagoditvami.[5] Pri zamejevanju kategorij in definiranju tega, kaj naj zajemajo, smo se poskušali v čim večji meri držati jezikovnega predpisa, za podporo sorodnim raziskavam in vpogled v sistem označevanja pri naših raziskavah smo smernice za označevanje objavili na spletu.[6] Pri sestavi kategorizacije smo se držali načela, da naj bo ta čim bolj univerzalna (da torej lahko s čim bolj podobnimi kategorijami opiše čim več primerov tako standardne kot tudi nestandardne rabe vejice, tj. odvečne vejice, manjkajoče vejice itd.). To ni bilo povsem

5 Osnovne značilnosti tipologije, vezane na nestandardno stavo vejice, so podane v Popič in Fišer (2015) in Popič et al. (2016).

6 Smernice za označevanje so na voljo na spletni strani http://nl.ijs.si/janes/wp-content/uploads/2014/09/vejice-smernice.pdf.

izvedljivo, zato tipologija vsebuje tudi kategorije, vezane predvsem na specifične primere (tako denimo kategorija Stavčni člen opisuje predvsem nestandardno »stavčnočlensko« rabo vejice, ki je vezana zgolj na odvečno vejico), povečini pa so kategorije vendarle univerzalne in namenjene opisu katerega koli segmenta jezika – tipologija torej ni namenjena zgolj opisu (ne)standardne stave vejice v spletni rabi, temveč je bil namen sestavljavcev prav ta, da lahko zadosti opisu katerega koli žanra ali besedilnega tipa.

Poleg jezikovnosistemskega opisa stave vejice smo se pri pripravi tipologije oprli tudi na že opravljene empirične, izgradivne preglede stave vejice v različnih okoljih, in sicer v spletnem (Popič in Fišer 2015), šolskem (Logar in Popič 2015, Kosem et al. 2012) ter v okolju poklicnih piscev (Popič 2014). Na podlagi pridobljenih spoznanj smo lahko v določeni meri predvideli problematična mesta pri stavi vejice in pripravili začetno različico tipologije. Z njo smo označili testni nabor tvitov in na podlagi sprotnih opažanj tipologijo dopolnjevali. Glavne razlike med začetno in končno tipologijo so se nanašale predvsem na nestandardno odvečno vejico, pri kateri tipična razdelitev glede na jezikovnosistemski opis odpove, saj temelji na skladenjski razčlenitvi – odvečnost vejice pa temelji ravno na tem, da uporabnik skladenjske razčlenitve ne (pre)pozna. Glede na to, da slovenski jezikovni predpis daje specifične napotke o tem, kje vejica mora stati, manj pozornosti pa namenja temu, kje vejice ne bi smelo biti, je za predvidevanje odvečnih vejic v okviru tipologije za označevanje potrebna predhodna empirična raziskava.[7]

Kot prikazuje Tabela 1, smo pri vseh segmentih skladenjske rabe vejice uvedli še podkategoriji levo- in desnosmerne vejice (Korošec 2003), saj v tem prepoznavamo različne (potencialne) vzgibe in razloge za nestandardno stavo vejice, v vsakem primeru pa smo želeli vsako posamezno stavo vejice opisati kar najnatančneje.

Kot lahko vidimo, kategorije v veliki meri temeljijo na razdelitvi, podani v jezikovnem predpisu, dodani sta kategoriji Stavčni člen in Besedna zveza, pri neskladenjski rabi vejice pa smo želeli čim podrobneje razčleniti, za katero vrsto nestandardne stave gre, zato so vključene še štiri podkategorije. Za neskladenjsko rabo vejice je rezerviran tudi parameter »napačen znak«, saj v tovrstnih primerih kategoriji manjkajoče in odvečne vejice nista (nujno) relevantni (npr. nestandardna raba pike namesto vejice). V določeni meri smo jezikovnosistemske kategorije tudi združevali (glej predvsem kategorijo 1.6), in sicer na mestih, kjer drobljenje informacij po naših pričakovanjih ne bi bistveno izboljšalo informacij o nestandardni rabi vejice, bi pa močno povečalo število kategorij. Pri združevanju kategorij smo bili pozorni predvsem na to, da smo združevali »skladenjsko sorodne« kategorije, tako da so informacije, ki jih z analizo stave vejice dobimo, čim bolj posplošljive na vse elemente znotraj kategorije.

7 Verjetno je utemeljeno pričakovati, da bo slovenska preskripcija v prihodnje zajemala tudi napotke o najpogostejših napakah, osnovane na empiričnem pregledu gradiva.

Tabela 1: Pregled skladenjskih in neskladenjskih oznak po kategorijah.

1 Skladenjska
1.1 Priredje
1.1.1 Vezalno
1.1.1.1 Desnosmerna
1.1.1.2 Levosmerna
1.1.2 Stopnjevalno
1.1.2.1 Desnosmerna
1.1.2.2 Levosmerna
1.1.3 Ločno
1.1.3.1 Desnosmerna
1.1.3.2 Levosmerna
1.1.4 Protivno
1.1.4.1 Desnosmerna
1.1.4.2 Levosmerna
1.1.5 Vzročno
1.1.5.1 Desnosmerna
1.1.5.2 Levosmerna
1.1.6 Posledično
1.1.6.1 Desnosmerna
1.1.6.2 Levosmerna
1.1.7 Pojasnjevalno
1.1.7.1 Desnosmerna
1.1.7.2 Levosmerna
1.2 Odvisnik
1.2.1 Osebkov
1.2.1.1 Desnosmerna
1.2.1.2 Levosmerna
1.2.2 Predmetni
1.2.2.1 Desnosmerna
1.2.2.2 Levosmerna
1.2.3 Krajevni
1.2.3.1 Desnosmerna
1.2.3.2 Levosmerna
1.2.4 Časovni
1.2.4.1 Desnosmerna
1.2.4.2 Levosmerna
1.2.5 Načinovni
1.2.5.1 Desnosmerna
1.2.5.2 Levosmerna
1.2.6 Vzročni
1.2.6.1 Desnosmerna
1.2.6.2 Levosmerna
1.2.7 Namerni
1.2.7.1 Desnosmerna
1.2.7.2 Levosmerna
1.2.8 Pogojni
1.2.8.1 Desnosmerna
1.2.8.2 Levosmerna
1.2.9 Dopustni
1.2.9.1 Desnosmerna
1.2.9.2 Levosmerna
1.2.10 Prilastkov
1.2.10.1 Desnosmerna
1.2.10.2 Levosmerna
1.3 Polstavek
1.3.1 Desnosmerna
1.3.2 Levosmerna
1.4 Stavčni člen
1.5 Besedna zveza
1.6 Pri-, pa- in dostavek ter izpostavek[8]
1.6.1 Desnosmerna
1.6.2 Levosmerna
1.7 Soredje
2 Neskladenjska
2.1 X → vejica
2.2 Vejica → X
2.3 Tipkarska napaka
2.4 Drugo

8 Kategorija je bila zastavljena kot zbir formalno sorodnih skladenjskih podkategorij, ki bi posamezno močno povečale število skupnih kategorij, a brez posebnega doprinosa k razumevanju (ne)standardne stave vejice. Nekoliko problematično je združevanje različnih kategorij, ki denimo ne morejo biti desno- ali levosmerne, vendarle pa je sama *smernost* v tovrstnih primerih izjemno koristna informacija, prepričani smo, da v večji meri kot podatek o tipu strukture.

Izdelana klasifikacija se je izkazala za uporabno, tako da je za označevanje standardne stave vejice nismo spreminjali, seveda pa to pomeni, da so bile določene kategorije, uvedene predvsem za opisovanje napačne stave vejice, tokrat izpuščene. Tudi pri označevanju nestandardne stave se je klasifikacija izkazala za uporabno, kar je bilo tudi pričakovano, saj temelji na jezikovnem predpisu (ravno tako kot tudi standardna stava). V tem oziru lahko rečemo, da je klasifikacija univerzalna, da pa terja razširitev, predvsem s podrobnejšimi jezikovnosistemskimi informacijami (denimo besednovrstnimi), če bi želeli stavo vejic (zlasti nestandardno) opisati natančneje.

3.2 Označevanje

Med procesom označevanja je bilo dvojno označenih 500 tvitov, in sicer smo najprej označili nestandardno stavo vejic, rezultate (glej Popič in Fišer 2015 ter Popič et al. 2016) pa smo želeli obogatiti dopolniti še s podatki o tem, katera mesta pri stavi vejice so neproblematična, zato smo dvojno označili še vse ustrezno stavljene vejice. To smo izvedli tako, da smo dvojno označili v istem naboru 500 tvitov (z vsebovanimi oznakami po prvem sklopu označevanja).

Pri označevanju je bilo mogoče za posamezni primer izbrati le eno kategorijo, večkategorialnih oznak – npr. ko je utemeljitev za stavo vejice lahko tako levokot desnosmerna – nismo dopuščali. Če je bilo mogočih več interpretacij, smo izbrali kategorijo, ki je bila v danem kontekstu bolj povedna ali pa verjetnejši razlog za (ne)stavo vejice. Po koncu označevanja (ne)stave vejice je kuratorka pregledala vseh 500 dvojno označenih tvitov in pri neskladjih med označevalcema sprejela dokončno odločitev. Po podatkih orodja WebAnno je Cohenov koeficient ujemanja med označevalcema kappa znašal 0,57, kar pomeni srednje dobro ujemanje. Zelo dobrega ujemanja zaradi težavnosti problema, sprotnega vzpostavljanja tipologije in smernic ter neizkušenosti označevalcev ni bilo mogoče pričakovati. Neskladja med označevalcema so bila najpogostejša predvsem pri naslednjih segmentih:

- ločevanje posameznih medmetov znotraj pastavka;
- prepoznavanje vezalnosti, protivnosti in posledičnosti veznikov *pa* ter *in*;
- obravnava smeškov in drugih metajezikovnih struktur, tipičnih za uporabniške vsebine;
- prepoznavanje vrst odvisnikov, zlasti v primerih, pri katerih so vezniške strukture zapisane nestandardno.

Pri prvih treh problematičnih mestih smo se odločili za princip minimalne intervencije, zato se je kuratorka izogibala posegov v izvorno stavo vejic, razen če je poseg nedvoumno predvidevalo določeno pravopisno pravilo ali kontekst, medtem ko je bila pri zadnjem problematičnem torišču potrebna vsebinska odločitev. Za čim ustreznejšo obravnavo težavnih segmentov smo sprejeli natančne smernice, s čimer bo označevanje v prihodnje enostavnejše in bolj usklajeno, s tem pa smo poskrbeli tudi za sledljivost odločitev, ki smo jih sprejemali med procesom označevanja. Skupno je bilo pri 500 tvitih označenih 405 mest z nestandardno rabo vejice, kar pomeni v povprečju 0,8 oznake na posamezni tvit.

Da je bilo v tvitih več ustreznih vejic kot neustreznih, kaže že podatek o številu oznak za standardno rabo vejic, in sicer je bilo označenih 628 mest, tj. 1,26 oznake (in s tem ustrezno stavljene vejice) na tvit. Tudi pri označevanju standardne stave vejice smo naleteli na nekaj težav, a bistveno manj kot pri označevanju nestandardne stave; težave pa so bile vezane predvsem na prepoznavanje odvisnikov v primerih, ko je bil zapis (tudi samih veznikov) izrazito nestandarden ali pa je bilo treba določene dele besedila preprosto inferirati. Po zaključenem postopku označevanja je kurator[9] pri neskladjih med označevalcema sprejel končno odločitev, pri čemer je bila večina neskladij vezana na identifikacijo posamezne jezikovne enote, v določenih primerih pa je zaradi zgoščenosti oznak znotraj označevalskega vmesnika prišlo tudi do pomot.

4 ANALIZA STAVE VEJICE V OZNAČENEM KORPUSU TVITOV

V nadaljevanju predstavimo izsledke analize nestandardne rabe vejic v tvitih. Najprej predstavljamo kumulativne rezultate po posameznih kategorijah, ki jih podajamo v Tabeli 2, v nadaljevanju pa se osredotočimo na posamezne (pod)kategorije, zlasti tiste, ki so glede na rezultate empirične raziskave v uporabniških spletnih vsebinah še posebno problematične.

4.1 Nestandardna stava

V Tabeli 2 so podane frekvence oznak za nestandardno stavo vejic po posameznih kategorijah.

9 Z vidika metodologije je nujno izpostaviti, da označevalska ekipa v obeh postopkih ni bila (povsem) enaka, prav tako se je zamenjal kurator, oboje iz drugotnih, pragmatičnih razlogov.

Tabela 2: Pregled skladenjskih in neskladenjskih oznak po kategorijah. Pri nadkategorijah so podani deleži oznak znotraj skladenjskega in neskladenjskega segmenta.

Kategorija	Odveč	Manjka	Nap. znak
1 SKLADENJSKA	**19 (4,7 %)**	**382 (95,3 %)**	**0**
1.1 Priredja	**8 (2 %)**	**35 (8,7 %)**	**0**
1.1.1 Vezalno	**6**	8	**0**
1.1.2 Stopnjevalno	0	2	**0**
1.1.3 Ločno	2	0	**0**
1.1.4 Protivno	0	**13**	**0**
1.1.5 Vzročno	0	1	**0**
1.1.6 Posledično	0	10	**0**
1.1.7 Pojasnjevalno	0	1	**0**
1.2 Odvisniki	**2 (0,5 %)**	**230 (57,4 %)**	**0**
1.2.1 Osebkov	0	26	**0**
1.2.2 Predmetni	1	**89**	**0**
1.2.3 Krajevni	0	1	**0**
1.2.4 Časovni	0	20	**0**
1.2.5 Načinovni	1	11	**0**
1.2.6 Vzročni	0	16	**0**
1.2.7 Namerni	0	0	**0**
1.2.8 Pogojni	0	25	**0**
1.2.9 Dopustni	0	1	**0**
1.2.10 Prilastkov	0	41	**0**
1.3 Polstavek	**0**	**2 (0,5 %)**	**0**
1.4 Stavčni člen	**3 (0,7 %)**	**0**	**0**
1.5 Besedna zveza	**6 (1,5 %)**	**0**	**0**
1.6 Pa-, pri- do- in izpostavek	**0**	**115 (28,7 %)**	**0**
2 NESKLADENJSKA	**0**	**3 (75 %)**	**1 (25 %)**

Kot lahko razberemo iz Tabele 2, so oznake med kategorijami razporejene zelo heterogeno, zelo neenakomerno pa so posejane tudi oznake glede na vrsto skladenjskega razmerja. Ta heterogenost je značilna za obe ravni tehnične standardnosti tvitov,[10] ki sta zajeti v analiziranih podatkih, je pa treba omeniti, da je v tvitih, ki so tehnično nestandardni, vejica stavljena bistveno manj standardno (v tvitih z oznako T1L3 lahko denimo najdemo 118 manjkajočih vejic, v tvitih T3L3 pa 266 manjkajočih vejic). To lahko pripišemo tudi motiviranosti uporabnikov za stavljenje ločil, saj je pri odvečnih vejicah razmerje precej bolj

10 V korpusu Janes imajo vsa besedila pripisano stopnjo tehnične (zapis presledkov, ločil, velikih in malih začetnic) in jezikovne standardnosti (zapis besed, raba nestandardnega besedišča). Uporabljena je trostopenjska lestvica, pri čemer 1 označuje zelo standardna, 3 pa zelo nestandardna besedila (glej Ljubešić et al. 2018).

enakovredno – celo obratno (12 odvečnih vejic v tehnično standardnih tvitih, 7 pa v tehnično nestandardnih tvitih).

Kot lahko vidimo, je bilo v 500 tvitih zanemarljivo malo (4) primerov napačno stavljene neskladenjske vejice, prav tako pa se je pri izbranem vzorcu pokazalo, da je v uporabniških spletnih vsebinah skorajda zanemarljiv pojav odvečna vejica (19 primerov), saj je velika večina (382 primerov) oznak vezana na manjkajočo vejico. Pri tem je treba poudariti, da zaradi majhnega vzorca te ugotovitve ni mogoče posploševati. Velika večina popravkov se nanaša na skladenjsko rabo vejice (401 od skupno 405 primerov oz. 99 %). Glede na rezultate empirične analize lahko vidimo, da sta še posebno akutni kategoriji manjkajočih vejic pri odvisnikih (230 primerov oz. 57 % vseh skladenjskih oznak) in pastavčnih strukturah (115 primerov oz. 28,7 % vseh skladenjskih oznak).

4.1.1 Manjkajoča vejica

Kot lahko vidimo v Tabeli 2, je stava vejice z vidika manjkajoče vejice problematična predvsem pri odvisniških in pristavčnih strukturah, zato se v nadaljevanju nekoliko podrobneje posvetimo prav tema kategorijama. V Tabeli 3 je podan prikaz najpogostejših odvisniških vrst z označeno manjkajočo vejico.

Tabela 3: Pregled manjkajočih vejic pri odvisnikih po frekvenci in deležu znotraj skladenjske kategorije.

Odvisnik	Število oznak
Predmetni (D)	**80 (19,8 %)**
Prilastkov (D)	29 (7,1 %)
Osebkov (D)	24 (5,9 %)
Vzročni (D)	16 (4,0 %)
Pogojni (D)	13 (3,2 %)
Časovni (D)	13 (3,2 %)
Pogojni (L)	12 (2,9 %)
Prilastkov (L)	12 (2,9 %)
Načinovni (D)	10 (2,5 %)
Predmetni (L)	9 (2,2 %)
Časovni (L)	7 (1,7 %)
Osebkov (L)	2 (0,5 %)
Načinovni (L)	1 (0,2 %)
Krajevni (D)	1 (0,2 %)
Dopustni (D)	1 (0,2 %)

Kot lahko vidimo, je od skupno 230 manjkajočih vejic pri odvisnikih edina zares problematična kategorija desnosmerne vejice pri predmetnih odvisnikih, pri katerih najdemo več kot tretjino primerov z manjkajočo vejico. Tudi na splošno se je izkazalo, da je desnosmerna vejica precej bolj problematična kot levosmerna, saj zaseda šest najpogostejših mest z manjkajočo vejico (175 od skupno 230 manjkajočih vejic). Edini odvisniški vrsti, pri katerih smo zaznali nekaj več manjkajočih levosmernih vejic, sta kategoriji prilastkovih in vzročnih odvisnikov. Za slednje lahko sklepamo, da pogosto začenjajo povedi, pri prilastkovih odvisnikih pa uporabniki pogosto najverjetneje pozabijo skleniti oziralni odvisnik.

Tabela 4: Pregled oznak pri pa-, pri-, do- in izpostavčnih strukturah po frekvenci in deležu skladenjskih oznak.

	Odvečna	Manjkajoča
Levosmerna	/	79 (19,7 %)
Desnosmerna	/	36 (9,0 %)

Pri pristavkih, dostavkih, izpostavkih in pastavkih je problematična predvsem vejica na začetku, relativno pogosta pa je tudi nestandardna stava brez vejice na koncu stavka. V številnih primerih gre za nestandardno stavo vejice pri denimo medmetih (npr. hahaha), ki jih – tako domnevamo – uporabniki spletnih omrežij ne dojemajo kot pastavčne tvorbe, temveč kot metajezikovno pojavnost oz. verbalizirane emotikone. Zelo pogosta je tudi pojavnost nestandardne rabe vejice pri členkovnih pastavkih na začetku in koncu stavka, zlasti v navezavi s sklicem na uporabniško ime. Tovrstne rabe v korpusu nismo označevali kot odmik od standarda, predvsem zaradi pogostosti in zaradi tehničnih zahtev Twitterja, saj lahko stavljenje vejice v sklice povzroči težave s sklicem. V vsakem primeru pa način uporabe sklicev potrjuje domnevo, da uporabniki izražajo metajezikovne prvine z uporabo jezikovnih sredstev.

4.1.2 Odvečna vejica

Vseh primerov odvečne vejice od skupno 401 skladenjske oznake v zbirki podatkov je bilo 19, kar pomeni manj kot 5 odstotkov vseh oznak, njihova porazdelitev po kategorijah pa je naslednja:

Tabela 5: Pregled odvečnih vejic glede na kategorijo po frekvenci in deležu skladenjskih oznak.

Oznaka	Število
Vezalno priredje	6
Besedna zveza	6
Stavčni člen	3
Ločno priredje	2
Predmetni odvisnik	1
Načinovni odvisnik	1

Kot lahko vidimo, je bila vejica največkrat odvečna v besednih zvezah (denimo med sestavljenimi vezniki), med strukturami v vezalnem ali ločnem priredju, za stavčnim členom in med enakovrednima odvisnikoma. Sodeč po skupnem številu označenih nestandardnih mest, bi pričakovali večje število odvečnih vejic, predvsem stavčnočlenskih, zlasti glede na raziskave na standardnojezikovnem gradivu (prim. Popič 2014). Preden lahko zaključimo, da je neskladenjska raba vejice pretežno manjkajoča, ker je to ena od strategij krajšanja sporočil v tovrstnem načinu komuniciranja in ker je vejico psihološko in tehnično lažje izpustiti kot pa napisati po nepotrebnem, bi bilo nujno treba opraviti analizo na večjem vzorcu besedil tega žanra.

4.1.3. Neskladenjska vejica

Kot smo pokazali v predhodnih poglavjih, so bili v celotnem naboru podatkov označeni zgolj štirje primeri nestandardne neskladenjske vejice, ki so se pojavili v treh tvitih. Te izpisujemo v celoti v Tabeli 6:

Tabela 6: Pregled primerov z označeno nestandardno rabljeno neskladenjsko vejico.

Setamo po Lj in pride Zoki mim, se ustavi pa da Emanuelu petko. Tamalmu nc jasn. Recem**(,)** to je Zoki kralj**(,)** in se tamal zadere: Zoki kralj! :)
@user Ja bi mogla tud jst naletet na dobrega :) ampak ima res ogromno folka sfaljenega (ne ostri do **2.8**). Je, kar je.
@user »Maybe yes, maybe no, maybe go home**(,)**« so radi rekli Šerpe na odpravi na Annapurno, kadar smo jih vprašali, če bo vreme OK

Kot lahko vidimo, so od tega tri vejice v resnici skladenjske, in sicer gre za tri primere soredja (1. in 3. tvit; v oklepaju so podane manjkajoče vejice na mestih, ki so bila označena tudi med označevanjem), ki ga namenoma nismo vključili v tipologijo, saj v tem mediju nismo pričakovali primerov premega govora. Pri drugem primeru (krepko je tiskana pojavnica, ki je bila označena med označevanjem) pa je dejansko neuporabljena neskladenjska vejica, vendar lahko v tem primeru oznako problematiziramo, saj gre – kolikor lahko razberemo iz sobesedila – za navedbo odprtosti zaslonke fotografske leče, ki se tipično podaja s piko. Podobno kot pri nekaterih drugih kategorijah smo glede na dosedanje raziskave, izvedene na standardnem gradivu (glej npr. Popič 2014), pričakovali bistveno več primerov nestandardne rabe neskladenjske vejice.

4.2 Standardna stave vejice

Kot prikazuje Tabela 7, lahko 628 oznak za standardno stavljeno vejico v korpusu tvitov razporedimo v 19 kategorij, in medtem ko so oznake povečini homogeno razporejene po korpusu, lahko vidimo, da štiri kategorije izrazito izstopajo, in sicer gre za (levosmerno) pristavčno vejico ter (desnosmerno) vejico pri vezalnem priredju, predmetnem odvisniku in protivnem priredju.

Tabela 7: Pregled oznak standardne stave vejice v korpusu Janes-Vejica.

Kategorija	Št. oznak	
Pa-, pri- do- in izpostavek (L)	127	20,22%
Vezalno priredje (D)	99	15,76%
Predmetni odvisnik (D)	88	14,01%
Protivno priredje (D)	70	11,15%
Prilastkov odvisnik (D)	34	5,41%
Pa-, pri- do- in izpostavek (D)	31	4,94%
Pojasnjevalno priredje (D)	28	4,46%
Vzročni odvisnik (D)	24	3,82%
Pogojni odvisnik (D)	23	3,66%
Načinovni odvisnik (D)	18	2,87%
Časovni odvisnik (D)	17	2,71%
Pogojni odvisnik (L)	14	2,23%
Posledično priredje (D)	14	2,23%
Namerni odvisnik (D)	10	1,59%
Časovni odvisnik (L)	8	1,27%
Osebkov odvisnik (D)	8	1,27%
Prilastkov odvisnik (L)	5	0,80%
Stopnjevalno priredje (D)	5	0,80%
Predmetni odvisnik (L)	5	0,80%

Kot lahko vidimo, je kategorija z najvišjo frekvenco pri standardni stavi vejice obenem tudi ena najbolj problematičnih pri nestandardni stavi, s primerljivo frekvenco (127 : 115), kar pomeni, da je kljub zelo frekventni standardni stavi vejice pri pri-, pol-, do- in izpostavčnih strukturah v polovici primerov pri tovrstnih strukturah stava vejice še vedno nestandardna. Če združimo levo- in desnosmerno standardno stavo vejic pri tovrstnih strukturah (158 oz. več kot četrtina vseh standardno stavljenih vejic), vidimo, da je to relativno akutna zapisovalna težava, saj je 42 % vseh vejic stavljenih nepravilno (v vseh primerih gre za manjkajočo vejico), vendarle pa drži, da je zgolj 30 primerov nestandardne vejice.

Vezalno priredje zaradi svoje narave v normativnem smislu ni problematično, veliko težav pa povzročajo predmetni odvisniki. V označenem korpusu lahko tako najdemo 88 primerov standardne stave (desnosmerne) vejice pri predmetnih odvisnikih, medtem ko je primerov manjkajoče (desnosmerne) vejice celo več (90), kar nakazuje na to, da gre pri tovrstnih odvisnikih za pereč problem. To lahko v večji ali manjši meri pripišemo tudi dejstvu, da gre za pogost pojav v jeziku, obenem pa ugotovitve pritrjujejo dosedanjim raziskavam, ki kažejo, da uporabniki vejico pred predmetnim odvisnikom izpuščajo, zlasti v primerih, ko je predmetni veznik uveden z manj pogostim (ali formaliziranim) veznikom. Na to nakazujejo tudi frekvence posameznih pojavnic, pred katerimi v korpusu Janes-Vejica najdemo največ oznak za standardno stavo vejic in ki so podane v Tabeli 8 (zajete so pojavnice z vsaj 10 oznakami v korpusu).

Tabela 8: Frekvenca pojavnic za standardno stavljenimi vejicami v korpusu Janes-Vejica.

Pojavnica	Št. oznak
da	105
pa	31
če	27
ko	20
ne	16
ki	16
ampak	16
ker	14
je	12

Kot lahko vidimo, gre pri vseh pojavnicah za izrazito predvidljive slovnične besede, pri katerih raba vejice načeloma ne omahuje in ne peša. Seveda pa to ne pomeni, da je raba tudi vedno pravilna, zlasti v luči tega, da zaradi kompleksne sintakse vejica pred njimi pogosto mora biti izpuščena. Da pa bi lahko ugotovili, katere pojavnice so najbolj problematične, v Tabeli 9 podajamo pojavnice, pred

katerimi vejica najpogosteje umanjka (ponovno so zajete zgolj pojavnice, ki se v korpusu pojavijo vsaj 10-krat; tiste, ki jih najdemo med najpogostejšimi tudi pri standardni stavi vejice, so pisane krepko).

Tabela 9: Pojavnice z največ oznakami za manjkajočo vejico v korpusu Janes-Vejica.

Pojavnica	Št. oznak
da	**71**
k	**26**
pa	**17**
in	12
če	**10**
ker	**10**

Kot prikazuje Tabela 9, so najpogostejše oznake v veliki meri prekrivne, kar nakazuje na to, da je opuščanje vejic vsaj pred določenimi pojavnicami (denimo bazičnimi vezniki, kot so *da*, *ki*, *ko*, *če* in *ker*) zavestna odločitev, ne napaka. Vendar pa je pri tem treba omeniti, da lahko to sklepamo predvsem na podlagi raziskav jezikovne rabe poklicnih oz. izobraženih piscev, ne pa tudi šolarjev, kot prikazuje Tabela 10.

Tabela 10: Pojavnice z največ oznakami za manjkajočo vejico v korpusih Lektor in Šolar.

Lektor	Frekvenca	Šolar	Frekvenca
in	269	da	1524
kot	73	je	668
je	69	saj	585
ali	49	kako	394
ne	46	se	370
»	36	kar	325
da	28	kaj	312
v	26	kot	292
so	23	in	291
pa	22	ki	280
naj	22	ko	266
se	20	ker	222
«	19	naj	221
s	18	v	220
ter	16	vendar	192

Kot lahko vidimo, se težave z vejico pri šolarjih in poklicnih piscih in prevajalcih precej razlikujejo. Medtem ko rezultati iz Lektorja kažejo, da uporabniki vejico pozabijo/izpustijo v bolj zapletenih stavčnih strukturah, rezultati iz Šolarja kažejo, da šolarji vejico pogosto izpuščajo tudi pred povsem predvidljivimi in bazičnimi strukturami, kar je zaradi šolske narave seveda razumljivo, obenem pa lahko vidimo, da se najpogostejše pojavnice za izpuščenimi vejicami v bistveno večji meri ujemajo s podatki iz Šolarja kot iz Lektorja.

4 SKLEP

V prispevku smo predstavili tipologijo za označevanje rabe vejice v slovenščini in empirično raziskavo o rabi vejice, temelječo na korpusu Janes-Vejica. Temeljne rezultate raziskave lahko v grobem povzamemo z naslednjimi točkami:

1) Tudi v jezikovno nestandardni računalniško posredovani komunikaciji je standardna vejica pogostejša od nestandardne.
2) Točki 1 navkljub lahko vidimo, da uporabniki vejico pogosto tudi izpuščajo oz. ne stavijo pravilno, kot to velja za večino pisanja v slovenščini, ne glede na medij. Vsaj za določen del gradiva lahko utemeljeno trdimo, da gre za nameren izpust vejice, to pa je v spletnem okolju vsaj za del pisne produkcije tudi pričakovano in samoumevno.
3) Ko gre za nestandardno vejico, je bistvena manjkajoča vejica, odvečna vejica je v našem vzorcu malodane neproblematična.
4) Z vidika normativnosti imajo uporabniki težave predvsem s prepoznavanjem skladenjskih enot, kar bi moralo biti vodilo pri prihodnjem delu slovenske preskripcije. Sistem, opisan v pričujočem prispevku, se ponuja kot ustrezen za empirično raziskovanje jezikovne rabe (za širok razpon besedilnih zvrsti).
5) Tipologija, razvita za namene pričujoče raziskave, se je izkazala za ustrezno oz. vsaj zadostno za tovrstno raven jezikovnega opisa. Vsakršna globlja analiza bi terjala nadaljnje drobljenje kategorij.

Kot kažejo rezultati študije, je standardna vejica v jezikovno nestandardnih tvitih pogostejša od nestandardne. Pri tem seveda govorimo zgolj o frekvenci, kar pomeni, da je nujno izvesti še statistično-kvalitativno raziskavo, s katero bomo lahko natančno prikazali trende rabe, nerabe in neznanja uporabe vejice. Rezultati pričujoče raziskave prav tako pritrjujejo preteklim analizam, da je nestandardna stava vejice v slovenščini precejšen problem, ne glede na medij ali besedilno zvrst, kar pomeni, da je tako z raziskovalnega kot tudi preskriptivističnega (najbrž pa tudi pedagoškega) vidika ta tematika potrebna čim obširnejše obravnave, po

možnosti temelječe na realnem in sodobnem gradivu. Pričujoča razprava predstavlja eno od možnosti, kako tovrstno analizo zastaviti.

Zahvala

Avtorja se za pomoč pri anotaciji, kuriranju in pripravi tipologije zahvaljujeva Katji Zupan, Poloni Logar in Teji Kavčič.

Literatura

Arhar Holdt, Špela in Kaja Dobrovoljc, 2016: Vrednost korpusa Janes za slovensko normativistiko. *Slovenščina 2.0*, 4 (2): 1–37.

Dobrovoljc, Helena, 2004: *Pravopisje na Slovenskem*. Ljubljana: Založba ZRC.

Dobrovoljc, Helena, 2008: Jezik v e-poštnih sporočilih in vprašanja sodobne normativistike. Košuta, Miran (ur.): *Slovenščina med kulturami. Zbornik Slavističnega društva Slovenije* 19. Celovec: Slavistično društvo Slovenije. 295–314.

Eckart de Castilho, Richard, Chris Biemann, Iryna Gurevych in Sied Muhie Yimam. 2014. WebAnno: a flexible, web-based annotation tool for CLARIN. *Proceedings of the CLARIN Annual Conference (CAC) 2014*. Soesterberg, Netherlands. https://www.clarin.eu/sites/default/files/cac2014_submission_6_0.pdf

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

Fišer, Darja, Tomaž Erjavec, Ljubešić, Nikola, 2017: The compilation, processing and analysis of the Janes corpus of Slovene user-generated content. Ciara, R. Wigham in Gudrun Ledegen (ur.): *Corpus de communication médiée par les réseaux: construction, structuration, analyse*. Collection Humanités Numériques. Pariz: L'Harmattan (v tisku).

Goli, Teja, Osrajnik, Eneja in Darja Fišer, 2016: Analiza krajšanja slovenskih sporočil na družbenem omrežju Twitter. *Proceedings of the Language Technologies and Digital Humanities Conference*. Ljubljana, Slovenia. 77–82.

Jakop, Nataša, 2008: Pravopis in spletni forumi – kva dogaja? Košuta, Miran (ur.): *Slovenščina med kulturami. Zbornik Slavističnega društva Slovenije 19*. Celovec: Slavistično društvo Slovenije. 315–327.

Korošec, Tomo, 2003: K pravilom za skladenjsko vejico v Slovenskem pravopisu 2001. *Slavistična revija* 5/2: 247–266.

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

Levec, Fran, 1899: *Slovenski pravopis*. Dunaj: Cesarska kraljeva zaloga šolskih knjig.

Ljubešić, Nikola, Darja Fišer, Tomaž Erjavec, Čibej, Jaka, Marko, Dafne, Pollak, Senja in Škrjanec, Iza, 2015: Predicting the level of text standardness in user-generated content. *Proceedings of Recent Advances in Natural Language Processing*, 7.–9. september 2015. Hisar, Bolgarija. 371–378.

Ljubešić, Nikola, Tomaž Erjavec in Darja Fišer, 2018: Orodja za procesiranje nestandardne slovenščine. Fišer, Darja (ur.): *Viri, orodja in metode za analizo spletne slovenščine*. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani. 74–99.

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

Logar, Nataša in Damjan Popič, 2015: Vejica: rezultati anketne raziskave med dijaki in študenti. *Jezikoslovni zapiski* 2/2. 45–59.

Michelizza, Mija, 2014: Slovenščina v elektronskih medijih. *Razpotja* 15. http://www.razpotja.si/slovenscina-v-elektronskih-medijih/

Nebeská, Iva, 1999: *Jazyk. Norma. Spisovnost*. Praga: Karolinum.

Osrajnik, Eneja, Darja Fišer in Damjan Popič, 2015: Primerjava rabe ekspresivnih ločil v tvitih slovenskih uporabnikov in uporabnic. Fišer, Darja (ur.): *Zbornik konference Slovenščina na spletu in v novih medijih*. Ljubljana: Znanstvena založba Filozofske fakultete, 50–74.

Popič, Damjan, 2014: *Korpusnojezikoslovna analiza vplivov na slovenska prevodna besedila*. Doktorska disertacija. Ljubljana: Filozofska fakulteta.

Popič, Damjan in Darja Fišer, 2015: Vejica je mrtva, živela vejica. Žele, Andreja (ur.): *OBDOBJA 34: Slovnica in slovar – aktualni jezikovni opis*. Ljubljana: Znanstvena založba Filozofske fakultete. 609–618.

Popič, Damjan in Nataša Logar, 2015: Med dvema ognjema: kje stoji vejica v slovenskih gimnazijah? *OBDOBJA 34: Slovnica in slovar – aktualni jezikovni opis*. Ljubljana: Znanstvena založba Filozofske fakultete. 619–627.

Popič, Damjan, Darja Fišer, Katja Zupan in Polona Logar, 2016: Raba vejice v uporabniških spletnih vsebinah. Erjavec, Tomaž in Darja Fišer (ur.): *Zbornik konference Jezikovne tehnologije in digitalna humanistika*, 29. september–1. oktober 2016, Filozofska fakulteta, Univerza v Ljubljani, Ljubljana, Slovenija. 106–110.

Popič, Damjan in Darja Fišer, 2017: Fear and Loathing on Twitter: Attitudes towards Language. *Proceedings of the 5th Conference on CMC and Social Media Corpora for the Humanities (cmccorpora 17)*, oktober 2017, Eurac Research, Italija. 61–64.

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

Slovenski pravopis, 2001. Pravila. Ur. Toporišič, J. et al. Ljubljana: Založba ZRC, ZRC SAZU.

Verovnik, Tina, 2003: Vejica premalo, vejica preveč (2). *Pravna praksa* 22/21. 51.

Žibert, Živa, 2006: Slovenska vejica: balast ali skladenjska nujnost slovenskega knjižnega jezika? Diplomsko delo. Ljubljana: Fakulteta za družbene vede.

Regionalne jezikovne različice v slovenski računalniško posredovani komunikaciji: korpusni pristop z ročno označenim korpusom Janes-Geo

Jaka Čibej

Izvleček

V poglavju predstavljamo gradnjo in analizo ročno označenega korpusa Janes-Geo, ki predstavlja prvi korak h korpusnemu proučevanju slovenskih regionalnih jezikovnih različic v spletni slovenščini. Korpus Janes-Geo vsebuje približno 64.000 pojavnic, ki jih je prispevalo približno 270 uporabnikov Twitterja, ki glede na avtomatsko pripisane metapodatke o regionalni pripadnosti spadajo v eno od devetih regij (primorska, gorenjska, rovtarska, ljubljanska, dolenjska, štajerska, koroška, mariborska in panonska). V korpusu so bile ročno označene nestandardne jezikovne prvine v skladu z izdelano tipologijo. Namen korpusa Janes-Geo je dvojni: ugotoviti, v kakšnih oblikah se (najpogosteje) izraža jezikovna nestandardnost v spletni slovenščini, in primerjati razlike v rabi nestandardnih jezikovnih prvin med uporabniki iz različnih regij. Poleg postopka avtomatskega pripisovanja metapodatkov o regionalni pripadnosti uporabnikov opišemo tudi označevanje korpusa, njegovo sestavo in nekatere poglavitne razlike med njegovimi regionalnimi podkorpusi, npr. pogostost izpustov soglasnikov in samoglasnikov, različne nestandardne oblikoslovne prvine, najpogostejše nestandardno besedje in najpogostejše transformacije grafemov.

Ključne besede: regionalne jezikovne različice, slovenščina, tviti, geolokacija, računalniško posredovana komunikacija

1 UVOD

Z vzponom spleta in računalniško posredovane komunikacije (RPK) 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 so Facebook, Twitter, WhatsApp in Snapchat. Jezik v RPK (še posebej v klepetih in v drugih podobno neformalnih kontekstih) se od standarda precej razlikuje (Crystal 2011, Baron 2010, Myslin in Gries 2010), ena od njegovih ključnih značilnosti pa so tudi regionalno specifične jezikovne prvine (Ueberwasser 2013, Huang et al. 2016), kar velja tudi za slovensko RPK. Slovenščina je kljub relativno majhnemu številu govorcev in geografskemu ozemlju zelo razčlenjena: Ramovš (1931) je npr. govorce slovenščine razdelil v 7 narečnih skupin s skupno več kot 40 narečji in podnarečji (glej tudi Škofic et al. 2011: 11). Temu primerno so tudi regionalne jezikovne različice[1] slovenščine precej obširno raziskane, a le v govoru. Sistematičnih raziskav o tem, kako se slovenska regionalna jezikovna členjenost odseva v spletnem sporazumevanju, ki je za razliko od tradicionalne govorjene narečne rabe najpogosteje pisno, pa še ni na voljo, čeprav pisna spletna komunikacija že dolgo več ne zajema zanemarljivega deleža: kot poroča Valicon (2016), ima v Sloveniji profil na Facebooku že več kot 830.000 oseb v starosti od 15 do 75 let, skoraj 600.000 (oz. 70 %) oseb pa ga uporablja vsak dan. Podobno je tudi na Twitterju, kjer je profilov več kot 200.000, dnevnih uporabnikov 33.000, tedenskih pa 100.000.

V pričujočem poglavju povzemamo in nadaljujemo prve korake (Čibej in Ljubešić 2015, Čibej 2016) h korpusnemu proučevanju slovenskih regionalnih jezikovnih različic v spletni slovenščini in še dodatno razširimo nabor raziskav značilnosti slovenske RPK, ki so bile v okviru projekta JANES opravljene npr. o nestandardni skladnji (Arhar Holdt 2018), rabi vejic (Popič in Fišer 2018), pojavih krajšanja (Goli et al. 2016) ter preklapljanju med jeziki (Reher in Fišer 2018). Namen naše raziskave je predstaviti eno od metod za proučevanje regionalnih jezikovnih različic na Twitterju, ponuditi sistematičen uvid v načine, na katere se nestandardnost kaže v slovenski spletni komunikaciji, in ugotoviti, ali se izražanje nestandardnosti razlikuje med različnimi regijami.

Poglavje začnemo s kratkim pregledom sorodnih raziskav o regionalni jezikovni variantnosti v računalniško posredovani komunikaciji (razdelek 2). Nato v razdelku 3 opišemo postopek avtomatskega pripisovanja metapodatkov o regionalni pripadnosti uporabnikom, način vzorčenja korpusa in izdelavo tipologije nestandardnih jezikovnih prvin v slovenskih tvitih ter smernic za označevanje. V razdelku 4 predstavimo poglavitne razlike med regionalnimi

1 V članku uporabljamo termin regionalne jezikovne različice v smislu jezikovnega sistema, ki je odvisen od geografskega oz. regionalnega izvora jezikovnega uporabnika.

podkorpusi glede na šest glavnih kategorij nestandardnih jezikovnih prvin iz tipologije: izpusti, transformacije, nestandardno besedje, različice pogostih besed, nestandardno oblikoslovje in drugo. V zaključku strnemo ugotovitve, orišemo uporabnost korpusa za jezikoslovne raziskave in navedemo načrte za prihodnje delo.

2 REGIONALNA JEZIKOVNA VARIANTNOST V GOVORU IN RAČUNALNIŠKO POSREDOVANI KOMUNIKACIJI

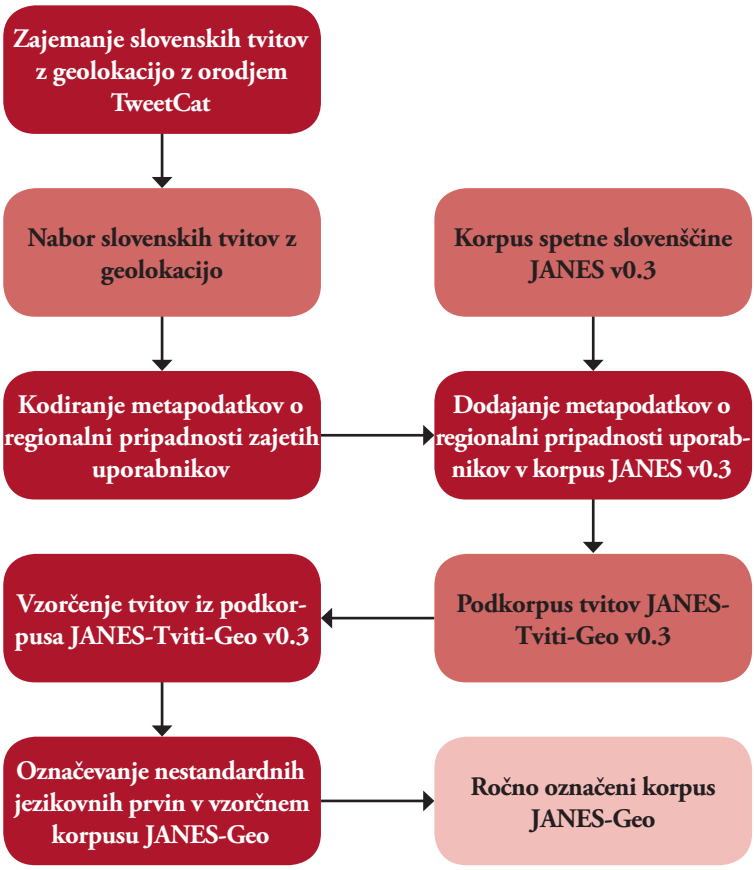
Raziskava, opisana v tem prispevku, se umešča na področje sociolingvistike, natančneje v disciplini korpusne dialektologije in korpusne dialektometrije (Szmrecsanyi 2011), ki temelji na analizi korpusnih besedil za sistematično merjenje razdalj in razlik med jezikovno rabo uporabnikov iz različnih geografskih regij. Za številne tuje jezike so bile že opravljene korpusne dialektološke raziskave, najpogosteje na podlagi transkripcij govora v govornih korpusih, kot je npr. korpus DynaSAND za nizozemska narečja (Kunst in Wesseling 2010), Nordic Dialect Corpus za nordijske jezike (Johanessen et al. 2009) in Freiburg Corpus of English Dialects za regionalne jezikovne različice angleščine (Hernández 2006). Slovenska dialektologija se je doslej zanašala predvsem na terenske raziskave z informanti (glej npr. Logar 1981, Kenda Jež 2002), ustrezen dialektološki korpus, ki bi omogočal korpusni pristop k problemu, pa še ni bil zgrajen. Korpus govorjene slovenščine GOS (Verdonik in Zwitter Vitez 2011) sicer vsebuje posnetke govorcev iz vseh regij, a primarno ni bil zasnovan za dialektološke namene. Na tej točki je treba omeniti tudi to, da so se slovenske dialektološke raziskave do zdaj opirale na t. i. idealnega narečnega govorca (Bitenc 2016: 180), pri katerem ni prišlo do prilagajanja drugim jezikovnim različicam. Ta pristop pa zanemarja širok nabor govorcev in jezikovnih različic, ki v različnih vidikih odstopajo tako od standarda kot od t. i. čistega narečja. V okviru raziskave v tem prispevku se ne osredotočamo na določen tip govorca, temveč jezikovno variantnost opisujemo empirično in brez predpostavk, tudi zato, ker gre pri računalniško posredovani komunikaciji za drug medij (pisni).

Raziskovanje regionalnih jezikovnih prvin v uporabniških spletnih vsebinah je tudi v tujini še precej sveže. S tem področjem so se do zdaj ukvarjali pretežno jezikovni tehnologi, jezikoslovci pa v mnogo manjši meri. Raziskave so se osredotočale predvsem na gradnjo novih orodij npr. za avtomatsko prepoznavanje regionalnih jezikovnih različic (Harrat et al. 2013, Cotterell in Callison-Burch 2014 za arabščino; Eisenstein et al. 2010, Eisenstein et al. 2015 za ameriško angleščino; Ljubešić in Kranjčić 2014 za hrvaščino, srbščino, bosanščino in

črnogorščino), za strojno prevajanje (Harrat et al. 2014 med regionalnimi jezikovnimi različicami arabščine in sodobno standardno arabščino; Haddow et al. 2013 med dunajsko regionalno jezikovno različico nemščine in standardno avstrijsko nemščino) in oblikoskladenjsko označevanje (Khakimov et al. 2015 za tatarsko narečje mišar; Ruef in Ueberwasser 2013 za švicarsko nemščino; Bernhard in Ligozat 2013 za alzaščino). Gre torej za raziskovalno področje, ki pokriva zelo raznovrsten nabor jezikov, tudi neinstitucionalnih in takšnih z majhnim številom govorcev. To kaže na svetovni trend, ki potrjuje, da bi bilo tudi slovenska jezikovnotehnološka orodja smiselno prilagoditi, da bodo dovolj robustna za obdelavo regionalnih jezikovnih različic na spletu, slovenščini pa za nadaljnji korak v to smer manjka neobremenjen jezikovni opis spletne slovenščine in rabe regionalnih jezikovnih različic v njej. Raziskovanje značilnosti regionalnih jezikovnih različic na spletu se je začelo v jezikoslovni skupnosti razraščati prav zdaj: tudi za angleščino so bila namreč šele pred kratkim objavljena obširnejša dela s tega področja. Grieve (2016) npr. predstavi sodobni korpusni in statistično podprti pristop k dialektologiji in dialektometriji na primeru regionalnih jezikovnih različic pisne ameriške angleščine, o proučevanju regionalne jezikovne členjenosti na družbenih medijih na splošno pa pišejo Eisenstein (2015) in Jørgensen et al. (2015). Za raziskavo v tem prispevku je še posebej relevantnen prispevek Huang et al. (2016), ki obravnava regionalne jezikovne različice ameriške angleščine na Twitterju s pomočjo tvitov z geolokacijo.

3 IZDELAVA ROČNO OZNAČENEGA KORPUSA JANES-GEO

V naslednjih podrazdelkih opisujemo izdelavo ročno označenega korpusa Janes-Geo, ki je bil vzorčen iz korpusa slovenske računalniško posredovane komunikacije Janes v0.3 (Fišer et al. 2015). Izdelava je potekala v več stopnjah, ki jih prikazuje delotok na Sliki 1. Zeleni okvirčki predstavljajo postopke, modri uporabljene podatkovne zbirke, oranžni okvirček pa končni rezultat. Najprej smo zajeli tvite s podatki o geolokaciji in na njihovi podlagi kodirali metapodatke o regionalni pripadnosti zajetih uporabnikov. Metapodatke smo nato dodali v že obstoječi korpus Janes v0.3, na njihovi podlagi pa smo iz njega nato vzorčili besedila za ročno označeni korpus Janes-Geo. Temu sta sledila še izdelava tipologije nestandardnih jezikovnih prvin in ročno označevanje korpusa.

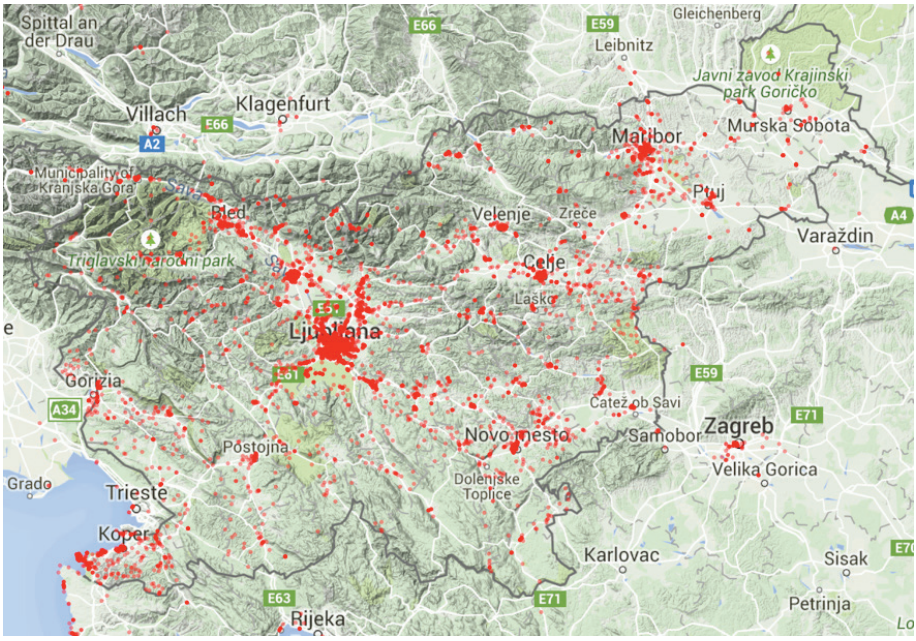


Slika 1: Postopek izdelave ročno označenega korpusa Janes-Geo.

3.1 Zbiranje tvitov s podatki o geolokaciji

Na prvi stopnji smo zbirali slovenske tvite s podatki o geolokaciji, tj. s podatki o zemljepisni širini in dolžini, s katerih je bil tvit poslan. Tvite smo začeli zajemati januarja 2015 z orodjem TweetCat (Ljubešić et al. 2014) in v približno pol leta (do avgusta 2015) zajeli 130.143 tvitov, ki jih je objavilo 1.661 uporabnikov. Razporeditev zajetih tvitov glede na njihove koordinate prikazuje Slika 2.

Iz razporeditve je razvidno, da zajeti tviti pokrivajo dobršen del Slovenije. Največ tvitov je bilo sicer poslanih iz mestnih središč in njihove neposredne okolice, npr. iz Ljubljane, Maribora, Celja in Kranja, a so zastopana tudi manj urbana območja.



Slika 2: Razporeditev zajetih tvitov z geolokacijo.

Na tej stopnji smo upoštevali samo zasebne uporabnike, ki so bili že vključeni v korpus Janes v0.3, ostale pa smo izločili. Za ta poseg smo se odločili ob predpostavki, da lahko od uporabniških računov organizacij, kot so agencije, medijske hiše in podjetja, s precejšnjo verjetnostjo pričakujemo, da na Twitterju v prevladujoči meri objavljajo tvite v standardni slovenščini, obenem pa objavljajo mnogo več avtomatsko generiranih tvitov, kar bi v naš korpus vneslo šum.

Po izločitvi nezasebnih uporabniških računov je ostalo 119.236 tvitov (približno 92 % vseh zajetih), ki jih je napisalo 1.524 uporabnikov. V korpusu Janes v0.3 je zasebnih uporabnikov 5.806, torej smo z zbiranjem tvitov z geolokacijo do avgusta 2015[2] zajeli približno četrtino (26 %) v korpus vključenih uporabnikov.

3.2 Kodiranje metapodatkov o regionalni pripadnosti

V naslednjem koraku smo z orodjem Google Maps API v3 Tool[3] območje Slovenije (vključno z zamejskimi regijami v Furlaniji, na avstrijskem Koroškem in v

2 Z raziskavo smo začeli avgusta 2015 in takrat prvič izvozili zajete tvite z geolokacijo, tvite pa smo zajemali še naprej in novopridobljene podatke (do aprila 2016) uporabili za preverjanje zanesljivosti metode avtomatskega pripisovanja metapodatkov o geolokaciji (več o tem v razdelku 3.2.1).

3 Google Maps API v3 Tool: http://www.birdtheme.org/useful/v3tool.html.

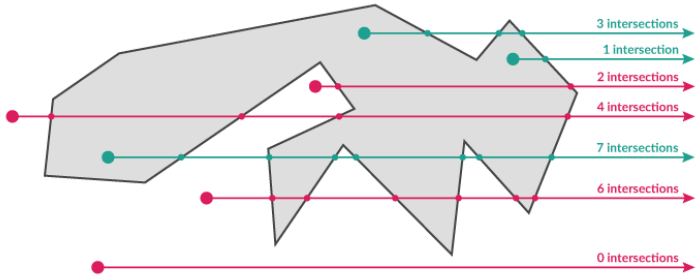
Porabju) razdelili na koordinatne poligone, za kar smo uporabili orodje Google Maps API v3 Tool. Glede na dosedanje raziskave smo imeli na voljo več načinov delitve, ki odsevajo bodisi narečne skupine (Ramovš 1931; Logar in Rigler 1986; Toporišič 2000: 23) ali statistične regije (Zemljarič Miklavčič 2008). Za namene te raziskave smo izbrali delitev na regije[4] v skladu s sedmimi glavnimi narečnimi skupinami po Ramovšu (1931), saj je bila ta kategorizacija med vsemi najbolj robustna, podrobnejša delitev pa bi zaradi neenakomerne razporeditve zbranih podatkov povzročala dodatne težave pri vzorčenju. Slovensko govoreče območje smo tako razdelili na skupno devet koordinatnih poligonov. Prvih sedem predstavlja narečne skupine (gorenjsko, dolenjsko, štajersko, panonsko, koroško, rovtarsko in primorsko), dodatna poligona pa predstavljata Ljubljano in Maribor, ki smo se ju 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) pri zasnovi govornih korpusov. Tako nastale koordinatne poligone predstavlja Slika 3 (ljubljanski in mariborski poligon zaradi majhnosti nista prikazana).

Slika 3: Razdelitev slovensko govorečega področja na koordinatne poligone.

S pomočjo metode metanja žarka (ang. *ray-casting method*; Sparks in Krishnan 2012) smo za vsak zajeti tvit z geolokacijo preverili, iz katerega koordinatnega

4 V članku se pri gradnji in analizi korpusa Janes-Geo s terminom »regije« ne nanašamo na narečne skupine, temveč na koordinatne poligone, na podlagi katerih smo določili metapodatke o regionalni pripadnosti uporabnikov. Za ločevanje terminov narečna skupina in regija smo se odločili, ker metapodatki o regionalni pripadnosti uporabnikov ne odsevajo nujno njihovega jezikovnega ozadja v skladu s kategorizacijo po narečnih skupinah, obenem pa se naša členitev ne prekriva povsem s členitvijo na narečne skupine, saj Ljubljano in Maribor obravnavamo posebej.

poligona je bil poslan. Metoda metanja žarka iz podane točke (v našem primeru so to koordinate tvita) zariše premico oz. žarek ter preveri število presečišč med žarkom in robovi podanega poligona. Če je število liho, točka leži v notranjosti poligona (Slika 4).



Slika 4: Prikaz metode metanja žarka.

Tabela 1 prikazuje razporeditev zajetih tvitov po regijah.[5] V razporeditvi tvitov se že v tem pogledu kaže precejšnja razlika med regijami, saj je bilo daleč največ tvitov (36 %) poslanih iz Ljubljane, najmanj pa iz rovtarske (slaba 2 %) in panonske regije (dobra 2 %), ki sta tudi po površini med najmanjšimi (če ne štejemo Ljubljane in Maribora). Zanimivo je, da poligona Maribora in Ljubljane po količini tvitov nista primerljiva, saj je bilo kar desetkrat več tvitov poslanih iz Ljubljane kot iz Maribora. Dobro zastopana je tudi Gorenjska (slabih 19 %), precejšen delež tvitov pa je bil poslan tudi iz tujine (slabih 16 %).

Tabela 1: Razporeditev tvitov z geolokacijo po regijah.

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

5 Tabela 1 prikazuje tudi tvite, ki so bili poslani iz tujine. Ti ne vključujejo tvitov, poslanih iz Furlanije, avstrijske Koroške in Porabja – te smo dodali primorski, koroški in panonski regiji. Tvitov, ki so bili poslani iz tujine, pri izdelavi in vzorčenju korpusa za ročno označevanje nismo upoštevali.

V naslednjem koraku smo za vsakega od uporabnikov izračunali, kolikšen delež svojih tvitov je poslal iz vsakega od devetih poligonov, glede na deleže pa smo vsakemu uporabniku pripisali metapodatek o regionalni pripadnosti. Ob predpostavki, da so uporabniki pogosto tudi mobilni in objavljajo iz vsaj dveh regij (zlasti če upoštevamo, da so bili tviti zajeti v polletnem obdobju, ki vključuje tudi poletne počitniške mesece), smo določili hevristiko, s katero smo zmanjšali medregionalni šum in se osredotočili samo na podatke, značilnejše za regijo. Metapodatke smo tako pripisali samo uporabnikom, ki so več kot 90 % tvitov poslali iz ene same regije in so obenem poslali vsaj 3 tvite.

Uporabnikov, ki so izpolnjevali oba kriterija, je bilo skupno 269, razrez čistih uporabnikov[6] po regijah pa prikazuje Tabela 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.

Tabela 2: Razporeditev uporabnikov po regijah.

Regija	Število vseh zasebnih uporabnikov	Delež (%)	Število čistih uporabnikov	Delež (%)	Razmerje med čistimi in vsemi zasebnimi uporabniki v regiji
Gorenjska	208	13,65	50	12,95	0,24
Dolenjska	92	6,04	23	5,96	0,25
Štajerska	170	11,15	46	11,92	0,27
Panonska	43	2,82	17	4,40	0,40
Koroška	33	2,17	6	1,55	0,18
Primorska	99	6,50	33	8,55	0,33
Rovtarska	37	2,43	7	1,81	0,19
Ljubljana	506	33,20	125	32,38	0,25
Maribor	59	3,87	14	3,63	0,24
Tujina	277	18,18	65	16,84	0,23
Skupno	1524	100,00	386	100,00	0,25

Skoraj tretjina vseh čistih uporabnikov je spadala v ljubljansko regijo, sledijo pa ji gorenjska (13 %), štajerska (12 %) in primorska regija (8,5 %). Najmanj čistih uporabnikov imata koroška in rovtarska regija (manj kot 2 % odstotka).

V zadnjem stolpcu je podan količnik razmerja med čistimi uporabniki in vsemi zasebnimi uporabniki znotraj regije. Večji količnik pomeni večji delež čistih uporabnikov (in posledično manjšo mobilnost uporabnikov znotraj regije). Pri večini regij je čistih uporabnikov približno četrtina. Izstopajo panonska regija z nekoliko

6 V prispevku uporabnike, ki izpolnjujejo kriterije za pripis metapodatka o regionalni pripadnosti (tj. 90-odstotni prag za delež tvitov iz dominantne regije in najmanj 3 poslani tviti), imenujemo *čisti uporabniki*.

manjšo mobilnostjo ter panonska in koroška regija, kjer je čistih uporabnikov nekoliko manj.

Metapodatke o regionalni pripadnosti smo nato vnesli v korpus Janes v0.3 in izdelali podkorpus Janes-Tweet-Geo v0.3.4, iz katerega smo v naslednjem koraku vzorčili besedila za vzorčni korpus Janes-Geo (glej razdelek 3.3).

3.2.1 Ponovitev kodiranja metapodatkov

Tvite z geolokacijo smo ponovno izvozili aprila 2016. Na tej točki je bilo vseh zajetih tvitov 160.888, kar je 20 % več kot v prvi fazi zbiranja.[7] Na novozajetih tvitih smo ponovili avtomatsko pripisovanje metapodatkov, razporeditev tvitov in uporabnikov po regijah pa se je med fazama večinoma ohranila, kot je razvidno iz Tabele 3 in Tabele 4.

Tabela 3: Razporeditev in porast tvitov po regijah v prvi in drugi fazi zajemanja.

Regija	Število tvitov (1. faza)	Delež (%)	Število tvitov (2. faza)	Delež (%)	Porast tvitov (%)
Gorenjska	22.070	18,51	27.399	17,03	+24 %
Dolenjska	6.922	5,81	9.864	6,13	+43 %
Štajerska	9.284	7,79	15.989	9,94	+72 %
Panonska	2.512	2,11	3.873	2,41	+54 %
Koroška	4.203	3,52	5.170	3,21	+23 %
Primorska	5.748	4,82	8.383	5,21	+46 %
Rovtarska	2.348	1,97	2.873	1,79	+22 %
Ljubljana	43.018	36,08	57.008	35,43	+33 %
Maribor	4.340	3,64	6.116	3,80	+41 %
Tujina	18.791	15,76	24.213	15,05	+29 %
Skupno	119.236	100,00	160.888	100,00	+35 %

V večini regij se je število tvitov povečalo za približno 25–30 %, največji porast pa so zabeležile štajerska (72 %), panonska (54 %) in primorska regija (46 %).

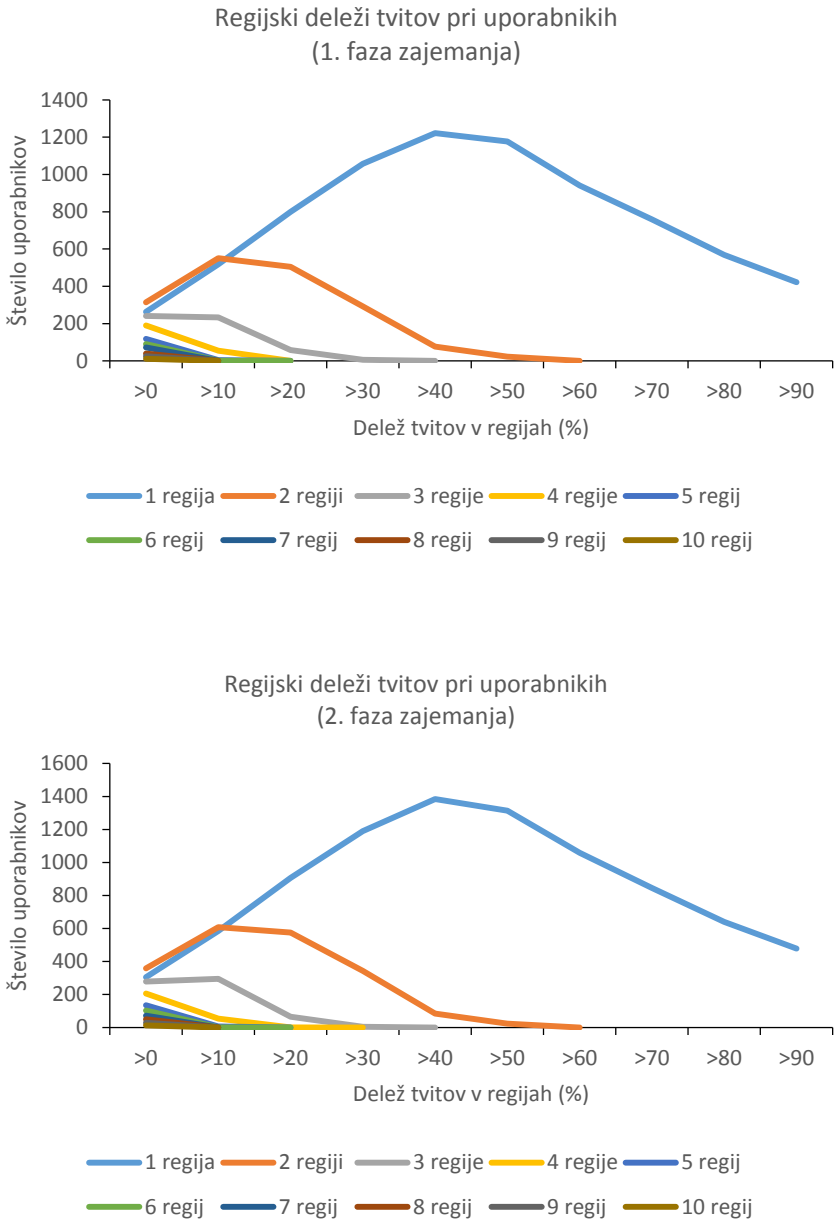
7 Na tej točki je treba omeniti, da so nekateri uporabniki med prvo in drugo fazo zajemanja svoje tvite (ali celo uporabniške račune) izbrisali.

Tabela 4: Razporeditev in porast uporabnikov po regijah v prvi in drugi fazi zajemanja.

Regija	Število čistih uporabnikov (1. faza)	Delež (%)	Število čistih uporabnikov (2. faza)	Delež (%)	Porast uporabnikov (%)
Gorenjska	50	12,95	50	13,89	0 %
Dolenjska	23	5,96	21	5,83	-9 %
Štajerska	46	11,92	43	11,94	-7 %
Panonska	17	4,40	15	4,17	-12 %
Koroška	6	1,55	4	1,11	-33 %
Primorska	33	8,55	30	8,33	-9 %
Rovtarska	7	1,81	7	1,94	0 %
Ljubljana	125	32,38	109	30,28	-13 %
Maribor	14	3,63	15	4,17	+7 %
Tujina	65	16,84	66	18,33	+2 %
Skupno	386	100,00	360	100,00	-7 %

Da bi preverili, do kolikšne mere so avtomatsko pripisani metapodatki stabilni in zanesljivi, smo primerjali spremembe v razporeditvi deležev tvitov pri uporabnikih med prvo in drugo fazo. Rezultati primerjave kažejo, da je populacija zelo stabilna, saj so uporabniki med fazama v veliki večini ohranili svojo dominantno regijo: od čistih uporabnikov sta namreč samo dva (0,12 % vseh čistih uporabnikov v prvi fazi) ob prehodu v drugo fazo spremenila dominantno regijo, omeniti pa je treba tudi, da sta se iz tujine premaknila v ljubljansko regijo, kar pomeni, da ju v raziskavi kot uporabnikov iz tujine nismo upoštevali. 32 čistim uporabnikom iz prve faze se je delež v dominantni regiji nekoliko zmanjšal in padel pod 90-odstotni prag, zato v drugi fazi niso bili več obravnavani kot čisti uporabniki. Iz tega razloga je število čistih uporabnikov med fazama nekoliko upadlo, kar smo pričakovali, saj so nekateri uporabniki v prvi fazi poslali le majhno količino tvitov (med 3 in 10), ki po vsej verjetnosti niso ustrezno predstavljali njihove regionalne razporeditve.

Slika 5 prikazuje število uporabnikov glede na število regij, iz katerih so poslali določen delež svojih tvitov. Vsaka črta predstavlja število regij, v katerih so uporabniki prisotni. Kot vidimo, je zelo malo uporabnikov (okrog 150), ki so tvite pošiljali iz več kot treh regij, obenem pa je tudi malo uporabnikov (okrog 50), ki imajo v več kot dveh regijah delež tvitov nad 20 %. Večina jih ima eno dominantno regijo, iz katere so poslali 50–90 % tvitov, preostali delež pa je razdeljen med druge regije. Med grafoma prve in druge faze ni bistvenih razlik, iz česar lahko sklepamo, da podatki precej zanesljivo predstavljajo regionalno dinamiko uporabnikov, vključenih v pričujočo raziskavo.



Slika 5: Primerjava regijske razporeditve uporabnikov v prvi in drugi fazi.

3.3 Vzorčenje korpusa Janes-Geo

Janes-Tweet-Geo v0.3.4 je podkorpus korpusa Janes v0.3, ki zajema približno 268.000 tvitov in 4,5 milijona pojavnic, vanj pa je vključenih vseh 386 uporabnikov, ki so bili med kodiranjem metapodatkov o regionalni pripadnosti opredeljeni kot čisti (glej opombo 5). Tabela 5 prikazuje prerez podkorpusa Janes-Tweet-Geo v0.3.4 po regionalnih metapodatkih uporabnikov.

Tabela 5: Sestava podkorpusa Janes-Tweet-Geo v0.3.4 po regionalnih podkorpusih.

Regija	Tviti	Delež (%)	Pojavnice	Delež (%)	Pojavnice L1	Pojavnice L2	Pojavnice L3	Delež L1 (%)	Delež L2 (%)	Delež L3 (%)
Gorenjska	39.961	15	620.386	14	337.157	180.769	102.460	54	29	17
Dolenjska	18.204	7	312.783	7	220.718	69.227	22.838	71	22	7
Štajerska	43.787	16	751.658	17	497.089	202.650	51.919	66	27	7
Panonska	5.198	2	76.048	2	55.814	18.253	1.981	73	24	3
Koroška	6.518	2	129.104	3	74.382	47.256	7.466	58	37	6
Primorska	14.211	5	241.023	5	176.326	53.370	11.327	73	22	5
Rovtarska	5.215	2	81.844	2	52.660	22.767	6.417	64	28	8
Ljubljana	96.217	36	1.569.946	36	1.113.182	369.921	86.843	71	24	6
Maribor	5.001	2	81.276	2	60.417	18.013	2.846	74	22	4
Tujina	33.632	13	547.655	12	358.477	151.928	37.250	65	28	7
Skupno	267.944	100	4.411.723	100	2.946.222	1.134.154	331.347	67	26	8

Tudi v korpusu Janes-Tweet-Geo v0.3.4 se je ohranila podobna razporeditev tvitov po regijah kot pri zajetih tvitih z geolokacijo, na podlagi katerih smo pripisali metapodatke. Korpus Janes-Tweet-Geo v0.3.4 je bil avtomatsko označen tudi s stopnjo jezikovne nestandardnosti (Ljubešić et al. 2015). V povprečju je 67 % vsakega regionalnega podkorpusa označenega kot standardnega (L1), okrog 26 % delno nestandardnega (L2) in le okrog 7 % nestandardnega (L3). Pojavnic, ki so se pojavljale v nestandardnih tvitih (L3), je torej le okrog 330.000.

V večini tvitov, ki so vključeni v korpus Janes-Tweet-Geo v0.3.4, torej ne pričakujemo velike mere nestandardnih jezikovnih prvin, zato smo kot primarni vir gradiva za vzorčni korpus Janes-Geo vzeli tvite z najvišjo stopnjo jezikovne nestandardnosti (L3). V vzorčni korpus smo hoteli vključiti skupno 4.500 tvitov, tj. po 500 tvitov iz vsake od devetih regij (tvite iz tujine smo izključili). Iz korpusa Janes-Tweet-Geo v0.3.4 smo najprej izvozili vse tvite z oznako L3, nato pa glede na število uporabnikov v regiji določili maksimalno količino tvitov (med 25 in 50), ki jih je uporabnik lahko prispeval v vzorec. Na ta način smo poskrbeli, da je vzorec karseda uravnotežen, saj so razlike v produktivnosti med uporabniki lahko

zelo velike (nekateri so poslali le po 3 tvite, drugi pa tudi po 2000). Pri regijah z zelo majhno količino podatkov (Koroška, Rovtarska, Maribor in Panonska), v katerih po tovrstnem vzorčenju ni bilo mogoče zbrati 500 nestandardnih tvitov, smo vključili tudi tvite s stopnjo jezikovne standardnosti L2. Končno sestavo vzorčnega korpusa Janes-Geo prikazuje Tabela 6.

Tabela 6: Sestava vzorčnega korpusa Janes-Geo.

Regija	Število tvitov	Delež tvitov (%)	Število pojavnic	Delež pojavnic (%)
Gorenjska	500	12,58	8.502	13,28
Dolenjska	500	12,58	7.957	12,43
Štajerska	500	12,58	8.209	12,82
Panonska	500	12,58	7.222	11,28
Koroška	258	6,49	4.630	7,23
Primorska	500	12,58	8.560	13,37
Rovtarska	383	9,64	5.948	9,29
Ljubljana	500	12,58	7.702	12,03
Maribor	333	8,38	5.281	8,25
Skupno	3.974	100,00	64.011	100,00

Vzorčni korpus Janes-Geo vsebuje približno 4.000 tvitov oz. 64.000 pojavnic, kar predstavlja okrog 1,5 % celotnega korpusa Janes-Tweet-Geo v0.3.4 oziroma slabo petino (19 %) vseh nestandardnih tvitov (L3) v njem. Večina regij zajema po približno 12–13 % korpusa, izjeme pa so Koroška (7 %), Rovtarska (9 %) in Maribor (8 %). V vseh treh regijah je bilo uporabnikov premalo, da bi po kriterijih vzorčenja (tudi z vključitvijo tvitov L2) prispevali 500 tvitov.

3.4 Izdelava tipologije in smernic za označevanje nestandardnih jezikovnih prvin v tvitih

Na podlagi ročnega pregleda 200 tvitov iz vsake regije (skupno torej 1800 tvitov) smo zabeležili vse za raziskavo relevantne nestandardne jezikovne prvine. Te so večinoma zajemale nivoja zapisa in besedišča, v manjši meri pa tudi oblikoslovje in nekatere druge, bolj priložnostne in manj sistematične spremembe.[8] Vse zabeležene nestandardne prvine smo nato hierarhično kategorizirali ter izdelali tipologijo in smernice za označevanje nestandardnih jezikovnih prvin v tvitih (Čibej

8 V raziskavi nismo upoštevali rabe ločil, šumnikov in velike začetnice, saj se v računalniško posredovani komunikaciji pogosto opuščajo, kar je lahko tudi posledica naprave, s katere uporabnik pošilja tvit (npr. telefonski zapisi brez šumnikov). V prvotni različici tipologije smo predvidevali tudi skladenjski nivo, a smo v vzorčnih tvitih odkrili zanemarljivo malo nestandardnih skladenjskih pojavov, zato smo kategorijo odstranili.

2017).[9] Tipologija v trenutni različici (v1.0) vsebuje 6 glavnih kategorij (izpuste, transformacije, nestandardno oblikoslovje, nestandardno besedje, nestandardne različice pogostih besed z variantnimi oblikami, drugo) in skupno 292 različnih oznak, podrobnejši prerez s številom oznak na kategorijo pa prikazuje Tabela 7. Več kot polovico (58 %) različnih oznak zavzemajo izpusti, skoraj petino pa transformacije.

Tabela 7: Število oznak na kategorijo v tipologiji nestandardnih jezikovnih prvin v slovenskih tvitih.

Kategorija	Število različnih oznak	Delež (%)
Izpusti	170	58,22
Transformacije	58	19,86
Nestandardno oblikoslovje	9	3,08
Nestandardno besedje	13	4,45
Nestandardne različice pogostih besed z variantnimi oblikami	27	9,25
Drugo	15	5,14
Skupno	292	100,00

3.4.1 Pregled glavnih kategorij tipologije

V tem razdelku na kratko predstavimo glavne kategorije tipologije nestandardnih jezikovnih prvin in njihove oznake.

Kategorija, v katero spada največ nestandardnih jezikovnih prvin v tvitih, so **izpusti**, ki jih v kontekstu te raziskave definiramo kot izpuščanje grafemov pri zapisu v primerjavi z neposredno standardno različico besede, delimo pa jih na dve podkategoriji, in sicer na izpuste soglasnikov (*glej* → *lej*) ter izpuste samoglasnikov (*sovražim* → *sovražm*). Oznake pri izpustih lahko vsebujejo naslednje podatke:

a) ali gre za izpust soglasnika (*Ik*) ali samoglasnika (*Iv*),

b) ali gre za izpust končnega (*Ikk*, *Ivk*) ali nekončnega grafema (*Ikn*, *Ivn*),

c) katera je besedna vrsta besede, v kateri je prišlo do izpusta (*G* – glagol, *S* – samostalnik, *P* – pridevnik, *R* – prislov in *D* – drugo),

d) oblikoskladenjske značilnosti besede, v kateri je prišlo do izpusta (npr. število, spol, sklon),

e) kateri grafem je bil izpuščen.

9 Označevalne smernice in tipologija so prosto dostopne na uradni spletni strani projekta JANES: http://nl.ijs.si/janes/viri/.

Oznaka *IvnSmei.e* npr. označuje izpust (*I-*) nekončnega samoglasnika *e* (*-vn-* in *-.e*) v samostalniku (*-S-*) moškega spola (*-m-*) v ednini (*-e-*) in imenovalniški obliki (*-i-*), npr. *teden* → *tedn*, *konec* → *konc*.

Pri kategoriji **transformacij** smo zabeležili vse grafeme, ki jih je uporabnik v besedi zapisal drugače, kot bi to zahtevala standardnoslovenska različica besede (*všeč* → *ušeč*, *mislila* → *mislala*). Oznake transformacij vsebujejo podatke o izvirnem grafemu (ali sklopu grafemov) in o ciljnem grafemu (ali sklopu grafemov). Oznaka *Ta.o* npr. označuje transformacijo grafema *-a* v grafem *-o*, npr. *prav* → *prov*.

Jezikovne prvine, ki smo jih uvrstili v kategorijo **nestandardnega oblikoslovja**, so se v vzorčnem korpusu pojavljale zelo sporadično in mnogo redkeje v primerjavi z ostalimi prvinami, a smo za razliko od skladnje kategorijo v tipologiji ohranili, saj so bili pojavi v njej bolj sistematični. V kategorijo smo uvrstili nestandardna obrazila pri glagolih (*greva* → *grema*, *bova* → *boma*, *morava* → *morve*) in samostalnikih (*Bučku* → *Bučkotu*, *s penziči* → *s penzičmi*). Pri ostalih pregibnih besednih vrstah oblikoslovnih posebnosti nismo zaznali.

V kategorijo **nestandardnega besedja** smo dodali vse besede, ki smo jih v kontekstu dojemali kot nestandardne,[10] pri čemer smo se opirali predvsem na kvalifikatorje (npr. *pogovorno*, *narečno*) v referenčnih virih, kot so Slovar slovenskega knjižnega jezika, Slovar novejšega besedja, Sprotni slovar slovenskega jezika, Slovenski pravopis in Slovenski etimološki slovar).[11] Omeniti je treba, da je bilo ocenjevanje nestandardnosti pri pojavnicah, ki niso bile opisane v referenčnih virih, do določene mere nujno subjektivno – to še zlasti velja za besede, privzete iz tujih jezikov. Kategorijo nestandardnega besedja smo razdelili na devet podkategorij: pogovorne/slengovske/narečne/žargonske besede (*NSB.Pog*, npr. *gujdek*, *kafič*, *spizditi*), germanizmi (*NSB.Ger*, npr. *cajt*, *zihr*), anglizmi[12] (*NSB.Ang*, npr. *kjut*, *appov*), kroatizmi/srbizmi (*NSB.Srb*, npr. *svašta*, *rukohvatskim*), italianizmi (*NSB.Ita*, npr. *mona*, *birca*), hispanizmi (*NSB.Špa*, npr. *el clasico*), francizmi (*NSB.Fra*, npr. *passé*), besede iz spletnega jezika (*NSB.Net*, npr. *jbg – jebiga*, *s5 – spet*) ter priložnostne/ustvarjalne tvorjenke (*NSB.Kre*, npr. *butljezik*, *paradajzkomunajzar*).

V kategorijo **različic pogostih besed** smo uvrstili omejen nabor besed, ki so se med označevanjem v besedilih pogosto pojavljale v več različicah zapisa, npr. osebni zaimek *jaz*, ki se je v vzorčnem korpusu pojavljal v oblikah *jst*, *js*, *jz*, *jest*, *ist* in *ject*, ali prislov *zdaj*, ki se je pojavljal kot *zdj*, *zj*, *zdej*, *zej* in *zaj*. Označili smo samo

10 Pri tej kategoriji nismo upoštevali besed, ki imajo neposredne standardne ustreznice, njihova nestandardnost pa se izkazuje na način, ki spada pod druge kategorije te tipologije – to so npr. besede z izpuščenimi samoglasniki (*tudi* → *tud*) ali besede, v katerih je prišlo do transformacij samoglasnikov (*utrgalo* → *ftrgalo*) ali soglasnikov (*bog* → *boh*).

11 http://fran.si/

12 Pri anglizmih smo označevali tudi stopnjo podomačitve, za kar smo uporabili oznake *NSB.Ang.N* (povsem nepodomačeno, npr. *shopping*), *NSB.Ang.P.F* (podomačitev, razvidna iz fonetiziranega zapisa, npr. *imidž*), *NSB.Ang.P.K* (podomačitev, razvidna iz končnice, npr. dumplingi) in *NSB.Ang.P.O* (podomačitev, razvidna tako iz fonetiziranega zapisa kot iz končnice).

nestandardne oblike. V končni različici označenega korpusa nam te oznake omogočajo, da opazujemo, ali so uporabniki pri rabi različic dosledni oz. ali izbirajo med več različicami, obenem pa lahko preverimo, ali so nekatere različice značilnejše za uporabnike iz določene regije. Oznake iz te kategorije vsebujejo samo normalizirano obliko označene besede, npr. *V.zakaj* za besede *zakej*, *zaka*, *zakva* ipd.

V kategorijo **drugo** smo vključili vse ostale nestandardne jezikovne prvine, ki jih nismo mogli uvrstiti v nobeno od prej naštetih kategorij, npr. pisanje skupaj (*Dskupaj*, npr. *ne bi* → *nebi*), sklapljanje besed (*Dsklop*, npr. *to je* → *toj*), vrivanje dodatnega grafema (*Dvrivanje*, npr. *zajtrk* → *zajterk*), raba nestandardnih besed, ki delujejo kot vezniki (*Dk*, *Dki*, *Dka* za *k*, *ki*, *ka*) in nestandardna raba kategorije živosti *(Dživost,* npr. *[matram] iphone* → *[matram] iphona*).

Poudariti je treba, da tipologija sicer vsebuje oznake za vse proučevane nestandardne prvine, ki smo jih zaznali v korpusu, a je kljub temu ne moremo obravnavati kot izčrpno, saj je vzorčni korpus, na podlagi katerega je bila izdelana, relativno majhen. S precejšnjo zanesljivostjo pa lahko trdimo, da vsebuje primere vseh najpogostejših kategorij nestandardnih jezikovnih pojavov, ki se pojavljajo v slovenskih tvitih, obenem pa je zastavljena tako, da jo je mogoče nadgraditi z novimi oznakami, zaradi česar je primerna tudi za morebitno prihodnje označevanje večjih vzorcev.

3.5 Označevanje korpusa Janes-Geo

V tvitih smo označevali samo pojavnice, ki smo jih po vsaj enem od kriterijev v smernicah dojemali kot nestandardne. Vsaki nestandardni pojavnici smo pripisali oznake za vse prisotne nestandardne jezikovne prvine, razen če ni bilo v smernicah določeno drugače. Pojavnici, v kateri se je npr. poleg dveh izpustov pojavila še ena transformacija, smo pripisali tri ustrezne oznake iz tipologije.

```
@user     jah, men so ble ušeč, zato sm jih tut kupu. pojamram
pa zato k niso ble lih zastonj, pa sm pač probu mal informacijo
razširt :/

@user       [jah]{NSB.Pog},   [men]{IvkD.i}   so   [ble]{IvnGd.i}
[ušeč]{Tv.u},   zato   [sm]{IvnGs1e.e}   jih   [tut]{IvkD.i}{Td.t}
[kupu]{Tl.u}{IvnGd.i}. [pojamram]{NSB.Ger} pa zato [k]{Dk} niso
[ble]{IvnGd.i} [lih]{NSB.Ger} zastonj, pa [sm]{IvnGs1e.e} pač
[probu]{NSB.Ger}{Tl.u}{IvnGd.a}     [mal]{IvkR.o}     informacijo
[razširt]{IvnGn.i}{IvkGn.i} :/
```

Slika 6: Primer tvita brez oznak in z oznakami za nestandardne jezikovne prvine.

Slika 6 prikazuje primer tvita, v katerem so bile označene vse nestandardne jezikovne prvine v skladu s smernicami. Vsaki nestandardni pojavnici je pripisana ena ali več ustreznih oznak. Pojavnica *lih* je npr. označena kot germanizem (*NSB.Ger*), pojavnica *probu* pa ima tri oznake (*NSB.Ger* za germanizem, *Tl.u* za transformacijo *-l* v *-u* (*probal* → *probau*) in *IvnGd.a* za izpust nekončnega samoglasnika *-a* v preteklem deležniku glagola (*probau* → *probu*).

4 ANALIZA OZNAČENEGA KORPUSA JANES-GEO

Iz končne različice ročno označenega korpusa Janes-Geo smo s pomočjo regularnih izrazov izvozili oznake in besede, ki so bile z njimi označene, in sicer na več nivojih: po regionalnih podkorpusih, po posameznih uporabnikih in za vsak posamezen tvit. V tem prispevku se osredotočamo samo na analizo razlik med posameznimi regionalnimi podkorpusi in v manjši meri na razlike med posameznimi uporabniki.

Tabela 8 prikazuje število besed,[13] ki so bile označene kot nestandardne, ter število oznak v korpusu Janes-Geo. Tretji stolpec podaja delež nestandardnih besed v primerjavi z vsemi besedami, zadnji stolpec pa delež oznak glede na število vseh oznak v korpusu.

Tabela 8: Števila in deleži nestandardnih besed ter oznak v korpusu Janes-Geo.

Regija	Število besed	Število nestandardnih besed	Delež (%)	Število oznak	Delež (%)
Gorenjska	7.834	1.836	23,44	2.552	20,56
Dolenjska	7.447	1.610	21,62	2.195	17,69
Štajerska	7.516	1.273	16,94	1.659	13,37
Panonska	6.539	436	6,67	499	4,02
Koroška	4.232	454	10,73	569	4,59
Primorska	7.823	1.325	16,94	1.758	14,17
Rovtarska	5.547	868	15,65	1.169	9,42
Ljubljana	6.930	1.148	16,57	1.504	12,12
Maribor	4.820	428	8,88	505	4,07
Skupno	58.688	9.378	15,98	12.410	100,00

Od približno 59.000 besed v celotnem korpusu jih je bilo okrog 9.300 označenih kot nestandardnih, kar predstavlja približno 16 % vseh besed. Že na tem nivoju lahko identificiramo razlike med regijami: z nekoliko višjo nestandardnostjo

13 Kot besede smo obravnavali vse pojavnice, ki niso ločila, številke, emotikoni, URL-naslovi, sklici na uporabniška imena (@avtor) ali ključniki (#kljucnik).

(21–23 %) izstopata gorenjska in dolenjska regija, manj nestandardne (7–10 %) pa so manjše regije, tj. panonska, mariborska in koroška. To je pričakovano, saj smo zaradi pomanjkanja podatkov vanje vključili tudi tvite z nižjo stopnjo jezikovne nestandardnosti (L2).

Tudi po številu in deležu oznak najbolj izstopata gorenjska in dolenjska regija, zanimivo pa je, da gorenjska v primerjavi z dolenjsko vsebuje nekoliko večji delež oznak (20,5 % in 17,5 %), čeprav je delež nestandardnih besed pri obeh regijah primerljiv. Ker se na eni besedi najpogosteje verižijo oznake za izpuste in transformacije, razlika v deležih nakazuje, da lahko v gorenjski regiji pričakujemo več izpustov in transformacij.

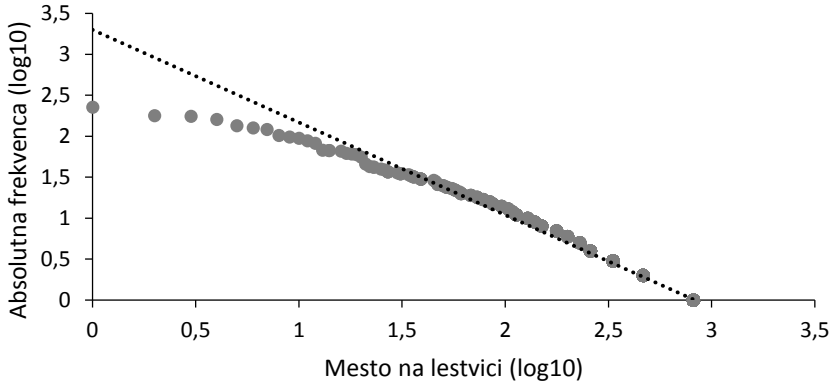
Tabela 9 prikazuje prvih 60 najpogostejših označenih besed iz korpusa Janes-Geo, ki imajo absolutno frekvenco večjo od 20. Da bi zmanjšali razpršenost podatkov, smo besede pri izvozu pretvorili v zapis z malimi začetnicami in brez šumnikov. Absolutna frekvenca za besedo *drgac* npr. zajema tudi različice *Drgac*, *Drgač*, *drgač* ipd. Relativna frekvenca je izračunana glede na število vseh besed v korpusu in normalizirana na 1.000 besed. Omeniti je treba tudi, da izvoz najpogostejših besed ni upošteval oznak, zato so nekatere oblike besed prekrivne (npr. *ka* kot različica vprašalnega zaimka *kaj* ali kot veznik; *ma* kot členek ali kot sedanjiška oblika glagola *imeti* v tretji osebi; *dobr* kot pridevnik moškega spola *dober*, kot pridevnik srednjega spola *dobro* ali kot prislov *dobro*).

Tabela 9: Prvih 60 najpogostejših nestandardnih besed iz korpusa Janes-Geo.

Mesto	Nestandardna beseda	Absolutna frekvenca	Relativna frekvenca	Mesto	Nestandardna beseda	Absolutna frekvenca	Relativna frekvenca
1	sm	227	3,87	31	cist	35	0,60
2	tud	179	3,05	32	fajn	35	0,60
3	a	176	3,00	33	itak	35	0,60
4	blo	161	2,74	34	tolk	34	0,58
5	sam	135	2,30	35	u	33	0,56
6	kr	127	2,16	36	lol	32	0,55
7	sej	121	2,06	37	lahk	32	0,55
8	jst	103	1,76	38	morm	32	0,55
9	k	98	1,67	39	dost	30	0,51
10	ma	95	1,62	40	tist	30	0,51
11	mal	88	1,50	41	skor	30	0,51
12	pol	82	1,40	42	jz	30	0,51
13	tko	68	1,16	43	evo	30	0,51
14	al	67	1,14	44	nism	30	0,51
15	dobr	67	1,14	45	zihr	29	0,49
16	mam	66	1,12	46	nevem	28	0,48
17	nc	62	1,06	47	bli	26	0,44

Mesto	Nestandardna beseda	Absolutna frekvenca	Relativna frekvenca	Mesto	Nestandardna beseda	Absolutna frekvenca	Relativna frekvenca
18	zdej	61	1,04	48	omg	26	0,44
19	js	60	1,02	49	bit	26	0,44
20	kej	56	0,95	50	tak	25	0,43
21	zdj	46	0,78	51	drgac	25	0,43
22	pr	43	0,73	52	zarad	24	0,41
23	ful	42	0,72	53	dej	24	0,41
24	bla	42	0,72	54	kul	24	0,41
25	neki	40	0,68	55	kva	23	0,39
26	tut	39	0,66	56	tok	23	0,39
27	ka	37	0,63	57	vidla	22	0,37
28	mas	37	0,63	58	okol	22	0,37
29	dons	37	0,63	59	wtf	22	0,37
30	men	36	0,61	60	vseen	21	0,36

Kot je razvidno iz Tabele 9, so tudi najpogostejše nestandardne besede v korpusu relativno redke, saj se pojavljajo v povprečju manj kot enkrat na 1.000 besed (povprečna relativna frekvenca prvih 60 najpogostejših besed je 0,95). Med najpogostejšimi nestandardnimi besedami najdemo nekatere pogoste glagole večinoma v sedanjiški obliki (*sm*, *nism*, *mam*, *mas*, *morm*) oz. kot pretekle deležnike ali nedoločnike (*bla*, *bli*, *vidla*, *bit*), osebne (*jst*, *js*, *jz*, *men*) in vprašalne zaimke (*ka*, *kej*, *kva*). Pogosti so tudi členki (*tud*, vprašalni členek *a*, *ma*, *evo*), vezniki (*sam*, *sej*, *k*, *ka*) ter prislovi (*zdej*, *zdj*, *zj*, *pol*, *dons*). Nestandardnost slovenskih uporabnikov Twitterja se torej pogosto in najbolj sistematično izkazuje v omejenem naboru zaprtih besednih vrst, pri odprtih besednih vrstah, kot so samostalniki, pridevniki in glagoli, pa je raba



Slika 7: Frekvenčna razporeditev nestandardnih besed v korpusu Janes-Geo.

nestandardnih oblik bolj sporadična in razpršena. Samostalniški besedi z najvišjo relativno frekvenco sta npr. *dnarja* (0,15; 149. mesto) in *cajt* (0,14; 151. mesto).

Slika 7 prikazuje razporeditev nestandardnih besed v korpusu Janes-Geo glede na njihovo absolutno frekvenco in mesto na lestvici. Razporeditev je razmeroma linearna in se v večjem delu dobro sklada z idealno Zipfovo distribucijo (koeficient trendne premice je -1,14), iz česar lahko sklepamo, da korpus Janes-Geo kljub majhnemu številu pojavnic relativno dobro predstavlja raznolikost nestandardnega besedišča uporabnikov Twitterja. Razporeditev od idealne Zipfove distribucije odstopa le pri približno 10 najpogostejših besedah, pri katerih je frekvenca nekoliko manjša od pričakovane, in pri besedah na zadnjem mestu lestvice, ki se v korpusu pojavijo zgolj enkrat. Takšnih besed je 2.686, kar znaša slabih 29 % vseh nestandardnih besed.

4.1 Pregled glavnih kategorij nestandardnih jezikovnih prvin

V tem razdelku predstavljamo analizo označenega korpusa z vidika vseh šestih glavnih kategorij oznak iz tipologije. Tabela 10 prikazuje število in delež vseh oznak znotraj različnih kategorij glede na regijo.

Tabela 10: Oznake v korpusu Janes-Geo po glavnih kategorijah in regijah.

Regija	Izpusti	Transformacije	Različice pogostih besed	Nestandardno besedje	Nestandardno oblikoslovje	Drugo
Gorenjska	1.321	348	302	435	5	141
	51,76 %	13,64 %	11, 83%	17,05 %	0,20 %	5,53 %
Dolenjska	1.152	300	214	415	8	106
	52,48 %	13,67 %	9,75 %	18,91 %	0,36 %	4,83 %
Štajerska	786	212	174	427	3	57
	47,38 %	12,78 %	10,49 %	25,74 %	0,18 %	3,44 %
Panonska	151	51	32	224	3	38
	30,26 %	10,22 %	6,41 %	44,89 %	0,60 %	7,62 %
Koroška	230	65	69	167	4	34
	40,42 %	11,42 %	12,13 %	29,35 %	0,70 %	5,98 %
Primorska	741	288	221	419	9	80
	42,15 %	16,38 %	12,57 %	23,83 %	0,51 %	4,55 %
Rovtarska	566	191	120	228	0	64
	48,42 %	16,34 %	10,27 %	19,50 %	0,00%	5,47 %
Ljubljana	695	198	128	394	8	81
	46,21 %	13,16 %	8,51 %	26,20 %	0,53 %	5,39 %
Maribor	217	66	29	178	7	8
	42,97 %	13,07 %	5,74 %	35,25 %	1,39 %	1,58 %
Skupno	5.859	1.719	1.289	2.887	47	609
	47,21 %	13,85 %	10,39 %	23,26 %	0,38 %	4,91 %

Najpogostejša kategorija nestandardnih jezikovnih prvin v korpusu so izpusti, ki zajemajo skoraj polovico oz. 47 % vseh oznak. Izpustom sledijo kategorije nestandardnega besedja (23 %), transformacij (14 %), različic pogostih besed (10 %) in drugo (5 %). Najmanj oznak je v kategoriji nestandardnega oblikoslovja (le 0,38 %).

4.1.1 Izpusti

Kot prikazuje Tabela 11, v korpusu močno prevladujejo izpusti samoglasnikov, ki v posamezni regiji zajemajo do 95 % vseh izpustov (oz. v celotnem korpusu približno 92 % vseh izpustov). Izpusti soglasnikov se pojavljajo mnogo redkeje (od 6 do 8 %). Z nekoliko večjim deležem izpustov soglasnikov izstopajo Panonska, Koroška in Maribor, a je to po vsej verjetnosti predvsem posledica vzorčenja.

Tabela 11: Izpusti samoglasnikov in soglasnikov v korpusu Janes-Geo.

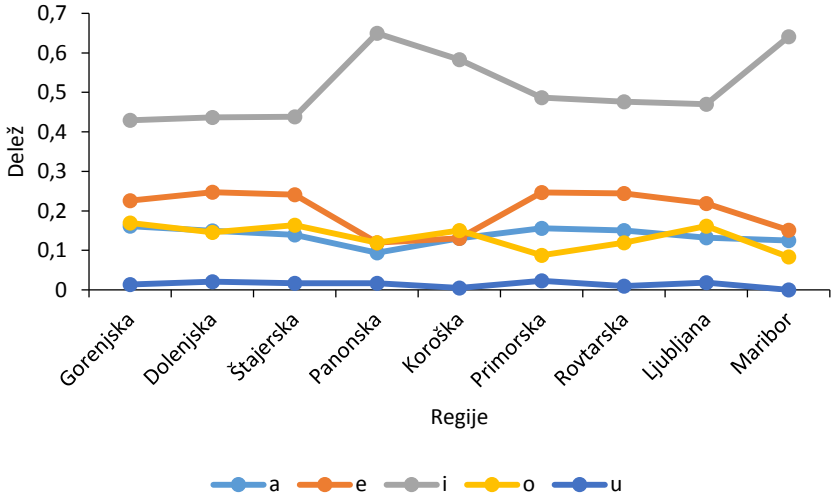
Regija	Izpust soglasnika	Delež (%)	Izpust samoglasnika	Delež (%)
Gorenjska	105	7,95	1216	92,05
Dolenjska	76	6,60	1076	93,40
Štajerska	61	7,76	725	92,24
Panonska	34	22,52	117	77,48
Koroška	31	13,48	199	86,52
Primorska	43	5,80	698	94,20
Rovtarska	54	9,54	512	90,46
Ljubljana	50	7,19	645	92,81
Maribor	25	11,52	192	88,48
Skupno	479	8,18	5380	91,82

Iz Tabele 12 lahko razberemo, da so izpusti nekončnih samoglasnikov v korpusu nekoliko pogostejši (okrog 60 %) od izpustov končnih samoglasnikov (okrog 40 %). Pri porazdelitvi najbolj izstopa Koroška (56 % izpustov končnih samoglasnikov in le 43 % nekončnih), a glede na zelo majhno število uporabnikov številke po vsej verjetnosti niso reprezentativne.

Kot prikazuje Slika 8, se v korpusu Janes-Geo najpogosteje izpušča samoglasnik *-i*, najredkeje pa samoglasnik *-u*. Pri nekaterih regijah so opazne razlike v deležih – Panonska, Koroška in Maribor v primerjavi z drugimi samoglasniki mnogo pogosteje izpuščajo *-i* (okrog 60 % vseh izpustov samoglasnikov), precej redkeje pa samoglasnik *-e* (okrog 10 % v primerjavi s približno 25 % na Gorenjskem, Dolenjskem in Štajerskem). Na Primorskem, Rovtarskem in v Mariboru je opazno tudi redkejše izpuščanje samoglasnika *-o* (manj kot 10 %).

Tabela 12: Izpusti končnih in nekončnih samoglasnikov v korpusu Janes-Geo.

Regija	Izpusti končnih samoglasnikov	Delež (%)	Izpusti nekončnih samoglasnikov	Delež (%)
Gorenjska	485	39,88	731	60,12
Dolenjska	411	38,20	665	61,80
Štajerska	312	43,03	413	56,97
Panonska	45	38,46	72	61,54
Koroška	113	56,78	86	43,22
Primorska	213	30,52	485	69,48
Rovtarska	192	37,50	320	62,50
Ljubljana	269	41,71	376	58,29
Maribor	94	48,96	98	51,04
Skupno	2134	39,67	3246	60,33



Slika 8: Izpusti samoglasnikov v korpusu Janes-Geo.

4.1.2 Transformacije

Tabela 13 prikazuje razporeditev pojavitev 10 najpogostejših transformacij v korpusu Janes-Geo po regijah.

Tabela 13: Najpogostejše transformacije v korpusu Janes-Geo po regijah.

Regija	Tl.u	Taj.ej	Tv.u	Ta.o	Tj.i	Tpri.pr	Tz.s	To.u	Td.t	Tl.o
Gorenjska	79	62	26	58	24	16	11	6	12	0
Dolenjska	89	42	28	22	24	26	7	18	9	0
Štajerska	46	35	9	4	14	14	15	5	4	12
Panonska	3	2	3	3	0	1	1	3	0	7
Koroška	0	30	2	1	3	2	6	2	3	2
Primorska	80	64	22	5	18	7	11	17	6	1
Rovtarska	24	36	26	12	13	6	17	9	1	0
Ljubljana	51	29	10	14	17	11	10	13	8	1
Maribor	7	6	0	2	5	2	3	2	3	7
Skupno	379	306	126	121	118	85	81	75	46	30

Najpogostejša je transformacija *-l* v *-u*, do katere največkrat pride pri moški obliki preteklega deležnika glagola (*naredil* → *naredu*, *spomnil* → *spomnu*). Najredkejša je v panonski, mariborski in koroški regiji, zanimivo pa je, da se sorodna transformacija *-l* v *-o* (*naredil* → *naredo*, *spomnil* → *spomno*) najpogosteje pojavlja prav v teh regijah (in na Štajerskem).

Sledita transformaciji *-aj* v *-ej* (*daj* → *dej*, *kaj* → *kej*, *zdaj* → *zdej*), ki jo prav tako redko zasledimo v panonski in mariborski regiji (a je pogostejša v koroški), in *-v* v *-u* (*všeč* → *ušeč*, *v* → *u*), ki je najpogostejša v gorenjski, dolenjski, primorski in rovtarski regiji. Precejšnje razlike se kažejo tudi pri transformaciji *-a* v *-o* (*prav* → *prov*, *gledal* → *gledov*), ki je zelo pogosta na Gorenjskem in (v nekoliko manjši meri) na Dolenjskem, v drugih regijah pa mnogo redkejša. Nasprotno je na Dolenjskem in Primorskem nekoliko pogostejša transformacija *-o* v *-u* (*blo* → *blu*, *okno* → *oknu*).

4.1.3 Nestandardno besedje

Besed, ki so bile uvrščene v kategorijo nestandardnega besedja (NSB), je bilo v celotnem korpusu 2.887, kar predstavlja slabo tretjino (31 %) vseh označenih besed. Tabela 14 prikazuje števila in deleže NSB po regijah v primerjavi z vsemi označenimi besedami in vsemi besedami.

Zanimiv podatek je, da je pri vseh regijah delež nestandardnega besedja v primerjavi z vsemi besedami primerljiv (giblje se med približno 3 in 6 %), nasprotno pa so razlike v deležih nestandardnega besedja glede na vse označene besede precej višje: nestandardno besedje ima najmanjši delež v gorenjski regiji (24 %), najvišjega pa v panonski (51 %). Na tem mestu je treba znova upoštevati, da so bili pri regijah z manj podatki (Panonska, Koroška, Maribor in Rovtarska) uporabljeni tudi tviti s stopnjo jezikovne nestandardnosti L2, zaradi česar je lahko višji delež nestandardnega besedja tudi posledica manjše količine izpustov, ki jih avtomatska

klasifikacija nestandardnosti lažje zazna in jih je zato v tvitih z nižjo stopnjo jezikovne nestandardnosti manj. Kljub temu vseh razlik ne moremo pripisati le vzorčenju: podobnost med Panonsko in Mariborom, ki sta tudi po geografski legi blizu, po vsej verjetnosti ni naključna. To še dodatno podpirajo manj izrazite razlike med Koroško in Rovtarsko (ki vsebujeta tudi tvite L2) na eni ter regijami z več podatki (ki vsebujejo samo tvite L3) na drugi strani. V mariborski in panonski regiji se torej na prvi pogled kaže tendenca, da se nestandardnost v jeziku pogosteje izraža z besediščem kot z izpusti. Razlika med Gorenjsko (24 % NSB) in Panonsko (51 % NSB) je tudi statistično veljavna (χ^2 (2, N = 2.272) = 131,12; p < 0,01). Podobno razmerje najdemo tudi med npr. Dolenjsko (25 % NSB) in Mariborom (41 % NSB) – χ^2 (2, N = 2.038) = 40,98; p < 0,01.

Tabela 14: Števila in deleži nestandardnega besedja po regijah v korpusu Janes-Geo.

Regija	NSB	Vse označene besede	NSB/Vse označene besede (%)	Vse besede	NSB/Vse besede (%)
Gorenjska	435	1.836	23,69	7.834	5,55
Dolenjska	415	1.610	25,78	7.447	5,57
Štajerska	427	1.273	33,54	7.516	5,68
Panonska	224	436	51,38	6.539	3,43
Koroška	167	454	36,78	4.232	3,95
Primorska	419	1.325	31,62	7.823	5,36
Rovtarska	228	868	26,27	5.547	4,11
Ljubljana	394	1.148	34,32	6.930	5,69
Maribor	178	428	41,59	4.820	3,69
Skupno	2.887	9.378	30,78	58.688	4,92

Tabela 15: Najpogostejše nestandardno besedje v korpusu Janes-Geo.

Pojavnica	f_A	Pojavnica	f_A
ma	60	btw	17
ful	42	cool	17
fajn	35	jao	17
itak	35	skos	17
lol	32	glih	16
evo	30	brezveze	15
zihr	29	jap	15
omg	26	un	12
kul	24	app	11
wtf	22	lih	11
kao	20	matr	11
folk	19	wow	11
jp	19	ziher	11
sorry	18		

Tabela 15 prikazuje vse besede NSB v korpusu, ki so se pojavljale z absolutno frekvenco, večjo od 10. Takšnih besed je skupno 28, kar znaša le 1,7 % od 1.745 različnih besed NSB v korpusu. Podatki torej izkazujejo visoko stopnjo razpršenosti, kar je posledica majhnega števila pojavnic v korpusu nasploh, v manjši meri pa tudi dejstva, da med analizo še nismo imeli na voljo normalizirane in lematizirane različice korpusa. Na podlagi preliminarnega pregleda seznama nestandardnega besedja smo sicer sklepali, da lematizacija razpršenosti ne bi znatno izboljšala, saj se le manjši del besed pojavlja v več oblikah. Približno 1.440 (83 %) vseh različnic NSB se v korpusu pojavi le enkrat, zato je korpus premajhen za zanesljivo medregionalno statistično primerjavo posameznih besed, omogoča pa primerjavo različnih podkategorij nestandardnega besedja, kot prikazuje Tabela 16.[14]

Tabela 16: Najpogostejše podkategorije nestandardnega besedja v korpusu Janes-Geo.

Regija	NSB.Pog	NSB.Ang	NSB.Ger	NSB.Srb	NSB.Ita	NSB.Net
Gorenjska	126	153	103	14	3	26
Dolenjska	137	153	80	18	3	23
Štajerska	133	151	96	24	8	11
Panonska	85	63	37	20	2	10
Koroška	69	37	33	9	1	16
Primorska	148	119	64	32	26	25
Rovtarska	73	94	37	7	0	13
Ljubljana	112	160	59	26	0	26
Maribor	57	68	27	8	0	14
Skupno	940	998	536	158	43	164

Najpogostejši kategoriji v korpusu sta *NSB.Pog*, ki vsebuje mdr. pogovorne, narečne in slengovske besede, in *NSB.Ang*, ki vsebuje anglizme z različnimi stopnjami podomačitve na ravni zapisa in oblikoslovja. Sledijo germanizmi (*NSB.Ger*), kroatizmi in srbizmi (*NSB.Srb*) ter italianizmi (*NSB.Ita*), ki pa jih je opazno manj, a je zanimivo, da so večino prispevali uporabniki iz primorske regije. Besede iz spletnega jezika (*NSB.Net*) so razporejene nekoliko bolj enakomerno.

4.1.4 Variantne različice pogostih besed

V korpusu so kot variantne označene različice naslednjih besed: *da*, *danes*, *domov*, *jaz*, *kaj*, *kako*, *koliko*, *kolikokrat*, *kolikor*, *kot*, *lahko*, *nekaj*, *potem*, *prav*, *saj*, *tako*,

14 Tabela 16 prikazuje samo podkategorije z največjimi frekvencami. Dodatne kategorije so še francizmi, hispanizmi in priložnostne tvorjenke, a jih je v korpus vključenih le peščica, zato jih tu ne navajamo.

takoj, *takole*, *toliko*, *tukaj*, *tule*, *včeraj*, *zakaj*, *zdaj*, *zdajle* in *zjutraj*. Seznam je bil med označevanjem sproti dopolnjevan z novimi različicami. Za primer si oglejmo regionalno razporeditev različic besede *koliko*, ki je prikazana v Tabeli 17.

Tabela 17: Različice besede koliko v korpusu Janes-Geo.

Regija	kok	kolk	kolko	kuk
Gorenjska	6	4	0	0
Dolenjska	7	3	0	0
Štajerska	0	0	0	0
Panonska	0	0	4	0
Koroška	0	0	0	0
Primorska	0	2	7	0
Rovtarska	2	1	1	2
Ljubljana	2	4	0	1
Maribor	0	0	3	0
Skupno	17	14	15	3

V korpusu smo zabeležili štiri različice: *kok*, *kolk*, *kolko* in *kuk*. Kljub majhnemu številu pojavitev že lahko opazimo nekatere vzorce. Različici *kok* in *kolk* se pojavljata na Gorenjskem in Dolenjskem ter v osrednjem delu Slovenije (Rovtarska in Ljubljana), različica *kolko* pa je skupna uporabnikom iz primorske, panonske in mariborske regije.

4.1.5 Nestandardno oblikoslovje

V korpusu je bilo v kategorijo nestandardnega oblikoslovja uvrščenih le nekaj nestandardnih prvin, skupaj le 47 (od 0 do največ 9 v vsaki regiji oz. pri skupno 36 uporabnikih, kar znaša dobrih 7 % vseh uporabnikov v korpusu). Tovrstni jezikovni pojavi so torej v našem vzorcu razmeroma redki – kar 125-krat redkejši od izpustov. Kar 37 primerov oz. 79 % vsega nestandardnega oblikoslovja smo zaznali pri glagolih, preostalih 10 primerov oz. 21 % pa pri samostalnikih.

Tabeli 18 in 19 prikazujeta vse zaznane nestandardne jezikoslovne posebnosti pri glagolih in samostalnikih ter podajata frekvenco, regije, v katerih so bile prvine prisotne, in ponazoritvene primere.

Tabela 18: Nestandardno oblikoslovje pri glagolih v korpusu Janes-Geo.

Nestandardno glagolsko obrazilo	f_A	Regije	Primer
Podaljšava kratkega nedoločnika na *-č* s *-t*	21	Gorenjska, Dolenjska, Ljubljana, Štajerska, Primorska	*reči* → *rečt, teči* → *tečt, obleči* → *obležt*
Obrazilo *-ma* v 1. osebi dvojine	8	Maribor, Koroška, Panonska	*greva* → *grema, zmeniva* → *zmenma, bova* → *boma*
Podaljšava s *-s* v 2. osebi množine	2	Primorska	*morate* → *moreste, imate* → *maste*
Obrazilo *-te* v 2. osebi množine	2	Primorska, Dolenjska	*boste* → *bote*
Obrazilo *-ta* v 2. osebi dvojine	2	Maribor	*bosta* → *bota*
Obrazilo *-ve* v 1. osebi dvojine	1	Dolenjska	*morava* → *morve*
Obrazilo *-ava* namesto *-uje* v 3. osebi ednine	1	Koroška	*opravičuje* → *opravičava*

Tabela 19: Nestandardno oblikoslovje pri samostalnikih v korpusu Janes-Geo.

Nestandardno samostalniško obrazilo	f_A	Regije	Primer
Podaljšava samostalnika moškega spola s *-t*	8	Primorska, Dolenjska, Panonska, Maribor	*deme* → *demote, psiha* → *psihota, nona* → *nonota*
Podaljšava samostalnika moškega spola z *-m*	2	Dolenjska, Primorska	*pri bogu milemu* → *pr bogmi milmi, s penziči* → *s penzičmi*

Zaradi nizkih frekvenc korpus Janes-Geo natančnejše medregionalne statistične primerjave na nivoju nestandardnega oblikoslovja ne dopušča, a ponuja zanimivo izhodišče za nadaljnje morfološke raziskave na večji količini podatkov.

4.1.6 Drugo

Pri kategoriji Drugo se v tem prispevku osredotočamo le na dve najpogostejši kategoriji, in sicer na pisanje skupaj (*Dskupaj*) ter sklapljanje besed (*Dsklop*).

Pisanje skupaj smo zabeležili pri 82 različnicah (oz. 175 pojavnicah), le šest pa se jih v korpusu pojavi več kot petkrat (oz. le 17 več kot enkrat). Najpogostejši zapisi skupaj in njihove absolutne frekvence so prikazani v Tabeli 19.

Tabela 20: Najpogostejši zapisi skupaj v korpusu Janes-Geo.

Zapis skupaj	f_A
nevem	28
pomoje	15
ane	11
ubistvu	11
nebi	10
vredu	6

Sklapljanje besed se je pojavljalo pri omejenem naboru besed (18 različnic in 33 pojavnic). Sklope, ki se v korpusu pojavijo več kot enkrat, prikazuje Tabela 21. Najpogosteje gre za kombinacijo naslonk. Zanimivo je, da so večino (20 pojavnic oz. 61 %) prispevali uporabniki z Gorenjske, preostanek pa uporabniki iz Dolenjske (4), Štajerske (5), Rovtarske (1) in Ljubljane (3).

Tabela 21: Sklapljanje besed v korpusu Janes-Geo.

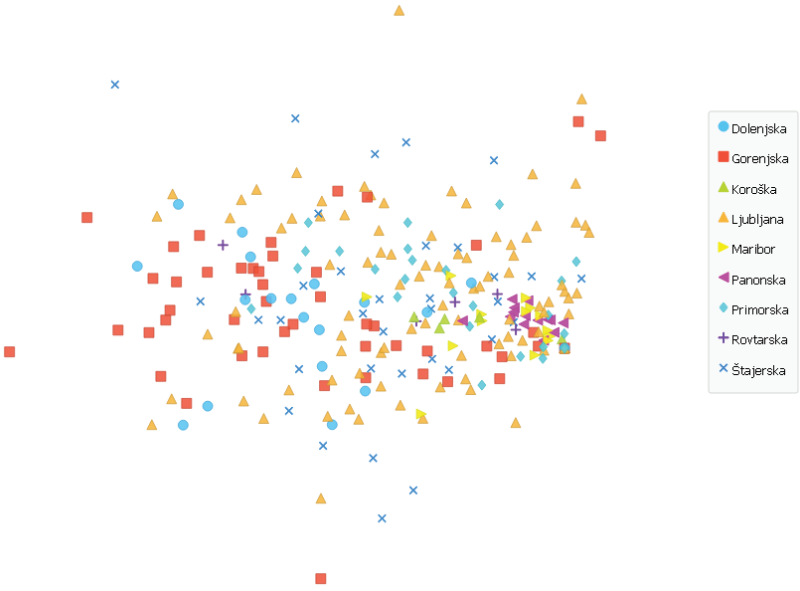
Sklop	Normalizirana oblika	f_A
nau	ne bo	6
sej	saj je	6
toj	to je	4
daj	da je	2
as	a si	2

4.2 Vizualizacija uporabnikov glede na uporabljene nestandardne jezikovne prvine

Da bi preverili, ali je uporabnike iz korpusa Janes-Geo mogoče razvrstiti v skupine tudi na podlagi njihove (ne)rabe nestandardnih jezikovnih prvin (in ne zgolj na podlagi geolokacije njihovih tvitov oz. pripisanih metapodatkov o regionalni pripadnosti), smo rabo nestandardnih jezikovnih prvin vsakega uporabnika vizualizirali. Za vsakega uporabnika smo izvozili relativne frekvence $(f)_r$[15] vseh nestandardnih jezikovnih prvin, ki jih je uporabil v tvitih, in jih nanizali v *n*-dimenzionalne vektorske reprezentacije $\vec{v}_u = (f_{r1}, f_{r2}, f_{r3},...,f_{rn})$, pri čemer je *n* število vseh kategorij in podkategorij nestandardnih jezikovnih prvin iz tipologije.

15 Relativne frekvence smo izračunali tako, da smo absolutno frekvenco določenega nestandardnega jezikovnega pojava iz tipologije (npr. IvkGn.i, izpust končnega samoglasnika *-i* v nedoločniku, *delati* → *delat*) delili s številom pojavnic, ki jih je uporabnik prispeval v korpus Janes-Geo.

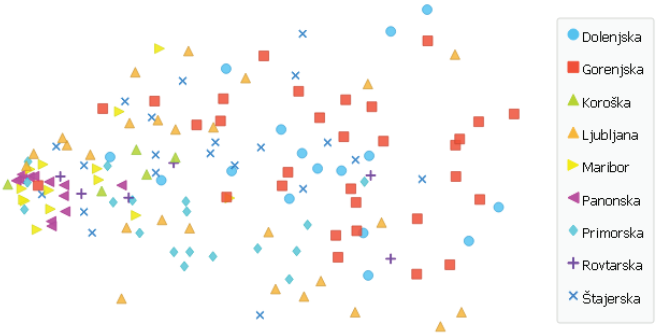
Vektorjem smo nato s pomočjo metode večdimenzionalnega skaliranja (Borg in Groenen 2005) zmanjšali dimenzionalnost ter jih s pomočjo paketa za podatkovno analizo Orange (Demšar et al. 2013) vizualizirali v dvodimenzionalni razsevni grafikon, ki v čim večji možni meri ohrani razdalje med izvirnimi vektorji. Slika 9 prikazuje vizualizacijo vseh uporabnikov v korpusu Janes-Geo glede na nestandardne jezikovne prvine v njihovih tvitih.



Slika 9: Vizualizacija uporabnikov v korpusu Janes-Geo glede na rabo nestandardnih jezikovnih prvin.

Čeprav je korpus relativno majhen in vsebuje zelo kratka besedila, so pri vizualizaciji uporabnikov glede na njihovo rabo nestandardnih jezikovnih prvin že opazne gruče. Največji vpliv na položaj uporabnikov v grafikonu ima količina izpustov, ki so najpogostejša in najbolj razširjena kategorija nestandardnih jezikovnih pojavov v korpusu. Najbolj razpršeni so ljubljanski uporabniki, ki jih je tudi največ, njihovo razpršenost pa gre pripisati več razlogom: nekateri uporabniki so v korpus prispevali manj gradiva, zaradi česar tudi relativne frekvence njihovih nestandardnih prvin ne predstavljajo ustrezno njihove dejanske jezikovne rabe. Obenem je mogoče sklepati, da se v ljubljansko regijo glede na njen centralni kulturni in gospodarski položaj uvrščajo uporabniki z zelo raznorodnimi jezikovnimi ozadji. Vizualizacija torej potrjuje ustreznost naše odločitve, da ljubljansko regijo obravnavamo posebej. Najbolj očitne so razlike med Gorenjsko in Dolenjsko na eni strani ter Panonsko in Mariborom na drugi, vmes pa prihaja tudi do nekoliko manj očitnega gručenja štajerskih, koroških in primorskih uporabnikov. Da bi

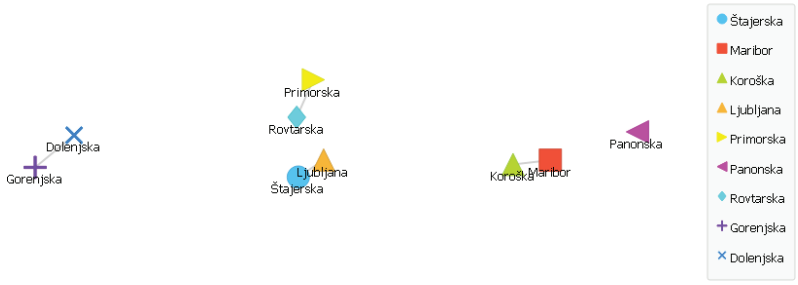
nekoliko zmanjšali šum, smo vizualizacijo ponovili samo z uporabniki, ki so v korpus Janes-Geo prispevali vsaj 100 pojavnic. To mejo smo določili na podlagi mediane števila pojavnic vseh uporabnikov (108 pojavnic). Rezultat vizualizacije prikazuje Slika 10.



Slika 10: Vizualizacija uporabnikov z več kot 100 pojavnicami v korpusu Janes-Geo.

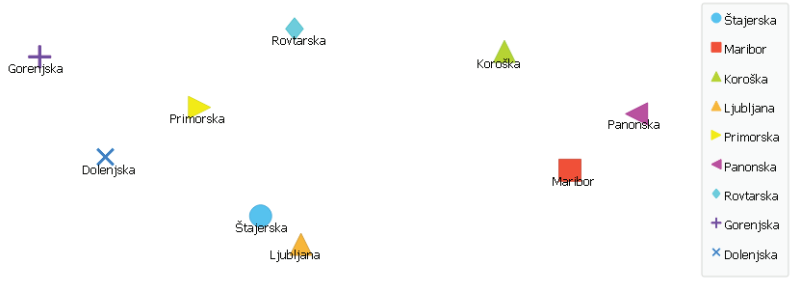
Z določitvijo minimalnega praga v pojavnicah smo nekoliko zmanjšali razpršenost ljubljanskih uporabnikov, ki so zdaj nekoliko bolj zgoščeni, a je zanimivo, da se ohranjata dve jasni veji (v spodnjem in zgornjem delu grafikona). Večjo zgoščenost lahko opazimo tudi pri uporabnikih iz primorske regije. Kot je pričakovano, bi še jasnejšo in natančnejšo sliko dobili z več podatki o jezikovni rabi posameznega uporabnika.

Preverili smo tudi, ali so razlike v jezikovni rabi podobne tudi med celotnimi regionalnimi podkorpusi, ne samo med posameznimi uporabniki. Vizualizacijo prikazuje Slika 11.



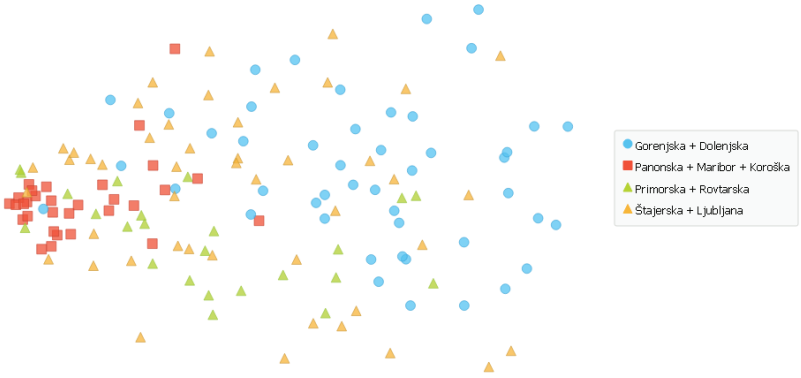
Slika 11: Vizualizacija regionalnih podkorpusov glede na vsebovane nestandardne jezikovne prvine.

Glede na podobnost se kaže jasna delitev na tri oz. štiri skupine, ki so skladne z ugotovitvami iz prejšnjih vizualizacij in primerjav. Ker je ta delitev v veliki meri odvisna od količine izpustov kot najpogostejše kategorije nestandardnih prvin (v gorenjski in dolenjski regiji je bilo izpustov največ, v mariborski in panonski pa najmanj), smo preverili podobnost regionalnih podkorpusov samo ob upoštevanju vseh drugih kategorij iz tipologije. Rezultat prikazuje Slika 12.



Slika 12: Vizualizacija regionalnih podkorpusov glede na vsebovane nestandardne jezikovne prvine (brez izpustov).

Kot vidimo, se razdalje in položaji tudi v primeru neupoštevanja izpustov v precejšnji meri ohranjajo. Nekoliko večje razlike je mogoče opaziti le med rovtarsko in primorsko ter med koroško in mariborsko regijo. Zanimivo je, da podobnosti med regijami v grobem sledijo tudi geografski razporeditvi. Na podlagi podobnosti med podkorpusi lahko regije torej združimo v štiri makroregije: Panonska + Maribor + Koroška, Štajerska + Ljubljana, Gorenjska + Dolenjska ter Primorska + Rovtarska. Slika 13 prikazuje vizualizacijo vseh uporabnikov, ki so v korpus prispevali več kot 100 pojavnic, ob upoštevanju novih, makroregionalnih metapodatkov.



Slika 13: Vizualizacija jezikovne rabe uporabnikov glede na makroregionalno delitev.

Najbolj razpršeni so še vedno uporabniki iz ljubljanske in štajerske regije, pri ostalih makroregijah pa se že nekoliko jasneje kažejo razmejitve. Glede na podobnosti med podkorpusi bi torej lahko posplošili, da so glede na kategorije, ki smo jih upoštevali pri označevanju, v korpusu Janes-Geo v grobem prisotne štiri jezikovne različice nestandardne spletne slovenščine, ki predstavljajo štiri makroregije: severovzhodno Slovenijo (Panonska + Maribor + Koroška), vzhodni del z Ljubljano (Štajerska + Ljubljana), osrednji del v smeri severozahod-jugovzhod (Gorenjska + Dolenjska) ter zahodni in jugozahodni del (Primorska + Rovtarska).

5 SKLEP

V poglavju smo predstavili poglavitne vidike ročno označenega korpusa Janes-Geo, ki omogoča korpusni pristop k proučevanju regionalnih jezikovnih različic v spletni slovenščini in je prosto dostopen pod licenco Creative Commons – Priznanje avtorstva (CC BY-SA 4.0) na repozitoriju CLARIN.SI (Čibej et al. 2018). Poleg postopka avtomatskega pripisovanja regionalnih metapodatkov na podlagi tvitov z geolokacijo smo predstavili gradnjo in označevanje korpusa Janes-Geo ter prvi vpogled v regionalne jezikovne različice v slovenski računalniško posredovani komunikaciji. Poleg samega korpusa kot pomemben rezultat raziskave velja omeniti tudi označevalne smernice in tipologijo nestandardnih jezikovnih prvin v spletni slovenščini, ki je bila sicer izdelana na podlagi tvitov, a lahko predpostavljamo, da je uporabna tudi za označevanje drugih besedilnih tipov v spletni komunikaciji (npr. forumska sporočila, blogovski zapisi in komentarji na novice).

Čeprav je korpus po številu pojavnic zelo majhen in ga zato ne moremo obravnavati kot povsem reprezentativnega, dobro prikazuje regionalno jezikovno variantnost v spletni slovenščini, kar potrjuje, da je uporabljena metoda uspešna. V prihodnje bi bilo raziskavo smiselno ponoviti na večjem vzorcu, ki vsebuje več zajetih uporabnikov in obenem ponuja čim večje število pojavnic na uporabnika, saj je le na ta način slika rabe nestandardnih jezikovnih prvin realna. Poleg tega zdajšnji vzorec za številne kategorije nestandardnih prvin zaradi pomanjkanja podatkov oz. prevelike razpršenosti ne omogoča učinkovite primerjave, vendar ponuja dobro izhodišče za nadaljnje raziskave tudi na drugih pisnih besedilnih tipih, v katerih lahko pričakujemo nestandardno slovenščino.

Kot zamisel za prihodnje delo velja izpostaviti tudi evalvacijo metode avtomatskega kodiranja metapodatkov o regionalni pripadnosti s pomočjo anketiranja uporabnikov, rabo nestandardnih jezikovnih prvin v spletni slovenščini pa bi bilo dobro raziskati tudi z bolj sociolingvističnega vidika, npr. v katerih situacijah uporabniki uporabljajo bolj regionalno obarvano jezikovno različico, v kolikšni

meri se jezikovno prilagajajo sogovorcem in v kolikšni meri na to vpliva tip komunikacije (javno, zasebno).

Zaznane medregionalne razlike bodo uporabljene tudi kot značilke za razvoj klasifikatorja, ki bo uporabnika uvrstil v ustrezno regijo. Klasifikator bo nato, če bo uspešen, omogočil gradnjo večjega korpusa, v katerem bo mogoča tudi statistična primerjava pojavov, ki so bili v korpusu Janes-Geo preredki.

Zahvala

Za tehnično podporo pri zasnovi raziskave se zahvaljujem Nikoli Ljubešiću in Tomažu Erjavcu.

Literatura

Arhar Holdt, Špela, 2018: Korpusni pristop k skladnji računalniško posredovane slovenščine. Fišer, Darja (ur.): *Viri, orodja in metode za analizo spletne slovenščine*. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani. 228–253.

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

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

Bitenc, Maja, 2016: *Z jezikom na poti med Idrijskim in Ljubljano*. Ljubljana: Znanstvena založba Filozofske fakultete.

Borg, Ingwer in Patrick Groenen, 2005: *Modern Multidimensional Scaling: theory and applications (2nd ed.)*. New York: Springer-Verlag. 207–212.

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

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

Čibej, Jaka in Nikola Ljubešić, 2015: "S kje pa si?" – Metapodatki o regionalni pripadnosti uporabnikov družbenega omrežja Twitter. Fišer, Darja (ur.): *Zbornik konference Slovenščina na spletu in v novih medijih*. Ljubljana: Znanstvena založba Filozofske fakultete. 10–14.

Čibej, Jaka, 2016: Framework for an Analysis of Slovene Regional Language Variants on Twitter. *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities*. Ljubljana: Znanstvena založba Filozofske fakultete. 17–21.

Čibej, Jaka, 2017: *Tipologija in smernice za označevanje nestandardnih jezikovnih prvin v slovenskih tvitih*. https://nl.ijs.si/janes/viri

Čibej, Jaka, Tomaž Erjavec in Darja Fišer, 2018: *Tweet corpus of Slovene regional language variants Janes-Geo v1.0*. http://hdl.handle.net/11356/1174

Čibej, Jaka, Špela Arhar Holdt, Tomaž Erjavec in Darja Fišer, 2018: Ročno označeni korpusi Janes za učenje jezikovnotehnoloških orodij in jezikoslovne raziskave. Fišer, Darja (ur.): *Viri, orodja in metode za analizo spletne slovenščine*. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani. 44–73.

Demšar, Janez, Tomaž Curk, Aleš Erjavec, Črt Gorup, Tomaž Hočevar, Mitar Milutinovič, Martin Možina, Matija Polajnar, Marko Toplak, Anže Starič, Miha Štajdohar, Lan Umek, Lan Žagar, Jure Žbontar, Marinka Žitnik in Blaž Zupan, 2013: Orange: Data Mining Toolbox in Python. *Journal of Machine Learning Research* 14/2013. 2349–2353. http://eprints.fri.uni-lj.si/2267/1/2013-Demsar-Orange-JMLR.pdf

Eisenstein, Jacob, 2015: Written dialect variation in online social media. Boberg, Charles, John Nerbonne in Dominic Watt (ur.): *Handbook of Dialectology*. New York: Wiley.

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

Fišer, Darja, Ljubešić, Nikola, Erjavec, Tomaž. 2015. The JANES corpus of Slovene user generated content: construction and annotation. *International Research Days: Social Media and CMC Corpora for the e-humanities: Book of Abstracts. 23.-24. October 2015*. Rennes, France. 11.

Goli, Teja, Eneja Osrajnik in Darja Fišer, 2016b: Analiza krajšanja slovenskih sporočil na družbenem omrežju Twitter. Erjavec, Tomaž in Darja Fišer (ur.): *Proceedings of the Conference on Language Technologies and Digital Humanities*. Ljubljana: Znanstvena založba Filozofske fakultete. 77–82.

Grieve, Jack, 2016: *Regional Variation in Written American English*. Cambridge: Cambridge University Press.

Haddow B., A. Hernandez-Huerta, F. Neubarth, H. Trost, 2013: Corpus Development for Machine Translation between Standard and Dialectal Varieties. Zbornik konference *Workshop 'Adaptation of Language Resources and Tools for Closely Related Languages and Language Variants' of the 9th Int. Conf. on Recent Advances in Natural Language Processing (RANLP 2013)*. Hissar, Bulgaria. 7–14.

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

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

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

Huang, Yuan, Diansheng Guo, Alice Kasakoff in Jack Grieve, 2016: Understanding US regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems 59*. 244–255.

Johanessen, Janne Bondi, Joel Priestley, Kristin Hagen, Tor Anders Åfarli in Øystein Alexander Vangsnes, 2009: The Nordic Dialect Corpus – an Advanced Research Tool. Jokinen, K. in E. Bick (ur.): *Zbornik konference 17th Nordic Conference of Computational Linguistics NODALIDA 2009. NEALT Proceedings Series Volume 4*.

Jørgensen, Anna Katrine, Dirk Hovy in Anders Søgaard, 2015: Challenges of studying and processing dialects in social media. *Proceedings of the ACL 2015 Workshop on Noisy User-generated Text*. Peking, Kitajska. 9–18.

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

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

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

Ljubešić, Nikola in Denis Kranjčić, 2014: Discriminating between VERY similar languages among Twitter users. *Zbornik konference Language technologies: 17th International Multiconference Information Society IS2014*. Ljubljana: Institut »Jožef Stefan«.

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

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

Logar, Tine, 1981: Govor kraja Podlešče na Banjšicah (Glasoslovna študija). *Goriški letnik* 8/1981. 275–283.

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

Popič, Damjan in Darja Fišer, 2018: (Ne)normativnost računalniško posredovane komunikacije v slovenščini: merilo vejice. Fišer, Darja (ur.): *Viri, orodja in metode za analizo spletne slovenščine*. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani. 140–159.

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

Reher, Špela in Darja Fišer, 2018: Kodno preklapljanje v objavah slovenskih uporabnikov Twitterja. Fišer, Darja (ur.): *Viri, orodja in metode za analizo spletne slovenščine*. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani. 294–323.

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

Sparks, Evan in Sanjay Krishnan, 2012: *TweetLocalize: Inferring Author Location in Social Media*. University of Berkeley. http://bid.berkeley.edu/cs294-1-spring13/images/c/c2/TweetLocalizeReport.pdf

Szmrecsanyi, Benedikt, 2011: Corpus-based dialectometry: a methodological sketch. *Corpora* 6/1. 45–76.

Škofic, Jožica in drugi, 2011: *Slovenski lingvistični atlas 1: Človek (telo, bolezni, družina)*. Ljubljana: Založba ZRC.

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

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

Valicon, 2016: *Raziskava MEDIA+. Maj/junij 2016*. http://www.valicon.net/files/Sporocilo%20za%20javnost%202016-06-23%20(1).pdf

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

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

Logar, Tine in Jakob Rigler, 1986: Karta slovenskih narečij. Avtotehna Zveza organizacij za tehnično kulturo Slovenije. https://www.dlib.si/?URN=URN:NBN:SI:IMG-VSVHWWS9

Tviti kot leksikografski vir za analizo pomenskih premikov v slovenščini

Darja Fišer, Nikola Ljubešić

Izvleček

V poglavju predstavimo potencial družbenega omrežja Twitter za spremljanje leksikalnih novosti s poudarkom na proučevanju sprememb v rabi že ustaljenega besedišča. Pristop temelji na primerjavi semantičnih profilov besed, zgrajenih iz referenčnega korpusa in iz korpusa tvitov, z metodo distribucijskega modeliranja besed. Rezultate pristopa ovrednotimo z ročno, korpusno podprto leksikografsko analizo, ki jo nadgradimo s predlogom tipologije za samodejno zaznane pomenske premike. Poleg šuma, do katerega prihaja zaradi napak pri predprocesiranju uporabljenih korpusov, predstavljeni pristop izpostavi velik delež dragocenih kandidatov za pomenske premike, med katerimi prednjačijo tisti, ki se v tvitih pojavijo zaradi dnevnih dogodkov in neformalnih sporočanjskih okoliščin.

Ključne besede: leksikalna semantika, distribucijsko modeliranje besed, pomenski premiki leksikografije, družbena omrežja

1 UVOD

Besede skozi rabo ves čas dobivajo nove pomene ali pomenske odtenke ali pa izgubljajo tiste, ki niso več v rabi (Mitra et al. 2015). Pomen besed se tipično spreminja sistematično (Campbell 2004) in se po navadi širi ali oža (Sagi idr. 2009), velikokrat pa besede dobivajo tudi nove pozitivne ali negativne konotacije, procesa, ki ju v leksikalni semantiki imenujemo amelioracija in pejoracija (Cook in Stevenson 2009). Klasičen primer širjenja pomena v slovenščini predstavlja *miška*, ki je tradicionalno pomenila poljsko žival, danes pa z njo poimenujemo tudi računalniško miško. Nasprotni proces se je zgodil s pomenom besede *gas*, ki se je v 19. stoletju uporabljal tudi za plinske ulične svetilke, danes pa ga v nestandardnem jeziku uporabljamo le še za plin. Amelioracijo lahko npr. opazimo pri prislovu *hudo*, ki je v standardni slovenščini negativno konotirana, medtem ko ima v nestandardni rabi izrazito pozitiven naboj. Obratni primer semantičnega premika, pejoracije, opazimo pri samostalniku *blondinka*, ki je v standardni slovenščini nevtralen, v nestandardni pa rabljen skoraj izključno slabšalno.

Detekcija novih pomenov je velik, a pomemben izziv za leksikografijo, ki je nujen za posodabljanje slovarskih gesel glede na sodobno rabo, uporabniške spletne vsebine in družbena omrežja pa so zaradi množične priljubljenosti in živahne jezikovne rabe idealen vir za tovrstne informacije. Aktualen popis semantičnega inventarja potrebujejo tudi različne jezikovnotehnološke aplikacije, kot sta npr. odgovarjanje na vprašanja in strojno prevajanje.

Zaradi obsežnosti jezikovnega gradiva, ki jih je za to nalogo potrebno analizirati, so lahko učinkoviti le avtomatizirani pristopi. V tem prispevku predstavljamo pristop za samodejno detekcijo pomenskih premikov, ki smo ga prvič predstavili že v Fišer in Ljubešić (2016), kjer smo ga preizkusili na manjšem eksperimentu. Ker so bili prvi rezultati spodbudni, smo za namene pričujoče publikacije bistveno nadgradili jezikoslovno analizo kandidatov, ki je sedaj opravljena na obsežnejšem korpusnem gradivu, poenotenih protokolih gradnje besednih skic v obeh korpusih ter z izboljšano tipologijo in smernicami za razvrščanje kandidatov. Zaradi boljšega razumevanja razvite metode in prizadevanj za njene nadaljnje izboljšave smo za to poglavje nekoliko spremenili tipologijo pomenskih premikov. V kategorijo kandidatov, pri katerih z ročno analizo ni bilo zaznanega pravega pomenskega premika, smo uvedli podkategoriji za kandidate, ki so se na vrh seznama uvrstili zaradi napak pri predprocesiranju korpusa, in dejanske lažne kandidate (glej razdelek 4.3). Zato se rezultati analize, predstavljene v tem poglavju, razlikujejo od tistih, ki smo jih objavili v predhodnih publikacijah, kar se kaže predvsem v zmanjšanem deležu identificiranih večjih premikov proti povečanemu deležu identificiranega šuma v predprocesiranju korpusov.

2 SORODNE RAZISKAVE

Medtem ko so samodejno prepoznavanje pomenov besed temeljito proučevali že številni raziskovalci (Sparck Jones, 1986; Ide in Véronis 1998, Schütze 1998, Navigli 2009), so avtomatski pristopi k detekciji pomenskih premikov še vedno razmeroma slabo raziskani, čeprav so zelo pomembno teoretično-analitično vprašanje v leksikalni semantiki pa tudi dragocen aplikativen izziv v sodobni leksikografiji, ki bi lahko prispeval k lažjemu in hitrejšemu posodabljanju slovarskih gesel, kar je cilj, ki z vse večjo dostopnostjo diahronih, tematsko in žanrsko specifičnih korpusov postaja vse bolj dosegljiv.

Večina raziskav na področju samodejnega prepoznavanja pomenskih premikov se osredotoča na diahrono sledenje sprememb v rabi in pomenu besed na podlagi zelo obsežnih zgodovinskih korpusov, ki zajemajo besedila izpred nekaj desetletij ali celo stoletij (Mitra et al. 2015; Tahmasebi et al. 2011; Hamilton et al. 2016). Drugi priljubljeni pristop je primerjava besednih pomenov v dveh ali več korpusih, ki vsebujejo besedila iz različnih časovnih obdobij ali žanrov. Cook et al. (2013), na primer, nove besedne pomene identificirajo s primerjavo t. i. *ciljnega korpusa* z *referenčnim korpusom*, za kar uporabijo metode tematskega modeliranja za indukcijo besednih pomenov. Preprostejši in potencialno robustnejši pristopi ne zahtevajo predhodnega razlikovanja med specifičnimi pomeni, temveč merijo kontekstualno razliko leksema v dveh ali več korpusih. Gulordava in Baroni (2011), na primer, pomenske premike merita s pomočjo distribucijske podobnosti med besednimi vektorji, zgrajenimi iz dveh različnih korpusov. Podoben pristop uporabimo tudi v tem poglavju, v katerem uporabimo distribucijsko modeliranje za prepoznavanje novih pomenov v jeziku slovenskih tvitov.

3 METODA

Pristop, uporabljen v pričujočem poglavju, temelji na distribucijskem modeliranju pomena besed. Za zaznavanje pomenskih premikov je ključna gradnja in primerjava dveh distribucijskih semantičnih modelov za vsako iztočnico iz dveh korpusov: prvega za rabo iztočnice v splošnem jeziku slovenščini, za kar smo uporabili referenčni korpus Gigafida,[1] drugega pa za iztočnico, kot se uporablja v spletni slovenščini, za kar smo uporabili korpus Janes-Tweet (glej Erjavec et al. 2018).

1 http://www.gigafida.net

Za gradnjo in primerjavo distribucijskih modelov smo uporabili orodje *word2vecf* (Levy in Goldberg 2014b).[2] Kot kontekstne značilke smo upoštevali površinske oblike in se tako izognili pogostemu šumu, do katerega prihaja pri označevanju in lematizaciji nestandardnih besedil. Značilke smo zajeli iz kontekstnega okna dveh besed na vsaki strani iztočnice, pri čemer ločil nismo upoštevali. Relativnega položaja iztočnic nismo kodirali.

Na ta način smo dobili vektorske predstavitve 200 dimenzij za vsako od 5425 lem, ki se v korpusu Janes-Tweet pojavijo vsaj 500-krat. Z znižanjem tega precej strogega frekvenčnega praga bi sicer zlahka pridobili večji nabor besed, a smo ta kriterij uporabili, ker se v pričujoči raziskavi osredotočamo na splošno besedišče, ki je pogosto v različnih žanrih.

Za izračun pomenskih premikov smo uporabili kosinusno podobnost, pretvorjeno v mero razdalje (kosinusno podobnost odštejemo od 1) med vektorjema za iztočnico, zgrajenima iz standardnega in nestandardnega korpusa. Pri tem smo izhajali iz predpostavke, da je razdalja med vektorjema iztočnice, ki se v obeh korpusih uporablja v istem pomenu (npr. *banana*), mnogo manjša med vektorjema iztočnice, ki se pojavlja v različnih pomenih (npr. *miška*).

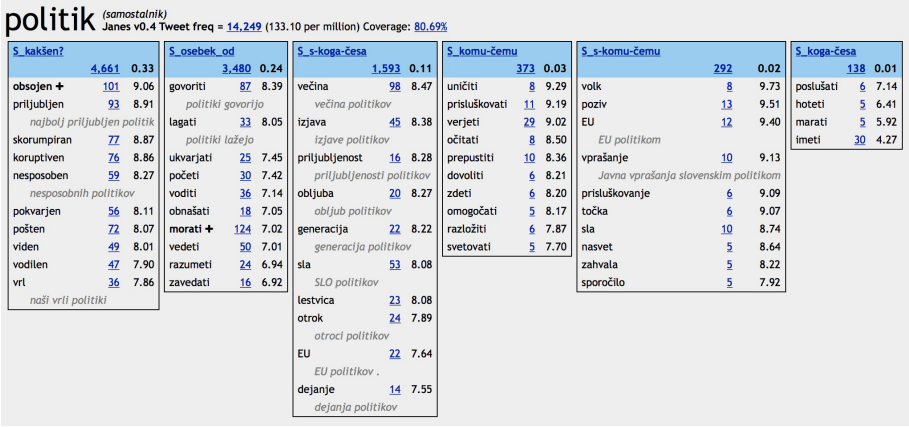
Predstavljena metoda je razmeroma preprosta in ne upošteva dejstva, da je večina besed v korpusu uporabljenih v številnih pomenih, prav tako pa tudi ne ločuje med različnimi vrstami pomenskih premikov. Vendar je za alternativni pristop potrebno predhodno razdvoumljanje večpomenskih besed, ki je že sama na sebi zelo zahtevna naloga, tako da bi v naš postopek v podatke vnašala precej šuma, še posebej, ker delamo z nestandardnim jezikom. Dodatna omejitev razdvoumljanja je, da ne zmore prepoznati novih pomenov, ki so eden naših glavnih ciljev. Ne glede na to smo prepričani, da je predlagani preprost pristop lahko neposredno uporaben za leksikografsko delo, saj izpostavi lekseme, ki so bodisi (1) uporabljeni v različnih pomenih bodisi (2) imajo drugačno frekvenčno distribucijo pomenov v obeh korpusih, kar je oboje pomembno za opis rabe besed. Tako preprost in robusten pristop bi bilo enostavno integrirati v leksikografski delotok (Gantar et al. 2015).

4 ANALIZA

Metodo smo preizkusili z ročno analizo 200 lem, katerih konteksti se v referenčnem korpusu in korpusu tvitov najbolj razlikujejo. Za to smo uporabili primerjavo besednih skic iste leme v obeh korpusih v orodju Sketch Engine (Kilgarriff et. al. 2014), ki temeljijo na slovničnih vzorcih za slovenščino, ki sta jih razvila

2 https://bitbucket.org/yoavgo/word2vecf/

Krek in Kilgarriff (2006). Besedne skice so povzetki slovničnega in kolokacijskega vedenja iztočnice. Prikazujejo kolokatorje iztočnice in so razvrščene glede na slovnična razmerja, na primer na besede, ki so predmet glagolu, besede, ki služijo kot osebek glagola itd., kot ponazarja Slika 1.



Slika 1: Besedna skica za besedo »politik« v korpusu Janes-Tweet. Oznake na vrhu vsakega stolpca so imena slovničnih relacij, npr. S_kakšen? (pridevnik + samostalnik). Sivo obarvane fraze prikazujejo, kako se iztočnica povezuje s svojimi kolokatorji, npr. »najbolj priljubljen politik«. Kolokacije v krepkem tisku in znakom + ponujajo nadaljnje besedne skice za večbesedno zvezo, npr. »pravnomočno obsojen politik«. Število pojavitev pri vsaki kolokaciji vsebuje povezave na konkordance.

S primerjavo besednih skic v korpusih Janes-Tweet in Gigafida smo izvedli analizo pomenskih premikov, kot ponazarja Tabela 1. V obeh korpusih smo izračunali besedne skice za besedo »pirat«, pri čemer smo v obeh korpusih upoštevali le tiste kolokatorje, ki se z iztočnico pojavijo vsaj petkrat, in število kolokatorjev v posamezni semantični relaciji omejili na 25. Za razvrščanje kolokatorjev smo uporabili asociativno mero logDice. Čeprav zaradi prostorskih omejitev v tabeli navajamo samo pet najmočnejših kolokatorjev treh najbolj produktivnih besednih skic za iztočnico, smo v analizo zajeli vse slovnične relacije in kolokatorje v obeh korpusih. Po potrebi smo analizirali tudi konkordance kolokatorjev.

Tabela 1: Najpogostejših pet kolokatorjev za tri najproduktivnejše slovnične relacije besede »pirat« v korpusih Janes in Gigafida. Besede v krepkem tisku zaznamujejo nov pomen v korpusu Janes, ki v korpusu Gigafida ni izkazan.

Janes			**Gigafida**		
iztočnica: pirat pogostost: 1.034 (9,65 na milijon) pokritost leme:[3] 69,05 %			iztočnica: pirat pogostost: 9.941 (7,05 na milijon) pokritost leme: 86,37 %		
Relacija	Kolokator	Frekv. / logDice	Razmerje	Kolokator	Frekv. / logDice
Pridevnik + iztočnica	somalski	7 / 10,48	Pridevnik + iztočnica	somalijski	203 / 11,19
	somalijski	5 / 9,61		somalski	48 / 9,10
	islandski	**6 / 8,80**		zdelan	28 / 8,29
	vesoljski	6 / 7,70		karibski	60 / 8,18
	spleten	8 / 3,97		novodoben	38 / 6,71
Glagol + iztočnica v rodilniku	**voliti**	**6 / 7,45**	Glagol + iztočnica v rodilniku	preganjati	20 / 6,65
	podpreti	**5 / 6,11**		kaznovati	6 / 5,12
	/	/		snemati	15 / 5,07
	/	/		loviti	12 / 4,56
	/	/		prehiteti	6 / 4,44
Samostalnik + iztočnica v rodilniku	**sestanek**	**5 / 8,05**	Samostalnik + iztočnica v rodilniku	bitka	18 / 7,15
	/	/		zatočišče	9 / 7,05
	/	/		preganjanje	18 / 6,69
	/	/		jahta	6 / 6,68
	/	/		prekletstvo	8 / 6,53

Kot je razvidno iz kolokatorjev v Tabeli 1, lahko iz korpusa Gigafida razberemo tri pomene, kolokatorji nekaterih se lahko prekrivajo:

1) oseba, ki ropa ladje (npr. »somalijski«, »jahta«, »zatočišče«),
2) oseba, ki nezakonito razmnožuje zaščitene vsebine (npr. »novodoben«, »preganjati«, »kaznovati«) in
3) metaforično / knjiga, film, naslov TV-oddaje (npr. »zdelan«, »prekletstvo«, »snemati«).

V korpusu Janes na podlagi kolokatorjev v besedni skici zaznamo podobne pomene:

1) »somalski«
2) »spleten« in
3) »vesoljski«.

3 Pokritost leme je mera, ki ponazarja delež vseh pojavitev obravnavane leme v uporabljenem korpusu, ki so bile upoštevane pri gradnji besedne skice.

Poleg njih pa analiza konkordanc za kolokatorje, kot so »islandski«, »voliti«, »podreti « in »sestanek«, ki se v besedni skici iz korpusa Gigafida ne pojavijo, kažejo na nov pomen besede v tvitih, objavljenih v letih 2014 in 2015, in sicer:

4) osebe, ki so člani novih političnih strank iz Slovenije in drugih držav Evropske unije.

Ta pomen v korpusu Gigafida ni izkazan, saj korpus vključuje samo besedila, ki so bila objavljena do leta 2011, politično gibanje pa je dobilo zagon po uspehu na volitvah; v Nemčiji leta 2011, na Islandiji leta 2013 in v EU leta 2014.

Seveda pa vse razlike med korpusoma še zdaleč ne pomenijo nujno novih pomenov, temveč izkazujejo tudi subtilnejše razlike v rabi, kot sta oženje pomena in redistribucija pogostosti pomenov zaradi razlik v temah, žanrih in registrih, ki so v korpusih zastopani. Zato v ročni analizi razlikujemo med večjimi in manjšimi pomenskimi premiki, ki jih nadalje delimo tudi v podkategorije:

1) Večji pomenski premiki
   a. Vezani na dnevne dogodke
   b. Vezani na razlike v registru[4]
   c. Vezani na razlike v mediju
2) Manjši pomenski premiki
   a. Vezani na distribucijo pomenov
   b. Vezani na omejenost rabe na določene ustaljene vzorce
   c. Vezani na oženje pomena
3) Napake
   a. Zaradi šuma pri predprocesiranju korpusa
   b. Lažni kandidati

Rezultate analize večjih in manjših pomenskih premikov predstavljamo v razdelkih 4.1 in 4.2. Kot pri vsakem samodejnem postopku je tudi pri detekciji pomenskih premikov pričakovati določeno stopnjo šuma, ki se lahko pojavi v kateri koli fazi predprocesiranja korpusa ali pa zaradi pomanjkljivosti predlagane metode. Te primere obravnavamo v razdelku 4.3.

4.1 Večji pomenski premiki

Za boljši uvid v pomenski odtis besedišča, ki glede na referenčni korpus prikazuje največje razlike v rabi na Twitterju, ločujemo med pomenskimi premiki, ki so

4 Z izrazom register opredeljujemo rabo konvencionalizirane rabe jezika, skladne s specifičnimi sporočanjskimi funkcijami in družbenimi okoliščinami (Lee 2001).

vezani na dnevne dogodke, takšnimi, do katerih prihaja zaradi razlik v registru, in tistimi, ki so značilni za medij.

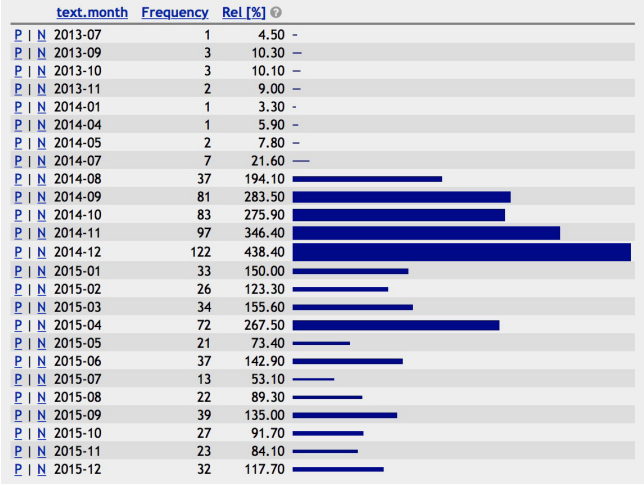
4.1.1 Pomenski premiki, vezani na dnevne dogodke

V to kategorijo uvrščamo novo rabo besed, do katere je prišlo zaradi dnevnih dogodkov, političnih razmer, naravnih katastrof in družbenih okoliščin. Eden takšnih primerov je omenjeni primer »pirat«, ki je bil včasih povezan izključno z morjem, zadnje čase pa je povezan tudi z internetom in celo s politiko, saj označuje člane nove politične stranke, in sicer v izrazito pozitivnem kontekstu (glej Tabelo 1). Drug tak primer je samostalnik »vztrajnik«, ki se v Gigafidi pojavlja precej redko in vedno s pomenom 'del stroja', medtem ko se na Twitterju uporablja precej pogosteje in se skoraj izključno nanaša na 'protestnike'. Najbolj produktivne slovnične relacije in najmočnejši kolokatorji za to iztočnico iz obeh korpusov so prikazani v Tabeli 2.

Tabela 2: Najpogostejših pet kolokatorjev za tri najproduktivnejše slovnične relacije za besedo »vztrajnik« v korpusih Janes in Gigafida. Besede v krepkem tisku označujejo nov pomen v korpusu Janes, ki v korpusu Gigafida ni izkazan.

Janes			**Gigafida**		
iztočnica: vztrajnik pogostost: 816 (7,62 na milijon) pokritost leme: 72,79 %			iztočnica: vztrajnik pogostost: 423 (0,30 na millijon) pokritost leme: 87,47 %		
Relacija	Kolokator	Frekv. / logDice	Pridevnik + iztočnica	Kolokator	Frekv. / logDice
Pridevnik + iztočnica	**drag**	**7 / 4,53**	Pridevnik + samostalnik Iztočnica v imenovalniku + glagol	dvomasen	12 / 11,41
	pravi	**5 / 2,24**		vrteč	5 / 6,24
	/	/		magneten	16 / 5,37
	/	/		lahek	15 / 2,45
	/	/		težek	7 / 0,50
Iztočnica v imenovalniku + glagol	**zapeti**	**7 / 8,81**	Osebek + glagol	skrbeti	6 / 1,97
	vztrajati	**5 / 8,05**	Samostalnik + iztočnica v rodilniku	/	/
	/	/		/	/
	/	/		/	/
	/	/		/	/
Samostalnik + iztočnica v rodilniku	**Viktor**	**8 / 10,78**	Samostalnik + samostalnik v rodilniku	motor	13 / 3,43
	Odbor	**10 / 7,33**		sistem	5 / 0,30
	/	/		/	/
	/	/		/	/
	/	/		/	/

Zanimivo je, da redki najzgodnejši primeri rabe besede »vztrajnik« iz korpusa Janes-Tweet, ki so nastali še pred obdobjem političnih in družbenih nemirov v letih 2013–2014, pripadajo tehničnemu pomenu besede, kot v referenčnem korpusu Gigafida. Na Sliki 2 lahko opazimo izrazito povečanje rabe te besede v korpusu tvitov leta 2014, ko je od začetnih 3 pojavitev od oktobra 2013 do oktobra 2014 narasla na 83, višek pa dosegla decembra 2014 (122) in nato decembra 2015 ponovno upadla na 32. Porast rabe besede sovpada z obdobjem protestov proti obsodbi Janeza Janše, ko opazimo tudi porast protestniškega pomena, ki se najprej pojavi skupaj s tehničnim in nato popolnoma prevlada, tako da med zadetki iz decembra 2015, ki je zadnji mesec, zajet v korpusu, ni zaslediti nobene rabe prvotnega pomena več.



	text.month	Frequency	Rel [%]
P | N	2013-07	1	4.50
P | N	2013-09	3	10.30
P | N	2013-10	3	10.10
P | N	2013-11	2	9.00
P | N	2014-01	1	3.30
P | N	2014-04	1	5.90
P | N	2014-05	2	7.80
P | N	2014-07	7	21.60
P | N	2014-08	37	194.10
P | N	2014-09	81	283.50
P | N	2014-10	83	275.90
P | N	2014-11	97	346.40
P | N	2014-12	122	438.40
P | N	2015-01	33	150.00
P | N	2015-02	26	123.30
P | N	2015-03	34	155.60
P | N	2015-04	72	267.50
P | N	2015-05	21	73.40
P | N	2015-06	37	142.90
P | N	2015-07	13	53.10
P | N	2015-08	22	89.30
P | N	2015-09	39	135.00
P | N	2015-10	27	91.70
P | N	2015-11	23	84.10
P | N	2015-12	32	117.70

Slika 2: Mesečna frekvenčna distribucija besede »vztrajnik« v korpusu Janes-Tweet.

Analiza računov kaže na izrazito lokalizirano uporabo te iztočnice, ki je vezana na majhen krog uporabnikov. Med korporativnimi računi (glej Erjavec et al. 2018) se skoraj vse pojavitve (157 ali 67 %) pojavijo v objavah političnega gibanja Odbor 2014, ki se je zavzemal za izpustitev Janeza Janše, ostale pa v objavah osrednjih in lokalnih pisarn stranke SDS ter časopisov, revij in portalov, ki podpirajo SDS (Demokracija, Politikis.si, Reporter). Med zasebnimi računi je najpogostejših 10 uporabnikov besede, podpornikov Janeza Janše oz. stranke SDS, poskrbelo za 24 % vseh pojavitev v korpusu.

Podrobnejši pregled čustvene zaznamovanosti tvitov, ki vsebujejo besedo *vztrajnik*, nam pokaže zelo zanimivo sliko. Izkaže se, da ima izrazito pozitivno konotacijo. Primer:

»Mnogoštevilni vztrajnice in **vztrajniki** so ponosno zapeli slovensko in evropsko himno. #svobodaJJ«

Po drugi strani pa je konotacija na videz podobnega izraza »vstajnik«, ki ga večinoma isti uporabniki uporabljajo mnogo pogosteje (frekvenca 1767; 16,51 na milijon) v pogovorih o protivladnih protestih, ki so se začeli v času, ko je bil Janša še vedno premier, izrazito negativna.

Vstajniki ste zombiji, vstajniki ste placanci, vstajniki ste levi fasisti, vstajniki ste ovce, ampak nikoli pa ni bilo receno, da ste drhal.

4.1.2 Pomenski premiki, vezani na register

V korpusu tvitov najdemo veliko pomenov, ki v referenčnem korpusu niso izkazani, saj se preko Twitterja odvija veliko neformalne komunikacije, kar prav tako vpliva na semantični odtis besed. Takšen primer je samostalnik »penzion«, ki v standardni slovenščini pomeni *gostinski obrat*, vendar se v nestandardnem jeziku uporablja tudi v pomenu *upokojitev*, kar prikazuje izvleček iz besednih skic v Tabeli 3. Zaradi minimalnega frekvenčnega kriterija petih pojavitev besedne skice iz korpusa Janes niso zelo informativne, vendar vzorec predlog + »penzion« jasno kaže, da v njem poleg pomena *gostišče* zasledimo tudi pomen *upokojitev*.

Tabela 3: Primeri konkordanc za edino produktivno slovnično razmerje za besedo »penzion« v korpusih Janes in Gigafida. Besede v krepkem tisku označujejo nov pomen v korpusu Janes, ki v besednih skicah iz Gigafide ni izkazan.[5]

Janes			**Gigafida**		
iztočnica: penzion pogostost: 1.073 (10,02 na milijon) pokritost leme: 97,48 %			iztočnica: penzion pogostost: 4.898 (3,47 na milijon) pokritost leme: 92,81 %		
Slovnična relacija	Kolokator	Freq / logDice	Slovnična	Kolokator	Freq / logDice
Predlog + iztočnica	**iz**	**127 / 5,04**	Predlog + iztočnica	izpred	107 / 8,22
	Primer: moja mam bi šla *iz Penzion* paketa na Enostavni 300			Primer: odhod bo ob 17. uri izpred *penziona* Špik	
	v	**589 / 4,25**		pred	53 / 0,72
	Primer: ker sem se odloču da se mi ne da več, grem z naslednjim letom *v penzion*			Primer: koncert narodnozabavne skupine Gašperji na plaži pred *penzionom* Tiha dolina	
	do	**9 / 1,73**		v	757 / 0,44
	Primer: vsako leto mi manjka več *do penziona*			Primer : Prenočiti je možno le v *penzionih* ali najeti počitniško hišico.	

5 Se pa pojavi v posameznih konkordancah, npr. zvezah »v penzion«.

Ko kriterij najmanj petih sopojavitev sprostimo, podoben primer rabe prikazujeta tudi naslednja vzorca:

- pridevnik + »penzion«: »(ne)zaslužen«, »priviligiran«, »invalidski«, »zaslužen«, »prisilen«, »predčasen« in
- glagoli, pri katerih iztočnica »penzion« nastopa kot predmet: »izplačevati«, »prislužiti«, »uživati«, »dočakati«, »zaslužiti«.

Tovrstna neformalna jezikovna raba je še posebej dragocena, ker ni zadostno pokrita s tradicionalnimi leksikalnimi viri, zastopana ni niti v večini obstoječih korpusov za slovenščino. Z naraščanjem obsega in pomena komunikacije na družbenih omrežjih postaja vse pomembnejše tudi proučevanje tega segmenta jezika, prav tako bi ga bilo potrebno ustrezno vključiti v sodobne leksikalne vire. Zagotavljanje pokritosti nestandardnega jezika je nujno tudi za robustno avtomatsko procesiranje šumnih spletnih besedil.

4.1.3 Pomenski premiki, vezani na medij

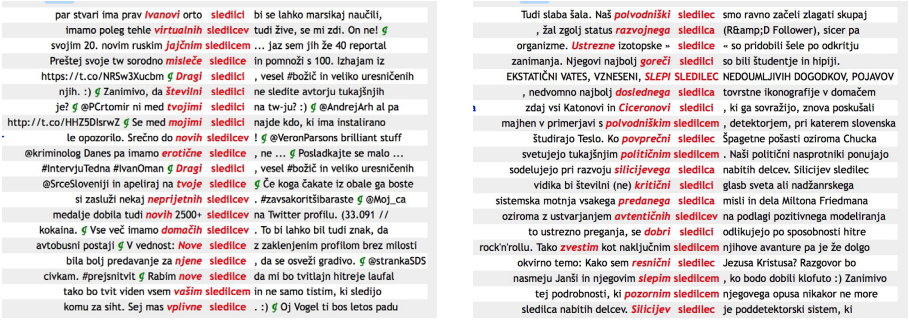
Zadnja skupina večjih pomenskih premikov so nove sporazumevalne konvencije, ki so se pojavile na družbenih omrežjih in so si nekaj obstoječega besedišča prisvojila za nove, specializirane namene. Primer tega pojava je samostalnik »sledilec« (angl. *follower*), pri katerem lahko jasno vidimo spremembo v rabi. Najprej opazimo, da se je njegova raba izrazito povečala (601 zadetkov ali 0,43 na milijon v referenčnem korpusu, ki zajema 1,2 milijarde pojavnic, v primerjavi z 2854 zadetki ali 26,65 v 10-krat manjšem korpusu tvitov). Podrobnejši pregled besednih skic iztočnic v obeh korpusih pokaže specializacijo pomena besede na mikroblogerski platformi Twitter iz enega izmed naslednjih pomenov:

1) sledilec prepričanja in dela vplivnih politikov, verskih voditeljev ali umetnikov (npr. »predan «, »zvest«, »nauk«, »ideja«, »gibanje«);
2) oseba ali organizacija, ki posnema vedenje in govorjenje drugih in ki sam ni vodja (npr. »slep«, »podrejen«, »trend«, »četica«, »prepisovalec«); in
3) sledilna naprava ali medij (npr. »izotopski«, »satelitski«, »vgrajen«, »radiokativen«, »silicijski«);

v

1a) uporabnik, ki spremlja objave drugih uporabnikov na Twitterju in drugih družbenih omrežjih (npr. »nov «, »število «, »nabirati «, »meja «, »milijon«).

Ta pomen v korpusu Janes-Tweet močno prevladuje, kar je razvidno iz 20 naključnih konkordančnih vrstic za razmerje pridevnik + »sledilec« v obeh korpusih na Sliki 3. V korpusu tvitov samo primera 1 in 17 pripadata pomenu 1) – sledilec prepričanja in dela vplivnih politikov, verskih voditeljev ali umetnikov, vsi drugi pa so primeri novega specializiranega pomena – uporabnik, ki spremlja objave drugih uporabnikov na Twitterju in drugih družbenih omrežjih. Na drugi strani pa v Gigafidi 11 (55 %) od vseh prikazanih primerov pripada pomenu 1) – sledilec prepričanja in dela vplivnih politikov, verskih voditeljev ali umetnikov, 5 (25 %) pomenu 3) – sledilna naprava ali medij in 4 (20 %) pomenu 2) oseba ali organizacija, ki posnema vedenje in govorjenje drugih in ki ni vodja. Primeri s pomenom z družbenih omrežij se v Gigafidi ne pojavijo.



par stvari ima prav *ivanovi* orto	sledilci	bi se lahko marsikaj naučili,
imamo poleg tehle *virtualnih*	sledilcev	tudi žive, se mi zdi. On ne! ↯
svojim 20. novim ruskim *jajčnim*	sledilcem	... jaz sem jih že 40 reportal
Preštej svoje tw sorodno *misleče*	sledilce	in pomnoži s 100. Izhajam iz
https://t.co/NRSw3Xucbm ↯ *Dragi*	sledilci	, vesel #božič in veliko uresničenih
njih. :) ↯ Zanimivo, da *številni*	sledilci	ne sledite avtorju tukajšnjih
je? ↯ @PCrtomir ni med *tvojimi*	sledilci	na tw-juž :) ↯ @AndrejArh al pa
http://t.co/HHZSDlsrwZ ↯ Se med *mojimi*	sledilci	najde kdo, ki ima instalirano
le opozorilo. Srečno do *novih*	sledilcev	! ↯ @VeronParsons brilliant stuff
@krimilolog Danes pa imamo *erotične*	sledilce	, ne ... ↯ Posladkajte se malo ...
#IntervjuTedna #IvanOman ↯ *Dragi*	sledilci	, vesel #božič in veliko uresničenih
@SrceSloveniji in apeliraj na *tvoje*	sledilce	↯ Če koga čakate iz obale ga boste
si zaslužil nekaj *neprijetnih*	sledilcev	. #zavsakoritšibaraste ↯ @Moj_ca
medalje dobila tudi *novih* 2500+	sledilcev	na Twitter profilu. (33.091 //
kokaina. ↯ Vse več imamo *domačih*	sledilcev	. To bi lahko bil tudi znak, da
avtobusni postaji ↯ V vednost: *Nove*	sledilce	z zaklenjenim profilom brez milosti
bila bolj predavanje za *njene*	sledilce	, da se osveži gradivo. ↯ @strankaSDS
civkam. #prejsnitvit ↯ Rabim *nove*	sledilce	da mi bo tvitlajn hitreje laufal
tako bo tvit viden vsem *vašim*	sledilcem	in ne samo tistim, ki sledijo
komu za siht. Sej mas *vpiivne*	sledilce	. :) ↯ Oj Vogel ti bos letos padu

Tudi slaba šala. Naš *polvodniški*	sledilec	smo ravno začeli zlagati skupaj
, žal zgolj status *razvojnega*	sledilca	(R&D Follower), sicer pa
organizme. *Ustrezne* izotopske »	sledilce	« so pridobili šele po odkritju
zanimanja. Njegovi najbolj *goreči*	sledilci	so bili študentje in hipiji.
EKSTATIČNI VATES, VZNESENI, *SLEPI*	SLEDILEC	NEDOUMLJIVIH DOGODKOV, POJAVOV
, nedvomno najbolj *doslednega*	sledilca	tovrstne ikonografije v domačem
zdaj vsi Katonovi in *Ciceronovi*	sledilci	, ki ga sovražijo, znova poskušali
majhen v primerjavi s *polvodniškim*	sledilcem	, detektorjem, pri katerem slovenska
študirajo Teslo. Ko *povprečni*	sledilec	Špagetne pošasti oziroma Chucka
svetujejo tukajšnjim *političnim*	sledilcem	. Naši politični nasprotniki ponujajo
sodelujejo pri razvoju *silicijevega*	sledilca	nabitih delcev. Silicijev sledilec
vidika bi številni (ne) *kritični*	sledilci	glasb sveta ali nadžanrskega
sistemska motnja vsakega *predanega*	sledilca	misli in dela Miltona Friedmana
oziroma z ustvarjanjem *avtentičnih*	sledilcev	na podlagi pozitivnega modeliranja
to ustrezno preganja, se *dobri*	sledilci	odlikujejo po sposobnosti hitre
rock'n'rollu. Tako *zvestim* kot naključnim	sledilcem	njihove avanture pa je že dolgo
okvirno temo: Kako sem *resnični*	sledilec	Jezusa Kristusa? Razgovor bo
nasmeju Janši in njegovim *slepim*	sledilcem	, ko bodo dobili klofuto :) Zanimivo
tej podrobnosti, ki *pozornim*	sledilcem	njegovega opusa nikakor ne more
sledilca nabitih delcev. *Silicijev*	sledilec	je poddetektorski sistem, ki

Slika 3: Naključne konkordančne vrstice za razmerje pridevnik + »sledilec« iz korpusa Janes-Tweet (levo) in korpusa Gigafida (desno).

4.2 Manjši pomenski premiki

Med besedami, ki v korpusu tvitov glede na referenčni korpus v svojem semantičnem odtisu izkazujejo manjše razlike, razlikujemo med naslednjimi tremi kategorijami: spremembe v frekvenčni distribuciji rabljenih pomenov, omejenost rabe na določene ustaljene vzorce in semantično oženje.

4.2.1 Spremembe v distribuciji pomenov

V prvo vrsto manjših premikov spadajo tisti primeri, pri katerih smo v obeh korpusih prepoznali enake pomene, vendar z drugačno razporeditvijo po njihovi pogostosti. Dober primer je samostalnik »sesalec«, ki v obeh korpusih pomeni tako žival kot tudi gospodinjski pripomoček, vendar v referenčnem korpusu

prevladuje pomen 'žival', v korpusu tvitov pa pomen 'gospodinjski pripomoček'. Kot prikazuje Tabela 4, samostalnik bolj izstopa v tvitih in samo dve kolokaciji (»edin« in »vrsta«) od najpogostejših petih v vseh treh najbolj produktivnih slovničnih relacijah se v korpusu tvitov nanašata na pomen *žival*, medtem ko je razmerje v referenčnem korpusu ravno obratno (samo »globinski« in »prodajati« se nanašata na napravo, ostale kolokacije se vse nanašajo na žival).

Tabela 4: Najmočnejših pet kolokatorjev za tri najproduktivnejše slovnične relacije za besedo »sesalec« v korpusih Janes in Gigafida. Besede v krepkem tisku so znak prerazporeditve pomenov v korist nestandardnega pomena besede sesalec (tj. v pomenu gospodinjskega pripomočka), ki prevladuje v korpusu Janes.

Janes			**Gigafida**		
iztočnica: sesalec pogostost: 701 (6,54 na milijon) pokritost leme: 78,17 %			iztočnica: sesalec pogostost: 7.047 (4,99 na milijon) pokritost leme: 90,02 %		
Razmerje	Kolokacija	Frekv / logDice	Razmerje	Kolokacija	Frekv / logDice
Pridevnik + samostalnik	**robotski**	**10 / 9,91**	Pridevnik + samostalnik	kopenski	81 / 8,01
	globinski	**7 / 9,53**		rastlinojed	30 / 8,00
	voden	**8 / 6,58**		morski	502 / 7,71
	edin	6 / 3,70		**globinski**	**47 / 7,68**
	nov	**11 / 1,26**		kloniran	26 / 7,49
Glagol + predmet	**imeti**	**7 / 8,81**	Glagol + predmet	klonirati	9 / 8,57
	kupiti	**5 / 8,05**		pleniti	9 / 8,52
	/	/		loviti	16 / 4,98
	/	/		napadati	6 / 4,88
	/	/		**prodajati**	**10 / 3,27**
Samostalnik + samostalnik v rodilniku	**vrečka**	**7 / 9,01**	Samostalnik + samostalnik v rodilniku	kloniranje	53 / 8,89
	zvok	**10 / 8,24**		samica	18 / 7,61
	vrsta	5 / 5,52		tkivo	19 / 7,18
	/	/		genom	10 / 7,09
	/	/		mladič	12 / 7,07

4.2.2 Omejenost na ustaljene vzorce

V to skupino uvrščamo primere, pri katerih zaznamo opazna neskladja v ustaljenih vzorcih, v katerih je iztočnica redno uporabljena in ki opredeljujejo njen pomen. Tipičen primer te kategorije je samostalnik »eter«. Glede na njegovo

besedno skico, ki je povzeta v Tabeli 5, je njegova uporaba v Gigafidi pogosta in raznolika (6.264 ali 4,44 na milijon) ter uporabljena v:

1) dobesednem (kemijskem) pomenu (npr. »molekula«, »element«, »alkohol«, »ester«, »dietil«);
2) metaforičnem pomenu, vključno z imeni podjetij, filmov, itd. (npr. »moralni«, »brazilski«, »svoboden«, »življenje«, »svetloba«); in
3) pomenu oddajanja (npr. »radio«, »televizijski«, »postaja«, »oddaja«, »v«), ki je od naštetih najmanj pogost.

Prav zadnji pomen, torej pomen oddajanja, v tvitih močno prevladuje (npr. »v«, »proti«, »iz/from«, »radijski«), zgolj peščica kolokacij pa pripada pomenu, rabljenemu v kemiji (npr. »borov«, »teorija«), prav tako identificiramo nekaj imen podjetij, ki so pravzaprav lematizacijske napake (npr. »Etra«, »eTRI«). Tudi v tem primeru je v korpusu tvitov iztočnica prominentnejša kot v referenčnem korpusu (8,12 proti 4,44 na milijon), in sicer izrazito pogosto v zvezi »v etru«. V Gigafidi je ta raba sicer zabeležena, vendar velika večina primerov sodi v kemijski ali metaforični pomen.

Tabela 5: Najpogostejših pet kolokatorjev za dve najbolj produktivni slovnični relaciji za besedo »eter« v korpusih Janes in Gigafida. Besede v krepkem tisku kažejo na prerazporeditev pomenov v korist nestandardnemu vzorcu »v etru«, ki prevladuje v korpusu Janes.

Janes			**Gigafida**		
iztočnica:	eter		iztočnica:	eter	
frekvenca:	870 (8,12 na milijon)		frekvenca:	6.264 (4,44 na milijon)	
pokritost leme:	95,86 %		pokritost leme:	85,25 %	
Relacija	Kolokacija	Frekv. / logDice	Relacija	Kolokacija	Frekv. / logDice
Pridevnik + samostalnik	85,25%	85,25%	Pridevnik + samostalnik	škroben	7 / 7,18
	Petričev	5 / 11,17		**radijski**	**279 / 6,71**
	radijski	**5 / 6,21**		toploten	11 / 4,17
	/	/		svetloben	9 / 3,71
	/	/		kemičen	9 / 3,47
Predlog + samostalnik	**izven**	**8 / 6,77**	Predlog + samostalnik	**proti**	**627 / 5,27**
	v	**707 / 4,52**		**izven**	**9 / 3,80**
	proti	**17 / 3,85**		**preko**	**18 / 3,33**
	iz	**6 / 0,63**		**zunaj**	**8 / 3,16**
	/	/		**skozi**	**29 / 2,67**

4.2.3 Semantično oženje

Tretji tip manjših pomenskih premikov, ki smo jih opazili, je oženje semantičnega repozitorija besed, ki najverjetneje ni znak zamiranja določenih pomenov, temveč je posledica omejenega števila tem, ki se pojavljajo v diskusijah na Twitterju v primerjavi s številom tem v referenčnem korpusu. Takšen primer je glagol »posodobiti«, ki se glede na besedne skice v korpusu Gigafida uporablja z zelo različnimi predmeti, kot so »infrastruktura«, »proizvodnja«, »park«, »oprema« in »flota«. Na drugi strani pa je v korpusu tvitov glagol omejen na predmete, ki se nanašajo na informacijske tehnologije: »aplikacija«, »stran«, »seznam«, »sistem«.

Še en primer oženja semantičnega odtisa v korpusu tvitov je večpomenski samostalnik »faks«, ki se v korpusu Gigafida pojavlja bodisi v pomenu naprave bodisi v pomenu ustanove, pri čemer je dominanten prvi, kar je razvidno že iz najmočnejših kolokacij v relaciji glagol + predmet: »pošiljati«, »oddajati«, »sprejemati«, »dokončati«, »vpisati«. V tvitih je raba samostalnika precej prominentnejša (frekvenca 45,78 proti 16,76 na milijon), prav tako močno prevladuje pomen ustanove: »končati«, »pustiti«, »pogrešati«, »narediti«, »imeti«, kar ni presenetljivo, saj je telefaks praktično že zastarela tehnologija.

4.3 Analiza napak

Poleg podrobne analize zaznanih pomenskih premikov analiziramo tudi napačno prepoznane kandidate. Pomenskega premika nismo zaznali pri 103 (51 %) od 200 najpogostejših kandidatov, pri čemer ločujemo dve vrsti napačno prepoznanih kandidatov. V prvo kategorijo sodijo kandidati, ki so se na vrh seznama uvrstili zaradi napak pri predprocesiranju korpusa. Če je bila npr. iztočnici v enem od korpusov pripisana napačna lema, je njen vektor za ta korpus razumljivo povsem drugačen od vektorja za isto iztočnico v drugem korpusu, saj si pravzaprav ne delita semantičnih lastnosti. Pričakujemo lahko, da se bo ta kategorija z razvojem orodij za procesiranje postopoma zmanjševala. V drugo kategorijo pa uvrščamo napačne kandidate, ki so se visoko na lestvici pojavili zaradi predlagane metode, vendar ročni pregled besednih skic zanje ni razkril pomenskih premikov. To so dejanske napake, ki razkrivajo omejitve predlaganega pristopa in jih je nujno potrebno reševati v prihodnjih izboljšavah metode.

4.3.1 Napake zaradi predprocesiranja

Z analizo lažnih kandidatov za pomenske premike, do katerih je prišlo zaradi šuma, ki ga ustvarjajo orodja za jezikoslovno procesiranje slovenščine, smo identificirali 90

(45 %) takšnih primerov. Raven šuma ne preseneča, saj se v korpusu Janes soočamo z izrazito nestandardnimi podatki, ki jih je težko procesirati z visoko stopnjo natančnosti. Drug pogost vir napak se je pojavil zaradi dejstva, da smo v raziskavi uporabili korpusa, ki sta bila označena in lematizirana z dvema različnima orodjema. Obenem pa naša analiza kaže, da je tovrsten tip šuma najvišji na vrhu seznama in nato vztrajno pada. Kar se tiče vzroka za napačno predprocesiranje, so daleč najpogostejši vir napak tuje besede iz tujejezičnih tvitov, ki so bili napačno prepoznani kot slovenski, in slovenski tviti, ki so delno napisani v tujem jeziku (34 %). Drugi najpogostejši tip napak so nestandardne ali nestandardno zapisane besede, ki jih orodje za oblikoskladenjsko označevanje in lematizacijo ni pravilno analiziralo (33 %). Na tretjem mestu so težave, povezane z osebnimi imeni (17 %), ki so še posebej izrazite, če se osebno ime prekriva z občnim. To kaže na težavnost avtomatskega procesiranja uporabniško generiranih vsebin in na to, kako pomembno je zagotoviti, da bodo orodja NLP postala robustnejša tudi za fenomene jezika družbenih medijev.

Tabela 6: Distribucija napak zaradi predprocesiranja.

Vrsta napake	Frekv.	%	Primeri
tuja	27	34 %	duda (namesto ang. »dude/frajer«) danka (namesto nem. »Danke/hvala«) kad (namesto hrv. »kad/ko«)
nestandardna	26	33 %	ajda (namesto »ajde« v pomenu gremo) bol (namesto »bol/bolj«) hod (namesto »hodu/hodil«)
ime	15	18 %	Tanko (priimek, prekriven s s pridevnikom in prislovom »tanek, tanko«) Nedelo (časopis, prekriven s samostalnikom »nedelo«) Rim (mesto, prekriven s samostalnikom »rima«)
lema	13	15 %	meni (namesto »jaz«) moa (izumrla ptica, namesto »moj«) novic (redovniški pripravnik, namesto »novica«)
besedna vrsta	6	6 %	dobro (prekrivna pridevnik in prislov) fin (prekrivna pridevnik in prislov)
diakritike	1	2 %	sel (kdor prinaša sporočilo, namesto »šel«)
orodje	2	2,5 %	nazadnje (normalizirano kot »ne nazadnje«)
Skupno	90	100 %	

4.3.2 Lažni kandidati

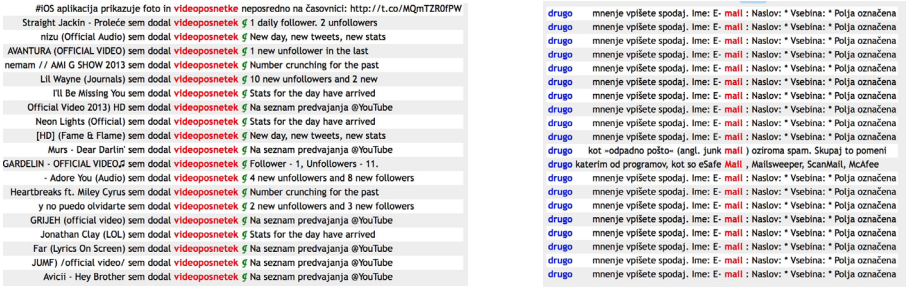
Še posebej nas zanima 13 (6,5 %) kandidatov, pri katerih glede na besedne skice pomenskega premika nismo opazili in ki niso napake, ki so se pojavile med

predprocesiranjem. S pomočjo analize konkordanc jih razvrstimo v dve skupini: besede, ki so uporabljene v samodejno ustvarjenih tvitih ali v ponavljajočih oglasnih besedilih v Gigafidi (9 oz. 69 %) in besede, ki so uporabljene v telegrafskem tviterskem ali tipičnem novinarskem slogu (4 oz. 31 %). Ocenjujemo, da je glede na kompleksnost naloge dobljeni delež lažnih kandidatov dober rezultat. Neposredno izboljšanje rezultata za prvi tip napak bi lahko dosegli z ustavljenimi postopki prepoznavanja in filtriranjem popolnih in delnih duplikatov v korpusih. Za celovitejše vrednotenje razvitega pristopa pa bi bilo treba analizirati večje število kandidatov, predvsem tistih, ki so glede na stopnjo kontekstualnih razlik med modeloma iz obeh korpusov razvrščene nižje na seznamu.

Tabela 7: Distribucija lažnih primerov.

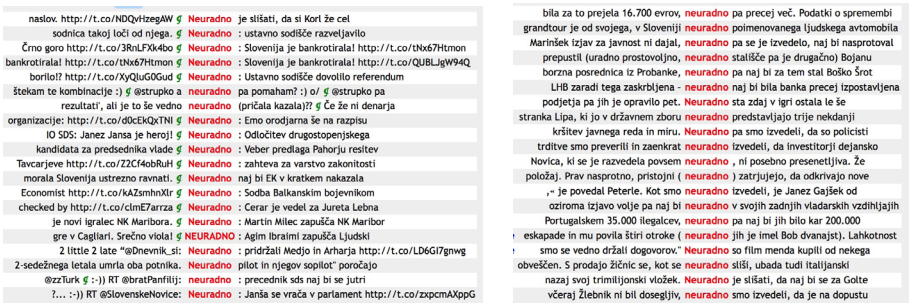
Vrsta napake	Frekv.	%	Iztočnica
Samodejno generirano besedilo	9	69 %	aktualno barometer frizerka kontakten kviz magazin mail veterinar videoposnetek
Slog	4	31 %	dopoldan ekskluzivno izjemoma neuradno
Skupaj	13	100 %	

Primere besed, ki se pojavijo v samodejno ustvarjenih tvitih ali ponavljajočih se oglasnih besedilih iz korpusa Gigafida, prikazuje Slika 4.



Slika 4: Primeri samodejno ustvarjenih besedil v korpusih Janes in Gigafida.

Na Sliki 5 podajamo primer slogovno različne rabe besede »neuradno« v primerjanih korpusih. Čeprav se beseda v Gigafidi večinoma uporablja v časopisnih člankih in revijah, je še vedno skladenjsko integrirana v stavek. Po drugi strani pa jo uporabniki – tudi zasebni uporabniki, ki niso novinarji ali predstavniki medijev – v tvitih uporabljajo v izrazito telegrafskem slogu.



Slika 5: Primeri slogovno različne rabe besede »neuradno« v korpusih Janes in Gigafida.

4.4 Rezultati in diskusija

Rezultati analize so povzeti v Tabeli 8. Iz nje je razvidno, da je bila določena vrsta pomenskega premika zaznana v nekaj manj kot polovici vseh primerov analiziranega vzorca, na podlagi česar sklepamo, da bi predlagani pristop – čeprav ni dovolj natančen za uporabo v popolnoma avtomatiziranem scenariju – lahko bil uporaben kot polavtomatski postopek, ki bi leksikografom postregel z avtomatsko generiranimi predlogi, ti pa bi potem rezultate ročno pregledali. Če upoštevamo razmeroma velik delež napak, do katerih je prišlo zaradi napak pri predprocesiranju korpusov (45 %), bi lahko z uporabo robustnejših orodij za nestandardno slovenščino rezultate še znatno izboljšali: boljša identifikacija jezika, v katerem je besedilo napisano, robustnejše procesiranje nestandardnih različic zapisa besed in nestandardnega besedišča, natančnejše prepoznavanje lastnih imen in boljše oblikoskladenjsko označevanje.

Večji in manjši pomenski premiki so bili skoraj enako pogosti (približno četrtina vzorca pri vsakem). Nepresenetljivo lahko večino pomenskih premikov pripišemo manj formalnemu registru in značilnostim tem, ki se pojavljajo v pogovorih na Twitterju (ki skupaj predstavljajo za četrtino analiziranega vzorca), kar sistematično izpostavi razlike v osredotočenosti in razponu tem v obeh korpusih. Dejstvo, da je bilo zaznanih mnogo več novih primerov rabe (novi pomeni zaradi registra, družbenega konteksta in medija) kot pa oženj (pomeni in vzorci, omejeni na

tematiko) (25 % proti 13 %) nakazuje na to, da bi lahko referenčni korpus še izboljšali s sodobnejšimi besedili in besedili z družbenih medijev ter drugimi manj formalnimi in manj standardnimi komunikacijskimi praksami, saj vsebujejo bogato in dragoceno jezikovno gradivo, ki je iz referenčnega korpusa zaenkrat skoraj popolnoma izključeno.

Še posebej zanimivi so zaznani novi pomeni kot rezultat širšega družbenega konteksta, v katerem se uporabniki izražajo na Twitterju, s čimer se prilagodijo tako nastajajočim novicam (9 %) kot tudi dinamičnim konvencijam komuniciranja na družbenih medijih (4 %). To leksikografsko gradivo je izjemno dragoceno, saj bi lahko služilo kot osnova za posodobitev obstoječih leksikosemantičnih virov slovenščine, s čemer se potrjuje potreba po spremljevalnem korpusu, ki za slovenščino zaenkrat še ni na voljo.

Tabela 8: Porazdelitev tipov pomenskih premikov v slovenski tviterščini. Vrednosti v oklepajih so izračunani za podkategorijo, odstotki pa za kategorijo.

Kategorija	Podkategorija	Frekv.	%
Večji premiki		51	25 %
	Register	(26)	(51 %)
	Dogodki	(18)	(35 %)
	Medij	(7)	(14 %)
Manjši premiki		46	23 %
	Distribucija pomenov	(24)	(52 %)
	Ožanje pomenov	(20)	(43 %)
	Ustaljeni vzorci	(2)	(4 %)
Brez premika		103	51 %
	Šum zaradi predprocesiranja	(90)	(87 %)
	Lažni kandidati	(13)	(13 %)
Skupaj		200	100 %

Z ročno analizo smo identificirali tudi nekaj kreativnega pripisovanja novih pomenov sicer vsakdanjim besedam (npr. »kahla«, ki se nanaša na politika Karla Erjavca, ki zelo prepoznavno izgovarja črko r; »pingvin«, ki se nanaša na »zamrznjenega« vodjo politične stranke Zares Zorana Jankoviča; in šaljivo rabo besede »modrec« za »modrc«).

Rezultati izvedene jezikovne analize kažejo, da lahko pristop, ki ga smo ga predstavili v pričujočem poglavju, močno pripomore k rednim polavtomatskim posodobitvam tako splošnih kot tudi specializiranih na korpusu temelječih leksikalnih virov.

5 SKLEP

V poglavju smo predstavili potencial distribucijskega simuliranja za avtomatizacijo leksikografskega dela, ki smo ga preizkusili na problemu zaznavanja pomenskih premikov v slovenščini na družbenih omrežjih. Pomenski premik besede smo izmerili kot razdaljo med reprezentacijo njenega semantičnega odtisa, naučeno iz referenčnega korpusa, in njeno ustreznico iz korpusa tvitov. Za 200 besed z izkazanimi največjimi razlikami smo opravili ročno analizo, ki je pokazala, da je – z izjemo zlahka prepoznanega šuma zaradi napak, do katerih je prišlo pri predprocesiranju (45 % analiziranih primerov) – pristop prinesel veliko dragocenih kandidatov z izkazanimi pomenskimi premiki, med katerimi so še posebej zanimivi novi pomeni, ki so posledica jezikovne rabe ob odzivanju na dnevno dogajanje ter neformalne komunikacije. Zanimivo bi bilo spremljati, ali so zaznani novi pomeni in semantični premiki kratkotrajni in kateri izmed njih bodo postali trajni del leksikosemantičnega repozitorija.

Prispevek v slovenski prostor vnaša model detektiranja pomenskih premikov, ki je neposredno uporaben pri razvoju slovarja slovenske tviterščine (Gantar et al. 2016), prav tako pa tudi pri nadgradnji in posodabljanju obstoječih splošnih slovarskih priročnikov in baz za slovenščino (Gorjanc et al. 2015). Vendar prispevek predstavljene metode sega še mnogo dlje, saj demonstrira uporabnost distribucijskih pristopov za številne leksikografske naloge za slovenščino ter tudi druge jezike, pri čemer potrebujemo le dva dovolj obsežna jezikoslovno označena korpusa; referenčni korpus in ciljni korpus, ki pokriva tisti segment jezikovne rabe, ki nas za konkretno raziskavo zanima, npr. korpus govorjenega jezika, historični korpus, različni področnospecifični korpusi ipd.

V nadaljevanju raziskav se nameravamo osredotočiti na (1) razširitev ročne analize na kandidate s spodnjega dela seznama, (2) razširitev pristopa na redkejše besedišče, (3) primerjavo predlaganega pristopa z alternativnimi metodami, kot je npr. učenje reprezentacij besed s pomočjo besednih skic oz. skladenjskih vzorcev in (4) uporabo nadzorovanega učenja za prepoznavanje pomenskih premikov, razločevanje med različnimi vrstami pomenskih premikov in filtriranje napak, do katerih pride pri predprocesiranju.

Zahvala

Zaradi mednarodne avtorske zasedbe je bil rokopis tega poglavja delno napisan v angleškem jeziku. Za pomoč pri prevodu se zahvaljujemo Lei Anžur

Literatura

Agres Kat, Stephen McGregor, Matthew Purver in Geraint Wiggins, 2015: Conceptualizing creativity: From distributional semantics to conceptual spaces. *Proceedings of the Sixth International Conference on Computational Creativity*. 118–125.

Bengio Yoshua, Aaron Courville in Pascal Vincent, 2013. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35/8: 1798–1828.

Blei, David M., Andrew Y. Ng in Michael I. Jordan, 2003: Latent Dirichlet Allocation. *The Journal of Machine Learning Research* 3: 993–1022.

Campbell, Lyle, 2013. *Historical linguistics.* Edinburg: Edinburgh University Press.

Church, Kenneth. W. in Hanks, Patrick, 1990: Word association norms, mutual information, and lexicography. *Computational Linguistics* 16/1. 22–29.

Cook, Paul in Suzanne Stevenson, 2010: Automatically Identifying Changes in the Semantic Orientation of Words. *Proceedings of the 7th LREC Conference*. 28–34.

Cook, Paul, Jey Han Lau, Michael Rundell,Diana McCarthy in Timothy Baldwin: 2013: A lexicographic appraisal of an automatic approach for detecting new word senses. *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex conference*. Tallinn, Estonia. 49–65. http://eki.ee/elex2013/proceedings/eLex2013_04_Cook+etal.pdf

Čibej, Jaka in Nikola Ljubešić, 2015: »S kje pa si?«. *Proceedings of the conference Slovene on the web and in the new media. Ljubljana: Znanstvena založba Filozofske fakultete.* 10–14.

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

Firth, John R, 1957: A Synopsis of Linguistic Theory. *Studies in Linguistic Analysis*: 1-32.

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

Fišer, Darja in Nikola Ljubešić, 2016: Detecting Semantic Shifts in Slovene Twitterese. Proceedings of the Tenth Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2016. Brno, Czech Republic.

Gantar, Polona, Iza Škrjanec, Darja Fišer in Tomaž Erjavec, 2016: Slovar tviterščine. *Proceedings of the Conference on Language Technologies and Digital Humanities*, Ljubljana, Slovenia. 71–76.

Fodor, Imola, 2002: *A Survey of Dimension Reduction Techniques*. Center for Applied Scientific Computing, Lawrence Livermore National Laboratory. https://e-reports-ext.llnl.gov/pdf/240921.pdf

Gantar, Polona, Iztok Kosem in Simon Krek, 2015: Leksikografski proces pri izdelavi spletnega slovarja sodobnega slovenskega jezika. Gorjanc, Vojko, Polona

Gantar, Iztok Kosem in Simon Krek (ur.): *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete. 280–297.

Golub, Gene H. in Christian Reinsch, 1970: Singular value decomposition and least squares solutions. *Numerische mathematik* 14/5: 403–420.

Grčar, Miha, Simon Krek in Kaja Dobrovoljc, 2012: Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. *Proceedings of the 8th Language Technologies Conference*. Ljubljana: Institut »Jožef Stefan«.

Hamilton William L., Jure Leskovec in Dan Jurafsky, 2016: Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin, Germany. 1489–1501.

Tomaž Erjavec, Darja Fišer, Jaka Čibej in Špela Arhar Holdt, 2016: CMC training corpus Janes-Norm 1.2. *Slovenian language resource repository CLARIN.SI*.

Gulordava, Kristina in Marco Baroni, M. 2011. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. *Proceedings of the Workshop on GEometrical Models of Natural Language Semantics*. Edinburgh, Scotland. 67–71.

Ide, Nancy in Jean Véronis, 1998: Introduction to the special issue on word sense disambiguation: the state of the art. *Computational linguistics* 24/1: 2–40.

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

Krek, Simon in Adam Kilgarriff, 2006: Slovene Word Sketches. *Proceedings of the 5th Slovenian/First International Languages Technology Conference*. https://www.kilgarriff.co.uk/Publications/2006-KrekKilg-Ljub-SloveneWS.pdf

Lenci, Allessandro, 2008: Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics* 20/1: 1–31.

Lee, David, 2001: Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology* 5/3. 37–72.

Levy, Omer in Yoav Goldberg, 2014a: Neural word embedding as implicit matrix factorization. *NIPS'14 Proceedings of the 27th International Conference on Neural Information Processing Systems*. 2177–2185.

Levy, Omer in Yoav Goldberg, 2014b: Dependency-Based Word Embeddings. *Proceedings of ACL*. 302–308.

Levy, Omer in Yoav Goldberg, 2014: Linguistic Regularities in Sparse and Explicit Word Representations. *Proceedings CoNLL:* 171–180.

Ljubešić, Nikola, Darja Fišer, Tomaž Erjavec, Jaka Čibej, Dafne Marko, Senja Pollak in Iza Škrjanec, 2015: Predicting the Level of Text Standardness in User-generated Content. *Proceedings of Recent Advances in Natural Language Processing*. 371–378.

Ljubešić, Nikola in Tomaž Erjavec, 2016: Corpus vs. Lexicon Supervision in Morphosyntactic Tagging: The Case of Slovene. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. 1527–1531.

Nikola Ljubešić, Zupan, K., Darja Fišer and Tomaž Erjavec Normalising Slovene data: historical texts vs. user-generated content. *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*. 146–155.

Logar, Nataša, Miha Grčar, 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.

McCulloch, Warren. S. in Pitts, Walter, 1943: A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics* 5/4. 115–133.

Mikolov, Tomas, Kai Chen, Greg Corrado in Jeffrey Dean, 2013a: Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, Tomas, Kai Chen, Greg Corrado in Jeffrey Dean, 2013b: Linguistic Regularities in Continuous Space Word Representations. *Proceedings of HLT-NAACL 2013*. 746–751.

Mitchell, T.M. 1997. *Machine Learning*. McGraw-Hill, Inc. New York, NY, USA.

Mitra, Sunny, Ritwik Mitra, Suman Kalyan Maity, Martin Riedl, Chris Biemann, Pawan Goyal In Animesh Mukherjee, 2015: An automatic approach to identify word sense changes in text media across timescales. *Natural Language Engineering* 21/5. 773–798.

Navigli, Roberto, 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)* 41/2.

Schütze, Hinrich, 1998: Automatic word sense discrimination. *Computational linguistics* 24/1. 97–123.

Sagi, Eyal, Stefan Kaufmann in Brady Clark, 2009: Semantic Density Analysis: Comparing Word Meaning across Time and Phonetic Space. *Proceedings of the EACL 2009 Workshop on GEometical Models of Natural Language Semantics*. 104–111.

Sparck Jones, Karen, 1986: *Synonymy and Semantic Classification.* Edinburg: Edinburgh University Press.

Kanerva, Pentti, Jan Kristoferson in Anders Holst, 2000: Random Indexing of Text Samples for Latent Semantic Analysis. *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*. 1036.

Salton, Gerard, Wong, Andrew, Yang, Chungshu, 1975: A vector space model for automatic indexing. *Communications of the ACM* 18/11. 613–620.

Tahmasebi, Nina, Thomas Risse in Stefan Dietze, 2011: Towards automatic language evolution tracking, a study on word sense tracking. *Proceedings of the Joint Workshop on Knowledge Evolution and Ontology Dynamics*. http://ceur-ws.org/Vol-784/evodyn2.pdf

Sparck Jones, K., 1972: A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation* 28/1. 11–21.

Zipf, George K., 1936: *The psycho-biology of language*. New York, London: Routledge.

Priloge:
Seznami besed z označenim tipom pomenskega premika

Priloga 1: Seznam besed, pri katerih smo zaznali večje pomenske premike.

Lema z besedno vrsto	Vrsta pomenskega premika	Mera razdalje
fiskalen#Ag	dogodek	0,580353749510
zombi#Nc	dogodek	0,570473162536
pirat#Nc	dogodek	0,563599872708
arhivski#Ag	dogodek	0,543735200555
bojevnik#Nc	dogodek	0,538704741919
nevtralnost#Nc	dogodek	0,528946112923
komisarka#Nc	dogodek	0,526273297109
pingvin#Nc	dogodek	0,517542639943
risa#Nc	dogodek	0,500310188561
malomaren#Ag	dogodek	0,490241581560
vstajnik#Nc	dogodek	0,476038560358
astronomski#Ag	dogodek	0,471662874089
prepih#Nc	dogodek	0,470011599897
kahla#Nc	dogodek	0,451749782663
tisa#Nc	dogodek	0,383248742949
dl#Nc	dogodek	0,367170101908
supervizor#Nc	dogodek	0,285819575689
vztrajnik#Nc	dogodek	0,143540587272
opomnik#Nc	medij	0,580180651285
sledenje#Nc	medij	0,578872731644
sledilec#Nc	medij	0,556588211346
q#Nc	medij	0,474845868899
ms#Nc	medij	0,467078140408
nm#Nc	medij	0,402784956150
rt#Nc	medij	0,211168972262
bus#Nc	register	0,579057574473
bio#Ag	register	0,578177532133
profi#Nc	register	0,577361896651
modrec#Nc	register	0,574217906729
dostavljati#Vm	register	0,571193498689
ultra#Ag	register	0,567350218824
noro#Rg	register	0,566595894804
skozi#Rg	register	0,564839986817
penzion#Nc	register	0,558843681744
jaz#Nc	register	0,557462156560
carski#Ag	register	0,553271554077

Lema z besedno vrsto	Vrsta pomenskega premika	Mera razdalje
optika#Nc	register	0,542834517374
sesalec#Nc	register	0,537911484878
depresiven#Ag	register	0,530165847517
misterij#Nc	register	0,506956166729
takisto#Rg	register	0,500157740660
karma#Nc	register	0,498301378493
spin#Nc	register	0,497109940644
info#Ag	register	0,496657400098
ajd#Nc	register	0,493257015790
glavno#Rg	register	0,490177276439
gotov#Ag	register	0,465296448441
smrkec#Nc	register	0,451190591675
fakin#Nc	register	0,391605134818
meh#Nc	register	0,361707010382
top#Ag	register	0,343414605450

Priloga 2: Seznam besed, pri katerih smo zaznali manjše pomenske premike.

Lema z besedno vrsto	Vrsta pomenskega premika	Mera razdalje
obsojenec#Nc	oženje	0,606742377344
drsati#Vm	oženje	0,589984006409
mavrica#Nc	oženje	0,585132147897
asociacija#Nc	oženje	0,569884071932
kozolec#Nc	oženje	0,563829321520
produktiven#Ag	oženje	0,563099861782
posodobiti#Vm	oženje	0,562595017512
kongresen#Ag	oženje	0,558579997685
enka#Nc	oženje	0,555646649303
ponovitev#Nc	oženje	0,544808051012
kvadrat#Nc	oženje	0,526933310826
podnapis#Nc	oženje	0,517714981669
posodobitev#Nc	oženje	0,517572595909
agregat#Nc	oženje	0,493994603878
faks#Nc	oženje	0,466049923982
malezijski#Ag	oženje	0,449709928416
album#Nc	oženje	0,439074282021
beta#Nc	oženje	0,429867341906
mol#Nc	oženje	0,420191219414
predvajanje#Nc	oženje	0,349996691863
panda#Nc	vrstni red	0,581181767675
odklop#Nc	vrstni red	0,581121382368

Lema z besedno vrsto	Vrsta pomenskega premika	Mera razdalje
burka#Nc	vrstni red	0,578919014186
domena#Nc	vrstni red	0,574300502153
tesla#Nc	vrstni red	0,566861095077
testen#Ag	vrstni red	0,565518883875
kompas#Nc	vrstni red	0,559512599605
izpiten#Ag	vrstni red	0,546894441609
odmev#Nc	vrstni red	0,530065326675
agenda#Nc	vrstni red	0,527949021220
vezava#Nc	vrstni red	0,525247135181
cifra#Nc	vrstni red	0,523292962123
recenzija#Nc	vrstni red	0,519103247498
kopitar#Nc	vrstni red	0,516488872572
x#Nc	vrstni red	0,514193345981
ciganski#Ag	vrstni red	0,512854059519
pogovoren#Ag	vrstni red	0,509486755021
krogec#Nc	vrstni red	0,494365095953
konzorcij#Nc	vrstni red	0,484658137741
profilen#Ag	vrstni red	0,480269943843
greznica#Nc	vrstni red	0,460715813930
bolha#Nc	vrstni red	0,451526694121
opeka#Nc	vrstni red	0,381554496304
tajfun#Nc	vrstni red	0,353925432855
eter#Nc	stalna besedna zveza	0,521370091208
biti#Vm	stalna besedna zveza	0,518225790812

Priloga 3: Seznam besed, pri katerih smo zaznali napake v predprocesiranju.

Lema z besedno vrsto	Vrsta napake
sel#Nc	manjkajoči diakritiki
a#Nc	tujejezična beseda
ata#Nc	tujejezična beseda
biga#Nc	tujejezična beseda
danka#Nc	tujejezična beseda
duda#Nc	tujejezična beseda
era#Nc	tujejezična beseda
gama#Nc	tujejezična beseda
jak#Nc	tujejezična beseda
kad#Nc	tujejezična beseda
let#Nc	tujejezična beseda
lik#Nc	tujejezična beseda
lina#Nc	tujejezična beseda

Lema z besedno vrsto	Vrsta napake
lot#Nc	tujejezična beseda
market#Nc	tujejezična beseda
mina#Nc	tujejezična beseda
nada#Nc	tujejezična beseda
navoj#Nc	tujejezična beseda
nem#Ag	tujejezična beseda
oda#Nc	tujejezična beseda
runa#Nc	tujejezična beseda
sita#Nc	tujejezična beseda
som#Nc	tujejezična beseda
sura#Nc	tujejezična beseda
talka#Nc	tujejezična beseda
tim#Nc	tujejezična beseda
uriti#Vm	tujejezična beseda
y#Nc	tujejezična beseda
deti#Vm	pripisana napačna lema
h#Nc	pripisana napačna lema
kola#Nc	pripisana napačna lema
logo#Nc	pripisana napačna lema
meni#Nc	pripisana napačna lema
meniti#Vm	pripisana napačna lema
mesti#Vm	pripisana napačna lema
mik#Nc	pripisana napačna lema
moa#Nc	pripisana napačna lema
novic#Nc	pripisana napačna lema
pestiti#Vm	pripisana napačna lema
stan#Nc	pripisana napačna lema
tuje#Rg	pripisana napačna lema
gin#Nc	ime
hoc#Nc	ime
hrt#Nc	ime
intelekt#Nc	ime
lek#Nc	ime
lump#Nc	ime
marka#Nc	ime
nedelo#Nc	ime
osa#Nc	ime
pikati#Vm	ime
rima#Nc	ime
tanko#Rg	ime
uk#Nc	ime

Lema z besedno vrsto	Vrsta napake
veber#Nc	ime
ajda#Nc	nestandardna raba
bat#Nc	nestandardna raba
beka#Nc	nestandardna raba
boja#Nc	nestandardna raba
bol#Nc	nestandardna raba
butati#Vm	nestandardna raba
dila#Nc	nestandardna raba
dob#Nc	nestandardna raba
hod#Nc	nestandardna raba
ke#Rg	nestandardna raba
kea#Nc	nestandardna raba
klas#Nc	nestandardna raba
koka#Nc	nestandardna raba
luk#Nc	nestandardna raba
maja#Nc	nestandardna raba
mona#Nc	nestandardna raba
plata#Nc	nestandardna raba
pona#Nc	nestandardna raba
sejati#Vm	nestandardna raba
skupiti#Vm	nestandardna raba
tele#Nc	nestandardna raba
tkati#Vm	nestandardna raba
treti#Vm	nestandardna raba
vesti#Vm	nestandardna raba
veti#Vm	nestandardna raba
vod#Nc	nestandardna raba
dobro#Nc	pripisana napačna besedna vrsta
drug#Nc	pripisana napačna besedna vrsta
fin#Ag	pripisana napačna besedna vrsta
garant#Nc	pripisana napačna besedna vrsta
halo#Nc	pripisana napačna besedna vrsta
pod#Nc	pripisana napačna besedna vrsta
nazadnje#Rg	pripisana napačna besedna vrsta
obresti#Vm	pripisana napačna besedna vrsta

Priloga 4: Seznam besed, ki so bili lažni kandidati za pomenske premike.

Lema z besedno vrsto	Razlog za napako
aktualno#Rg	avtomatsko generirane vsebine
barometer#Nc	avtomatsko generirane vsebine
frizerka#Nc	avtomatsko generirane vsebine
kontakten#Ag	avtomatsko generirane vsebine
kviz#Nc	avtomatsko generirane vsebine
magazin#Nc	avtomatsko generirane vsebine
mail#Nc	avtomatsko generirane vsebine
veterinar#Nc	avtomatsko generirane vsebine
videoposnetek#Nc	avtomatsko generirane vsebine
dopoldan#Nc	specifičen slog
ekskluzivno#Rg	specifičen slog
izjemoma#Rg	specifičen slog
neuradno#Rg	specifičen slog

Korpusni pristop k skladnji računalniško posredovane slovenščine

Špela Arhar Holdt

Izvleček

Poglavje predstavlja označevanje in analizo skladenjskih značilnosti računalniško posredovane slovenščine. Na ravni besednega reda najprej predstavimo pripravo korpusa Janes-Syn in prilagoditve označevalnega sistema specifikam računalniško posredovane slovenščine, nato pa preverimo, kako trije neodvisni označevalci razumejo besednoredno zaznamovanost v danem gradivu, kolikšno je ujemanje med njihovimi odločitvami, katere vrste besednorednih problemov se glede na pripisane oznake v podatkih pojavljajo in kako so ti problemi razporejeni glede na avtomatsko pripisano kategorijo jezikovne (ne)standardnosti. Raziskava pokaže, da besednoredno zaznamovanost označevalci opredeljujejo zelo različno, pogled na podatke z vidika pripisane (ne)standardnosti pa postavlja v razmerje koncepta zaznamovanosti in nestandardnosti, kar bi bilo v nadaljevanju smiselno raziskati z vključitvijo podatkov govorjenega jezika. Kategorizacija označenih besednorednih značilnosti osvetljuje segment skladnje računalniško posredovane slovenščine, ki je bil do sedaj neraziskan, in nakazuje naslednje korake za korpusnojezikoslovne raziskave besednega reda v slovenščini.

Ključne besede: računalniško posredovana slovenščina, korpusno jezikoslovje, skladenjsko označevanje, tviti, besedni red

1 UVOD

Uvodno zadržanost do jezikovnih realnosti, ki jih prinaša ali prvič v zgodovini širše javno izpostavlja računalniško posredovana komunikacija, v zadnjih letih nadomešča rastoč raziskovalni interes, tako na področju jezikoslovja kot tudi obdelave naravnih jezikov, podatkovnega rudarjenja in drugih disciplin, ki jim je v interesu opisati oz. uporabljati podatke sodobne jezikovne rabe. Študijam na večjih jezikih (npr. Crystal 2011; Myslin in Gries 2010; Storrer 2013; Chanier 2015) so se pridružile študije na slovenščini, večina kot rezultat projekta JANES.

Pregled literature ob začetku projektnih aktivnosti je razkril, da je skladnja slovenske računalniško posredovane komunikacije raziskovalno komajda dotaknjeno področje. Omeniti je mogoče nekatere pilotske (zdaj je že mogoče reči tudi pionirske) kvalitativne študije na avtentičnem gradivu različnih vrst, h katerim se vračamo v nadaljevanju prispevka: Kranjc (2003) analizira jezik spletnih klepetov, Dobrovoljc (2008) e-pošta sporočila, Jakop (2008) forumske zapise, Kalin Golob (2008) SMS-e, Michelizza (2015) članke Wikipedije in bloge. Zahtevnejši problem za načrtovanje raziskav je bilo dejstvo, da ob začetku projekta JANES virom navkljub tudi celovitejši korpusni opis značilnosti standardne slovenske skladnje še ni bil na voljo;[1] od obsežnejših korpusnih skladenjskih študij je mogoče izpostaviti monografijo Ledinek (2014), ki se pred analizo izbranega nabora glagolov in njihove vezljivosti ukvarja tudi z vprašanjem skladenjskega označevanja slovenščine, zasnovo FrameNeta za slovenščino (Može 2013) in razprave, vezane na pripravo Leksikalne baze za slovenščino (Gantar 2011; 2015). Odsotnost sintetičnih referenčnih informacij, ki so predpogoj za obsežnejše korpusnojezikoslovne primerjave in sistematično identifikacijo nestandardnih skladenjskih prvin v odnosu do standardnih,[2] je usmerila delo v razvoj metodologije za rabo obstoječih virov pri preučevanju specifičnih skladenjskih vprašanj, na drugi strani pa v pripravo novih virov, ki bodo celovitejši vpogled lahko omogočili v prihodnosti.

Prvi cilj projektne aktivnosti je bil tako izdelati, preizkusiti in evalvirati metodologijo, ki omogoča raziskave trendov rabe nestandardnih jezikovnih prvin za potrebe slovenske normativistike. Ker korpus računalniško posredovane

1 Vrzel naslavlja nacionalni projekt Nova slovnica sodobne standardne slovenščine: viri in metode (ARRS J6-8256, vodja Simon Krek), ki se pričenja v času priprave prispevka.

2 Računalniško posredovana slovenščina je zbirni pojem za različne načine komunikacije oz. raznovrstne besedilne vrste. Za slednje je mogoče reči, da *v splošnem* prinašajo opazen delež nestandardnih jezikovnih prvin, če slednje razumemo kot prvine, ki se razlikujejo od standar(dizira)nega dela jezika (Krek 2015), pogosto v smeri nenamernih ali namernih odstopov od obstoječih jezikovne norme. Ker se projekt JANES ciljno posveča nestandardnim prvinam v slovenščini (tudi z vidika identifikacije, kako »nestandardno« sploh opredeljevati, glej Stabej et al. 2016), se na slednje osredotočamo tudi v prispevku, pri čemer pa se je treba na vseh mestih zavedati, da je na ravni posameznih računalniško posredovanih besedil stopnja in vrsta nestandardnosti zelo različna.

slovenščine Janes (Erjavec et al. 2018) prinaša besedila, ki za razliko od večine gradiva v referenčnih korpusih večinoma niso jezikovno korigirana, realneje izkazuje tendence rabe oz. (ne)intuitivnost obstoječih jezikovnih pravil v širši jezikovni skupnosti. Metodološko premišljena primerjava podatkov korpusa Janes in referenčnega korpusa lahko razkrije tisti del jezikovnih sprememb, ki so širše oz. sistemske, in jih loči od redkejših, sporadičnih ali avtorsko/žanrsko vezanih odklonov od norme. Za preizkus metode smo izbrali zveze samostalnika z neujemalnim levim prilastkom (npr. *solo petje, RTV prispevek* proti *solopetje, RTV-prispevek*). Dosedanja slovenistična polemika o tem jezikovnem problemu se na eni strani dotika vprašanja, katere od tovrstnih zvez zapisovati skupaj in katere narazen, na drugi strani pa, kako v primeru zapisa narazen besednovrstno uvrščati prvi del besedne zveze. Študija se posveti tem vprašanjem, pri čemer je pomembno odkritje, da je praksa zapisovanja v korpusu Janes prepričljivo konsistentnejša od prakse zapisovanja v korpusu Kres, kar pomeni, da jezikovna regulacija obravnavanega problema krepi oz. povišuje variantnost v jezikovni rabi. To dejstvo je v nasprotju s pričakovanim ter odpira ključna vprašanja o namenu ter načinu lektoriranja v slovenskem prostoru, kot tudi stanju in vlogi jezikovnih priročnikov za slovenščino in obstoječih standardizacijskih teles ter praks. Opis metode in rezultati so bili predstavljeni v Arhar Holdt in Dobrovoljc (2015; 2016) ter diskutirani v Stabej et al. (2016).

V tem prispevku se osredotočamo na drugi korak projektnih aktivnosti, tj. pripravo skladenjsko označenega korpusa Janes-Syn, ki služi kot izhodišče za nadaljnje odvisnostno označevanje računalniško posredovane slovenščine, s fokusom na njenih nestandardnih značilnostih. Glavne označevalne odločitve so že bile opredeljene v kratkem prispevku (Arhar Holdt et al. 2016), ki ga na tem mestu nadgradimo z natančnejšo analizo podatkovnega seta in primeri gradiva. Nato predstavimo rezultate jezikoslovne analize skladenjskih značilnosti označenih podatkov s poudarkom na besednem redu. K vprašanju pristopimo z označevanjem zaznamovanega besednega reda, ki so ga neodvisno izvedli trije označevalci, oznake kategoriziramo ter prikažemo ujemanje med označevalci po kategorijah, v diskusijo pa pritegnemo tudi pregled identificiranih kategorij glede na njihovo pojavljanje v primerih, ki so bili v izhodiščnem korpusu avtomatsko označeni kot jezikovno standardni ali nestandardni. Prispevek zaključujemo s strnitvijo projektnih spoznanj in nalog za nadaljnje delo.

2 IZDELAVA IN OZNAČEVANJE KORPUSA JANES-SYN

Če so primerjave med korpusom Janes in korpusom Kres (ali Gigafida) dobro izhodišče za obravnavo posameznih skladenjskih problemov, zlasti če gre za

primere, kjer je podatke mogoče pridobiti z uporabo oblikoskladenjskih oznak, je za celovitejše primerjave potrebno zagotoviti označenost korpusov na skladenjskem nivoju. Za to nalogo smo pripravili pilotsko množico skladenjsko označenih tvitov, ki lahko služi za nadaljnje učenje razčlenjevanja slovenske računalniško posredovane komunikacije.[3]

V sklopu projekta je bilo veliko pozornosti namenjene razvoju oz. prilagoditvam postopkov jezikoslovnega označevanja slovenščine specifikam nestandardnega jezika. V te namene je bil razvit učni korpus Janes, v katerem so pojavnice ročno popravljene na ravni segmentacije in tokenizacije ter normalizirane na besedni ravni, ročno pa so pregledane tudi pripisane leme ter oblikoskladenjske oznake (Čibej et al. 2016a; 2016b). Iz korpusa smo vzorčili množico 200 tvitov (475 stavkov), in sicer na način, da vsebuje enakomerne deleže tvitov, ki so avtomatsko označeni kot jezikovno in tehnično (ne)standardni (Ljubešić et al. 2015), ter obenem vključuje primere, ki so daljši od 120 znakov in v avtorstvu zasebnih uporabnikov. Zadnja pogoja sta omogočila, da smo v korpus zajeli besedila, ki vsebujejo dovolj za označevanje relevantnih specifik. Rezultati so bili pripravljeni v tabeli (Slika 1), ki poleg besedila tvita vsebuje razpoložljive metapodatke (ID, čas nastanka, spol in ime uporabnika, reakcije uporabnikov na tvit (všečkanje, deljenje) ter avtomatsko pripisane kategorije jezikovne in tehnične standardnosti ter sentimenta sporočila.

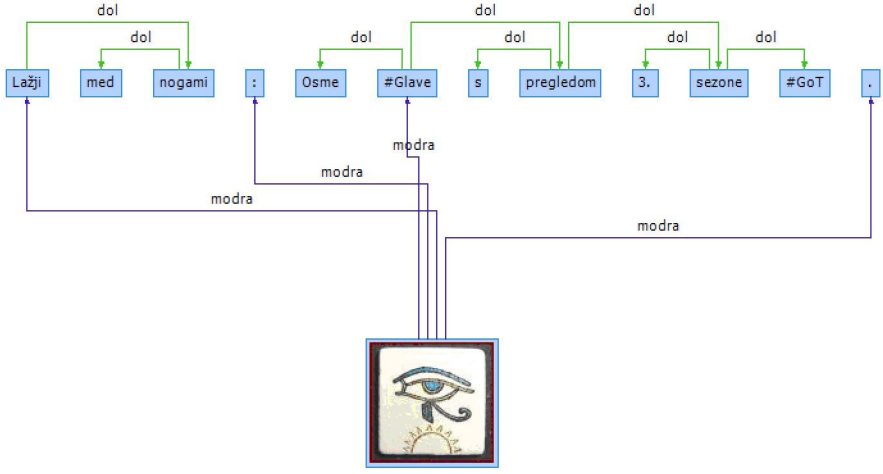
	A	B	C	D	E	F	G	H	I	J	K	L	M
1		id	name	sex	text	created	favorited	retweeted	std_tech	td_tech_n	std_ling	sentiment	source
2	*	tid.7487658297135104		male	Manj kot dan je trajalo, da smo dobili prvi jailbreak za Applovo najnovejšo različico operacijskega sistema iOS, 4.2.1. Kdo bi si mislil.	2010-11-24T17:36:11	0	0	T1	42095	L1	negative	private
3	*	tid.2661225364320624 64		male	Nekateri zvesti podporniki... Še vam ni jasno, da če bi želeli videti vsak tweet kandidatov, bi enostavno sledili njim? #predsednik12	2012-11-07T10:18:51	0	0	T1	42095	L1	negative	private
4	*	tid.3118386786055127 04		male	@petrasovdat V priemerjavi s svojim predhodnikom nedvomno. Okoliščine in pogoji dela pa so mu bili vse prej kot naklonjeni.	2013-03-13T13:58:29	0	0	T1	42095	L1	negative	private
5	*	tid.3360546606936432 64		female	Delam, delam, delam, odstranil bom plevel, prekopal bom vrtiček, prepeval ves vesel... #gardening inspired by Palček Primož #nowsinging	2013-05-19T09:44:09	1	0	T1	42095	L1	positive	private
6	*	tid.3427181956663992 32		male	Čeferin: sodišče podlega javnemu mnenju. Ni samo obsodilna sodba tista, ki kaže na to, da pravna država funkcionira. #pogledislovenije	2013-06-06T19:02:39	1	3	T1	42005	L1	negative	private
7	*	tid.3526913780205486 08		male	Lažji med nogami: Osme #Glave s pregledom 3. sezone #GoT. Gobcamo @anzet @BokiNachbar @WiC_HmR @matevzluzar http://t.co/LWHogSK9nj	2013-07-04T07:32:31	2	4	T1	1.0	L1	neutral	private

Slika 1: Vzorčenje korpusa Janes-Syn (anonimizirani prikaz).

Za označevanje vzorca smo izbrali sistem odvisnostne skladnje JOS (Erjavec et al. 2010), ki je bil razvit posebej za slovenski jezik in v slovenskem prostoru uspešno uporabljen za označevanje učnega korpusa ssj500k (Krek et al. 2015). Na podlagi slednjega je bil v sklopu projekta Sporazumevanje v slovenskem

3 Kot vsi drugi rezultati projekta JANES je tudi Janes-Syn prosto dostopen za uporabo, zaradi tehničnih težav s pretvorbo med formati sicer v nekoliko skrajšani različici (4.000 pojavnic oz. 170 besedil). Kot podatkovna množica je na voljo na repozitoriju CLARIN.SI (Arhar Holdt et al. 2017, http://hdl.handle.net/11356/1086), iskanje po korpusu pa je mogoče tudi v konkordančniku noSkE: http://nl.ijs.si/noske/sl.cgi/corp_info?corpname=janes.syn.

jeziku[4] razvit tudi razčlenjevalnik za slovenščino (Dobrovoljc et al. 2012).[5] Vzorec 200 tvitov je bil s tem programom avtomatsko razčlenjen in nato uvožen v program za vizualizacijo drevesnic SSJ (avtor J. Brank, glej Sliko 2). Pripisane skladenjske oznake oz. povezave so bile nato ročno popravljene skladno z označevalnimi smernicami (Holozan et al. 2008), prilagoditve sistema specifikam nestandardnega jezika pa so bile zabeležene v nadgrajeni različici smernic (Arhar Holdt 2016). Dopolnitve so na petih ravneh: označevanje žanrsko specifičnih elementov; raba tujejezičnih prvin; obravnava eliptičnosti in fragmentarnosti jezika; nestandardna raba ločil; in druge skladenjske posebnosti, h katerim se vračamo v nadaljevanju prispevka.



Slika 2: Primer označenega tvita v Označevalniku SSJ.

V procesu vzorčenja in pretvorbe smo iz obravnave izpustili 4 besedila, končni nabor, o katerem pišemo v prispevku, obsega torej 196 tvitov. Vsi v nadaljevanju navedeni primeri besedil so, kot rečeno, normalizirani na besedni ravni (Čibej et al. 2016b); poleg tega uporabljamo različice z odstranjenimi nerelevantnimi žanrsko specifičnimi elementi (glej razdelek 2.1), saj se s tem izognemo navajanju uporabniških imen, ki bi lahko razkrila identiteto pišočih.[6] Razlike med izvornim besedilom in različico v prispevku prikazuje naslednji primer:

4 Projektna stran: www.slovenscina.eu.

5 Alternativna izbira, prav tako že preizkušena za slovenščino (Dobrovoljc in Nivre 2016, Dobrovoljc et al. 2016), bi bil sistem Universal Dependencies, katerega prednost je medjezikovna primerljivost rezultatov. Odločitev za sistem JOS (v danem trenutku in za dano nalogo) utemeljujejo: dobra predhodna seznanjenost s sistemom in obstoj označevalnih smernic, na katerih je bilo mogoče osnovati dopolnitve za nestandardni jezik, ter obstoj zmogljivega programa za označevanje in pregledovanje drevesnic, v katerem je bilo mogoče med delom raziskovati odločitve, aplicirane v korpusu ssj500k.

6 V primerih, kjer imena služijo kot nujni del ponazoritve, smo ohranili imena inštitucij ali znanih oseb iz sveta politike ali športa.

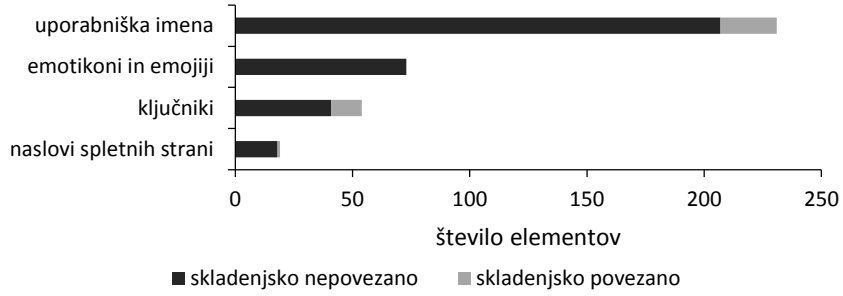
[A] izvorno besedilo: @union_pivo pizda, zakaj morm zmer uniona pit z laško kozarca? A to je taka politika al nimate kozarcev al vam je vseeno za kulturo piva??!!

[B] normalizacija na besedni ravni: @union_pivo pizda, zakaj moram zmeraj Uniona piti z Laško kozarca? A to je taka politika ali nimate kozarcev ali vam je vseeno za kulturo piva??!!

[C] izpust tviterskih elementov: pizda, zakaj moram zmeraj Uniona piti z Laško kozarca? A to je taka politika ali nimate kozarcev ali vam je vseeno za kulturo piva??!!

2.1 Žanrsko specifični elementi

Pri označevanju korpusa Janes-Syn je bila sprejeta odločitev, da bodo žanrsko specifični elementi (emotikoni, emojiji, naslovi spletnih strani, sklici na uporabniška imena in ključniki) povezani v strukturo drevesnic, kadar so del stavčne skladnje, sicer bodo izpuščeni iz obravnave, pri čemer se pri elementih na začetku in koncu tvitov nagibamo k nevključevanju. Rezultati kažejo, da je izpustljivih elementov bistveno več, v smislu povezovanja pa so relevantna predvsem uporabniška imena in ključniki, izjemoma tudi naslovi spletnih strani. Emotikoni in emojiji so v skladenjskem smislu v obravnavanem vzorcu vsi izpustljivi. Slika 3 prikazuje razmerje med povezanimi in nepovezanimi elementi v korpusu Janes-Syn.



Slika 3: Skladenjsko povezani in nepovezani tviterski elementi v Janes-Syn.

Med označevanjem je bilo v skladenjske strukture vključenih 24 uporabniških imen (10 % vseh). V Janes-Syn imena običajno nastopajo v vlogi osebka (14 primerov) ali predmeta (7 primerov), redkeje v prislovnih vlogah:

[1] Hmm, **@Delo** je zbrisalo razmislek ob Tugomerju da smo spet pod Franki, ker sta Bxl in Lux. frankovski središči, pa tudi Juncker je Frank ...?

[2] Zanimivo, da vaša **@strankaSDS** in njeni člani nabirajo točke z **@JJansaSDS** ter komunisti, ko jih sami toliko omenjate, več kot program.

[3] Ko berem komentarje pod tekstom o plebiscitu na **@rtvslo**, mi je žal, da večina njih, ne bo nikoli v rokah kakšnega polpismenega desetarja v JLA.

Kar se tiče ključnikov, je bilo skladenjsko povezanih 13 primerov (24 % vseh). Tudi ključniki v strukturah nastopajo samostojno ali kot del besednih zvez, pretežno v vlogi osebka (6 primerov) ali predmeta (5 primerov):

[4] @LakovicJaka hvala za vse kar si naredil za reprezentanco. Čeprav ti letos ni šlo brez tebe ne bi bili **#junaki.** Rečem ti lahko le SREČNO.

[5] Več kot očitno sta risa hotela ujeti **avion za #sochi.** Nista mogla zraven zaradi Massijeve prtljage.

[6] @bota112 reciva, da je **#krneki** … tako kot večina državne inf., z funkcionalno nepismenimi IT managerji - ki je **#btw** kazenska funkcija v **#du** :(

Primer skladenjsko vpetega naslova spletne strani je v obravnavanem vzorcu samo eden:

[7] jah saj za manj recimo tudi jaz ne bi peljal ;) samo dobro oni jih malo več peljejo :) probaj še **na http://t.co/YaVQdnaN5p** :)

Opisane označevalne odločitve so primerljive delu na angleščini pri Kaufmann in Kalita (2010), kjer so ločevanje skladenjsko relevantnih tviterskih elementov od nerelevantnih tudi avtomatizirali, in sicer z upoštevanjem besednorednih oznak konteksta: če uporabniškemu imenu denimo sledi veznik, predlog ali glagol, to s precejšnjo zanesljivostjo nakazuje njegovo skladenjsko relevantnost. Ključniki so v tem smislu nekoliko zahtevnejša naloga, saj so različnih besednih vrst.[7]

2.2 Tujejezični elementi

Kot ugotovljeno (Michelizza 2015: 161–65, Reher in Fišer 2018), vsebuje računalniško posredovana slovenščina (v splošnem) opazen delež tujejezičnih prvin. V označenem gradivu se slednji pojavljajo v 26 % tvitov, od tega 20 % iz angleščine in 6 % iz sorodnih južnoslovanskih jezikov. V primerih je mogoče opaziti različne stopnje prilagojenosti slovenskemu črkovanju in oblikoskladnji (Čibej et al. 2016b), potrdi se torej ugotovitev, da uporabniki tuje besedišče samoiniciativno »ne le oblikoslovno in skladenjsko, temveč tudi pisno podomačijo« (Jakop

7 Svojevrsten označevalni izziv, ki ga zaenkrat puščamo za prihodnost, predstavlja notranja struktura ključnikov, ki so lahko sestavljeni iz ene ali več besed v enem ali več različnih jezikih in z raznovrstnimi zapisovalnimi specifikami.

2008: 322). Kar se tiče dolžine tujejezičnih elementov, se pojavljajo tako posamezne besede (46 primerov) kot tudi besedne zveze (18 primerov) in daljše stavčne strukture (17 primerov).

[8] Optimiziran je tako, da čim manj porabi brez veze. Več kot pošiljaš, več bo porabil. Če **šeraš** slike in »**stickyje**« bo šlo orenk gor.

[9] Na Kongrescu smo, **meanwhile,** v popolnoma drugi dimenziji. Nek **hardcore band** se dere. sliši se boljše kot **basket, just so you know.**

[10] videla par sekund posnetka na Fb, ko ena (vstavi poljubno žaljivko) tepe parmesečnega dojenčka. **Da bog da joj majka ljubila sliku na banderi!**

Vpenjanje tujejezičnih elementov v skladenjsko označevanje je pri Janes-Syn potekalo tako, da se posamezne besede in tiste besedne zveze, kjer je odvisnost med elementi enostavno določljiva (in primerljiva s slovensko skladnjo, npr. zveza samostalnika in določujočega pridevnika), vpenja v skladenjsko drevo. Daljših struktur, zlasti stavčnih, ne vpenjamo oz. jih povezujemo kot fragmente neposredno na označevalno jedro. Označevalna izkušnja kaže, da je tovrstno ločevanje v praksi dovolj izvedljivo in smiselno, saj na tak način v drevesu ohranimo lastna imena tipa *Creative Commons, Candy Crush*, kot tudi občno besedišče različnih besednih vrst (npr. *fake, prpa, chatati*). Odločitev je v nadaljevanju treba preveriti na večji količini gradiva in v luči konkretiziranih označevalnih namenov po potrebi prilagoditi.

2.3 Nestandardna raba ločil

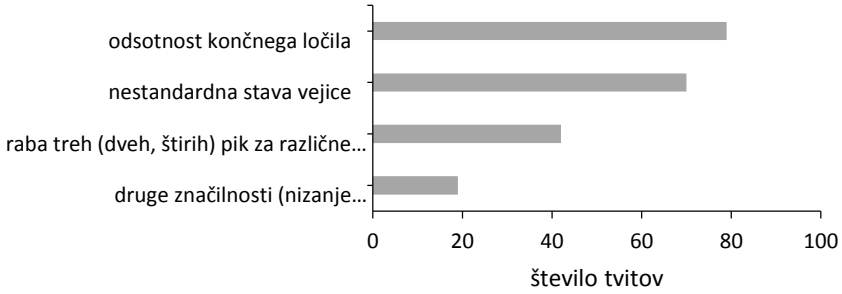
Tudi o nestandardni rabi ločil v računalniško posredovani slovenščini se je že pisalo, pogosto z normativističnega vidika (npr. Jakop 2008: 323, Dobrovoljc 2008: 309–311, Michelizza 2015: 133–140, Popič et al. 2016). V obravnavanem vzorcu je najti raznovrstne specifike rabe ločil v 69 % označenih tvitov. Pojavljajo se: izpuščanje končnih ločil oz. njihovo nadomeščanje z emotikoni, ključniki ipd.; odstopi od norme na ravni rabe vejice, predvsem njeno izpuščanje; raba treh (ali dveh, štirih) pik za nakazovanje premorov ali kot nadomestilo drugih ločil; in druge težave, npr. pri rabi ločil za poročani govor, upoštevanju razlik pri rabi vezajev in pomišljajev ter nizanju klicajev in/ali vprašajev za izražanje stopnje čustvene obarvanosti sporočila:

[11] nekje sem bral da obstajajo študije o smrtonosnosti cepljenj in hudi škodljivosti chemtrailov …

[12] težave z elektriko **…** pokličeš kolega ki te ne pusti na cedilu **…** spiješ enega ali dva **…** nice Saturday

[13] Janez.Janša se bo vrnil v velikem SLOGU, Z VSEM SLOVENSKIM NARODOM NA ČELU IN Z milijoni EVROV V ŽEPU ZA STORJENE MU KRIVICE, JA.!!

Številčni podatki za prisotnost navedenih elementov so prikazani na Sliki 4 (posamezni tvit lahko vsebuje eno ali več navedenih značilnosti).



Slika 4: Značilnosti glede rabe ločil v Janes-Syn.

Pri razčlenjevanju standardne slovenščine so ločila lahko koristna informacija za določanje stavčnih meja in razmerij med deli povedi. Kot je razvidno (tudi) iz obravnavanih podatkov, se je na ločila v računalniško posredovani komunikaciji mogoče zanašati v manjši meri. Značilnosti nestandardne rabe, ki so sistemske (npr. odsotnost končnih ločil pred emotikoni), je mogoče pri avtomatski predpripravi korpusnih besedil upoštevati, nekateri drugi načini rabe so manj predvidljivi. Glede na ugotovljeno se zdi, da je mogoče delo nadaljevati v dve smeri, bodisi s poskusi učenja razčlenjevanja brez upoštevanja ločil ali z vključitvijo koraka njihove normalizacije.[8]

2.4 Fragmentarnost jezika in izpusti

V podatkih korpusa Janes-Syn se pojavljajo tako krajšanja in izpusti kot drobljenje sporočila po vzoru govorjenega jezika (glej primer [12]). Tovrstne značilnosti, ki so jih identificirali mdr. tudi Kaufman in Kalita (2010) ter Schneider (2015) na angleških tvitih, pri nas pa Kranjc v spletnih klepetih (2003: 76), Kalin Golob v SMS-ih (2008: 292), Michelizza v blogih in Wikipediji (2015: 216–220), predstavljajo za skladenjsko označevanje poseben izziv.

Analiza označenega vzorca pokaže, da se elipsa pojavi v 20 primerih, večinoma kot izpust pomožnega glagola (11 primerov), redkeje izpust modalnega glagola

8 Druga možnost bi ponudila tudi možnost za izboljšavo orodij, kot so slovnični pregledovalniki, kot je bilo nakazano npr. v Holozan (2013) ter Kranjc in Robnik Šikonja (2015).

(2 primera), zaimka (2 primera), simbola (1 primer) ali polnopomenske besede, npr. v sklopu fraz, kot so *ni druge* ali *potem pa ti meni* (4 primeri). Rezultati so primerljivi izsledkom Goli et al. (2016), ki ugotavljajo, da je med krajšanji v 800 slovenskih tvitih na skladenjski ravni sicer bistveno manj posegov kot na drugih ravneh, prevladujejo pa predvsem izpusti pomožnika *biti*.

[14] Šele zdaj videl kakšna drama je bila v CONCACAF, ko so ZDA v zadnjih minutah priigrale Mehiki dvoboj z N. Zelandijo. Konec velikega rivalstva?

[15] tudi jaz .. zato pa prvo v glavi porihtati, da je šejk namesto obroka .. in ker je sladek, boš sčasoma zgubila sugar rush

[16] haha saj se mi je zdelo, premalo si na Štajerskem!!! potem pa ti meni o manjkajočih j-jih in i-jih v besedah. fff :D

Za označevanje skladnje in skladenjsko razčlenjevanje so izpusti večji problem, kadar vplivajo na strukturo drevesnice, tj. kadar gre za elemente, kot so npr. jedra besednih zvez ali stavčni povedek, na katere se tipično vežejo drugi elementi. Označevalni sistemi se z vprašanjem tovrstnih izpustov soočajo na dva načina, ena od možnosti je povišanje odvisnega elementa na mesto manjkajočega (Dobrovoljc in Nivre 2016), na drugi strani povezovanje fragmentov neposredno na označevalno jedro pohitri proces označevanja in odločanja (Kong et al., 2012), seveda ob določeni izgubi informacij. Ker sistem označevanja JOS to omogoča, se pri označevanju Janes-Syn odločamo za drugi način, v naslednjem koraku pa bi bilo smiselno primerjati rezultate obeh pristopov z vidika označevalnih stroškov v primerjavi s pridobitvijo oz. izgubo na ravni kakovosti rezultatov za določen raziskovalni namen.

3 BESEDNI RED RAČUNALNIŠKO POSREDOVANE SLOVENŠČINE

Do sedaj obravnavane skladenjske značilnosti imajo neposreden vpliv na proces označevanja, zato smo jih obravnavali v ločenem poglavju. Razen naštetega pa se v povezavi z računalniško posredovano slovenščino pojavljajo tudi druge ugotovitve. Med pogosteje obravnavanimi temami je vprašanje skladenjske kompleksnosti: Kranjc (2003: 76) denimo na jeziku spletnih klepetalnic opaža (tudi) strukturno zapletene večstavčne povedi, ob čemer so v rabi predvsem predmetni odvisniki. Dobrovoljc (2008: 309) zapiše, da je skladenjska zgradba e-poštnih sporočil »brez zapletenih povedi, veliko je kratkih stavkov, ki so povezani v poved brez veznikov«. Michelizza (2015: 213–234) zaključi, da skladnja blogov in wikipedijskih člankov ni okrnjena in da skladenjske specifike, kolikor jih je zaznati,

ne vplivajo na razumevanje pomena. Redkejši so izsledki o drugih skladenjskih lastnostih računalniško posredovane slovenščine, vključno z vprašanjem besednega reda, ki nas zanima v nadaljevanju prispevka.

Da je besedni red za dano raziskovalno področje relevantno vprašanje, so navedene študije sicer nakazale, niso pa tematike podrobneje analizirale. Kranjc (2003) v povzetku navaja, da se v jeziku spletnega klepeta kaže govorna forma v »spremenjenem besednem redu in neupoštevanju načela členitve po aktualnosti«, vendar se k vprašanju v sami razpravi ne vrne. Kalin Golob (2008: 292) med značilnostmi analiziranih SMS-ov omenja »spontani besedni red«, ki ob preostalih identificiranih značilnostih kaže »zapis po govoru«, vendar konkretni primeri niso diskutirani. Michelizza (2015: 228–230) se členitve po aktualnosti dotakne prek tipologije leksemov, ki se pogosto pojavljajo v remi, ne posveča pa se vprašanju samega besednega reda. Vprašanje besednega reda v slovenščini, tako nezaznamovanega kot zaznamovanega, sicer velja za zahtevno in slabo raziskano področje, na kar avtorji opozarjajo že od Breznika (1908) naprej. Kasnejša dela (npr. Toporišič 1967, Jug Kranjec 1981, Vidovič Muha 2000, Toporišič 2008, Toporišič 2004 – v nadaljevanju SS 2004) se osredotočajo na členitev po aktualnosti na eni strani in stalno stavo na drugi, predvsem v nizu določujočih pridevnikov ter naslonskem nizu. Kadar je v literaturi govor o zaznamovanosti besednega reda, se slednja obravnava predvsem na jeziku leposlovnih del, zato ugotovitve na obravnavo računalniško posredovane komunikacije niso neposredno prenosljive.

Sledeč Toporišičevi smernici, da je kriterij za ločevanje običajnega in zaznamovanega besednega reda »takrat, kadar ga tako ali drugače občutimo« (Toporišič 2008: 31), smo se odločili, da potencialno zaznamovanost besednega reda v korpusu Janes-Syn preverimo s sodelovanjem treh neodvisnih označevalcev, kot je razloženo v nadaljevanju prispevka. V raziskavi nas zanima, kako označevalci razumejo besednoredno zaznamovanost v računalniško posredovani slovenščini, kolikšno je ujemanje med njihovimi odločitvami, katere vrste besednorednih problemov se glede na pripisane oznake pojavljajo v podatkih in kako so ti problemi razporejeni glede na avtomatsko pripisano kategorijo jezikovne (ne)standardnosti (o metodologiji pripisa glej Ljubešić et al. 2018).

3.1 Označevanje in kategorizacija problemov

Označevalci so za svoje delo prejeli tabelo z besedili korpusa Janes-Syn in nalogo označiti tvite, v katerih se zdi besedni red z vidika standardne pisne slovenščine zaznamovan. Ob tem so primere morali tudi popraviti, kot bi jih, denimo, v procesu lektoriranja, vendar po principu minimalne intervencije, tj. samo s spremembo zaporedja besed (in po potrebi ločil, začetnic), ne pa s

spreminjanjem ostalih jezikovnih značilnosti. Smernice za označevanje problemov so nalogo namenoma opredeljevale ohlapno, saj je bil eden od ciljev raziskave preveriti intuitivnost »besednoredne zaznamovanosti« in ujemanje med označevalci glede slednje.

Za potrebe prispevka smo rezultate označevanja razvrstili v vsebinske skupine, in sicer od spodaj navzgor na osnovi gradiva. Kot je običajno za tovrstna razvrščanja, so se nekatere kategorije pokazale zelo hitro in so povsem jasno ločljive (npr. primeri z naslonkami na prvem mestu stavka, nesklonljivim levim prilastkom, vprašanja postavitve členka), medtem ko so druge težje določljive in deloma medsebojno prekrivne (npr. členitev po aktualnosti, vprašanja razporeditve stavčnih členov v stavku). Številčne rezultate v nadaljevanju je treba razumeti z upoštevanjem navedenih metodoloških značilnosti.

3.2 Rezultati

V procesu označevanja je bilo opravljenih 131 popravkov besednega reda v 94 različnih tvitih. Popravki so bili nato razvrščeni v 14 skupin. Pri kategorizaciji so bili upoštevani vsi primeri, ki so bili označeni vsaj enkrat, torej tudi tisti, ki jih je označil samo en od označevalcev, preostala dva ne. To dejstvo je pomembno, ker je ujemanje med označevalci za dano nalogo zelo nizko, k čemur se vrnemo v razdelku 3.3. Da je ponazoritev rezultatov jasnejša, v primere tvitov dodajamo podatek o tem, kateri od označevalcev je posamezni primer izpostavil kot zaznamovanega (*označ. 1, 2* ali *3*).

Rezultati označevanja prinašajo večji nabor primerov, kjer bi kot osrednji problem lahko izpostavili **členitev po aktualnosti,** tj. vprašanje razvrščanja informacij od znanega k manj znanemu ali postavljanje osrednje, najbolj bistvene informacije v remi na koncu stavka oz. povedi (SS 2004: 660). Brez poznavanja avtorjevega namena in besedilnega konteksta je v praksi (ne)zaznamovanost členitve po aktualnosti težje presojati. Tviti so v tem smislu dodatno problematični,[9] ker so izvorno lahko (ne pa nujno) del dialoške komunikacije oz. se nanašajo/sklicujejo na vsebine, ki pri označevanju gradiva niso bile več na voljo. Na drugi strani se vprašanje razporejanja delov povedi v temo, prehod in remo mestoma prekriva z drugimi skladenjskimi vprašanji, kot opisujemo v nadaljevanju. Kjer je bilo vzrok identificirane in popravljene težave mogoče pripisati v kako drugo skupino, so pri kategorizaciji le-te imele prednost. Na koncu je v skupini členitev po aktualnosti ostalo 23 primerov (17,6 % vseh), ki so primarno pomensko-poudarjalne narave:

9 V primerjavi s celovitimi in zaključenimi besedili, ki se običajno uporabljajo za zgled členitve po aktualnosti, prim. Trdinovo bajko o grofu in medvedu v SS 2004: 660.

[17] G: »Od kdaj je nama tako kul **hladno vreme?** po mojem zato, ker se lahko stiskava, brez da bi naju švic lepil en na drugega.« > (*Označ. 2*) G: »Od kdaj nama je hladno vreme **tako kul**? Po mojem zato, ker se lahko stiskava, ne da bi naju švic lepil enega na drugega.«

[18] glej, ni druge kot, da jo unfollowamo **vsi,** če bo še naprej toliko nesramna … P.s. jaz bi raje onega ta temnega > (*Označ. 1, 3*) glej, ni druge, kot da jo vsi **unfollowamo,** če bo še naprej tako nesramna … P.S. Jaz bi raje onega ta temnega.

Z omenjeno deloma prekrivna kategorija so popravki besednega reda **glagolskih besednih zvez s prislovno sestavino** (21 primerov, 16,0 % vseh); prekrivnost je pri primerih, kjer je prislovna sestavina v remi, torej bi izvorni besedni red bilo mogoče potencialno razumeti s stališča vsebinskega poudarjanja. Skupina vsebuje tako zveze glagola s prislovom (11 primerov) kot zveze glagola s predložno samostalniško zvezo (10 primerov), pri čemer v tem prispevku vprašanje (ne)obveznosti prislovnega dela (SS 2004: 592) puščamo ob strani:[10]

[19] začnite jih **spreobračati na polno** v tisto kar ne marajo. Ali pa kimate in jim date prav, naredite pa po svoje. > (*Označ. 1, 2, 3*) Začnite jih **na polno spreobračati** v tisto, česar ne marajo. Ali pa kimate in jim date prav, naredite pa po svoje.

[20] Kolega, čisti Ljubljančan, je to za joke zinil v lokalu v centru Lj.. Konec na urgenci, so se lj. idioti **med seboj sfajtali** > (*Označ. 1, 2, 3*) Kolega, čisti Ljubljančan, je to za joke zinil v lokalu v centru Lj. Konec na urgenci, lj. idioti so se **sfajtali med seboj.**

Z vsebinskim poudarjanjem je povezana tudi skupina, ki smo jo nekoliko pavšalno poimenovali **vrivanje stavčnih členov** (10 primerov, tj. 7,6 % vseh), ker gre za postavitev (pretežno) stavčnega osebka ali predmeta med dele sestavljenega povedka oz. med povedek in prislovna določila. Ta skupina prinaša primere zaznamovanega besednega reda, ki se v govoru izraža v stavčni intonaciji in poudarku (Jug Kranjec 1981):

[21] ne vem, da ni @RomanLeljak -a dobila **Udba** v kremplje? Že nekaj dni ne čivka. Kaj mislite? Bi bilo treba tiralico? > (*Označ. 1, 3*) Da ni @RomanLeljak -a dobila v kremplje **Udba?** Že nekaj dni ne čivka. Kaj mislite? Bi bilo treba tiralico?

[22] šetamo po Lj. in pride **Zoki** mimo, se ustavi pa da Emanuelu petko. ta malemu nič jasno. rečem to je Zoki kralj in se ta mali zadere: Zoki kralj! > (*Označ. 1, 3*) Šetamo po Lj. in pride mimo **Zoki,** se ustavi in da Emanuelu petko. Malemu nič jasno. Rečem: »To je Zoki kralj,« in mali se zadere: »Zoki kralj!«

10 Zdi se, da označevalci prislove pogosteje premikajo na mesto desno od glagola, predložne zveze pa levo, vendar je podatkov za posplošitve premalo. Prav tako ni dovolj podatkov za ugotavljanje potencialne drugačne stave načinovnih določil oz. prislovov (prim. Toporišič 1967: 257–258).

Podoben poudarek si je mogoče misliti tudi pri primerih, kjer se po tovrstni stavi **povedek pojavlja na koncu stavka** (8 primerov, tj. 6,1 % vseh). Te primere smo obdržali kot ločeno skupino, ker je zanje značilno, da so za označevalce posebej opazni in da se obenem pojavljajo izključno v tvitih, avtomatsko označenih za jezikovno nestandardne (več o tem v razdelkih 3.3 in 3.4):

[23] Noben ISP ne ponuja dnevnega / tedenskega zakupa ? To bi bilo kul. jaz nimam TV-ja, zdajle bi pa plačala, da bi lahko **tekme** gledala. > (*Označ. 1, 2, 3*) Noben ISP ne ponuja dnevnega ali tedenskega zakupa ? To bi bilo kul. Jaz nimam TV-ja, zdajle bi pa plačala, da bi lahko gledala **tekme.**

[24] jaz : Bu. sodelavec : *krik, ki ga je **cela bajta** slišala * jaz :* nekontroliran izbruh smeha, ki ga še vedno cela bajta posluša* > (*Označ. 1, 2*) Jaz : Bu. Sodelavec : * Krik, ki ga je slišala **cela bajta.** * Jaz : *Nekontroliran izbruh smeha, ki ga še vedno posluša cela bajta.*

Ločeno so obravnavani tudi drugi primeri, vezani na mesto **povedka** oz. njegovih delov (7 primerov, tj. 5,3 % vseh). Gre za primere, kjer se del povedka, npr. modalni glagol, pojavlja na prvem mestu stavka, ali pa je popravek označevalca vezan na postavitev povedka v glavnem stavku, ki sledi odvisniku:

[25] Pa še 3 - 4 leta nazaj sem bil tako ponosen na svojo kondicijo (resda zlasti kar se tiče hoje) **Moram malo** spremeniti način življenja! * > (*Označ. 1*) Pa še 3-4 leta nazaj sem bil tako ponosen na svojo kondicijo (resda zlasti kar se tiče hoje). **Malo moram** spremeniti način življenja!

[26] če vprašate mene (in mislim, da se @anakobal_kobe strinja), Tini do res odličnega rezultata **manjka** le malo teže na spodnji smučki. > (*Označ. 3*) Če vprašate mene (in mislim, da se @anakobal_kobe strinja), **manjka** Tini do res odličnega rezultata le malo teže na spodnji smučki.

Za razliko od do sedaj naštetih kategorij, ki so deloma prekrivne, je enoznačna za identifikacijo kategorija problemov s **postavitvijo členka** (15 primerov oz. 11,5 % vseh). Popravki so v primerih, kjer členek v izvirniku (po presoji označevalca) ne stoji pred delom stavka, ki naj bi ga pomensko modificiral (SS 2004: 675):

[27] padanje se je **tudi** pričelo že kar nekaj časa nazaj in je dokaj konstantno od cca. začetka UA krize > (*Označ. 1*) **Tudi** padanje se je pričelo že kar nekaj časa nazaj in je dokaj konstantno od cca. začetka krize UA.

[28] Pri nas v Ustavi pa **seveda** vladavino imamo. Demokracija je ena od oblik vladavine nad državno - kapitalsko državo! > (*Označ. 2*) Pri nas v Ustavi pa vladavino **seveda** imamo. Demokracija je ena od oblik vladavine nad državno - kapitalsko državo!

Prav tako enostavno ločljiva skupina so primeri, kjer se **naslonka** pojavlja na prvem mestu stavka (15 primerov oz. 11,5 % vseh), kar se v standardnem jeziku utemeljuje v primerih izpusta vezniške ali naglašene sestavine pred njo (SS 2004: 676). V obravnavanem vzorcu se na atipičnem prvem mestu najpogosteje pojavita pomožni glagol (9 primerov) in povratni zaimek (5 primerov):

[29] Kolega, čisti Ljubljančan, je to za joke zinil v lokalu v centru Lj.. Konec na urgenci, **so se lj. idioti** med seboj sfajtali. > (*Označ. 1, 2, 3*) Kolega, čisti Ljubljančan, je to za joke zinil v lokalu v centru Lj. Konec na urgenci, **lj. idioti so se** sfajtali med seboj.

[30] so Ready To Go. **Se veselim** dobre letine pristankov. > (*Označ. 1, 2, 3*) So Ready To Go. **Veselim se** dobre letine pristankov.

Na vprašanje naslonskega niza se veže skupina primerov (8 primerov oz. 6,1 %), kjer označevalci popravijo **stavo naslonk** ob veznikih. Tipičen primer je veznik *pa*, ki naj bi nezaznamovano stal pred prostimi naslonkami (SS 2004: 676), v gradivu pa se pojavlja na mestu za pomožnikom (5 primerov):

[31] ma dva dedca čvekata na polno, vsi ostali **smo pa** hoteli dremati. potem se pa začneta meniti o kinih pa o filmih... > (*Označ. 1, 2, 3*) Ma, dva dedca čvekata na polno, vsi ostali **pa smo** hoteli dremati. Potem se pa začneta meniti o kinih pa o filmih ...

[32] Pomagajte mi sem sin odvisnika in odvisnice. oče je že 2 leti na Besedovnjaku mama **je pa** na Candy Crushu in Flappy Birdu. > (*Označ. 2*) Pomagajte mi, sem sin odvisnika in odvisnice. Oče je že 2 leti na Besedovnjaku, mama **pa je** na Candy Crushu in Flappy Birdu.

Ločeno smo obravnavali primere z **neujemalnim levim prilastkom** (5 primerov, 3,8 % vseh), in sicer tudi tiste, kjer je prilastek kratični. Pri slednjih gre sicer primarno za vprašanje zapisovanja z vezajem ali brez njega (Arhar Holdt in Dobrovoljc 2016), ker pa je bila označevalna naloga usmerjena v besedni red, so označevalci v tem okviru identificirani problem tudi reševali:

[33] pizda, zakaj moram zmeraj Uniona piti z **Laško kozarca**? A to je taka politika ali nimate kozarcev ali vam je vseeno za kulturo piva??!! > (*Označ. 1, 3*) Pizda, zakaj moram zmeraj Union piti z **kozarca Laško**? A je to taka politika ali nimate kozarcev ali vam je vseeno za kulturo piva??!!

[34] Plus, premikanje na **SD kartico** sem probala že stokrat. Tudi telefon to ponudi kot možnost. Rezultat? »Ni predmetov za premikanje« (*Označ. 1, 3*) Plus, premikanje na **kartico SD** sem probala že stokrat. Tudi telefon to ponudi kot možnost. Rezultat? »Ni predmetov za premikanje.«

Ločena skupina so še primeri (4 primeri, 3,1 % vseh), kjer je bila kot zaznamovana označena stava **osebnega zaimka,** ki je v izvornem besedilu v izpostavljeni poziciji:

[35] G : »Od kdaj je **nama** tako kul hladno vreme? po mojem zato, ker se lahko stiskava, brez da bi naju švic lepil en na drugega.« > (*Označ. 1, 2*) G : »Od kdaj **nama je** tako kul hladno vreme? Po mojem zato, ker se lahko stiskava, ne da bi naju švic lepil en na drugega.«

[36] Meni sta pa oba tako kjut v tem dialogu, da vama bom zdaj kar takole na daljavo rekla: jaz imam pa **vaju** oba rada! Eto! > (*Označ. 2*) Meni sta v tem dialogu oba tako kjut, da vama bom zdaj kar takole na daljavo rekla : Jaz **vaju** imam pa oba rada!

Redkeje sta zastopani skupini primerov, kjer je težava s **postavitvijo določujočega elementa** znotraj besedne zveze (2 primera oz. 1,5 % vseh) ali **zaporedja elementov znotraj glagolske zveze** (3 primeri ali 2,3 % vseh):

[37] ja no, meni se to tudi skozi dogaja! še večkrat pa s senčkami na Instagramu. naredim smokey zgleda pa **čisto nekaj nežnega** > (*Označ. 1, 2*) Ja no, meni se to tudi skozi dogaja! Še večkrat pa s senčkami na Instagramu. Naredim smokey, a zgleda **nekaj čisto nežnega.**

[38] Pripomba »PS« mi je všeč. O tej temi morava še kdaj kaj reči PS 2 pa ne, nisem videla. **Moram poguglati** > (*Označ. 1*) Pripomba »P. S.« mi je všeč. O tej temi morava še kdaj kaj reči. P. S. 2 pa ne, nisem videla. **Poguglati moram.**

Ostanejo še primeri (9 primerov, 6,9 % vseh), ki smo jih uvrstili v skupino **Drugo**. Gre za popravke posameznih označevalcev, izvirajočih iz različnega razumevanja vsebine obravnavanih tvitov in tudi zadane naloge, kot prikazuje spodnji popravek besednega reda v citirani pesmi:

[39] Kdo pozna to pesmico? "Takole se prične: Po belih in črnih tipkah tja v svet odjadralo skupaj **je prstkov deset** ..." > (*Označ.* 3) Kdo pozna to pesmico? "Takole se prične: Po belih in črnih tipkah **je** tja v svet odjadralo skupaj **deset prstkov** ..."

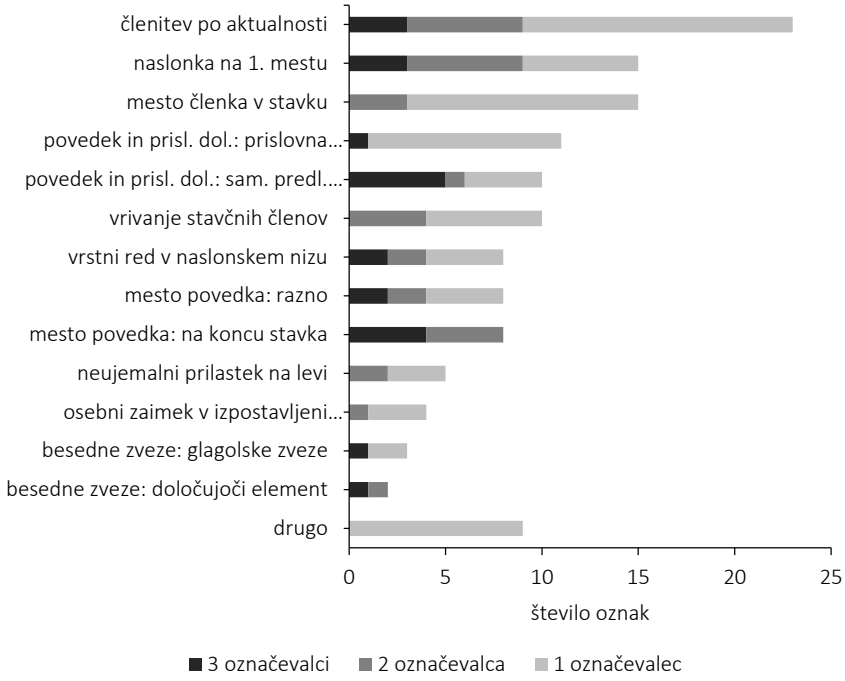
3.3 Ujemanje med označevalci

Ko govori o stilni vrednosti besednega reda, zapiše Toporišič (2008: 30) takole:

> Da si besede v govoru sledijo po določenem zaporedju, je znano. Za taka tipična besedna zaporedja imamo natančno razvit čut, ki nam pove, ali

> govoreči razvršča besede primerno navadam danega jezika ali ne. Če jih kdaj ne razvršča pravilno, nam navadno ni prav nič težko govorečemu besedni red popraviti.

Raziskava, ki jo predstavljamo v tem prispevku, prinaša povsem drugačne rezultate. Od 131 popravkov besednega reda v korpus Janes-Syn je samo 23 primerov (17,6 %) takih, da so jih primerljivo označili vsi trije označevalci. 32 primerov (24,4 %) sta primerljivo označila dva od treh označevalcev, preostalih 77 (58,8 %) označb pa je individualnih. Ujemanje med označevalci v posamezni kategoriji kaže Slika 5.



Slika 5: Ujemanje med označevalci po kategorijah problemov.

Zelo povedna ugotovitev je, da se pri nobeni od kategorij ne pojavljajo izključno primeri, kjer bi bilo označevanje enotno. Tudi če iščemo primere, ki sta jih enotno označila vsaj dva od treh, je mogoče izpostaviti samo dve kategoriji: primere, kjer se povedek pojavlja na koncu stavka (*zdajle bi pa plačala, da bi lahko* ***tekme gledala***) in manjšo skupino problemov pri postavitvi določujočega elementa znotraj besednih zvez (*zgleda pa* ***čisto nekaj nežnega***). V smislu same pogostnosti označenih primerov je nekoliko višjo skladnost opaziti še pri primerih, kjer je v ospredju mesto prislovnega določila v obliki predložne samostalniške besedne zveze (*začnite jih* ***spreobračati na polno*** *v tisto kar ne marajo*), pri členitvi po

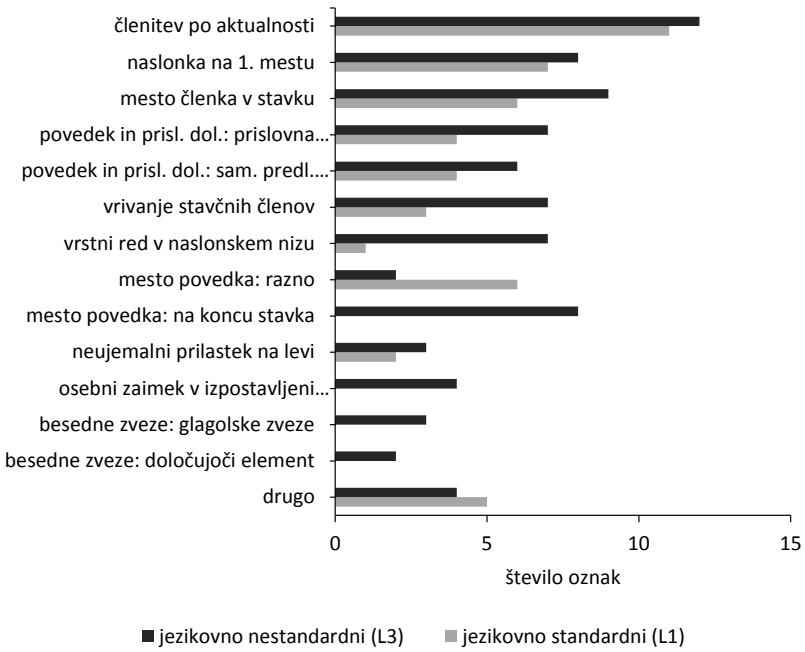
aktualnosti (*glej, ni druge kot, da jo unfollowamo **vsi***) in primerih, kjer se naslonka pojavlja na prvem mestu stavka (***so se** lj. idioti med seboj sfajtali*). Veliko individualnih oznak imajo na drugi strani – poleg kategorije členitve po aktualnosti, kjer so pričakovane – še kategorija mesta člena v stavku (*padanje se je **tudi** pričelo že kar nekaj časa nazaj*) in mesto prislovnega določila v obliki prislova (*jaz sem sicer bil tako zaspan, da sem skoraj **skup padel***).

Da ujemanje med označevalci za določene kategorije ne bo popolno, smo pričakovali, visoka raven in prisotnost odstopanj pri skoraj vseh kategorijah pa nas je vseeno presenetila. Neujemanje bi lahko deloma pripisali neizkušenosti označevalcev za tovrstno delo, čeprav bi glede na Toporišičev citat intuicija naravnega govorca (pri označevalcih gre dodatno za jezikoslovce) pri presojanju besednega reda morala voditi v primerljive rezultate.[11] Druga možna razlaga za dobljeni rezultat bi bila, da se pri obravnavi jezikovnega gradiva, ki prinaša opazen nabor specifičnih, tudi nestandardnih jezikovnih značilnosti, presojanje zaznamovanosti besednega reda prilagodi. Kar bi bilo vidno zaznamovano v sopostavitvi s standardnim jezikovnim gradivom, v množici tvitov, ki prinašajo drugačen tip jezika na vseh ravninah, postane manj opazno. Ali pa se označevalci, podobno kot učitelji pri popravljanju šolskih pisnih izdelkov, odločajo samo še za popravke, ki so res temeljni, manj moteče pa samodejno preskočijo. Samodejno prilagajanje kriterijev je vsekakor vznemirljivo vprašanje, ki bi se mu bilo smiselno v prihodnosti natančneje posvetiti. Na tem mestu pa je potrebno poudariti še, da je označevalna naloga v praksi zahtevnejša, kot se zdi po prebiranju že kategoriziranih rezultatov, kjer so za prikaz namenoma izbrani najbolj reprezentativni, jasni in nedvoumni primeri.

3.4 Nestandardno specifične kategorije

V okviru projekta JANES je bila razvita metodologija za avtomatski pripis tehnične in jezikovne (ne)standardnosti besedilom računalniško posredovane komunikacije. Značilke za določanje (ne)standardnosti posegajo na znakovno raven, npr. nestandardno rabo ločil in presledkov, ponovitev znakov, razmerje med abecednimi in neabecednimi znaki itd., in besedno raven, npr. rabo nestandardnega oz. atipičnega besedišča, zapis z velikimi ali malimi črkami itd. (Ljubešić et al. 2015). Skladenjska raven oz. raven besednega reda v metodologijo ni vključena, zato je zanimiva primerjava zastopanosti identificiranih besednorednih problemov glede na avtomatsko oceno (ne)standardnosti omenjenih dveh ravnin. Primerjave z jezikovno (ne)standardnostjo prikazuje Slika 6.

11 Za preverjanje tega izhodišče smo namenoma ohranili označevalne smernice ohlapne, brez povzetka besednorednih pravil oz. ugotovitev iz referenčnih jezikovnih priročnikov, na kaj naj bodo označevalci pozorni. Če bi označevalne poskuse nadaljevali, bi lahko smernice dopolnili, vendar se s tem proces spremeni v pripisovanje kategorij od zgoraj navzdol, s čimer tvegamo, da določenih pojavljajočih se jezikovnih realnosti v gradivu ne identificiramo.



Slika 6: Zastopanost kategorij glede na pripisani oznaki za jezikovno (ne)standardnost.

Avtomatski pripis jezikovne (ne)standardnosti seveda ni povsem zanesljiv, kljub temu pa rezultati kažejo zanimivo sliko, iz katere je mogoče sklepati o razmerju med zaznamovanimi in potencialno nestandardnimi značilnosti korpusa Janes-Syn. Kot nestandardni so denimo označeni vsi primeri, kjer se povedek pojavlja na koncu stavka (*zdajle bi pa plačala, da bi lahko* ***tekme gledala***), kjer se pojavlja izpostavljanje osebnih zaimkov (*od kdaj* ***je nama*** *tako kul hladno vreme?*) ter oznake razvrstitve elementov znotraj besednih zvez (***Moram poguglati***). K skupinam, kjer nestandardni primeri znatno prevladujejo, je mogoče prišteti še težave vrstnega reda v naslonskem nizu (*vsi ostali* ***smo pa*** *hoteli dremati*) in vrivanje stavčnih členov v sestavljeni povedek (*šetamo po Lj. in pride* ***Zoki*** *mimo*). Najbolj uravnoteženi glede na oznake (ne)standardnosti sta kategoriji členitve po aktualnosti (*glej, ni druge kot, da jo unfollowamo* ***vsi***) in naslonke na prvem mestu stavka (***so se*** *lj. idioti med seboj sfajtali*). Edina kategorija, kjer je prisotnih več označb iz jezikovno standardnih primerov, je atipično mesto povedka, predvsem na prvem mestu stavka (***Moram se danes zvečer*** *dol usesti*).

S primerjavo Slik 5 in 6 vidimo, koliko je določena identificirana besednoredna značilnost opazna v svoji zaznamovanosti ter kakšna je njena distribucija glede na

avtomatsko identificirane nestandardne prvine na drugih jezikovnih ravneh. Za najpogosteje zastopano kategorijo, členitev po aktualnosti, je značilno, da prinaša pretežno pomensko-poudarjalne probleme. To dejstvo pojasni visok delež individualnih označb ter enakomerno pojavljanje v jezikovno nestandardnih, kot tudi standardnih tvitih. Visoko število oznak na drugi strani priča o tem, da izbira centralne vsebinske točke pri upovedovanju ni enoznačna naloga oz. da je možnih več interpretacij, ki pa s stališča označevalcev niso enakovredno nevtralne. Rezultati so torej zelo zanimivi tudi v luči konceptov proste in stalne stave: v rezultatih se pojavlja veliko označb na ravni proste stave, na drugi strani pa neskladnosti z jezikoslovno identificiranimi »stalnostmi« niso označevane dosledno.

Zanimiva ugotovitev je, da se med najpogosteje označenimi značilnostmi pojavlja raba naslonke na prvem mestu stavka. Distribucija pojavitev na drugi strani kaže, da se takšna stava pojavlja enakomerno v nestandardnih ter standardnih tvitih. To jezikovno značilnost bi bilo torej mogoče (nekoliko provokativno) izpostaviti kot tisto, ki je v komunikaciji, ki skuša biti standardna, najboljši indikator, da temu ni tako. Prav tako zanimiva je kategorija postavljanja členkov, ki je (vsaj v teoriji) precej enoznačno, v podatkih pa se kaže precejšnja označevalna permisivnost, skupaj z visokim deležem pojavitev v standardnih besedilih ob sicer prevladujočih nestandardnih. Na tej osnovi lahko oblikujemo tezo, da v jezikovni rabi stava členka manj vpliva na razumevanje pomena, kot bi si morda mislili. Naj členek stoji ob delu stavka, ki ga modificira, ali nekje drugod v neposredni bližini, v sporočilu ga osmislimo po principu največje verjetnosti, skladenjsko neustrezne pozicije pa pri tem pogosto sploh ne opazimo.

Obe izpostavljeni vprašanji bi bilo v razpravi mogoče povezati z značilnostmi govorjenega jezika, za kar pa bi bilo treba vključiti podatke iz govornega korpusa, kar presega domet poglavja. Za primerjavo in nadaljnje analize so relevantne tudi identificirane kategorije, kjer si je mogoče misliti poseben stavčni poudarek v primeru, da bi bilo besedilo govorjeno, pri čemer se osebek in predmet izpostavljata na atipičnih pozicijah. Glede na rezultate raziskave se stava povedka na zadnje mesto stavka, tj. za (predvideno) intonacijsko izpostavljeni osebek ali predmet, pokaže kot najboljši indikator besednoredne nestandardnosti: pojavlja se izključno v besedilih, označenih za nestandardne, obenem pa so ti primeri za označevalce opazno zaznamovani.

4 SKLEP

V prispevku smo opisali pripravo korpusa skladenjsko označenih tvitov Janes-Syn in odločitve, ki smo jih pri tem sprejeli glede označevanja nestandardnih jezikovnih značilnosti. Korpus je mogoče uporabiti kot učno množico za učenje

razčlenjevanja računalniško posredovane slovenščine. Prvi poskus, na osnovi katerega bo mogoče pripraviti večjo količino ročno pregledanega gradiva, bo izpostavil morebitne pomanjkljivosti glede označevanja in omogočil optimizacijo smernic. Izbiro jezikovnospecifičnega označevalnega sistema JOS utemeljuje obstoj in lahka dostopnost potrebne infrastrukture za izvedbo dane naloge, v prihodnje pa bo treba več pozornosti posvetiti pretvorljivosti in izmenljivosti podatkov. Predvideno je, da se bodo v prihajajočem obdobju vprašanja sistema označevanja rešila na ravni obravnave standardnega jezika, nujno pa je zagotoviti, da se odločitve aplicirajo tudi na nestandardni ravni.

Raziskava besednega reda v korpusu Janes-Syn je potrdila, da gre za tematiko, ki brez dvoma potrebuje dodatno jezikoslovno pozornost. Izkazalo se je, da besednoredne zaznamovanosti označevalci ne razumejo enotno, da so njihovi zaznamki v več kot pol primerih individualni, da pogosto identificirajo zaznamovanost na ravni proste stave in obenem izpuščajo besednoredne atipičnosti na ravni stalne stave, vključno z jezikoslovno jasno definiranimi, npr. stava člena ob modificirani del stavka ali razvrstitev naslonk v nizu. Pogled na podatke z vidika avtomatsko pripisane oznake jezikovne (ne)standardnosti razkrije dodatne ugotovitve v razmerju zaznamovanosti in nestandardnosti, ki bi jih bilo treba v nadaljevanju raziskati z vključitvijo podatkov govorjenega jezika.

Kategorizacija označenih značilnosti osvetljuje področje skladnje računalniško posredovane slovenščine in je relevantno izhodišče za podrobnejše analize identificiranih problemov. Izsledki omogočajo premik razumevanja prototipske besednoredne zaznamovanosti in raziskovalnega fokusa od primerov, kot sta *domovina naša* in *dedec prebrisani* (Toporišič 2008: 31), in v splošnem iz konteksta literarnega jezika v realnost, ki jo izkazuje vsakdanja jezikovna produkcija širše populacije. V tem smislu je jasno, da je edini pristop, ki lahko v resnici stre oreh besednega reda v slovenščini, korpusnojezikoslovni. V nadaljevanju je treba zagotoviti skladenjsko označenost relevantnega korpusnega gradiva (referenčni korpus, govorni korpus, korpus računalniško posredovane komunikacije) ter podobnosti in razlike ugotoviti statistično, s premišljeno metodo, ki upošteva označevalne specifike in sestavo obravnavanih virov. Delo na projektu JANES, ki smo ga predstavili v tem prispevku, je korak v opisano želeno smer.

Zahvala

Priprava korpusa Janes-Syn je ekipno delo, pri katerem gre zahvala za izvedbo predvsem Tomažu Erjavcu. Pri označevanju besednega reda sta sodelovali Aleksandra Rajković in Polona Logar.

Literatura

Arhar Holdt, Špela in Kaja Dobrovoljc, 2015: Zveze samostalnika z nesklonljivim levim prilastkom v korpusih Janes in Kres. Fišer, Darja (ur.): *Zbornik konference Slovenščina na spletu in v novih medijih*. Ljubljana: Znanstvena založba Filozofske fakultete. 4–9.

Arhar Holdt, Špela in Kaja Dobrovoljc, 2016: Vrednost korpusa Janes za slovensko normativistiko. *Slovenščina 2.0* 4/2. 1–37.

Arhar Holdt, Špela, 2016: *Smernice za označevanje z odvisnostnim sistemom JOS: nestandardna slovenščina, v1.0*. Ljubljana: Specifikacije projekta Jezikoslovna analiza nestandardne slovenščine. Dostop: http://nl.ijs.si/janes/wp-content/uploads/2014/09/Janes-skladnja-v1.0.pdf

Arhar Holdt, Špela, Tomaž Erjavec in Darja Fišer, 2017*: CMC training corpus Janes-Syn 1.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1086.

Breznik Anton, 1908. Besedni red v govoru. *Dom in svet* 21. 258–267.

Chanier, Thierry, Céline Poudat, Benoit Sagot, Georges Antoniadis, Ciara R. Wigham, Linda Hriba, Julien Longhi in Djamé Seddah, 2014: The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. *Journal for Language Technology and Computational Linguistics* 29/2. 1–30.

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

Čibej, Jaka, Darja Fišer in Tomaž Erjavec, 2016a: Normalisation, Tokenisation and Sentence Segmentation of Slovene Tweets. *Proceedings of the Workshop on Normalisation and Analysis of Social Media Texts (NormSoMe)*. Portorož: ELRA. 5–10.

Čibej, Jaka, Špela Arhar Holdt, Tomaž Erjavec in Darja Fišer, 2016b: Razvoj učne množice za izboljšano označevanje spletnih besedil. Erjavec, Tomaž in Darja Fišer (ur.): *Proceedings of the Conference on Language Technologies and Digital Humanities.* Ljubljana: Znanstvena založba Filozofske fakultete. 40–46.

Dobrovoljc Helena, 2008: Jezik v e-poštnih sporočilih in vprašanja sodobne normativistike. Košuta, Miran (ur.): *Slovenščina med kulturami, Zbornik Slavističnega društva Slovenije 19.* Celovec: Slavistično društvo Slovenije. 295–314.

Dobrovoljc, Kaja in Joakim Nivre, 2016: The Universal Dependencies Treebank of Spoken Slovenian. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC '16).* Portorož. 1566–1573.

Dobrovoljc, Kaja, Tomaž Erjavec in Simon Krek, 2016: Pretvorba korpusa ssj500k v Univerzalno odvisnostno drevesnico za slovenščino. *Proceedings of the Conference on Language Technologies and Digital Humanities.* Erjavec, Tomaž in Darja Fišer (ur.): *Proceedings of the Conference on Language Technologies and Digital Humanities.* Ljubljana: Znanstvena založba Filozofske fakultete. 190–192.

Dobrovoljc, Kaja, Simon Krek in Jan Rupnik, 2012: Skladenjski razčlenjevalnik za slovenščino. Erjavec, Tomaž in Jerneja Žganec Gros (ur.): *Zbornik Osme konference Jezikovne tehnologije.* Ljubljana: Institut Jožef Stefan. 42–47.

Erjavec, Tomaž, Darja Fišer, Simon Krek in Nina Ledinek, 2010: The JOS linguistically tagged corpus of Slovene. *LREC 2010, 7th International Conference on Language Resources and Evaluations*. Valletta. 1806–1809.

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

Fišer, Darja, Tomaž Erjavec in Nikola Ljubešić, 2017: The compilation, processing and analysis of the Janes corpus of Slovene user-generated content. Wigham, Ciara R. in Gudrun Ledegen (ur.): *Corpus de communication médiée par les réseaux: construction, structuration, analyse*. Collection Humanités Numériques. Paris: L'Harmattan. V tisku.

Foster, Jennifer, Ozlem Cetinoglu, Joachim Wagner, Joseph Le Roux, Stephen Hogan, Joakim Nivre, Deirdre Hogan in Josef van Genabith, 2011: #hardtoparse: Pos tagging and parsing the twitterverse. Analyzing Microtext. *Papers from the 2011 AAAI Workshop.* 20–25.

Gantar, Polona, 2011: Slovnični in pomenski opisi v leksikalni bazi za slovenščino. Marušič, Franc in Rok Žaucer (ur.): *Zbornik prispevkov s simpozija 2011.* Nova Gorica: Univerza. 17–27.

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

Goli, Teja, Eneja Osrajnik in Darja Fišer, 2016: Analiza krajšanja slovenskih sporočil na družbenem omrežju Twitter. Erjavec, Tomaž in Darja Fišer (ur.): *Proceedings of the Conference on Language Technologies and Digital Humanities.* Ljubljana: Znanstvena založba Filozofske fakultete. 77–82.

Holozan, Peter, Simon Krek, Matej Pivec, Simon Rigač, Simon Rozman in Aleš Velušček, 2008: *Specifikacije za učni korpus.* Kamnik: Projekt »Sporazumevanje v slovenskem jeziku« ESS in MŠŠ. http://www.slovenscina.eu/Vsebine/Sl/Kazalniki/K2.aspx

Holozan, Peter. 2013: Uporaba strojnega učenja za postavljanje vejic v slovenščini. *Uporabna informatika* 21/4. 196–209.

Jakop, Nataša, 2008: Pravopis in spletni forumi – kva dogaja? Košuta, Miran (ur.): *Slovenščina med kulturami, Zbornik Slavističnega društva Slovenije 19.* Celovec: Slavistično društvo Slovenije. 315–327.

Jug Kranjec, Hermina, 1981: O pomenski in stilni vlogi besednega reda pri oblikovanju sporočilne perspektive povedi. *Jezik in slovstvo* 42/2-3. 37–42.

Kalin Golob, Monika, 2008: SMS-sporočila treh generacij. Košuta, Miran (ur.): *Slovenščina med kulturami, Zbornik Slavističnega društva Slovenije 19.* Celovec: Slavistično društvo Slovenije. 283–294.

Kaufmann, Max in Jugal Kalita, 2010: Syntactic normalization of twitter messages. *International conference on natural language processing, Kharagpur, India.*

Kranjc, Anja in Marko Robnik Šikonja, 2015: Postavljanje vejic v slovenščini s pomočjo strojnega učenja in izboljšanega korpusa Šolar. Fišer, Darja (ur.): *Zbornik konference Slovenščina na spletu in v novih medijih.* Ljubljana: Znanstvena založba Filozofske fakultete. 38–43.

Kranjc, Simona, 2003: Skladenjska analiza besedil, ki nastajajo v računalniških klepetih. Požgaj Hadži, Vesna (ur.): *Zbornik referatov z Drugega slovensko-hrvaškega slavističnega srečanja.* Ljubljana: Oddelek za slavistiko, Filozofska fakulteta. 69–82.

Kong, Lingpeng, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer in Noah A. Smith, 2014: A dependency parser for tweets. *Proc. of EMNLP*. Doha, Qatar. 1001–1012.

Krek, Simon, Kaja Dobrovoljc, Tomaž Erjavec, Sara Može, Nina Ledinek in Nanika Holz, 2015: *Training corpus ssj500k 1.4.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1052

Krek, Simon, 2015: Standardni in knjižni jezik – drugi poskus. Smolej, Mojca (ur.): *Slovnica in slovar – aktualni jezikovni opis (Obdobja 34)*. Ljubljana: Znanstvena založba Filozofske fakultete. 401–407.

Ledinek, Nina, 2014: *Slovenska skladnja v oblikoskladenjsko in skladenjsko označenih korpusih slovenščine*. Ljubljana: Založba ZRC.

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

Može, Sara, 2013: *FrameNet in večjezičnost: kontrastivna analiza glagolov premikanja v slovenščini in angleščini.* Doktorska disertacija. Ljubljana: Filozofska fakulteta.

Michelizza, Mija, 2015: *Spletna besedila in jezik na spletu.* Ljubljana: Založba ZRC.

Myslin, Mark in Stefan T. Gries, 2010: k dixez? A corpus study of Spanish Internet orthography. *Literacy and Linguistic Computing* 25/1. 85–104.

Popič, Damjan, Darja Fišer, Katja Zupan in Polona Logar, 2016. Raba vejice v uporabniških spletnih vsebinah. Erjavec, Tomaž in Darja Fišer (ur.): *Proceedings of the Conference on Language Technologies and Digital Humanities.* Ljubljana, Slovenia. 149–153.

Reher, Špela in Darja Fišer, 2018: Kodno preklapljanje v objavah slovenskih uporabnikov Twitterja. Fišer, Darja (ur.): *Viri, orodja in metode za analizo spletne slovenščine*. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani. 294–323.

Schneider, Nathan, 2015: What I've learned about annotating informal text (and why you shouldn't take my word for it). *The 9th Linguistic Annotation Workshop held in conjuncion with NAACL 2015*. 152–157.

Stabej, Marko, Helena Dobrovoljc, Simon Krek, Polona Gantar, Damjan Popič, Špela Arhar Holdt, Darja Fišer in Marko Robnik Šikonja, 2016: Slovenščina Janes: pogovorna, nestandardna, spletna ali spretna? *Slovenščina 2.0* 4/2. 100–126.

Storrer, Angelika, 2013: Sprachverfall durch internetbasierte Kommunikation? Linguistische Erklärungsansätze – empirische Befunde. *Sprachverfall? Dynamik – Wandel – Variation. Jahrbuch des Instituts für Deutsche Sprache 2013*. De Gruyter Mouton. 171–196.

Toporišič, Jože, 1967: Besedni red v slovenskem knjižnem jeziku. *Slavistična revija* 15/1-2. 251–274.

Toporišič, Jože, 2004: *Slovenska slovnica (SS). Četrta, prenovljena in razširjena izdaja. 2. natis.* Maribor: Založba Obzorja. 667–678.

Toporišič, Jože, 2008: *Stilnost in zvrstnost.* Ljubljana: Založba ZRC.

Vidovič Muha, Ada, 2000: *Slovensko leksikalno pomenoslovje: govorica slovarja.* Ljubljana: Znanstvena založba Filozofske fakultete.

Zupančič, Nataša, 2009: *Korpusna analiza slovenskega jezika na spletnih forumih.* Magistrsko delo. Ljubljana: Filozofska fakulteta.

Govorne prvine v nestandardni spletni slovenščini

Ana Zwitter Vitez, Darja Fišer

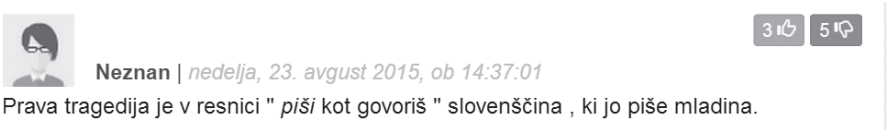
Izvleček

Komunikacija na forumih, družbenih omrežjih in novičarskih portalih je pogosto označena kot hibrid med govorom in standardnim pisnim jezikom. Da bi presegli ta stereotip, smo poskušali raziskati dejanske specifike govorjenega diskurza in računalniško posredovane komunikacije s pomočjo analize seznamov ključnih besednih oblik v korpusih Gos in Janes glede na korpus Kres. Rezultati kažejo, da so na besednovrstni ravni besedila računalniško posredovane komunikacije bliže pisnim besedilom kot govorjenemu diskurzu ter da se od standardnega zapisa najbolj odmikajo tviti in pozitivno naravnani komentarji na novičarskih portalih, najmanj pa besedila forumov. Na ravni besedišča so posebej zanimivi za govor specifični elementi interakcije z drugimi udeleženci, ki so najpogosteje prisotni v tvitih in komentarjih. Rezultate raziskave bi lahko uporabili pri razmisleku o teoriji zvrstnosti za slovenščino in jih vključili v snovanje novih jezikovnih priročnikov.

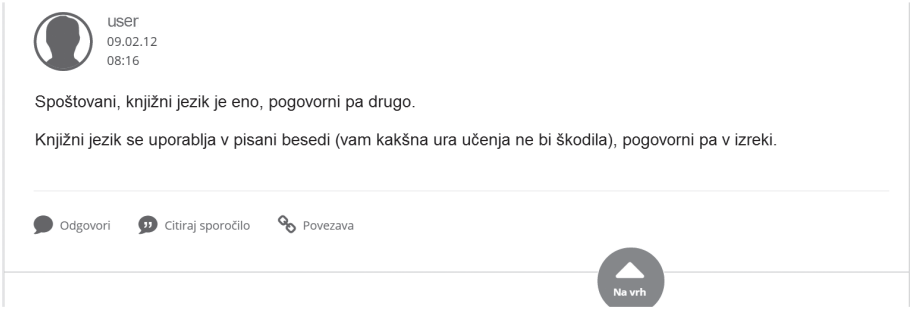
Ključne besede: govorjeni diskurz, računalniško posredovana komunikacija, neformalna komunikacija, korpusna analiza, elementi interakcije

1 UVOD

Računalniško posredovana komunikacija je v zadnjih desetletjih dobila pomembno vlogo na različnih raziskovalnih področjih zaradi vedno večje količine javnih, prosto dostopnih neformalnih besedil, s katerimi se uporabniki vsakodnevno odzivajo na aktualne družbene dogodke. Zaradi kombinacije javnosti besedil in dejstva, da jih lahko ustvarja slehernik in ne več poklicni pisec, kot je to veljalo nekoč, so tovrstna besedila pogosto deležna stereotipnih oznak (Sliki 1 in 2). Prvi stereotip sporoča, da je jezik družbenih medijev in računalniško posredovane komunikacije nasploh podoben govoru. Drugi stereotip sporoča, da je govorna komunikacija neformalna, pisna pa formalna.



Slika 1: Stereotip o podobnosti med govorom in jezikom družbenih medijev.[1]



Slika 2: Prepričanje, da morajo biti pisna besedila vedno standardna.[2]

Velik del neformalne komunikacije najverjetneje dejansko poteka v govorni obliki, vendar so njeni elementi vse bolj prisotni tudi v pisni slovenščini. To velja predvsem za besedila novih medijev (forumi, komentarji, blogi in družbena omrežja), ki jih tvorimo v interakciji z drugimi uporabniki. Čeprav se pri marsikateri prvini takih besedil ponuja misel »vpliva govora na pisni jezik« (Slika 1), velja tovrstne razlage vzeti z veliko mero previdnosti in uzavestiti misel D. Tannen (1982: 14), da ravnamo zmotno, »kadar primerjamo besedila različnih žanrov, razlike med njimi pa pripišemo razliki med govorjenim in pisnim jezikom.«

1 http://www.delo.si/kultura/knjizevni-listi/znanstvena-slovenscina-tragedija-v-dveh-dejanjih.html

2 https://med.over.net/forum5/viewtopic.php?t=1146255

Toporišičev model zvrstnosti (Toporišič 1991) sicer prevideva, da sta tako govorjeni kot pisni jezik lahko vpeta v bolj ali manj formalne okoliščine vsakokratnega sporazumevalnega položaja, zato vpeljuje kategorijo socialnih zvrsti, vendar favorizira pisni knjižni jezik, normativizacijo govorjenega jezika pa skuša reševati s splošnim pogovornim jezikom (Gruden 2013). Čeprav lahko govorjeni jezik znotraj Toporišičevega modela nastopa tako znotraj knjižnih in neknjižnih zvrsti, se je v slovensko zavest veliko bolj kot celotna slika zvrstnosti zakoreninila dialektika med pogovornim (tj. neformalnim, govorjenim) jezikom (Slika 1) in knjižnim (tj. formalnim, pisnim) jezikom (Slika 2), ki je Toporišič ni predvidel.

Tovrstne stereotipe je mogoče preseči z empiričnimi raziskavami. Zato smo poskušali raziskati stopnjo prekrivnosti med prvinami, ki so tipične tako za besedila računalniško posredovane komunikacije kot za govorjeni diskurz. S pomočjo frekvenčnih seznamov in seznamov ključnih besed smo identificirali specifike korpusov tvitov, komentarjev, forumov in korpusa Gos glede na referenčni pisni korpus Kres in nato zaznane specifike primerjali med seboj na besednovrstni in leksikalni ravni ter na ravni odstopanja od standardnega zapisa oziroma izgovorjave.

2 RAZISKAVE GOVORJENEGA DISKURZA IN RAČUNALNIŠKO POSREDOVANE KOMUNIKACIJE

Analize računalniško posredovane komunikacije so izjemno uporabne za področje marketinga (Tedeschi et al. 2015) in mnenjskih raziskav (Pang et al. 2008, Cambria et al. 2013, Smailović et al. 2013), saj lahko s tovrstnimi metodami ugotavljamo politično naravnanost volilcev, ugotavljamo zadovoljstvo strank s produkti ali zaznavamo centre nestrpnosti v družbi. Številne raziskave so že vzpostavile povezavo med specifikami spletnega in govorjenega diskurza. Jezikoslovno naravnane študije so precej široko zasnovane in pogosto zajemajo ravni zapisa, leksike, skladnje in pragmatike (Chovanec 2009, Nyström 2013), vendar je zaradi nereprezentativnosti vzorcev njihove zaključke težko posploševati. Kvantitativne analize so izvedene na velikih vzorcih besedil in večinoma usmerjene v posamezno jezikovno raven ali strukturo, kot so na primer frekvence besed (Leech in dr. 2001), raznolikost besedišča (Bamman in dr. 2014) ali segmentiranje diskurzivnih enot (Baron 2010). Nekatere študije podobnost med govorjenim in spletnim diskurzom vidijo tako očitno, da specifike jezika novih medijev uporabljajo kot izhodišče za raziskovanje govorjenega diskurza pri avtomatskem razpoznavanju govora (Vaufreydaz et al. 1999).

Vendar Crystal, ki je že pred leti preučeval specifike različnih zvrsti govorjenega in pisnega jezika (1995), opozarja (2001), da spletne komunikacije ne smemo

obravnavati kot govora, ki bi ga nekdo zapisal, saj gre za pisno komunikacijo, ki jo pogosto zaznamujejo neformalne komunikacijske okoliščine. Baron pa dodaja (2010), da tudi sodobne raziskave novih medijev svoje izsledke posplošujejo na celoten diskurz računalniško posredovane komunikacije, čeprav primerjajo zelo različne spletne besedilne zvrsti (blogi, spletne klepetalnice, forumi ipd.).

V slovenskem raziskovalnem prostoru govorjeni diskurz in pisna računalniško posredovana komunikacija do zdaj nista bili sistematično obravnavani skupaj, zato navajamo raziskave, ki so prispevale k poznavanju govorjenega jezika, in tiste, ki se ukvarjajo s pisnim spletnim diskurzom. Schlamberger Brezar (1998) v spontanem govorjenem diskurzu analizira retorične principe v argumentaciji, Smolej (2006) skladnjo in pragmatične prvine, Verdonik (2006) diskurzne označevalce in popravljanja, Zwitter Vitez (2009) realizacijo komunikacijskih strategij, Huber (2013) poudarek in pavzo, Tivadar (2015) razmerje med domačimi in tujimi prvinami v govorjenem jeziku.

Omenjene študije so izjemnega pomena za poznavanje produkcijskih razlik med govorjenimi in pisnimi besedili, ki močno vplivajo na zgradbo govornih oziroma pisnih enot diskurza, saj brez dvoma drži, da je »med govorjenim in pisnim jezikom celo vesolje«[3] (Morel in Danon Boileau 1998: 7). Ne obravnavajo pa specifik govorjenega diskurza brez vnaprej postavljenih hipotez. Korpusno zasnovane analize spontanega diskurza, ki temeljijo izključno na analizi gradiva, so izvedli Verdonik in Kosem (2012) ter Zwitter Vitez (2016), ki so na korpusu Gos izvedli raziskavo tipičnih prvin posameznih besedilnih zvrsti glede na formalnost sporazumevalnega položaja.

Specifike besedil novih medijev obravnavajo naslednje raziskave: Kalin Golob (2008) analizira kratka sporočila in v njih identificira številne pojave nestandardne rabe ločil, fonetični zapis, opuščanje pomožnega glagola in nedoločnika. Michelizza (2008) ugotavlja, da je jezik kratkih sporočil drugačen od pisnega jezika zaradi vpliva svetovnega spleta, slenga, angleščine in omejitev dolžine. Tudi rezultati raziskav o rabi slovenščine v elektronskih sporočilih (Dobrovoljc 2008) in na spletnih forumih (Jakop 2008) kažejo, da v tovrstni komunikaciji izstopa vpliv narečnih glasovnih posebnosti, prevzete besede iz tujih jezikov, čustveno zaznamovane besede, mašila in krajšave.

Predstavljene raziskave besedil novih medijev so zasnovane z vnaprej postavljenimi hipotezami o prisotnosti določenih jezikovnih struktur v analiziranih besedilnih žanrih, vendar med rezultati omenjajo tudi prisotnost govornih prvin. Zato se zdi smiselno izvesti celovito primerjavo specifik govorjenega diskurza in pisnih besedil računalniško posredovane komunikacije.

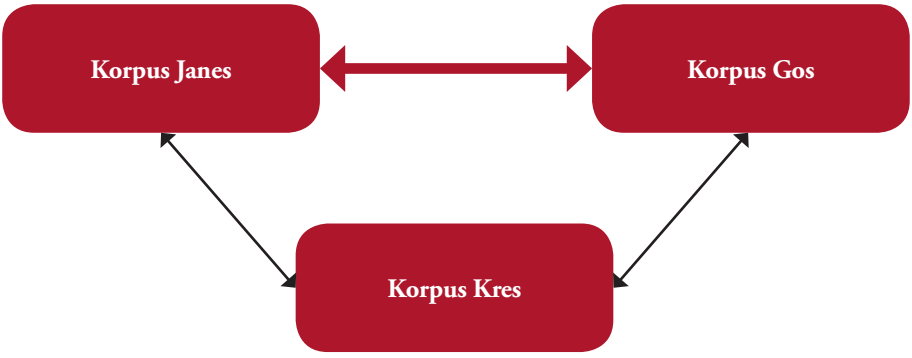
3 V izvirniku: »De l'oral à l'écrit, il y a un monde.«

3 METODOLOGIJA

Prvine govora v besedilih novih medijev smo analizirali s primerjavo treh korpusov slovenskega jezika:

- uravnoteženi 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 slovenskih spletnih uporabniških vsebin z 200 milijoni pojavnic (Erjavec et al. 2018).

Analiza je potekala v dveh fazah: najprej smo iz korpusa Gos izluščili tipične prvine govora na podlagi primerjave z uravnoteženim korpusom pisnih besedil Kres, potem pa smo pojavljanje identificiranih govornih prvin preverili v korpusu spletnih besedil Janes (Slika 3). Kvantitativno in kvalitativno analizo smo izvedli na ravni besednih vrst, besedišča in stopnje odmika od standarda .



Slika 3: Primerjava specifik korpusa Janes in korpusa Gos glede na korpus Kres.

Za pridobivanje podatkov iz korpusov smo uporabili orodje Sketch Engine[4] (Kilgarriff et al. 2004), ki omogoča vpogled v konkordance in primerjavo različnih korpusov s pomočjo relativnih frekvenčnih seznamov in seznamov ključnih oblik, ki se v določenem korpusu pojavijo zelo pogosto in obenem v primerjanem korpusu niso pogoste.

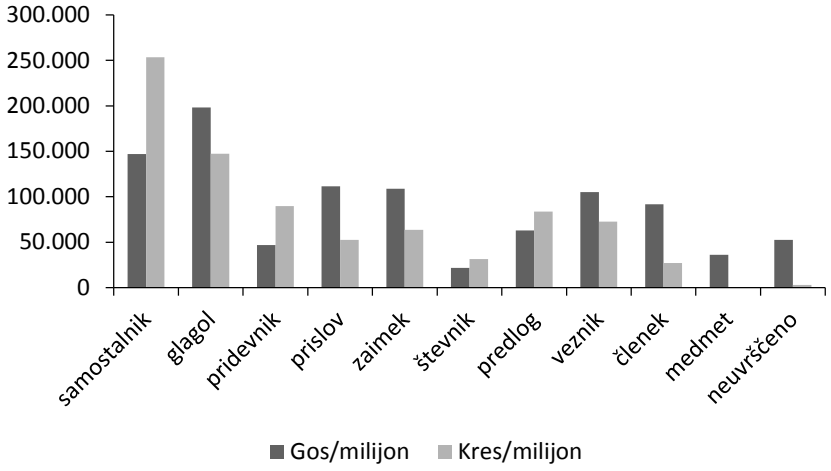
4 http://www.sketchengine.co.uk

4 ANALIZA

4.1 Besedne vrste

S pomočjo frekvenčnih seznamov smo analizirali zastopanost posameznih besednih vrst v korpusih Gos, Kres in Janes. Besedne vrste so bile določene avtomatsko v skladu s Priporočili za oblikoslovno označevanje JOS (Erjavec et al. 2010).

4.1.1 Specifike govora



Graf 1: Zastopanost besednih vrst v korpusu Gos in korpusu Kres.

Kvantitativna primerjava porazdelitve besednih vrst pokaže, da je v korpusu govorjene slovenščine glede na korpus Kres izrazito več glagolov, prislovov, zaimkov, veznikov, členkov in medmetov.

Da se bodo specifike govora pokazale skozi večjo zastopanost medmetov, členkov in zaimkov, smo pričakovali, saj te besedne vrste lajšajo sprotno tvorjenje in razumevanje govorjenih besedil. Nekoliko preseneča večja zastopanost glagolov, vendar podrobnejša analiza gradiva pokaže, da razliko v veliki meri tvorijo glagoli v vlogi diskurznih označevalcev, kot na primer *misliti* in *vedeti* (3):

(1) *zakaj kako* ***a veš mislim*** *eee poznaš eee [ime] od prej?*

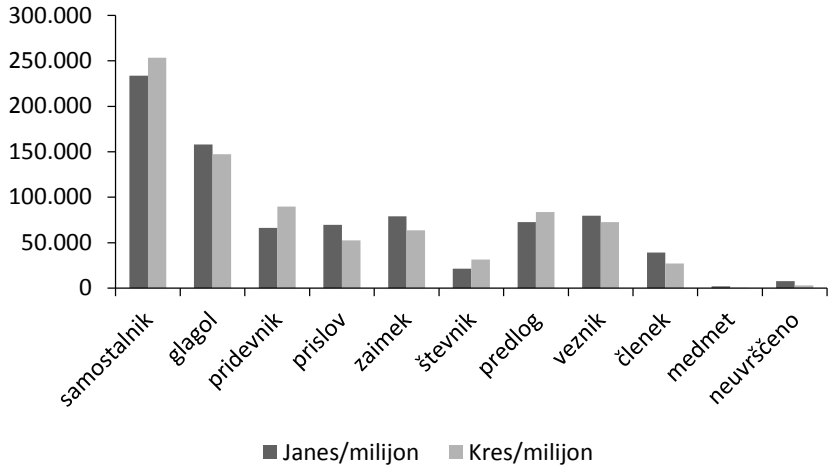
Poleg glagolov v vlogi diskurznih označevalcev prihaja v spontanem govoru do pogostih ponavljanj, ki so posledica zanimivega »uresničevanja paradigmatske osi

na sintagmatski osi« (Blanche Benveniste 1991, Smolej 2004), ko se elementi, med katerimi izbiramo znotraj svojega jezikovnega sistema, nakopičijo eden za drugim (*zakaj, kako, a veš, mislim, eee*). S pojavom kopičenja paradigmatskih elementov lahko razložimo tudi pogosto rabo veznikov (primer 2), ki sicer v pisnih besedilih kažejo na večjo skladenjsko kompleksnost besedil.

(2) zarad tega **ker** smo **ker** eee **ker** živimo oziroma **ker** so takrat že živeli v času ko so informacije dejansko z lahkoto krožle

Tako glagoli s pragmatično vlogo diskurznih označevalcev kot uresničevanje paradigmatske osi na sintagmatski omogočata sprotno tvorjenje govorjenih besedil, številni premori in ponovljene strukture pa pogosto olajšata tudi razumevanje govorjenih besedil.

4.1.2 Govorne prvine v spletnih besedilih



Graf 2: Besedne vrste v korpusu Janes in korpusu Kres.

Korpus Janes se po zastopanosti analiziranih besednih vrst sicer razlikuje od korpusa Kres, vendar se statistično umešča bliže pisnim kot govorjenim besedilom. V besedilih novih medijev so medmeti sicer sedemkrat bolj pogosti kot v korpusu Kres, vendar večinoma oponašajo neverbalne izraze in smeh (*haha, aja, hm*), ne odsevajo pa sprotnega procesa tvorjenja besedila kot v govoru (*eee, eem, mmm*), saj pisanje spletnih besedil vseeno omogoča več časa za tvorjenje koherentnih enot kot govor:

(3) srečkovička **haha** .. ti prvošm

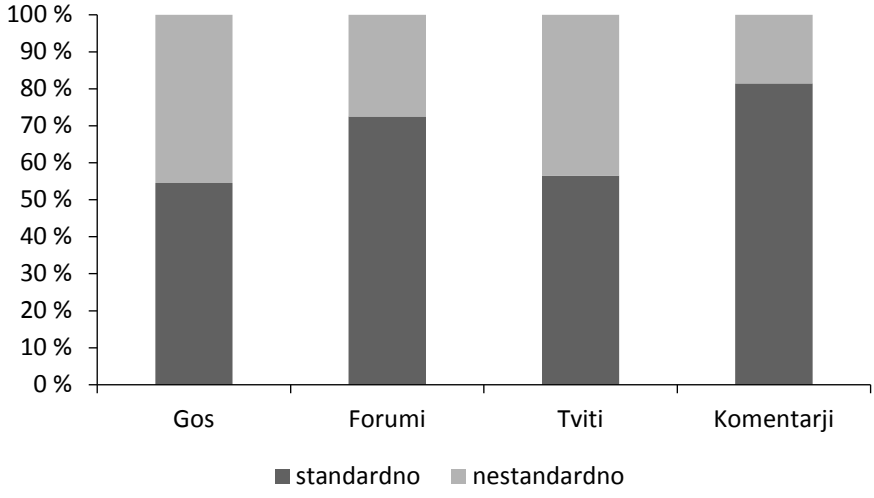
Podobnost med govorjenimi besedili in računalniško posredovano komunikacijo se kaže v manjši stopnji števnikov in pridevnikov glede besedila korpusa Kres. To bi lahko bila posledica tematske usmerjenosti posameznih besedilnih zvrsti korpusa Kres (primer 4) ali bistveno krajšega postopka tvorjenja spletnih in govorjenih besedil.

(4) Tudi **dvakratni olimpijski** zmagovalec Gianni Romme in »**srebrni**« Bob de Jong napovedala svoje slovo iz reprezentance

Analiza besednih vrst je pokazala, da je v govorjenih in spletnih uporabniških vsebinah bistveno več zaimkov, členkov in medmetov, manj pa pridevnikov in števnikov. Vendar kvalitativna analiza gradiva pokaže, da so v govoru prisotni drugi medmeti kot v računalniško posredovani komunikaciji, kjer prav tako ne opazimo povečanega števila glagolov in veznikov, ki so posledica sprotnega tvorjenja govorjenih besedil.

4.2 Odmik od standardne izgovorjave in zapisa

Med tipičnimi besednimi oblikami korpusov Gos in podkorpusov Janes smo želeli ugotoviti, v kolikšni meri besedišče, značilno za govor in računalniško posredovano komunikacijo, odstopa od priporočil izgovorjave (brez upoštevanja mesta naglasa in drugih prozodičnih lastnosti) oz. zapisa v SSKJ. Ključnim besednim oblikam v korpusih Gos in Janes smo pripisali oznako standardno/nestandardno (Graf 3).



Graf 3: Odmik od standarda v govorjenem korpusu in korpusih Janes.

Delež tipičnih besed v govoru in spletnih besedilih, ki imajo drugačno standardno obliko po priporočilih SSKJ, se giblje med 19 % in 46 %: največ takšnih besed najdemo v govoru (46 %) in v korpusu tvitov (43 %). Ko podrobneje analiziramo govorjeno gradivo, lahko zasledimo trend padanja standardne izgovarjave oz. naraščanja nestandardne izgovarjave od javnega informativnega diskurza proti nejavnemu zasebnemu diskurzu (Zwitter Vitez 2016). Pa vendar v prispevku Zwitter Vitez (2016) ugotavljamo, da je nestandardna izgovorjava precej prisotna tudi v formalnih besedilnih zvrsteh (32 %), kot kaže primer 5.

(5) smo policisti srečujemo se b reku z vsemi oblikami eee b reku bolečin stisk ampak men bi tuki zdaj največ pomenlo če bi recimo mmm ta punca ostala živa ne

Tudi računalniško posredovana komunikacija pogosto poteka v obliki neformalnih diskusij. Zato ne preseneča visoka stopnja nestandardnega zapisa na Twitterju (43 %), kar odseva prevladujočo sproščenost vzdušja med Twitteraši (primer 6).

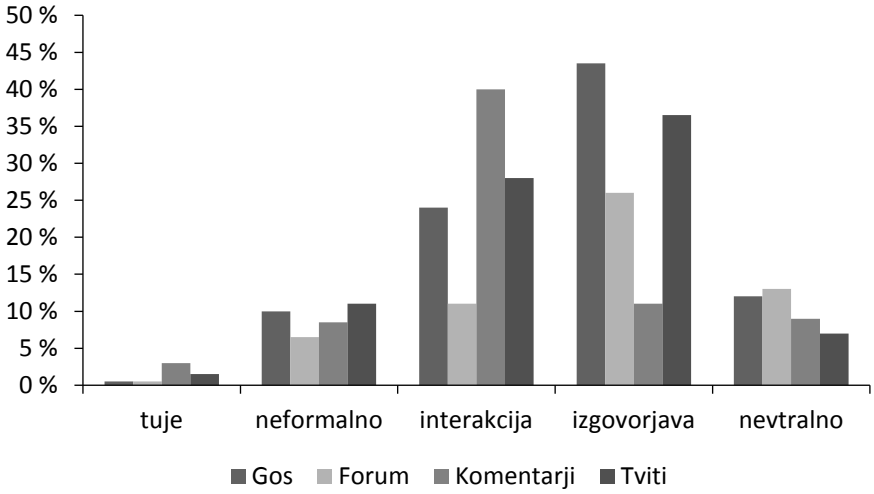
(6) - ma čak, a ne bi mogu bit model že v čuzi? koga se čaka?
- sodne počitnice so. al neki.
- a to tud za Furs vela?

Komunikacija na forumih je drugačne narave, saj avtorji besedila po navadi objavljajo, da rešijo konkreten problem z določenega področja (zdravstvo, gospodinjstvo, avtomobilizem ipd.), zato so tudi odgovori strukturirani bolj formalno. Komentarji na spletne novice so specifičen žanr, pri katerem smo v raziskavi Zwitter Vitez in Fišer (2016) zaznali zanimiv pojav, da naravnanost komentarja vpliva na stopnjo standardnosti sporočila. V pozitivnih komentarjih smo namreč v primerjavi z negativnimi zaznali veliko primerov zapisa, ki odstopa od standarda (*dejmo*), ter pogosto rabo velikih tiskanih črk (BO), nestandardno rabljenih ločil (*Dajmo klobasica!!!:)*) in emotikonov (*;)))*). Pri negativnih komentarjih pa je opaziti manj nestandardnega zapisa (*Če zaupaš našim medijem, so še skoraj vsako leto bile kritike glede naših pesmi pozitivne, ampak rezultata pa nobenega in isto bo letos)*. Možno razlago za zaznano razliko na ravni standardnosti komentarjev vidimo v dejstvu, da se avtorji pozitivnih komentarjev laže identificirajo s pripadnostjo družbeno-demografskim skupinam s tipičnimi jezikovnimi specifikami (*Dejmo, klobasica!!!*), pozicioniranje negativnega mnenja pa je bolj odvisno od sprejemanja skupnosti, zaradi česar avtorji uporabljajo manj zaznamovane jezikovne strukture.

4.3 Besedišče

Seznami ključnih besed (angl. *keywords*; Scott 1997) omogočajo preučevanje tipičnega konteksta, v katerem se te besede uporabljajo, na podlagi prevladujočega

konteksta pa naprej analiziramo njihovo vlogo v jezikovni rabi. Tipične rabe besed s seznamov ključnih besed smo kategorizirali v štiri skupine: nestandardna izgovorjava ali zapis (*lohk*), elementi interakcije (*hej*), neformalno besedišče (*spedenan*) in tuje besede (*jes*).



Graf 4: Kategorizacija besedišča v korpusu Gos in podkorpusi korpusa Janes.

Kot je pokazal že odmik od standardne izgovorjave oz. zapisa, v korpusu tvitov najbolj izstopa izrazita podobnost z govorjeno slovenščino glede na posebnosti izgovorjave oz. zapisa (43 % Gos in 36 % Twitter).

Kategorija interakcije je v vseh analiziranih korpusih dobro zastopana, posebej pa izstopa v korpusu komentarjev, ki mu sledi korpus Gos. Kvalitativna analiza pa pokaže, da so komentarji v opazovanih besedilnih zvrsteh precej različni. V komentarjih najdemo elemente, ki izražajo naravnanost oz. sodbo avtorja besedila (*bravo, verjetno, škoda, sramota*), medtem ko se v tvitih in v govoru pojavljajo drugačni elementi interakcije, ki vzdržujejo stik z naslovnikom (*hej*, *mhm, a veš*).

(7) **A veš**, **v bistvu** blokira sama sebe. (Twitter)
A veš in **v bistvu** na papirju izzveni kot da je to neka tuja firma **ne**. (Gos)

Neformalnih besed ni prav veliko (med 6 % in 11 %), so pa v različnih podkorpusih precej podobne (*ful, glih, kao*). Kot kaže primer 8, se v tvitih neformalne besede (*razirat, nažajfat*) prepletajo z elementi interakcije (*jah, sej veš*), in emotikoni (☺), kar ustvarja sproščeno vzdušje na Twitterju.

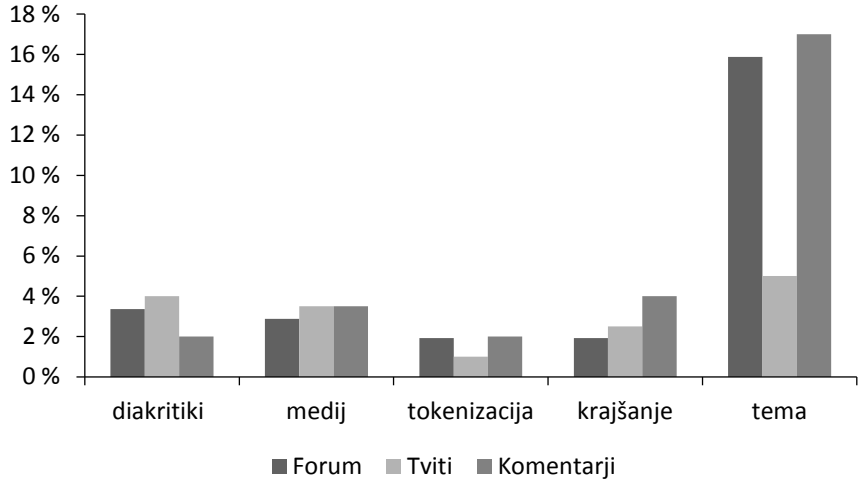
(8) jah sej veš.. za razirat se, morš bit nažajfan ☺

Tujih besed je zanemarljivo malo, vendar so v tem primeru rezultati analize ključnih besed vprašljivi, saj je zajem besedil zaradi zaznavanja jezika potekal s pomočjo izločanja gradiva z neslovenskimi besedami.

Na ravni besedišča je govorjenemu jeziku najmanj podoben podkorpus forumskih sporočil, kjer smo zaznali manj nestandardnega zapisa (25 %), neformalnih izrazov (6 %) in manj elementov interakcije kot v korpusu Gos. To verjetno korelira s formalnostjo forumskih tematik (višja stopnja pri zdravstvenih vprašanjih, nižja pri avtomobilističnih temah). Medtem ko uporabniki Twitterja uživajo v besednih igrah, so forumski uporabniki pragmatično usmerjeni v pridobivanje in izmenjavo informacij.

4.3.1 Elementi računalniško posredovane komunikacije

Tipične elemente računalniško posredovane komunikacije, ki zaradi narave komunikacijskega kanala ne morejo biti prisotni v korpusu Gos, smo analizirali posebej. To so kategorije izrazov, tipičnih za tematiko diskusije (*volitve*), odsotnosti diakritičnih znamenj (*a ves*), besede, vezane na medij (*lajkaš*), specifike v tokenizaciji (*neboš*) in postopki krajšanja besed (*slo*).



Graf 5: Tipični elementi računalniško posredovane komunikacije.

Izrazi, povezani s tematiko diskusije, so največkrat samostalniki in so najpogosteje prisotni v forumih (e.g. *avto, problem*) in v komentarjih (e.g. *tekma, volitve*). To ne preseneča, saj so bili forumi korpusa Janes zajeti glede na tematiko diskusije,

komentarji na novičarskih portalih pa so že po definiciji povezani z aktualnimi tematikami (npr. volitve). Besedila korpusa Gos in tviti pa so bili zajeti neodvisno od teme diskusije.

Vsi trije podkorpusi Janes kažejo veliko število ključnih besed, ki odsevajo jezikovno rabo družbenih omrežij (*com, všeč, videoposnetek*). Ti izrazi so bili prvotno specifični za eno platformo komunikacije (Youtube, Facebook), kasneje pa so se razširili na celotno računalniško posredovano komunikacijo (*všečkati, lajkati*).

Izpuščanje diakritičnih znamenj, krajšanje besed in nestandardna tokenizacija so se pri analizi sicer pojavile, vendar niso relevantne za kvantitativni del analize. Gre namreč za prvine, ki jih določeni uporabniki uporabljajo na različnih besedah, na seznamu ključnih 200 besed pa opazujemo le najpogostejše besede, na katerih smo te prvine zaznali (*veš/ves*). Tako ne moremo izvedeti, v kolikšnem deležu računalniško posredovane komunikacije je ta pojav dejansko prisoten. Podobno težavo opazimo pri pojavu nestandardne tokenizacije, ki ga prav tako lahko opazujemo le na primeru najpogostejših besednih oblik (*ne bi/nebi*). Prav tako opazimo, da v preučenih primerih nestandardne tokenizacije, ki je sicer bolj prisotna v forumskih sporočilih in komentarjih na spletne novice, ni opaziti znakov intencionalne igre z jezikom, kar morda odseva pomanjkanje jezikovne kompetence in ne jezikovne kreativnosti.

4.4 Elementi interakcije

Posebej zanimiva kategorija, ki smo jo zaznali tako v govorjenih kot v spletnih besedilih, so elementi interakcije z ostalimi udeleženci, prek katerih avtorji besedil utrjujejo svojo identiteto, gradijo samopodobo in razkrivajo svoj odnos do zunajjezikovne realnosti. Z dodatno analizo te kategorije smo želeli ugotoviti, kateri tipični elementi govorne interakcije so prisotni v besedilih računalniško posredovane komunikacije in kakšna je njihova vloga v vsakodnevni komunikaciji na spletu. Predvidevali smo, da lahko prek identifikacije elementov interakcije in njihovih funkcij sklepamo o komunikacijskem namenu tvorca besedila in posledicah njegovega jezikovnega udejstvovanja.

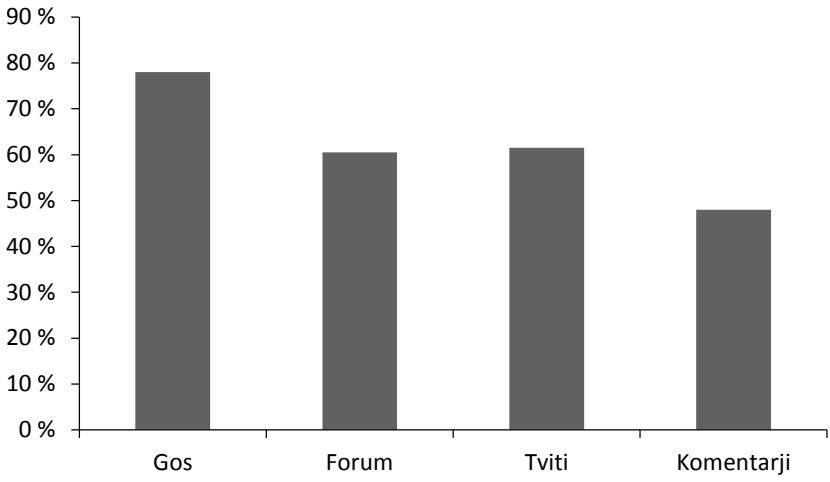
Ker je primarna raziskava besedišča pokazala, da so prvine interakcije med udeleženci prisotne tudi v sicer drugače razporejenih kategorijah (npr. pri izrazu *pejd*, pri katerem lahko zaznamo tako odstop od standardnega zapisa kot interakcijo z drugimi udeleženci), smo prvine interakcije v govoru in spletnih besedilih na novo kategorizirali in podrobneje preučili. Izluščili smo ključne besedne oblike (pojavnice) korpusa Gos in korpusa Janes. Nato smo označili vse elemente, ki označujejo interakcijo s sogovorci, in te elemente analizirali z metodologijo

Tracey in Robles (2013), razširjeno s kategorijami, ki sledijo shemi jezikovnih funkcij (Jakobson 1989). Dobili smo kategorije, ki jih ponazarjajo primeri v desnem stolpcu Tabele 1, pri katerih smo v primeru dvoma izbrali tisto jezikovno funkcijo, ki je za dani element dominantna.

Tabela 1: Kategorizacija elementov interakcije.

deiktika	*Jst*
vprašalnice	*Kaj*
modus	*Sigurno*
konativnost	*Gremo*
ekspresivnost	*uau, vidim*
fatičnost	*dobro jutro, evo*
performativ	*Obljubim*
metajezik	*Tvitam*
referenca	*glede, rekel*

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.

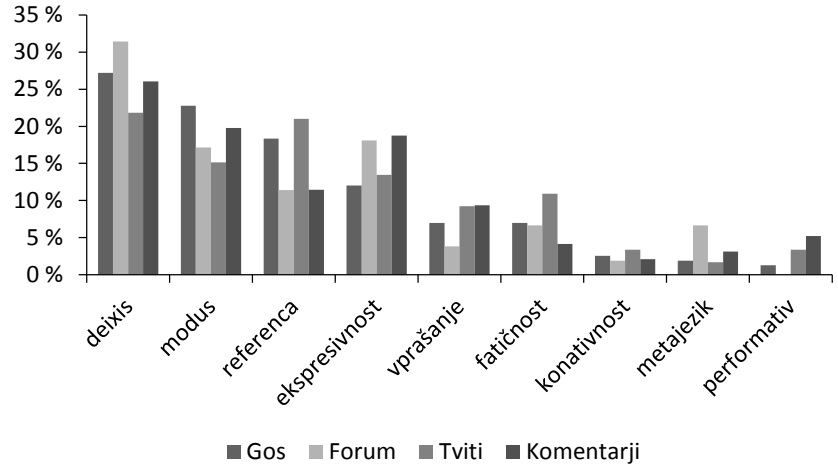


Graf 6: Elementi interakcije v korpusu Gos in podkorpusih Janes.

Analiza korpusov Gos, Janes-Forum, Janes-Tviti in Janes-Komentarji pokaže, da pri treh od štirih analiziranih besedilnih žanrov več kot 60 % analiziranih ključnih besednih oblik 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.

V naslednjem koraku smo zaznane elemente interakcije glede na prevladujočo pragmatično vlogo razvrstili v predstavljenih devet kategorij (Graf 7).



Graf 7: Analiza elementov interakcije v korpusu Gos in podkorpusih Janes.

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. Kvalitativna analiza gradiva pokaže, da so v korpusu Gos pogoste različne izgovorne variante osebnih in kazalnih zaimkov, kot so *jaz* (npr. *jst, jest, jez, jaz,*) ter časovnih in prostorskih prislovov, (npr. *zdej, zaj, zej, zdele*). V podkorpusih Janes pa deiktične izraze najdemo le v eni ali dveh variantah.

V korpusu Gos so zelo pogosti elementi modalnosti (*dejansko, itak, pač*), ki »izražajo stopnjo gotovosti, s katero govorec izraža svoje mnenje« (Morel in Danon Boileau 1998). Največ elementov modalnosti (20 %) najdemo v komentarjih (*očitno, sigurno, zgleda*), kar je skladno z namenom komentiranja na novičarskih

portalih. V tvitih je modalnih elementov manj, saj uporabnika tehnične okoliščine (omejena dovoljena dolžina, uporaba mobilnih naprav) in komuniciranje v realnem času spodbujajo k čim krajšim besedilom.

Podobno vlogo imajo ekspresivni izrazi, le da so bolj osredotočeni na naravnanost in razpoloženje samega avtorja (*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 sporazumevalne položaje, v Janesu pa pogosto predstavljajo ventil za izražanje mnenj, deljenje izkušenj in sproščanje frustracij uporabnikov. Poleg tega v analizo nismo zajeli neverbalnih prvin (npr. smeha), ki označujejo razpoloženje govorca, vendar so v korpusu Gos označeni kot dogodek in niso transkribirani kot besedilo.

Posebnost tvitov je izstopanje referencialnih elementov (*poznam, glede, tiče,*). To kaže na visoko stopnjo interaktivnosti tega medija, kjer udeleženci v pogovore vključujejo svoje izkušnje z zunanjim, realnim svetom. Tviti izstopajo tudi po deležu fatičnih elementov (*aja, btw, čao*), medtem ko komentarji niso namenjeni vzpostavljanju in vzdrževanju stika med sogovorci, temveč izražanju lastnega mnenja.

Po številu vprašalnic izstopajo tviti in komentarji, ki jih je celo več kot v korpusu Gos. V kategorijo metajezik smo zajeli izraze, ki se nanašajo na izrekanje samo, kar 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 (*gremo*, *morš*) in performativi (*obljubim, se strinjam*), ki jih v vsakem korpusu najdemo le po nekaj primerov.

Analiza elementov interakcije je pokazala, da najmočnejšo podobnost med govorjenim diskurzom in računalniško posredovano komunikacijo predstavljajo deiktični izrazi in referencialni elementi v tvitih, medtem ko so modalni elementi prisotni predvsem v komentarjih. V računalniško posredovani komunikaciji je variantnost zapisa manjša, opaziti pa je tudi določene omejitve zaradi medija samega, ki diktira ekonomičnost izražanja.

5 DISKUSIJA

Zdi se, da analiza na leksikalni in besednovrstni ravni ter na ravni odmika od standarda pove marsikaj o specifikah posameznih spletnih zvrsti v primerjavi z govorjenim diskurzom. Vendar predstavljena raziskava ne zajema skladenjske ravni analize, ki bi bistveno več povedala o tem, katere besedne zveze, tipi povedi in značilnosti upovedovanja so tipični za preučevana tipa diskurza in opazovane

besedilne zvrsti. Predvidevamo, da bi se s skladenjsko analizo dokončno pokazalo, kako specifičen je spontani govor zaradi sočasnega načrtovanja in izrekanja. Obenem pa verjamemo, da so programi za avtomatsko skladenjsko razčlenjevanje spontanega govora še v razvoju prav zato, ker je enota segmentiranja spontanega govora tako temeljno drugačna od tistih, ki smo jih vajeni analizirati v pisnih besedilih (Morel in Danon Boileau 1998).

Z analizo seznamov ključnih besednih oblik oziroma besednih vrst prav tako nismo zajeli prvin, ki jih lahko zaznamo s kvalitativnim vpogledom v gradivo spletnega in govorjenega diskurza. Hiter pogled na dogajanje na Twitterju namreč pokaže močno prisotno ironijo, ki je kvantitativne analize ne zaznajo (*Drugi tir? Ni važno. Važno, da bo referendum*). Podobne primere lahko zaznamo tudi v korpusu Gos (*eee kaj ti praviš bo dobr? ha? al ni to dobr? no dobr pol važno da gremo pol na čaj [smeh]*). Zato bi bilo v prihodnosti smiselno sistematično raziskati jezikovne prvine, s katerimi uporabniki izražajo svojo čustveno naravnanost (Martinc et al. 2017), nato pa s pomočjo identificiranih jezikovnih prvin raziskati različne tipe naravnanosti uporabnikov v govoru in v spletnih žanrih.

6 SKLEP

V raziskavi smo primerjali specifike govorjenega diskurza in računalniško posredovane komunikacije glede na standardna pisna besedila. Ugotovili smo, da je ortografska variantnost tvitov primerljiva z variantnostjo izgovorjave, kakršno najdemo v govorjenih besedilih, ker avtorji besedil v sproščenem vzdušju radi pokažejo svojo pripadnost določenim jezikovno prepoznavnim skupinam. Komentarje zaznamuje predvsem visoka stopnja za govor značilnih modalnih elementov, ki kaže avtorjevo naravnanost v zvezi z obravnavano tematiko. Vendar kvalitativna analiza gradiva pokaže, da so modalni elementi v govoru povezani z avtorjevo stopnjo gotovosti (*itak*), v komentarjih pa z njegovimi sodbami (*škoda*). Pri forumih je formalnost diskurza povezana s stopnjo formalnosti forumske tematike, zaznana odstopanja od pravopisnih pravil pa ne kažejo jezikovne ustvarjalnosti, ampak delujejo kot pomanjkanje jezikovne kompetence (*nebi*).

Vsak od analiziranih žanrov računalniško posredovane komunikacije ima seveda številne prvine, ki jih lahko zaznamo tudi v neformalnem govorjenem diskurzu, vendar je vloga prenosnika še vedno bistvenega pomena za razumevanje delovanja obeh analiziranih zvrsti. Sočasnega tvorjenja in načrtovanja besedil, ki temeljno zaznamuje spontani govorjeni diskurz, ne moremo opazovati v pisnih besedilih, kar se pokaže tudi prek osnovne primerjave zastopanosti besednih vrst. Verjetno pa razlog za pogosto izpostavljeno podobnost med računalniško posredovano komunikacijo in neformalnimi zvrstmi govorjenega diskurza predstavlja dejstvo, da

so specifike interakcije in neformalnih elementov na prvi pogled bolj vpadljive kot nekoliko ponotranjene, vendar vseprisotne prvine sprotnega tvorjenja govorjenega diskurza.

Smo z rezultati analize uspeli preseči stereotip o računalniško posredovani komunikaciji, ki je podobna govorjeni slovenščini, ta pa predstavlja sinonim za nestandardno slovenščino? Vsekakor lahko rečemo, da o računalniško posredovani komunikaciji, pa tudi o govorjenem diskurzu, vemo več. Vemo, katere specifike govora so povezane s procesom sprotnega tvorjenja besedil, katere specifike računalniško posredovane komunikacije s tehničnimi okoliščinami medija, katere z vzdrževanjem stika z ostalimi udeleženci in katere z neformalnostjo okoliščin, do katerih lahko pride tako v govorjenem kot v pisnem jeziku. Zasnova raziskave lahko prispeva k razmisleku o teoriji zvrstnosti za slovenščino in zaključi debato o pogovornem (govorjenem) in knjižnem (pisnem) jeziku, rezultate raziskave pa bi bilo smiselno vključiti tudi v snovanje novih jezikovnih priročnikov.

Literatura

Bamman, David, Jacob Eisenstein in Tyler Schnoebelen, 2014: Gender identity and lexical variation in social media. *Journal of Sociolinguistics* 1/2. 135–160.

Blanche-Benveniste, Claire, 1991: Les études sur l'oral et le travail d'écriture de certains poètes contemporains. *Langue française* 89. 52–71.

Cambria, Erik, Björn Schuller, Yunqing Xia, Catherine Havasi, 2013: New Avenues in Opinion Mining and Sentiment Analysis. *IEEE Intelligent Systems* 28/2. 15–21. http://sentic.net/new-avenues-in-opinion-mining-and-sentiment-analysis.pdf

Chovanec, Jan, 2009: Simulation of Spoken Interaction in Written Online Media Texts. *Brno Studies in English* 35/2. 109–129.

Crystal, David, 1995: *The Cambridge Encyclopedia of the English Language.* Cambridge: Cambridge University Press.

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

Dobrovoljc, Helena, 2008. Jezik v e-poštnih sporočilih in vprašanja sodobne normativistike. Košuta, Miran (ur.): *Slovenščina med kulturami*, Ljubljana: Slavistično društvo Slovenije. 295–314.

Erjavec, Tomaž, Simon Krek, Špela Arhar, Darja Fišer, Nina Ledinek, Amanda Saksida, Breda Sivec in Blaž Trebar, 2010: *Priporočila za oblikoslovno označevanje JOS*. http://nl.ijs.si/jos/msd/html-sl/msd.index.msds.html

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

Fišer, Darja, Tomaž Erjavec in Nikola Ljubešić, 2016: JANES v0. 4: Korpus slovenskih spletnih uporabniških vsebin. *Slovenščina 2.0* 4/2. 67–99. http://slovenscina2.0.trojina.si/arhiv/2016-2/2016-2-04/

Huber, Damjan, 2013: *Poudarek in pavza v standardnem slovenskem govoru.* Doktorska disertacija. Ljubljana: Filozofska fakulteta.

Gruden, Ana, 2013: *Pojmovanje socialne zvrstnosti v slovenskem prostoru s poudarkom na šolstvu*. Diplomsko delo. Ljubljana: Filozofska fakulteta.

Jakobson, Roman, 1989: *Lingvistični in drugi spisi*. Ljubljana: Studia Humanitatis.

Jakop, Nataša, 2008: Pravopis in spletni forumi - kva dogaja? Košuta, Miran (ur.): *Slovenščina med kulturami* – Slovenski slavistični kongres. Ljubljana: Slavistično društvo Slovenije. 315–327.

Kalin Golob, Monika, 2008: SMS-sporočila treh generacij. Košuta, Miran (ur.): *Slovenščina med kulturami*. Ljubljana: Slavistično društvo Slovenije. 283–294.

Kilgarriff, Adam, Pavel Richly, Pavel Smrz in David Tugwell, 2004: The Sketch Engine. *Zbornik EURALEX 2004*. Lorient. 105–116.

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

Morel, Mary Annick in Laurent Danon-Boileau, 1998: *Grammaire de l'intonation*. Pariz: Ophrys.

Nyström, Stefan, 2003: *Spoken Language Features in Internet Discussion Groups*. Diplomsko delo. Lund University.

Pang, Bo in Lillian Lee, 2008: Opinion Mining and Sentiment Analysis, *Foundations and Trends in Information Retrieval/2,* 1–135.

Scott, Mike, 1997: Pc Analysis of Key words – and Key Key words. *System* 25/1. 1–13.

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

Smolej, Mojca, 2004: Načini tvorjenja govorjenega diskurza – paradigmatska in sintagmatska os. Kržišnik, Erika (ur.): *Aktualizacija jezikovnozvrstne teorije na Slovenskem: členitev jezikovne resničnosti (Obdobja 22)*. Ljubljana: Center za slovenščino kot drugi/tuji jezik.

Smolej, Mojca, 2006: *Vpliv besedilne vrste na uresničitev skladenjskih struktur (primer narativnih besedil v vsakdanjem spontanem govoru). Doktorska disertacija*. Ljubljana: Filozofska fakulteta.

SSKJ – *Slovar slovenskega knjižnega jezika*. Ljubljana: Inštitut za slovenski jezik Frana Ramovša ZRC SAZU. http://www.fran.si/130/sskj-slovar-slovenskega-knjiznega-jezika

Tannen, Deborah, 1982: *Spoken and written language: exploring orality and literacy*. Norwood: Ablex.

Tedeschi, Antonio, Francesco Benedetto, 2015: A cloud-based big data sentiment analysis application for enterprises' brand monitoring in social media streams. *Research and Technologies for Society and Industry Leveraging a better tomorrow* Torino, Italija. 186–191. DOI: 10.1109/RTSI.2015.7325096

Tivadar, Hotimir, 2015: Razmerje med domačim in tujim v govorjenem knjižnem jeziku 3. tisočletja. Jesenšek, Marko (ur.): *Leopold Volkmer: prvi posvetni pesnik na slovenskem Štajerskem*. Maribor: Mednarodna založba Oddelka za slovanske jezike in književnosti. 139–149.

Toporišič, Jože, 1991: *Slovenska slovnica*. Maribor: Založba Obzorja.

Tracey, Karen in Jessica S. Robles, 2013: *Everyday talk. Second Edition. Building and Reflecting Identities*. New York: Guilford Press.

Vaufreydaz, Dominique, Mohamad Akbar, José Rouillard, 1999: Internet Documents: A Rich Source for Spoken Language Modeling. Keystone – Colorado. 277–281

Verdonik, Darinka, 2006: Popravljanja v spontano tvorjenih izjavah. *Slavistična revija* 54/2. 188–203.

Verdonik, Darinka in Iztok Kosem, 2012: Key word analysis of discourses in Slovene speech: differences and similarities. *Linguistica* 52/1. 309–321.

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

Zwitter Vitez, Ana, 2009: O spontanem govoru. Vitez, Primož (ur.): *Spisi o govoru*. Ljubljana: Znanstvena založba Filozofske fakultete. 241–254.

Zwitter Vitez, Ana in Darja Fišer, 2015: Novi mediji in govorjena slovenščina: zaznavanje, kodifikacija, analiza. Smolej, Mojca (ur.): *Slovnica in slovar – aktualni jezikovni opis (Obdobja 34)*. Ljubljana: Znanstvena založba Filozofske fakultete. 881–890. http://centerslo.si/wp-content/uploads/2015/11/34_2--Zwitter.pdf

Zwitter Vitez, Ana, 2016: Specifike govorjene slovenščine glede na formalnost sporazumevalnega položaja. Kržišnik, Erika in Miran Hladnik (ur.). *Toporišičeva obdobja*, Ljubljana: Znanstvena založba Filozofske fakultete. 351–359. http://centerslo.si/wp-content/uploads/2016/11/Zwitter.pdf

Zwitter Vitez, Ana in Darja Fišer, 2016: Linguistic Analysis of Emotions in Online News Comments-an Example of the Eurovision Song Contest. Fišer, Darja in Michael Beißwenger (ur.): *Proceedings of CMC and Social Media Corpora for the Humanities,* Ljubljana: Znanstvena založba Filozofske fakultete. 74–76. http://nl.ijs.si/janes/wp-content/uploads/2016/09/CMC-2016_Zwitter_Fiser_Linguistic-Analysis-of-Emotions.pdf

Raba ključnikov v slovenskih tvitih

Mija Michelizza

Izvleček

Prispevek obravnava ključnike v rabi glede na vlogo v slovenskih tvitih. Ključniki kot tip metapodatkov služijo za kategoriziranje, lahko pa opravljajo tudi različne komunikacijske vloge. Ključnike smo glede na vlogo v tvitih razporedili v osem kategorij, ki so se pokazale kot relevantne že v študiji Wikströma (2014): oznaka izbrane tematike, oznaka igre, metakomentar besedila tvita, dopolnilo pomena tvita, izražanje čustev, občutij ali razpoloženja, izražanje poudarjalnosti, izražanje humorja, igrivosti (z jezikom), popularna kultura in tradicija. Pri razdelitvi je treba upoštevati, da se kategorije med seboj ne izključujejo, isti ključnik lahko nastopa tako v vlogi kategoriziranja kot v kateri od komunikacijskih vlog. Pri slednjih opažamo, da se ključniki pogosteje vključujejo v skladnjo, vendar bodo s tega vidika potrebne nadaljnje raziskave. Ključniki pogosto izkazujejo povezanost pisanja tvitov s spremljanjem drugih medijev. Nekateri ključniki so lahko zelo dolgi, saj poleg besed vsebujejo tako besedne zveze kot tudi cele stavke, zelo redko pa v ključnikih poleg lojtre nastopajo drugi nečrkovni elementi. Čeprav predstavljajo ključniki novejši jezikovni element, ki v računalniško posredovani komunikaciji izstopa, ostajajo ti večinoma znotraj svojih vlog in se v večini tvitov iz naključnega vzorca analiziranega korpusa ne pojavijo.

Ključne besede: ključnik, lojtra, Twitter, računalniško posredovana komunikacija, slovenščina

1 UVOD

Twitter je ena izmed vodilnih in hkrati bolj uporabljanih[1] platform družbenih omrežij tudi v Sloveniji. Komunikacija na Twitterju je v primerjavi z drugimi družbenimi omrežji specifična, saj je besedilo v posameznih tvitih omejeno na 280 znakov,[2] zelo značilni za tovrstno komunikacijo so tudi ključniki. Navadno gre za besedo ali besedno zvezo, ki je brez presledkov zapisana za lojtro. Tovrstni zapis omogoči iskanje po izbranih ključnikih, med iskalnimi zadetki so tviti, ki vsebujejo iskani ključnik. Med večjimi in trenutno najbolj uporabljanimi plaformami družbenih omrežij je ta način iskanja mogoč na Facebooku, Twitterju, Instagramu in Snapchatu, uporabljajo pa ga tudi drugi mediji za komunikacijo z uporabniki Twitterja (na taisti platformi). Pogosto lahko namreč na televiziji vidimo oddajo v živo, ki v spodnjem delu ekrana predvaja tvite gledalcev. Zajem je mogoč s pomočjo ključnika, ki mora biti uporabnikom in gledalcem predhodno znan, npr. ime oddaje, pri čemer ni nujno, da je v ključniku izraz, ki bi bil tudi sicer splošno poznan, pogosto gre namreč za oznake, kratice, ki so lahko dogovorjene med uporabniki Twitterja (npr. *#sp14si*, ki je označeval tvite na temo svetovnega nogometnega prvenstva leta 2014 v slovenščini, ali *#plts*, ki je oznaka za Prvo ligo Telekoma Slovenije).

Kar nekaj raziskav je že bilo posvečenih jeziku slovenskih tvitov, vendar pa se redke dotaknejo tudi vloge ključnikov na Twitterju (npr. Erjavec in Fišer 2013; Arhar Holdt et al. 2016; Pertot et al. 2016), lahko so ključniki zaradi svoje oblikovne posebnosti pred začetkom analize tudi izločeni oz. označeni kot nerelevantni (npr. Arhar Holdt in Dobrovoljc 2015).

2 O IZRAZIH KLJUČNIK IN LOJTRA

Izraz *ključnik* (angl. *hashtag*), tudi *hešteg* in *hashtag*, je novejši leksem, ki še ni vključen v normativne slovarje slovenskega jezika, najdemo pa ga v nekaterih drugih spletnih slovarjih. V *Sprotnem slovarju slovenskega jezika* na portalu Fran.si je razlaga za izraz ključnik sledeča: »oznaka lojtra [#] skupaj z vsaj eno ključno besedo takoj za njo, ki na družbenih omrežjih služi zlasti za označevanje, razvrščanje vsebin po temi«.[3] *Slovar tviterščine*[4] ključnik razloži z razlago »*v spletni komunikaciji* beseda, ki

1 Po raziskavi Valicona o dosegu in uporabi družbenih omrežij v Sloveniji (v sklopu obširnejše raziskave o dosegih in načinih uporabe različnih vrst medijev in prodajnih kanalov) za leto 2016 je Twitter po številu odprtih profilov na drugem mestu (takoj za Facebookom), po številu dnevnih oz. tedenskih uporabnikov pa na četrtem (še za Instagramom in Snapchatom) (http://www.valicon.net/files/Sporocilo%20za%20javnost%202016-06-23%20(1).pdf; dostop: 28. julij 2017).

2 Konec leta 2017 je Twitter podvojil največje možno število znakov v tvitu, saj je bilo prvotno mogoče zapisati le do 140 znakov.

3 www.fran.si

4 Slovar, ki je dostopen na spletnem naslovu http://lexonomy.cjvt.si/slovar-tviterscine/ (Gantar et al. 2016: 71–76).

nosi ključno sporočilo besedila, zapisana neposredno za znakom #, npr. #poplave #kostanjevica #padavine – angl. hashtag«.[5] V *iPromovem slovarju*, ki prinaša »najobsežnejšo zbirko poslovenjenih izrazov s področja digitalnega marketinga in digitalne tehnologije«, je definicija nekoliko obširnejša:

> Beseda ali fraza, zapisana brez presledkov, ki ima na začetku simbol »hash« (#). Gre za obliko metapodatkovnega taga. Simbol se uporablja predvsem znotraj družbenih medijev in omogoča uporabnikom razvrščanje sporočil po želenih besedah ali frazah, pred katere postavijo znak #.[6]

Izraz *ključnik* je uvrščen tudi v *Razvezani jezik*, kjer najdemo navedena nekatera bolj ali manj ustrezna sinonimna poimenovanja: *ključna beseda*; *zbiraj*, *povezovalnica*, *sledilnica*, *povezava na tviterju*, *iskalnica*, *lojtrišče* in informacijo, da argentinski Slovenci z izrazom *ključnik* poimenujejo obesek za ključe,[7] pri čemer gre verjetno za kalkiranje iz španščine, saj je izraz za obesek za ključe, tj. *llavero*, nastal iz besede ključ (špan. *llave*). Tudi sicer oblika tega leksema ni nova. Zasledimo jo že v *Pleteršnikovem slovarju*, kjer je *ključnik* poimenovanje za kukavičjo lučco in lučnik[8], izraz pa je bil uvrščen tudi v *Besedišče slovenskega jezika z oblikoslovnimi podatki*, vendar očitno ni dovolj zastopan v besedilih, da bi se uvrstil tudi v *Slovar slovenskega knjižnega jezika* (dalje *SSKJ*).

Izraz *ključnik* je verjetno nastal s postopkom poenobesedenja iz stalne besedne zveze *ključna beseda*, vendar pa leksema nista sinonimna. Ključnik je lahko tudi ključna beseda, njegov inherentni del pa predstavlja znak lojtra, ki se piše desnostično. Pri zapisu ne gre za pravopisno pravilo, temveč za pravilo, zaradi katerega je komunikacija sploh mogoča. Le na tak način je namreč mogoče iskanje po izbranem ključniku znotraj neke spletne platforme.[9]

Izraz *lojtra* v pomenu ‘znak iz dveh poševnih in dveh prečnih črtic’ oz. ‘ tipka na računalniški tipkovnici s takim znakom’ se v slovarjih prvič pojavi šele leta 2012, z objavo *Slovarja novejšega besedja* (dalje *SNB*). Zaradi svoje pogovorne motiviranosti (v *SSKJ* je namreč osnovni pomen označen s kvalifikatorjem *nižje pogovorno*) je vzbujal dvome o ustreznosti poimenovanja pri nekaterih jezikoslovcih. V osnutku dokumenta *Slovenski izrazi za <@> in <#> v informacijski tehnologiji* iz leta 2003 je

5 http://lexonomy.cjvt.si/slovar-tviterscine/geslo-2287/

6 https://iprom.si/slovar/?q=hashtag

7 Za potrditev informacije se zahvaljujem Almi Kavčič.

8 Na lučnik se nanašajo tudi zgledi v Gigafidi: najdemo tri pojavitve, in sicer v časopisih Gorenjski glas, 2007; Savinjske novice, 2003; Novi tednik, 2007.

9 Lojtra se v elektronski komunikaciji uporablja še pri storitvah mobilne telefonije (npr. *123# oz. *448# za informacijo o stanju na predplačniškem računu pri obeh največjih mobilnih operaterjih v Sloveniji), v iskalniku Najdi.si se lojtra uporablja za iskanje po poljih, npr. po polju internetnega naslova (#url), besedila (#text), ključnih besed (#keywords), na IRC-u se lojtra uporablja za označevanje kanalov (Michelizza 2015: 129). Zappavigna (2015) meni, da bi se lahko raba ključnikov na Twitterju uveljavila po zgledu označevanja kanalov na IRC-u. Ima pa lojtra v besedilih lahko tudi povsem stilistično funkcijo, npr. kot parno ločilo lahko nastopi v poudarjenih delih besedila, npr. #Hamlet# (Crystal 2001: 35), v vlogi t. i. ekspresivnega ločila (Osrajnik et al. 2015: 50).

Janez Dular, takrat vodja Urada za slovenski jezik, za *afno* in *lojtro* predlagal poimenovanji *ajka* in *višaj*, ki pa se kasneje nista uveljavili. Čeprav v slovarskem delu *Slovenskega pravopisa* iz leta 2001 (dalje SP 2001) najdemo izraz *afna*, *lojtre* v pomenu znaka # ne najdemo. Tovrstna pisna znamenja tudi sicer doslej v pravopisu še niso bila obravnavana. Na »novo skupino pisnih znamenj, ki jih uporabljamo pri sporazumevanju prek spleta in sodobnih komunikacijskih orodij, t. i. pisna znamenja v informacijsko-komunikacijski tehnologiji« opozori *Sodobni pravopisni priročnik med normo in pravopisom* (Dobrovoljc in Jakop 2011: 56–57).

3 KLJUČNIKI ZA KATEGORIZIRANJE TVITOV IN NJIHOVE KOMUNIKACIJSKE VLOGE

Tradicionalni metapodatki so ločeni od glavnega besedila, kot ga vidi uporabnik informacijskega sistema, ključniki pa so vrsta metapodatkov, ki je uporabniku na očeh, vključeni so v glavno besedilo (Zappavigna 2015). Hkrati pa je s ključniki prvič izkazana odločitev uporabnikov, katera tema je po njihovem mnenju najbolj izpostavljena v izbranem besedilu (za tovrstno klasificiranje na spletu se uporabljata tudi izraza *družbeno klasificiranje* in *folksonomija*). Kot nov pojav v jeziku imajo ključniki tudi posebno vlogo v komunikaciji (ibid.).

Različne tuje študije, ki so analizirale ključnike, so pokazale, da lahko v grobem razdelimo funkcijo ključnikov v dve skupini: primarno, ki je namenjena kategoriziranju tvitov, saj s pomočjo ključnikov lahko enostavno iščemo tvite z isto tematiko, in sekundarno, ki je vezana na konverzacijo in pokriva vrsto komunikacijskih vlog (Wikström 2014: 128, Zappavigna 2015, Shapp 2014: 20).

Predvidevamo, da bodo v korpusu pogosteje rabljeni ključniki, torej tisti z višjo frekvenco pojavitev, večkrat v vlogi kategoriziranja tvitov, ključniki, ki se v korpusu pojavijo le redko, pa v kateri od komunikacijskih vlog. Ker želimo preveriti omenjeno tezo, bomo v prvem delu analizirali dva sklopa ključnikov, ki smo jih pridobili iz frekvenčnega seznama[10] v besedilnem korpusu Janes (Erjavec et al. 2018), in sicer v podkorpusu tvitov (Janes-Tweet). Prvi sklop so ključniki, ki imajo frekvenco pojavitev v korpusu višjo ali enako 2000 in jih je skupno 98, v drugem sklopu pa so tri podskupine[11] ključnikov s po 1 pojavitvijo v podkorpusu tvitov in jih je skupno 150.

Med 98 ključniki iz prve skupine jih ima prvih 12 najpogostejših funkcijo kategoriziranja. Gre torej za oznake, s katerimi uporabniki kategorizirajo svoje tvite,

10 V to skupino je bila uvrščena tudi lojtra brez ključnika (#), ki je bila iz nadaljnje analize izločena.

11 Prva skupina obsega 50 ključnikov, razvrščenih po abecedi, od #ajfonmakulcifokus do #ajekejzapomagat, v drugi skupini so ključniki od #koprnorcih do #kopgor, v tretji pa od #prijaznaSIlly do #prijateljiciAnžetaŠizGostilne.

da jih drugi uporabniki lažje najdejo. Ti ključniki so: #plts, #slonews, #junaki, #slochi, #PLTS, #Slovenia, #Ljubljana, #radiobattleSI, #ligaprvakov, #sp14si, #slovenia, #ljubljana.

> *tudi letos bo superpokal na igrišču pokalnega zmagovalca kot zmeraj do sedaj ali ne? #plts #SLOsuperpokal2016*
>
> *Ej, če pa tile naši #junaki niso Carji, pol pa ni nihče. Zakon ste: -) #EuroBasket2013*

Tudi sicer je v analizirani skupini najpogostejših ključnikov le 7 takih (kar znaša tudi približno 7 %), ki jasno izkazujejo komunikacijsko vlogo (#sampovem, #fail, #fact, #lol, #love, #fun, #wtf), vloga kategoriziranja se zdi pri teh ključnikih drugotnega pomena.

> *Ne reče se več, da spi kot dojenček ampak, da spi kot oči. #sampovem*
>
> *Mediji so že brez prvoaprilskih "šal" dovolj trapasti. #fail*

Med pogosteje uporabljanimi ključniki je tudi #Danes, s katerim so nekateri tviti jasno označeni z namenom kategoriziranja, saj iz sobesedila lahko ugotovimo, da gre za oznako oddaje Danes na Planet TV:

> *Več jutri v #Danes.*
>
> *Spet duper prispevki danes v #Danes. Thumbs up!*

V korpusu pa najdemo seveda tudi primere, kjer #Danes (tudi #danes) označuje časovno obdobje sedanjega dne. Tu vloga uporabe ključnika ni povsem jasna, zagotovo pa lahko rečemo, da vloga kategoriziranja ni tako očitna kot v prejšnjih primerih.

> *#Danes ob 10:30 #novosti COBISS+ #predstavitev #vzivo [URL] #COBISS #branje #knjige #IKT*
>
> *#Danes znani rezultati jesenskega roka #matura*

4 VLOGA KLJUČNIKOV V TVITIH

Pri analizi vlog ključnikov v tvitih se bomo oprli na delitev po Wikströmu (2014: 130), ki na podlagi analiziranega gradiva za angleščino razdeli ključnike v naslednjih osem kategorij: (1) oznaka izbrane tematike, (2) oznaka igre, (3) metakomentar besedila tvita, (4) dopolnilo pomena tvita, (5) izražanje čustev, občutij ali razpoloženja, (6) izražanje poudarjalnosti, (7) izražanje humorja, igrivosti (z jezikom), (8) memi in popularna kultura. Prvi dve kategoriji imata zlasti vlogo kategoriziranja, preostalih šest pa ima različne komunikacijske vloge. Za omenjene

kategorije sicer Wikström (ibid.) poudarja, da se med seboj ne izključujejo, temveč je njihova večfunkcionalnost prej pravilo kot izjema. Navaja še, da se pogosto ključniki pojavljajo v nepredvidljivih vlogah, skoraj vedno je prisotna neke vrste igra z jezikom, pogosto pa gre tudi za način izražanja pripadnosti skupnosti uporabnikov Twitterja (Wikström 2014: 149).

Omenjena klasifikacija se izkaže kot uporabna tudi za delitev vlog ključnikov v slovenščini, nekoliko spremenjena je zadnja skupina (Memi in popularna kultura), ki smo jo za potrebe analiziranega gradiva poimenovali Popularna kultura in tradicija. Kot nujno se pri analizi pokaže tudi opozorilo, da se kategorije med seboj ne izključujejo. Razvrščanje v posamezne kategorije je težavno tudi zaradi tega, ker so tviti pogosto del širše konverzacije med uporabniki in kot taki v korpusu pogosto iztrgani iz sobesedila.

ja ampak bi bore malo prodali na danasnji tekmi ;-) #ajetekmazaprtazajavnost

Nekateri tviti so sicer s pomočjo ključnikov najdljivi prek Twitterja, kot npr. zgornji tvit. S klikom nanj na Twitterju lahko ugotovimo, da gre za odgovor na idejo o uzakonjenju točenja piva na stadionih. Navedenemu tvitu pa sledi komentar, da bi na dogodek verjetno zaradi tega tudi kdo več prišel. Na tak način si lahko osmislimo širše sobesedilo, so pa v korpusu tudi tviti, kjer to ni mogoče, saj so jih uporabniki izbrisali oz. so zaprli ali zaklenili račune, na katerih so bili tviti objavljeni.

Pogosto je tvit komentar na dogajanje, ki se predvaja na televiziji, radiu in je v času objave razumljiv sledilcem na Twitterju, gledano s časovne distance pa se natančen pomen zakrije oz. lahko o njem le ugibamo.

Kako šest?!? Jaz vidim samo štiri... Kje sta še dva? #ajetoključ #sp14si

4. 1 Oznaka izbrane tematike

Za ta tip ključnikov je zelo značilno kategoriziranje. Predvidevamo, da se uporabnikom zdi pomembno, da so tviti, označeni s temi ključniki, najdljivi. Pogosto se uporabniki sami dogovorijo za enotno označevanje, včasih pa pobuda za označevanje tvitov z določenim ključnikom pride s strani medija, organizatorja dogodka ipd.

4.1.1 Športna tematika

Ključniki se pogosto uporabljajo za označevanje tvitov s športno tematiko. Pogoste so oznake za večja tekmovanja, svetovna prvenstva, olimpijske igre, npr.

#plts, #slochi, #sp14si, #Rio2016, #F1, #srcebije, ali za ekipe, npr. #junaki, #risi, #olimpija, #NKDomzale, #MojaUnionOlimpija.

Do #slochi imeli 7 medalj (pod SLO), zdaj pa na enih igrah 7! Noro, bravo [@per][12] #TvitajmoZaNase

Olimpijske igre so za nami. Pred vami pa Top 10 senzacij, ki smo jih videli v Riu de Janeiru. #Rio2016 [URL][13]

Še 25 minut do začetka tekme #NKDomzale - [@per]! Skupaj do vrha! [URL]

4.1.2 Televizijske in radijske oddaje

Med pogostejše sodijo tudi ključniki o bolj gledanih oddajah, npr. #soocenje, #odmevi, #tarca, #evrovizija, #ema2017, zlasti o resničnostnih šovih, npr. #BarPlanet, #bigbrotherslo, #slotalent, #znanobraz.

Maa, kolk nam bo lepo v Sloveniji. Nekoč... #odmevi

Isti ključnik lahko označuje, da je v tvitu izraženo mnenje, komentar (zlasti) o oddaji in je hkrati pogosto vključen v skladnjo besedila v tvitu (prvi spodnji primer), lahko pa gre za komentar dogajanja v oddaji (drugi primer), kjer običajno zasledimo ključnik (ali več njih) na koncu tvita.[14] Ta je izven skladnje besedila in opravlja zlasti vlogo kategoriziranja.

Ta #bigbrotherslo je en velik kulturni presežek. #majkemi

"Ta je bla šankovska. " Ja pa dobro sej sedite za šankom ženska. #bigbrotherslo

Podoben je še ključnik #zdajsevrti, ki ga pogosto uporabljajo radijske postaje za promocijo trenutnih vsebin ali uporabniki, ko želijo opozoriti na glasbo ali drugo vsebino, ki se trenutno predvaja in jim je običajno všeč.

#Zdajsevrti #nanananana #HeyJude [URL]

Minutke za dobro slovensko glasbo ... Bil sem daleč od ponorelega sveta: [URL] ♫ #zdajsevrti

12 [@per] v korpusu Janes označuje, da je na tistem mestu v tvitu navedeno uporabniško ime, oznaka [per] pa, da je na tistem mestu navedeno lastno ime.

13 [URL] v korpusu Janes označuje, da je na tistem mestu v tvitu naveden spletni naslov.

14 Raziskava Pertot et al. (2016: 242) je pokazala, da se večina ključnikov (72 %) pojavlja na koncu tvita.

4.1.3 Aktualno dogajanje

Posamezni pogosteje rabljeni ključniki se oblikujejo tudi za pomembne politične dogodke ali družbene teme, npr. #volitve14, #begunci.

> *Oddajo nocoj začnemo že ob 18:45. Ob 19:00 rezultati vzporednih volitev, potem 3 ure in pol vklopov, analiz. Da boste obveščeni. #volitve14*
>
> *V zbirni center na Sentilju ravnokar prispela dva nova avtobusa z begunci. #begunci [URL]*

Kot smo že omenjali, so tviti pogosto v vlogi komentiranja dogajanja, ki se predvaja na televiziji ali radiu. Zgornji primeri ključnikov v veliki meri izkazujejo povezanost pisanja tvitov s spremljanjem drugih medijev.

Zasledimo pa tudi ključnike, ki imajo zelo visoko frekvenco v korpusu, zaradi česar bi lahko predvidevali, da se kot pomembna kaže vloga kategoriziranja, vendar pa pregled gradiva na Twitterju[15] pokaže, da jih uporablja zgolj manjše število uporabnikov (navadno gre za generirane objave, namenjene drugim platformam ali medijem, ki se samodejno objavljajo še v obliki tvitov), npr. #Modrijani_SLO, #izvršba. Zelo pogosto je v tovrstnih tvitih na koncu objavljena povezava na Facebookov profil ali na domačo spletno stran.

> *(Y) če so tudi zate preplesane noči prekratke :) #Modrijani_SLO [URL]*
>
> *28.08.2013 #izvršba 1. javna dražba, lcd tv, omara, sedežna, monitor #Celje [URL]*

V to kategorijo so se uvrstili le redki primeri ključnikov iz druge skupine, torej z eno samo pojavitvijo. Tak primer je ključnik #kopije, za katerega lahko trdimo, da ni najbolje izbran, saj če ga skoraj nihče ne uporablja, po njem verjetno tudi nihče ne išče.

> *Kako zmanjšati nastalo škodo ob izgubi mobilnega telefona? #nasveti #varnost #kopije [URL] [URL]*

4.2 Oznaka igre

Med pogosteje rabljenimi so v korpusu tvitov tudi ključniki, ki označujejo različne igre na Twitterju. Pogosto gre za viralne akcije, ki trajajo lahko le kratek čas,

15 Korpus Janes namreč zakrije osnovne podatke o uporabniku (ime in uporabniško ime), zato smo v analizi ključnikov v različnih primerih, kot že omenjeno, besedila preverili tudi na Twitterju.

lahko pa tovrstni ključniki nastanejo kot pomoč pri zbirateljskih akcijah. Uporaba teh ključnikov je lahko tudi dobrodelna.

4.2.1 Primera #slochi in #TvitajmoZaNase

Ključnika #slochi in #TvitajmoZaNase sta bila dva izmed ključnikov, ki so se uporabljali za komentiranje zimskih olimpijskih iger v Sočiju leta 2014. Te igre so bile za slovenske športnike še posebej uspešne, posledično veliko je bilo zato tudi tvitov z omenjenimi ključniki. Poleg možnosti kategoriziranja oz. razvrščanja tvitov na izbrano tematiko pa je bila uporaba teh ključnikov tudi dobrodelna. Šlo je za akcijo Telekoma Slovenije in Olimpijskega komiteja Slovenije, ki je pod ključniki #TvitajmoZaNase in #slochi[16] zbirala tvite podpore, po koncu iger pa so bili tviti z omenjenimi ključniki preračunani v sredstva, ki so bila podarjena skladu za socialno ogrožene slovenske športnike (Štritof 2014).

> *Jakov Fak po zadnjem streljanju četrti. Za medaljo zaostaja 20 sekund. #slochi*
>
> *Obramba sramotna :-/ Dajmo stisnit v 2. polčasu! #TvitajmoZaNase #den #slo #handball #Rio2016 #Olympics*

4.2.2 Primer #radiobattleSI

Radio Battle je mednarodno radijsko tekmovanje, v katerem uporabniki Twitterja odločajo o tem, kdo vrti (naj)boljšo glasbo.[17] Z radiem Val 202 je na tem tekmovanju leta 2016 kot DJ sodeloval Andrej Karoli, in kdor je želel glasovati zanj, je moral uporabiti ključnik #radiobattleSI ali retvitati tvit, v katerem je bil ta ključnik uporabljen. Ker je bila pomembna samo uporaba ključnika (ta je namreč izbranemu DJ-ju prinašala točke), sama vsebina tvita pa je bila drugotnega pomena, že hiter prelet tvitov s tem ključnikom pokaže, da so tviti, ki bi jih ključnik vsebinsko dopolnjeval, zaznamoval, v manjšini. V nadaljevanju navajamo nekaj primerov tvitov s tega tekmovanja.

> *Mislim, da bi jedla čokolado! #radiobattleSI*
>
> *Loh bi več o rožah vedla! #radiobattleSI*

16 Poleg omenjenih dveh, ki sta bila tudi v korpusu Janes med najpogostejšimi uporabljenimi ključniki (#slochi s frekvenco 13.195, #TvitajmoZaNase pa s 4264 pojavitvami), so v dobrodelni akciji upoštevali še ključnik #slochi2014 in različne variacije zapisov z malo oz. veliko začetnico (Štritof 2014).

17 http://val202.rtvslo.si/radiobattle

Cimr mi teži, da sm skoz na telefonu! #radiobattleSI

Tašča tvita, rad jo imam! #radiobattleSI

sicer ne hodim na volitve, grem se pa #radiobattleSI

4.2.3 Primer #registrska

Pod ključnikom #registrska so zbrani tviti, navadno z objavljeno fotografijo nenavadnih, manj običajnih registrskih tablic. Pobudo za to zbirateljstvo je dal uporabnik Twitterja z uporabniškim imenom roni kordis (@had). Navajamo nekaj primerov tovrstne rabe: MS SERVER, GO LIVE, CE PIKA, KP IDEJ, KR NLP, LJ CAT, MB AJD, SG ZORAN, PO KER·AS, NM MAJA25. Primeri registrskih tablic so povzeti po tvitih s ključnikom #registrska na Twitterju.

4.2.4 Primer #nočnastraža

Ključnik #nočnastraža je rabljen v tvitih uporabnikov, ki dolgo v noč bedijo ali ponočujejo. Pri objavi gre za neke vrste igro, s katero »stražijo« Twitter, pogosto je v tvitih s tem ključnikom tudi povezava na izbrano glasbo.

Zadnje čase pretežno ne spim, pa se mi zdi da samo zato, ker se nadejam #nočnastraža vsakič sproti. :)

[URL] Ura za tango #nočnastraža

4.3 Metakomentar tvita

Kategorije ključnikov, ki sledijo, so manj osredinjene na dejansko kategoriziranje po izbrani tematiki, kot smo lahko videli pri kategorijah ključnikov za oznako izbrane tematike (4.1) in oznako igre (4.2). Še vedno pa gre za skupino ključnikov, ki je razmeroma pogosto rabljena, saj v to skupino sodita dva ključnika iz prve skupine po korpusni frekvenci analiziranih.

Pri ključniku #sampovem (7312 pojavitev) je izražena metajezikovna funkcija, gre za jezik o jeziku, sam ključnik ne prinaša novih informacij, tudi za iskanje se ne zdi ključnega pomena, na nek način deloma tudi poudarja vsebino, vendar se predvsem zadovolji z navedbo nekaterih dejstev, distancira se od nadaljnjega komentiranja.

Za vse,ki se danes odpravljate na morje s poč. prikolicami,opozorilo da se po Vipavski dolini ne smete voziti z njimi. #sampovem))

Ni nekega posebnega delovnega zagona danes. Nekaj je v zraku. #sampovem

Zelo podobna je raba ključnika #fact (v korpusu Janes ima 3719 pojavitev). V prvem primeru spodaj navedenih tvitov gre za željo po (znanstveni) potrditvi pred tem napisanega besedila, pogosto tudi nekega mnenja ali nepomembnega dejstva, drugi primer pa je zaradi vsebine tvita šaljiv, vendar vloga ključnika ostaja v obeh primerih enaka. Raba tega ključnika je lahko tudi ironična, pri čemer želi z njim uporabnik zanikati predhodno besedilo (tretji primer).

Za doječe mamice je brezalkoholno pivo priporočeno. #fact

V Afriki vsakih 60 sekund mine 1 minuta. #fact

vžigalnika ne bom izgubila približno nikoli več #fact [URL]

Iz druge skupine ključnikov, ki imajo v podkorpusu tvitov po eno pojavitev, ima podobno vlogo ključnik #ajetrebašesplohkejrečt. Tudi tu gre za metakomentar tvita, natančneje v nadaljevanju navedene hiperpovezave oz. vsebine, ki se tam prikaže.[18]

#ajetrebašesplohkejrečt ? :D [URL]

4.4 Dopolnilo pomena tvita

Naslednja skupina ključnikov ima vlogo dopolnjevanja napisanega v glavnem besedilu tvita. Vsi analizirani ključniki so bili v korpusu rabljeni samo enkrat, saj gre za kategorijo, pri kateri vloga kategoriziranja ni relevantna. Ključniki te skupine so lahko zelo dolgi, kar ponovno potrjuje domnevo, da v osnovi niso namenjeni temu, da bi uporabniki po njih iskali. Lahko so sestavljeni iz besednih zvez, npr. #prijateljičinaasociacijanatodebilnobesedo, #kopipejstsistem,[19] ali celih stavkov, npr. #kopridemdomov, #koplesescelonoc, #ajerepeževponatisu, #ajepunckaalfantek.

Včasih je tvit brez pojasnila v ključniku lahko povsem nerazumljiv, npr.:

bolanija in julijska krajina. #prijateljičinaasociacijanatodebilnobesedo

V večini primerov pa ključnik prinaša dopolnilno informacijo, pojasnjuje okoliščine, konkretizira predhodno informacijo ipd.

18 Iskanje po ključniku na Twitterju razkrije, da nas hiperpovezava pripelje do vabila oz. t. i. dogodka, objavljenega na Facebooku, za koncert Vlada Kreslina.

19 V tem primeru predvidevamo, da gre za besedno zvezo, saj je običajnejši zapis narazen, čeprav bi lahko bila tudi zloženka (zapis skupaj).

Nabavil... Sedaj pa samo še zaženem xbox... #kopridemdomov [URL]

Ob petih zjutraj sem Pepelka! /.../[20] #koplesescelonoc #clubcirkus #party #ljubljana @ Cirkus [URL]

prehajamo v zadnji stadij implementacije putinovskega razumevanja svobode medijev #kopipejstsistem

kdo bo stiskal toliko bestselerjev #ajerepeževponatisu

A se lahko zmenimo, da je kuzica z roza ovratnico zenskega spola? Mislim, a je res tko tezko? #ajepunckaalfantek #everyfuckingtime

4.5 Izražanje čustev, občutij ali razpoloženja

Za izražanje čustev se uporabljajo ključniki, ki so lahko zelo eksplicitni (#love, #happy) ali bolj posredni (#wtf). Pogosto so ključniki v angleščini oz. iz angleščine prevzeti, četudi je lahko besedilo tvita v celoti v slovenščini.

Pica in kruh sta bila odlična, mimogrede. #happy

evo ena, da si malo odpočijemo in se ne razburjamo #love [URL]

in so ljudje, ki po kampu sprehajajo zelvo. #lol

V korpusu je ta kategorija ključnikov zelo zastopana, saj je precej analiziranih ključnikov iz prve skupine, ki so v korpusu med najpogostejšimi: #lol (3209 pojavitev v podkorpusu tvitov), #love (2995 pojavitev), #fun (2769 pojavitev), #wtf (2484 pojavitev), #happy (2420 pojavitev).

V omenjeno skupino pa lahko uvrstimo tudi nekaj ključnikov, ki se v korpusu pojavijo po enkrat: #ajetomozno, #ajetkoteskobittocn, #ajepondelk.

Komentarji pa.... vauuu!! #wtf #ajetomozno

Pr dohtarjih so uro ze premaknil.. narocen ob 8:30 na vrsti ob 9:30 #ajetkoteskobittocn #banda

Omg. [@per] a mogoce veste kater vlak me je povozu? #fuckedup #ajepondelk ?

Pri interpretaciji oz. ugotavljanju, za katero čustvo, občutje ali razpoloženje (začudenje, zgražanje, negodovanje ipd.) gre v določenih primerih, pogosto pomagajo ključniki, ki so nanizani ob obravnavanem ali pa samo besedilo tvita.

V primeru ključnika #ajetobasketball imamo vsaj dve možnosti interpretacije: pri prvi možnosti se uporabnik sprašuje, ali je to (sploh) basketball oz. košarka, pri

20 Z oznako za izpust /.../ so iz tvitov izločeni emodžiji.

drugi možnosti pa lahko interpretiramo, da je igra tako slaba, da je uporabnika spomnila na risanko A je to.[21]

> *ne, superpokal nikakor ni bla najslabša tekma letos #ajetoBasketball #eurocup #olimpija*

Zapis je pri obeh možnostih enak, interpretacija precej podobna, v primeru, da gre za asociacijo na risanko, bi ključnik lahko uvrstili tudi v zadnjo kategorijo (prim. 4.8). So pa na splošno zaradi narave zapisa ključnikov mogoče težave pri segmentiranju tovrstnega besedila, saj gre običajno za zapis brez zaznamovanih presledkov. Reuter et al. (2016: 24), ki so se ukvarjali s samodejnim segmentiranjem in tokenizacijo teh jezikovnih posebnosti, navajajo primere, kjer lahko pride do težav: npr. #wordsoftheday. Interpretacija tega ključnika v angleščini je možna kot *words of the day* (bolj verjetna) ali *word soft he day* (manj verjetna). Podobno še: #brainstorm, *brainstorm* ali *bra in storm*.

4.6 Izražanje poudarjalnosti

V tej kategoriji so ključniki, ki poudarjajo besedilo tvita, ga ne dopolnjujejo tako kot v skupini 4.4, temveč z drugimi besedami poudarjajo, intenzificirajo že napisano in se tipično pojavljajo na koncu tvita oz. za njegovim glavnim besedilom, če sledi še kakšen ključnik ali hiperpovezava. Med bolj uporabljanimi se za izražanje poudarjalnosti uporablja ključnik #fail (5696 pojavitev v podkorpusu tvitov).

> *Kaj ko bi uporabili kakšen drugi font?! #fail [URL]*
>
> *Kdo se je zatipkal? Dikli[22] je vas v Latviji, kraj v Turčiji pa je Dikili. #fail #ajetakotezkopreveriti*

Ključnik #fail se sicer uporablja tudi znotraj glavnega besedila tvita in v teh primerih nima več tipične vloge poudarjanja predhodno povedanega, saj nastopa kot eden od stavčnih členov, in sicer kot prilastek ali kot povedkovo določilo (v pomenu 'napaka, neuspeh' oz. 'napačen'). V teh primerih je mogoč vpliv tipične poudarjalne rabe na tovrstno rabo ključnika, čeprav težko ugotovimo, kakšen je glavni namen uporabnika, ki se je za tak zapis odločil.

> *Iščem FURS in najdem #fail #url naslov.*
>
> *To pa je kar #fail s kuvertami. :) [URL]*

21 Asociacija na to risanko je namreč v ključnikih zaznana: *Ko se dva stara zobarj odlocita, da bosta zraven se ventilatorje popravljala #Hahahah #ajeto #zobar* (gre prav tako za primer iz korpusa Janes).

22 Krajevni imeni Dikli in Dikili sta v korpusu zabrisani z oznako [per], informacija je pridobljena s pomočjo iskalnika na Twitterju.

V nadaljevanju navajamo ključnike z vlogo poudarjalnosti, ki so imeli v podkorpusu tvitov po eno pojavitev. Ključnik #kopiramo, pri katerem zagotovo lahko opazujemo tudi druge vloge, povzame zapisano v tvitu, dodaja neodobravanje opisanega početja in z eksplicitnim izrazom še poudari prej zapisano.

> *Malo kisel okus, ko na spletni strani slo medija prebereš stavke, ki so enaki kot lastni včerajšnji tviti, ki pa tudi niso izvirni #kopiramo*

Sledita še dva ključnika (#kopijazacajtom, #kopija_zah._najboljsih_sosedov) s po eno pojavitvijo, ki prav tako poudarjata povedano z dodajanjem primerjave (z zahodnimi sosedi) oz. z eksplicitnim izrazom, ki poudari prej zapisano (kopija za cajtom). Ključnik #kopija_zah._najboljsih_sosedov je sicer eden redkih v sklopu pričujoče analize, ki poleg lojtre vsebuje tudi druge nečrkovne znake.

> *Zadnje predcasne #volitve14, pred naslednjimi predcasnimi, ki kot kaze bodo v naslednjih 12ih mesecih. #kopija_zah._najboljsih_sosedov*

> *Gledam #mojaslovenija in ne morem da se nebi spomnila na Mediaset zabavne showe iz 90' #kopijazacajtom*

Pogosto se ključniki, ki poudarjajo prej napisano, pojavljajo v večjem številu znotraj posameznega tvita in tudi z nizanjem izražajo intenzifikacijo. V primeru ključnika #kopimajstore (kjer je Kopi vzdevek za slovenskega hokejista Anžeta Kopitarja) je v glavnem besedilu izpostavljeno, da je igralec zelo dober, s prvim ključnikom uporabnik pove, da ga spoštuje, in z drugim ključnikom ponovi, da je zelo dober oz. da je mojster.

> *raztura od kr ma novo bajto... #respect #kopimajstore*

Sledi sklop ključnikov, ki z retoričnim vprašanjem (nekateri pa še z nizanjem več ključnikov zapovrstjo) poudarjajo sporočilo tvita. Ti ključniki so: #aježepetek, #ajezekonc, #ajevredno, #ajetočudn, #ajetonormalno, #ajetolktezkotoafnonapopatnatastaturo, #ajetoktežko, #ajetofinta, #ajetkotežko, #ajesplohšekdotočen.

> *Teden se mi ne bi mogel začeti "bolje"... Dobila napotnico za oralnega kirurga, ker mora 8. zapustiti mojo čeljust :/ #notcool #aježepetek*

> *Zakaj jih nihce ne povprasa o procesu sprejemanja zakonodaje v EU, recimo? Kajti bore malo je govora o sami EU. #EUvolitve #ajezekonc*

> *Kdaj, #MOL, kdaj? #sramota #zastoji #roska #ajevredno ? #brezveze..... [URL]*

> *Vsakič stehtam cedevito, da je perfekt razmerje. #ajetočudn ?*

> *Krško. 17.1.2013 ob 17 uri. 15 stopinj celzija! #ajetonormalno*

> *če bi me rad naredu živčnega me postavi pred Maca povej da se ful mudi in zahtevaj od mene da najdem @ #ajetolktezkotoafnonapopatnatastaturo*

Jaz sem šel zmenjat kune v kuvance, pa ni bilo problema. Nobenih listkov, ne bosta verjela. #ajetoktežko

#brutalnibob je že pol leta "... samo do konca meseca! " #čudno #ajetofinta ?

Če greš NA greš Z ali S, če greš V greš pa IZ. IN PIKA! #ajetkoteżko

Mislm, da se tud enkrat še ni zgodilo da bi mene nekdo čakal, če se zmenim za drink/meeting/karkoli. #ajesplohšekdotočen

4.7 Izražanje humorja, igrivosti (z jezikom)

V tej skupini so ključniki, vsi s po eno pojavitvijo v podkorpusu tvitov, ki izražajo humor, igrivost z jezikom in so kot del tvita, ki je humoren, šaljiv, z njim neločljivo povezani. Navajamo nekaj primerov: #prijazeljcki, #ajfonmakulcifokus, #ajekul.

Ker so moji prijateljcki zelo prijazni do mene jim pravim kar #prijazeljcki :)

Toj ta fokus /.../ #izolacijadnkizparadiznika #sejdrgacjetoprecejewww #ajfonmakulcifokus #nepomenskihashtagi #ok [URL]

In to še ni vse!! Prvih pet pošiljateljev prejme še čudežno krpico, ki lahko popivna vsaj 473l vode..... #ajekul ?

4.8 Popularna kultura in tradicija

Wikström (2014: 130) v svoji delitvi ključnikov zadnjo kategorijo poimenuje Memi in popularna kultura, saj je opazil, da so v ključnikih pogosti citati iz priljubljenih memov ali drugi citati iz popularne kulture.

Mem je videoposnetek ali fotografija z napisom, navadno šaljiv, ki se kopira in tako hitro razširi med internetnimi uporabniki, pogosto z manjšimi spremembami. Internetno poimenovanje mema je nastalo iz poimenovanja mem (angl. *meme*) v teoriji Richarda Dawkinsa, ki pravi, da se memi na podoben način kot geni prenašajo med ljudmi (Dawkins 2006: 219–220):

> Primeri memov so melodije, zamisli, fraze, moda, načini izdelovanja vrčev ali gradnje obokov. Geni se po genskem skladu širijo tako, da iz enega v naslednje telo potujejo s pomočjo semenčic in jajčec, memi pa se po memskem skladu širijo tako, da iz enih v naslednje možgane potujejo na način, ki bi mu na splošno lahko rekli oponašanje. Če znanstvenik izve za zanimivo zamisel, jo posreduje kolegom ali študentom. Omeni jo v

znanstvenem članku ali predavanju. Če se zamisel prime, lahko rečemo, da se razmnožuje in širi po populaciji od možganov do možganov.

Tipičnih internetnih memov v pričujoči raziskavi za slovenščino nismo zasledili, kar seveda ne pomeni nujno, da se ne uporabljajo. Z načrtnim iskanjem nekaterih memov, omenjenih v raziskavi Wikström (2014: 147–48), smo v podkorpusu tvitov našli le en primer mema #idonteven.[23]

V sejni sobi sedim in poslušam o že odpravljenih, ne popravljenih bugih, in kako naj se odprejo v jiri... #idonteven

Pričujočo kategorijo smo torej poimenovali Popularna kultura in tradicija, saj smo v ključnikih poleg nekaterih elementov popularne kulture zasledili pogoste elemente tradicije. Pri elementih tradicije imamo v mislih zlasti frazeološke enote ali del takih enot, ki so pravzaprav ideje, ki se – podobno kot Dawkinsovi memi – širijo med ljudmi.

V raziskavi smo našli primer, ko se del besedila priljubljene pesmi pojavi v ključniku: #kopridepolnoč, #kopridepolnoc z nadaljevanjem besedila v naslednjem ključniku #kogorijolesesvece. Gre seveda za refren skladbe Silvestrska noč, ki je nastala leta 1971, napisala sta jo Jože Privšek in Dušan Velkaverh, izvaja pa jo Alfi Nipič.[24]

Eni debilneži majo že par dni zapored generalko! #pirotehniki #kopridepolnoč

Celebrating new year in proper way. Happy new year everybody. #kopridepolnoc #kogorijolesesvece... [URL]

V nadaljevanju pa so primeri, ki v ključniku vsebujejo frazeološko enoto ali njen del. Primera kot pri norcih se pogosto uporablja za izražanje velike stopnje, mere česa (na spletu najdemo zapise, kot npr. *boli me grlo k pr norcih*; *slaba sezona se vleče kot pri norcih*; *A še kje uliva ko pr norcih?*), pri obravnavanem zgledu ključnika #koprnorcih iz podkorpusa tvitov sicer brez širšega sobesedila[25] težko razberemo pomen.

ja, kaj pa lahko naredi? Ama nista! #koprnorcih

Naslednja dva primera ključnikov (#kopljemsijamo, #kopljemsebijamo) sta varianti frazema *sam sebi jamo koplje* 'dela take stvari, ki ga spravljajo v nesrečo, so mu v pogubo'.[26]

23 Uporablja se, ko želimo izraziti šok ali ko ne vemo, kaj naj rečemo (http://www.urbandictionary.com/define.php?term=I%20don%27t%20even).

24 https://sl.wikipedia.org/wiki/Silvestrski_poljub

25 Tudi s pomočjo iskanja na Twitterju ne pridemo do širšega sobesedila pogovora. Lahko so predhodni in morebitni naslednji tviti v komunikaciji izbrisani ali pa ima uporabnik zaklenjen profil. (Dostop 3. avgust 2017).

26 Frazem je zabeležen že v SSKJ, od kjer je vzeta tudi razlaga.

Ves tisto k se sred predavanja, ob najbol neprimernem trenutku zacnes napolno smejat, ker si se necesa spomnu? No to. #kopljemsijamo

Kdor teorijo drajsa, prakse ne rabi. :) #kopljemsebijamo

Sledita dva primera ključnikov, katerih vsebina sodi na področje pragmatične frazeologije:[27] #ajeluna, #ajepolnaluna, pogosto kot vprašanje v obliki *Ali je danes polna luna?*, ki kot odziv na neobičajno dogajanje, ravnanje koga izraža začudenje, nejevoljo nad nenavadnim, neobičajnim, nevarnim dogodkom, dogajanjem ali nad veliko količino takih dogodkov v tekočem dnevu.

Haha ko se zaklenes ven iz bajte in potem moras po lojtro pa plezat skozi oken. Podoknicarji so mel tezek biznis #sampravm #ajeluna

Še na tržnici je danes mirnejše. Bela cesta... #rabim #mir #čebeljnjak #ajepolnaluna #stop #naglavne #slušalke

Zadnji primer ključnika se nanaša na znani citat iz Prešernove pesmi Apel in čevljar: »Le čevlje sodi naj Kopitar!«

Teli Slo vremenarji so ratal hudi čevljarji! #kopita

Kot smo že omenili, sta bila v analizo glede na pogostnost pojavljanja v korpusu vključena dva tipa ključnikov: prvi tip, za katerega je značilna visoka pogostnost (frekvenca pojavitev je višja ali enaka 2000), in drugi tip ključnikov, ki v korpusu nastopa s po eno pojavitvijo. Razlogov za redko zastopanost posameznih analiziranih ključnikov druge skupine je več:

- neobičajen podomačeni zapis besede oz. besedne zveze ali zapis v neosnovni obliki (zlasti v primerih sklanjanja), torej v oblikah, ki bi sicer bile običajne za zapis ključnikov in bi hkrati omogočale možnost iskanja,

  #Obrazec za #prijavo delovne dobe (M-4) oddajte do 31.5.! [URL] [@per] #m4 #zpiz

  Nekaj za vaše/naše razvajene otroke z #ajfoni in ostalimi top shit igračkami [URL]

- neobičajen zapis velike ali male začetnice,

  A kdo ve, če se da pri nas nabaviti kakšno revijo o hokeju? Poudarek na NHL-u, seveda! ;) #kopiStar #GoKingsGo

- zatipkana beseda.

  S pravimi #prijatleljicami in dobro #kavo je žilvljenje »IN«

  Hokejisti Los Angelesa igrajo svoj tretji finale Lige NHL (1993, 2012, 2014). Vselej so dobili prvo tekmo finala. #nhlslo #nhl #kopiar

27 Za razpravo o tem primeru se zahvaljujem doc. dr. Nataši Jakop in dr. Mateju Metercu.

5 RAZŠIRJENOST RABE KLJUČNIKOV V TVITIH

Za konec smo želeli še preveriti, v kolikšni meri se ključniki dejansko uporabljajo v tvitih. Izbrali smo 250 naključnih tvitov iz korpusa Janes (podkorpus Janes-Tweet) ter preverili, kako razširjena je raba ključnikov v besedilih tvitov.

Pokazalo se je, da kar 208 tvitov (83,2 %) ne vsebuje nobenega ključnika, kar si lahko razlagamo s tem, da so tviti pogosto del konverzacijskega niza, pri čemer navadno ni potrebe, da bi uporabniki odgovore na nek tvit prepogosto označevali s ključniki, ki bi bili najdljivi in so v rabi pogostejši kot ključniki z različnimi komunikacijskimi vlogami. Analizirani tviti, v katerih smo zasledili ključnike, so največkrat vsebovali enega (28 tvitov oz. 11,2 %), po dva ključnika je imelo 8 tvitov (3,2 %), tri ključnike so imeli 3 tviti (1,2 %), štiri ključnike en tvit (0,4 %) in pet ključnikov dva tvita (0,8 %).[28] Ključniki so torej zelo izstopajoč novejši element računalniško posredovane komunikacije, ki pa zaradi omejenosti na specifične vloge kategoriziranja in nekatere komunikacijske vloge v besedilih tvitov ni tako zelo zastopan, kot bi morda pričakovali.

5 SKLEP

Ključniki, ki s svojo izstopajočo obliko predstavljajo novejši jezikovni element, so v rabi glede na vlogo v tvitu dveh tipov. Na eni strani kot tip metapodatkov služijo za kategoriziranje, na drugi strani pa opravljajo različne komunikacijske vloge. Teza, da bodo v korpusu pogosteje rabljeni ključniki, torej tisti z višjo frekvenco pojavitev, večkrat v vlogi kategoriziranja tvitov, ključniki, ki se v korpusu pojavijo le redko, pa v kateri od komunikacijskih vlog, se ni potrdila v celoti. Ključnikov iz skupine pogosteje rabljenih nismo zasledili samo v kategorijah Dopolnilo pomena tvita, Izražanje humorja, igrivosti (z jezikom) in Popularna kultura in tradicija. V kategorijo Izražanje čustev, občutij ali razpoloženja pa je bilo vključenih kar 5 ključnikov iz skupine bolj frekventnih.

Ključnike smo glede na vlogo v tvitih razdelili v osem kategorij, pri čemer se kategorije med seboj ne izključujejo, isti ključnik lahko nastopa tako v vlogi kategoriziranja kot v kateri od komunikacijskih vlog. Pri slednjih opažamo, da se ključniki pogosteje vključujejo v skladnjo, vendar bodo s tega vidika potrebne nadaljnje raziskave. Nekateri ključniki so lahko zelo dolgi, saj poleg besed lahko vsebujejo tako besedne zveze kot tudi cele stavke, zelo redko pa v ključnikih poleg lojtre nastopajo drugi nečrkovni

28 Za primerjavo smo zastopanost ključnikov preverili še na 100 tvitih z osebne časovnice (dostop: 3. avgust 2017). Rezultati so zelo podobni, saj 80 % tvitov ni vsebovalo nobenega ključnika, po en ključnik je vsebovalo 11 % tvitov, po dva ključnika 4 % tvitov, po tri ključnike so imeli 3 % tvitov in po štiri ključnike 2 % tvitov.

elementi. Ključniki pogosto izkazujejo povezanost pisanja tvitov s spremljanjem drugih medijev, kjer jih tudi sicer lahko zasledimo. Ključniki se torej »selijo« iz svojega prvotnega okolja v druge medije, pogosto jih zasledimo tudi v reklamnih besedilih, vendar pa vseeno večinoma ostajajo znotraj svojih vlog. V večini tvitov iz naključnega vzorca analiziranega korpusa se ne pojavijo (83,2 %), čeprav statistika kaže, da imajo tviti s ključniki dvakrat večji ogled kot tisti brez njih (Reuter et al 2016: 23). Vsekakor bo zanimivo opazovati, kaj se bo z njimi dogajalo v prihodnje, saj se komunikacija na Twitterju in tudi na drugih platformah neprestano spreminja.

Literatura

Arhar Holdt, Špela in Kaja Dobrovoljc, 2015: Zveze samostalnika z nesklonljivim levim prilastkom v korpusih Janes in Kres. Fišer, Darja (ur.): *Zbornik konference Slovenščina na spletu in v novih medijih*, Ljubljana, 25.–27. november 2015. Znanstvena založba Filozofske fakultete, Ljubljana. 4–9. http://nl.ijs.si/janes/wp-content/uploads/2015/11/Konferenca2015.pdf

Arhar Holdt, Špela, Darja Fišer, Tomaž Erjavec in Simon Krek, 2016: Syntactic Annotation of Slovene CMC: First Steps. Fišer, Darja in Michael Beißwenger (ur.)*: Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities*, Ljubljana, Slovenija, 27.–28. september 2016. 3–6. http://nl.ijs.si/janes/wp-content/uploads/2016/09/CMC-2016_Arhar_et_al_Syntactic-Annotation-of-Slovene-CMC.pdf

Dawkins, Richard, 2006: *Sebični gen*. Ljubljana: Mladinska knjiga.

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

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

Erjavec, Tomaž, Darja Fišer in Nikola Ljubešić, 2015: Razvoj korpusa slovenskih spletnih uporabniških vsebin Janes. Fišer, Darja (ur.): *Zbornik konference Slovenščina na spletu in v novih medijih*, Ljubljana, 25.–27. november 2015. Ljubljana: Znanstvena založba Filozofske fakultete. 20–26. http://nl.ijs.si/janes/wp-content/uploads/2015/11/Konferenca2015.pdf

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

Gantar, Polona, Iza Škrjanec, Darja Fišer in Tomaž Erjavec, 2016: Slovar tviterščine. Erjavec, Tomaž in Darja Fišer (ur.): *Zbornik konference Jezikovne tehnologije in digitalna humanistika*. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani. http://nl.ijs.si/jtdh16/JTDH-2016-Proceedings.pdf

Michelizza, Mija, 2015: *Spletna besedila in jezik na spletu: primer blogov in Wikipedije v slovenščini*. Ljubljana: Založba ZRC, ZRC SAZU (Zbirka Lingua Slovenica 6).

Osrajnik, Eneja, Darja Fišer in Damjan Popič, 2015: Primerjava rabe ekspresivnih ločil v tvitih slovenskih uporabnikov in uporabnic. Fišer, Darja (ur.): *Zbornik konference Slovenščina na spletu in v novih medijih*, Ljubljana, 25.–27. november 2015. Znanstvena založba Filozofske fakultete, Ljubljana. 50–56. http://nl.ijs.si/janes/wp-content/uploads/2015/11/Konferenca2015.pdf

Pertot, Katerina, Petrovčič, Maja, Strojan, Nika, 2016: #Analiza novih komunikacijskih elementov na družbenem omrežju @Twitter. Erjavec, Tomaž in Darja Fišer (ur.): *Zbornik konference Jezikovne tehnologije in digitalna humanistika*, Ljubljana, 29. september–1. oktober 2016. Ljubljana: Znanstvena založba Filozofske fakultete. 239–244. http://www.sdjt.si/wp/wp-content/uploads/2016/09/JTDH-2016_Pertot-et-al_Analiza-novih-komunikacijskih-elementov.pdf

Reuter, Jack, Pereira-Martins, Jhonata, Kalita, Jugal, 2016: Segmenting Twitter Hashtags. *International Journal on Natural Language Computing (IJNLC)* 5/4. 23–36. http://aircconline.com/ijnlc/V5N4/5416ijnlc02.pdf

Shapp, Allison, 2014: Variation in the Use of Twitter Hastags. Qualifying Paper in Sociolinguistics. New York University. https://www.nyu.edu/projects/shapp/Shapp_QP2_Hashtags_Final.pdf

Štritof, Polonca, 2014: Kdo bi si mislil, da nas bo ponovno združila lojtra. *Metina lista*. *Mnenja*, *12. 3. 2014*. https://metinalista.si/kdo-bi-si-mislil-da-nas-bo-ponovno-zdruzila-lojtra/

Wikström, Peter, 2014: #srynotfunny: Communicative Functions of Hashtags on Twitter. *SKY Journal of Lingusitics* 27/2014. 127–152. http://www.linguistics.fi/julkaisut/SKY2014/Wikstrom.pdf

Zappavigna, Michele, 2015: Searchable talk: The linguistic functions of hashtags. *Social Semiotics* 25/3. https://www.academia.edu/18317926/Searchable_talk_The_linguistic_functions_of_hashtags?auto=download

Kodno preklapljanje v objavah slovenskih uporabnikov Twitterja

Špela Reher, Darja Fišer

Izvleček

V poglavju predstavimo kvantitativno in kvalitativno analizo kodnega preklapljanja v slovenskih tvitih. Analiza temelji na ročno označenem vzorcu tvitov iz korpusa Janes, za kar smo uporabili lastno označevalno shemo, ki je sestavljena iz petih ravni: jezik in vrsta preklopa, stopnja ortografske in morfološke prilagoditve preklopa v slovenščino ter slovnična vrsta preklopa. Kvantitativna analiza je pokazala, da preklapljanje ni redek pojav, da se pogosteje pojavlja znotraj stavkov, da so preklopi v enaki meri eno- in večbesedni in da vključujejo tudi slovnične besedne vrste. Med jeziki preklopov izrazito prevladuje angleščina, večina kodnih preklopov pa ohranja ortografske in morfološke značilnosti izvornega jezika. S kvalitativno analizo smo ugotovili, da kodno preklapljanje opravlja številne diskurzivne funkcije, kot so referenčna, ekspresivna in poudarjalna, glede na semantično polje pa so preklopi pogosto povezani s popularno kulturo, zlasti TV-oddajami, športom, hrano in Twitterjem. Med kodnimi preklopi smo identificirali številne idiome, frazne glagole, kolokacije in pregovore.

Ključne besede: kodno preklapljanje, prevzete besede, računalniško posredovana komunikacija, korpusno jezikoslovje

1 UVOD

Kodno preklapljanje (angl. *code switching*) je pojav, ko dvojezični govorec znotraj ene izjave ali stavka uporabi več kot en jezik (Gardner-Chloros 2009: 4), pri čemer vsaj enega od jezikov dobro obvlada, drugega pa bodisi zna uporabljati enako dobro kot prvega bodisi ga govori le deloma (Myers-Scotton 2002: 25). V najširšem pomenu pojem vključuje tudi preklapljanje med različicami istega jezika, s katerim govorci izrazijo več kot zgolj goli pomen besed, pri čemer izbirajo med dialekti, registri, stopnjami formalnosti ali intonacijami (Gardner-Chloros 2009: 4).

Poplack (1980: 583) kodno preklapljanje definira kot menjavanje dveh *jezikov* znotraj enega diskurza, stavka ali konstituenta (angl. *constituent*), Myers-Scotton (2000: 132) pa ga opredeli kot uporabo dveh ali več jezikovnih *različic* v enem pogovoru, pri čemer je različica lahko karkoli od genetsko nesorodnih jezikov do dveh slogov istega jezika.

V predstavljeni raziskavi uporabljamo ožjo opredelitev, saj nas zanima preklapljanje med različnimi jeziki, ne pa tudi med dialekti ali registri. S proučevanjem kodnega preklapljanja raziskovalci dobimo vpogled v jezikovne prakse govorcev na stičišču med jeziki. Ker se s preklapljanjem na različne načine srečuje večina ljudi na svetu (Gardner-Chloros 2009: 18), je pomembno in živahno raziskovalno področje.

1.1 Razmejitev kodnega preklapljanja od prevzemanja besed

Avtorji pogosto razmejujejo kodno preklapljanje od kodnega mešanja in rabe izposojenk (angl. *borrowing* ali *loanword*). Izpostaviti pa velja, da v tuji literaturi za razliko od slovenske ne razlikujejo med izposojenkami in tujkami. Toporišič (2004: 131) je kot skupno poimenovanje za besede, ki so v slovenščino prišle iz tujih jezikov in niso dediščina historične slovenščine ali praslovanščine, predlagal termin prevzeta beseda, kar kot krovni izraz za tujke in izposojenke v pričujoči raziskavi uporabljamo tudi mi.

Na splošno velja, da kot tujejezični elementi najpogosteje nastopajo posamične besede, običajno občnoimenski samostalniki (Myers-Scotton 1997: 180), vendar ni splošno sprejetega kriterija za razlikovanje med prevzetimi besedami in preklopi. Eden od predlaganih kriterijev za ločevanje med njimi je morfofonemska integracija besed (Gardner-Chloros in Weston 2015: 196). Izposojenke se prilagodijo morfologiji in vrstnemu redu besed v jeziku, v katerega so prevzete, od

izvornega jezika pa obdržijo zgolj etimologijo ali določene elemente fonologije. Drug kriterij je, da so prevzete besede razširjene v celotni skupnosti in so tako dostopne tudi enojezičnim govorcem (Poplack 2015: 920). Tretji kriterij upošteva sinonimnost besed, in sicer naj bi prevzete besede pogosteje zapolnjevale leksikalne vrzeli, kodni preklopi pa so zgolj dodatna možnost za izražanje poleg že obstoječega ekvivalenta (Gardner-Chloros in Weston 2015: 196). Myers-Scotton (1997: 191–204) kot merilo za ločevanje predlaga absolutno in relativno pogostost, pri čemer izpostavi dve težavi. Prva je absolutno število pojavitev (če se obe obliki pojavita le nekajkrat), druga pa arbitrarnost pri določanju relevantne relativne frekvence.

Z drugačnega zornega kota kodno preklapljanje in izposojanje nista strogo ločeni kategoriji, temveč del kontinuuma možnih sprememb zaradi stikov med jeziki (npr. Myers-Scotton 2002, Thomason 2001). Po tej teoriji se vsaka prevzeta beseda najprej pojavi kot kodni preklop, sčasoma pa se nekatere generalizirajo oz. intergrirajo. Po vzoru sorodnih raziskav (npr. Rosenberg 2013, Deuchar 2006) smo se odločili, da bomo v pričujočem poglavju uveljavljenost prevzetih besed določili pragmatično glede na njihovo (ne)vključenost v jezikovne priročnike (Slovar slovenskega knjižnega jezika, Slovenski pravopis in Sloleks). Razlikovanje tako temelji na predpostavki, da so prevzete besede integrirane in uporabljane do te mere, da so na voljo tudi enojezičnim govorcem. Čeprav se zavedamo, da s tem ne bomo zajeli tiste leksike, ki je med govorci sicer prevzeta, a v referenčnih priročnikih zaradi različnih konceptualnih in uredniških odločitev ni kodificirana, kar mdr. velja za številne pogovorne izraze, menimo, da je ob zavedanju omejitev to zadosten (in v tem trenutku metodološko najbolj transparenten) objektiven kriterij za predstavljeno raziskavo.

1.2 Vrste kodnega preklapljanja

Odvisno od obsega preklopa se kodno preklapljanje deli na znotrajstavčno (angl. *intrasentential*) in medstavčno (angl. *intersentential*) (Myers-Scotton 1997: 3). Pri medstavčnem preklapljanju se segmenti, tvorjeni v vrinjenem jeziku (glej razdelek 1.4), raztezajo čez cel stavek (ali več stavkov), do znotrajstavčnega preklapljanja med matičnim in vrinjenim jezikom pa pride znotraj enega stavka.[29]

Nekateri avtorji poleg znotrajstavčnega in zunajstavčnega ločujejo še t. i. pristavčno preklapljanje (angl. *tag switching*), ki pomeni vstavljanje tujejezičnih elementov (pristavkov, pojasnjevalnih dodatkov, vzklikov) na začetku ali koncu

29 V tem prispevku stavek razumemo v nadaljevanju kot najmanjšo samostojno enoto jezikovnega sporočila, ki bi se v standardni slovenščini začela z veliko začetnico in zaključila s končnim ločilom, kar je skladno tudi s Smernicami za označevanje korpusa slovenskih tvitov Janes (2016).

stavka, kar daje dvojezični karakter sicer enojezičnemu stavku, zaradi česar Poplack (1980) to vrsto preklapljanja poimenuje tudi simbolično (angl. *emblematic*) (Appel in Muysken 2005: 118). Za potrebe naše raziskave smo tovrstne preklope šteli v kategorijo znotrajstavčnega preklapljanja.

1.3 Kodno preklapljanje v računalniško posredovani komunikaciji

Raziskovalci so kodno preklapljanje sprva proučevali v okviru dvojezičnosti, bodisi pri posameznih govorcih bodisi v celotnih skupnostih, in sicer v govorjenih besedilih, najpogosteje v spontanem neformalnem govoru (Sebba 2012: 97). Raziskave je mogoče v grobem razdeliti v tri skupine: sociolingvistične/etnografske, pragmatične/analiza pogovora in slovnične. Zanimanje za preklapljanje v pisnih besedilih se je razvijalo počasneje (Gardner-Chloros in Weston 2015: 182–184). Najnovejše študije kodnega preklapljanja se osredotočajo na računalniško posredovano komunikacijo (Sebba 2012: 98, 100).

Ena najvidnejših raziskav s področja kodnega preklapljanja za slovenščino je sociolingvistična analiza jezika slovenske diaspore v Clevelandu (Šabec 1995). Čeprav je bila izvedena v obdobju pred razmahom računalniško posredovane komunikacije in zato tudi ne pokriva tovrstnega gradiva, je njen pristop, v katerem izstopa iz takratne izrazito močne strukturalistične slovenistične tradicije, in govorico ameriških Slovencev obravnava kot pojav jezika v stiku, relevanten tudi za raziskave spletne komunikacije.

Zaenkrat še ni splošno sprejete metodologije, ki bi upoštevala specifične značilnosti internetnega jezika. Raziskovalci tako črpajo iz različnih okvirov, razvitih za analizo govorjenega diskurza, pri čemer prevladujeta pragmatični in sociolingvistični pristop, ne toliko slovnični in lingvistični (Androutsopoulos 2013: 668). Raziskave zajemajo raznolik nabor žanrov in sociolingvističnih okoliščin ter uporabljajo različne metodološke pristope, vendar je bil zaenkrat v večini uporabljen (zgolj) kvalitativni pristop, saj so se raziskave osredotočale predvsem na izražanje družbene identitete, ne pa na empirično preverjanje hipotez na velikih korpusih (ibid.: 679).

Raziskave kodnega preklapljanja v računalniško posredovani komunikaciji je glede na družbene okoliščine mogoče razdeliti na dva glavna sklopa, in sicer tiste, ki se ukvarjajo z izseljenci, diasporami in etničnimi manjšinami, ter tiste, ki se ukvarjajo s preklapljanjem med mlajšimi udeleženci in globalno-lokalno (»glokalno«) rabo jezika, ki je povezana z glasbo in medijsko kulturo. Tu gre predvsem za stik med angleščino in nacionalnim jezikom, pri katerem avtorji opažajo

»minimalni bilingvizem«, tj. da govorci uporabljajo ustaljene angleške besedne zveze (npr. pozdrave, medmete in diskurzne povezovalce, slogane) (ibid.: 678).

Androutsopoulos (ibid.: 674) je zbral najpomembnejše raziskave, v katerih so preučevali kodno preklapljanje v spletnih klepetalnicah (angl. *Instant Relay Chat* oz. IRC), v elektronski pošti, na spletnih forumih in v SMS-ih, in sicer v javni in zasebni komunikaciji ter različnih jezikovnih kombinacijah, najpogosteje v paru z angleščino. Z razvojem novih medijev in družbenih omrežij se avtorji vse pogosteje odločajo tudi za preučevanje kodnega preklapljanja na Twitterju, Facebooku, blogih, v komentarjih na spletne novice in drugih spletnih okoljih.

1.4 Okvir matričnega jezika

Okvir matričnega jezika (angl. *Matrix Language Framework* oz. *MLF*), ki ga je razvila Carol Myers-Scotton, temelji na predpostavki, da jezika, ki sta v stiku, nista enakopravna, temveč ima eden (t. i. matrični oz. glavni) dominantno vlogo in zagotavlja morfosintaktični okvir, v katerega je mogoče vstavljati elemente iz drugega jezika (t. i. vrinjeni jezik oz. jezik, ki prispeva tuje prvine), bodisi kot posamične elemente ali daljše zveze, ki jih Myers-Scotton imenuje »otoki vrinjenega jezika« (angl. *Embedded Language Islands*). Izpostaviti velja, da za razliko od naše raziskave Myers-Scotton v ta model ni vključila medstavčnih preklopov.

Stavki, v katerih pride do kodnega preklapljanja, lahko tako vsebujejo tri vrste konstituentov:

1. otoke matričnega jezika (MJ), ki vsebujejo samo morfeme iz matričnega jezika,
2. otoke vrinjenega jezika (VJ),
3. elemente MJ + VJ, ki vsebujejo material iz obeh jezikov znotraj enega elementa.

Prvo in tretjo skupino ureja slovnica matričnega jezika, za otoke VJ pa je značilno, da so tvorjeni v skladu s slovničnimi pravili vrinjenega jezika (Myers-Scotton 1997: 6). Za kodno preklapljanje morajo govorci poznati vsaj nekaj morfemov iz vrinjenega jezika, sicer pa ni potrebno, da tuji jezik obvladajo (ibid.: 8). Vsak od jezikov ima tudi drugačen sociolingvistični status – vsaj v vrsti interakcije, kjer pride do kodnega preklopa, pogosto pa tudi v družbi na splošno (ibid.: 20).

Druga pomembna predpostavka modela je razlikovanje med sistemskimi in vsebinskimi morfemi. Termin vsebinski morfemi bolj ali manj ustreza polnopomenskim/leksikalnim besedam, sistemski morfemi pa funkcijskim/slovničnim besedam (ibid.: 48). Prvi namreč prispevajo semantično in pragmatično komponento

sporočila, medtem ko sistemski morfemi delujejo kot slovnični sistem, ki k sporočilu prispeva obliko (ibid.: 255). Myers-Scotton sicer opozarja, da kategorij ni mogoče vedno popolnoma ločiti, saj je lahko npr. predlog enkrat v prvi, drugič pa v drugi kategoriji (ibid.: 101–122).

Glavni cilj modela MLF je pojasniti, na kakšen način se lahko jezikovni material iz vrinjenega jezika pojavi v matričnem jeziku oziroma ali se sploh lahko pojavi. Odgovore na to naj bi dale slovnične informacije na abstraktni ravni jezikovne kompetence, ne zunanje okoliščine, v katerih pride do preklapljanja (ibid.: 242).

V pričujoči raziskavi smo se opirali na model MLF, nismo pa uporabili klasifikacije po Myers-Scotton, temveč smo sledili Smernicam za označevanje korpusa Janes (2016), kjer je predvidenih 11 besednih vrst in kategorija »neuvrščeno«.

1.5 Slovnične omejitve pri kodnem preklapljanju

Mnenje, da kodno preklapljanje urejajo univerzalna načela, ne partikularne omejitve, je sprejeto med mnogimi raziskovalci, toda med njimi ni soglasja, kakšna ta načela so. Nekateri vztrajajo, da je iskanje tovrstnih splošnih načel brezplodno (npr. Gardner-Chloros), drugi trdijo, da omejitve sploh ne obstajajo, tretji pa menijo, da se morebitne omejitve razlikujejo med različnimi skupnostmi (npr. Muysken) (Poplack 2015: 919–920).

V sistemu, ki ga je oblikoval Joshi (1982), večina omejitev temelji na splošni omejitvi preklapljanja slovničnih besednih vrst (določila, kvantifikatorji, predlogi, morfemi, ki izražajo glagolski čas, vezniki, zaimki, morfemi za izražanje svojine, pomožni glagoli) (Clyne 2009: 748). Podobno načelo z nekoliko drugačno terminologijo (t. i. *System Morpheme Principle* oz. načelo sistemskih morfemov) je predlagala Myers-Scotton, ki predpostavlja, da morajo sistemski morfemi (določila, števniki, svojilni pridevniki, kopule, vezniki, končnice za spol in sklon ter glagolski čas in vid), ki se pojavijo v elementih MJ + VJ, prihajati iz matričnega jezika (Myers-Scotton 1997: 120–121).

Vendar je teba poudariti, da so novejše raziskave in predvsem raziskave v različnih jezikovnih kombinacijah za vsako domnevno omejitev našle nasprotne primere, izkazalo pa se je tudi tudi, da omejitve niso univerzalne, temveč kvečjemu veljajo za določene jezikovne smeri (Myers-Scotton 1997: 34). Thomason (2001: 151) v zvezi s tem poudarja, da so odločilni dejavniki na strani posameznika, ki lahko vselej prevladajo nad morebitnimi slovničnimi in drugimi omejitvami. Zato je treba upoštevati, da vse predlagane omejite niso kategorične, temveč le odražajo tendence v določenih okoliščinah (Gardner-Chloros 2009: 154).

1.6 Funkcije kodnega preklapljanja

Klasifikacije funkcij kodnega preklapljanja so bile razvite v okviru raziskav, ki so preučevale preklapljanje v govorjenih (neformalnih) besedilih, a so uporabne tudi za opisovanje kodnih preklopov v računalniško posredovani komunikaciji.

Myers-Scotton trdi, da je motivacija za kodno preklapljanje predvsem družbena – govorci preklopijo v drug jezik, ker jim omogoča projiciranje druge dimenzije svoje osebnosti ali vplivanje na odnose z drugimi udeleženci v pogovoru. Tuje besedne zveze lahko izberejo tudi zato, ker bolje izpolnjujejo semantične/pragmatične namene, npr. ker se spremeni poudarek zaradi drugačnega besednega reda ali v primeru idiomov, katerih pomena ni mogoče natančno prenesti v drug jezik. Kodno preklapljanje brez očitne motivacije pa je po njenem del posebnosti jezikovne rabe, s katero govorec izraža svojo identiteto (Myers-Scotton 1997: 248–251).

Androutsopoulos (2013: 681) je diskurzne funkcije kodnega preklapljanja v računalniško posredovani komunikaciji opredelil na podlagi klasifikacij, ki pojasnjujejo preklapljanje v pogovorih, in sicer naj bi ljudje preklapljali iz naslednjih razlogov:

- v formulaične diskurzivne namene (pozdravi, poslovilni pozdravi),
- za tvorjenje kulturnospecifičnih žanrov (poezija, šale),
- za navajanje odvisnega oz. poročanega govora (ki je v nasprotju z avtorjevo izjavo),
- za ponovitev izjave z namenom poudarjanja,
- za naslavljanje določenega naslovnika, kot odziv na jezikovno izbiro predhodnika ali z namenom upreti se jezikovni izbiri drugih udeležencev,
- za kontekstualizacijo spremembe teme ali perspektive, poudarjanje razlike med dejstvi in mnenjem, informativno in čustveno izjavo ipd.,
- z namenom zaznamovati, kaj je izrečeno v šali in kaj zares,
- za izražanje strinjanja ali nestrinjanja.

2 ZASNOVA RAZISKAVE

Namen predstavljene raziskave je bil proučiti kodno preklapljanje v tvitih slovenskih uporabnikov, za kar smo uporabili kombinacijo kvantitativnih in kvalitativnih analiz na vzorcu ročno označenih preklopov. V nadaljevanju razdelka

predstavljamo raziskovalna vprašanja, način vzorčenja tvitov, označevalno shemo in potek označevanja.

2.1 Raziskovalna vprašanja

V kvantitativnem delu raziskave so nas zanimale naslednje splošne značilnosti preklapljanja: ali je delež preklopov odvisen od stopnje standardnosti besedil (glej poglavje Korpus slovenskih spletnih uporabniških vsebin Janes), kakšna je njihova distribucija glede na jezik preklopa ter razmerje med znotrajstavčnimi in medstavčnimi preklopi. Za boljše razumevanje slovničnih omejitev preklapljanja smo preverili razmerje preklopov na račun polnopomenskih (samostalnike, glagole, pridevnike) in slovničnih besed (predloge, veznike, zaimke). Pri proučevanju stopnje integracije preklopov v slovenščino smo analizirali delež preklopov, ki so slovenščini prilagojeni v zapisu in oblikoslovju. Ker korpus Janes vsebuje podatek o spolu avtorjev besedil (Erjavec et al. 2018), smo lahko preverili tudi, ali se pogostost in značilnosti kodnih preklopov razlikujejo glede na spol avtorja.

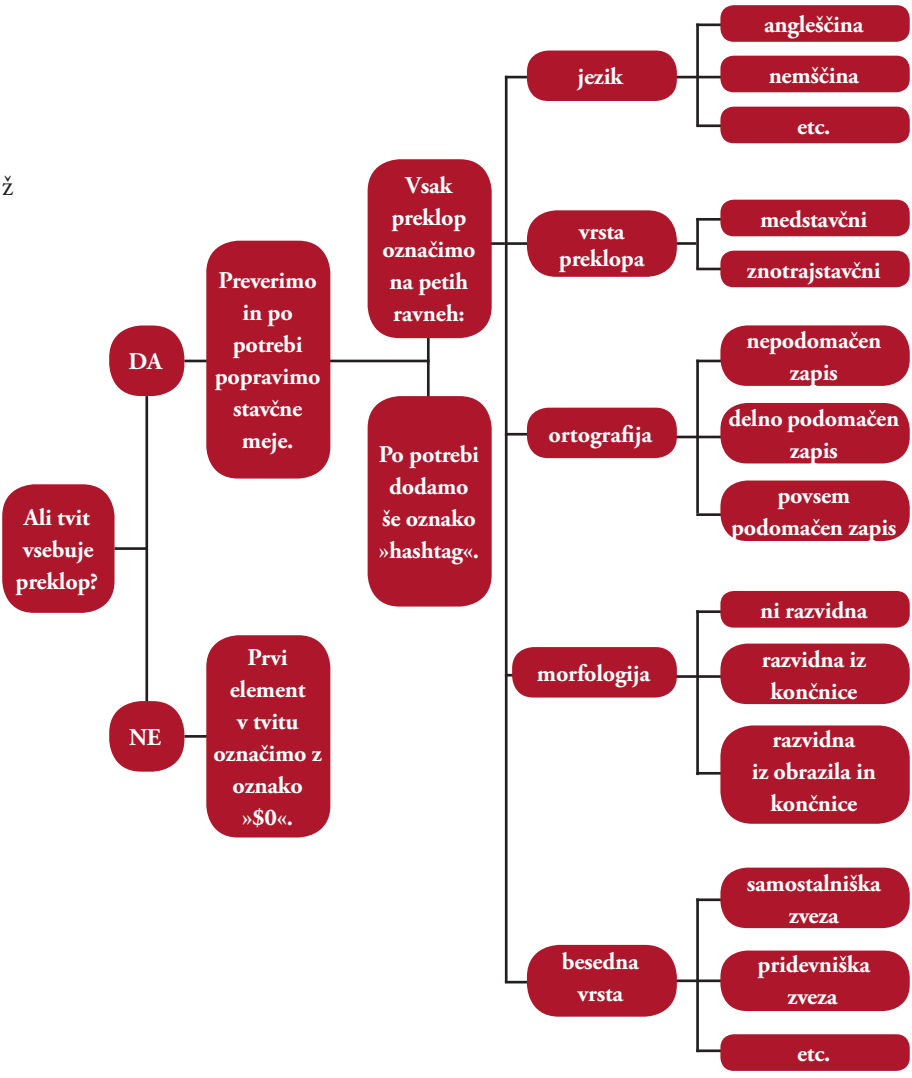
S kvalitativnimi analizami smo podrobneje proučili diskurzne funkcije kodnega preklapljanja, tematska polja, v katera preklopi sodijo, ter morebitne frazeološke posebnosti kodnih preklopov.

2.2 Označevalna shema

Pred izdelavo raziskovalnega korpusa smo na podlagi teoretičnih izhodišč oblikovali označevalno shemo in *Smernice za označevanje kodnega preklapljanja v korpusu slovenskih tvitov JANES* (Reher 2017), ki smo jo tudi preizkusili na manjšem vzorcu naključnih tvitov. Označevanje za kvantitativni del raziskave je potekalo na odprtokodni platformi WebAnno, ki je namenjena ročni anotaciji korpusov (Erjavec et al. 2016: 2), in sicer smo vsak kodni preklop ročno označili na petih ravneh. Kadar je bil preklop v ključniku, smo to označili z dodatno oznako. Potek označevanja in upoštevane kategorije so prikazane na Sliki 1.

Pri označevanju preklopov nismo upoštevali lastnih imen (uporabniških imen na Twitterju, osebnih imen, zemljepisnih imen, naslovov filmov in pesmi ipd.). Ta so v tvitih sicer zelo pogosta, a za jezikoslovno proučevanje preklopov niso zanimiva, saj so uporabljana skoraj izključno citatno. Za označevanje besedne

vrste preklopa smo uporabljali *Smernice za označevanje korpusa Janes*[30] ter upoštevali sobesedilo, v katerem se preklop pojavi.



Slika 1: Shema poteka označevanja kodnih preklopov v programu WebAnno.

Pri analizi razlik glede na jezik preklopa smo se omejlili na tri najpogostejše jezike, in sicer angleščino, nemščino in hrvaščino/srbščino/bosanščino (zaradi velike podobnosti med jeziki ločevanje med njimi v kratkih, pogosto celo enobesednih sklopih ni bilo smiselno). Pri analizi prilagoditev preklopa na slovenski sistem

30 http://nl.ijs.si/janes/wp-content/uploads/2014/09/JANES-jezikoslovne-smernice-v0.9.pdf

zapisovanja in pregibanja besed smo upoštevali le angleščino in nemščino, saj bi bilo zaradi visoke stopnje prekrivnosti oblik s slovenščino pri hrvaških/srbskih/bosanskih tvitih težko zanesljivo ugotavljati, ali je zapis prilagojen slovenščini ali je izvorno hrvaški/srbski/bosanski.

Za kvalitativni del raziskave smo si med označevanjem beležili še naslednje informacije:

- tematsko/semantično polje (npr. IT, Twitter, šport),
- ali je kodni preklop idiom, metafora, stalna besedna zveza ali pregovor,
- druga opažanja, ki bi se lahko izkazala kot zanimiva.

Razlogov, zakaj v tem delu nismo vnaprej pripravili zaključenega nabora kategorij tem in funkcij, je več. Med pregledom relevantne literature smo ugotovili, da zaenkrat še ni uveljavljene klasifikacije diskurznih funkcij kodnega preklapljanja v računalniško posredovani komunikaciji. Poleg tega je pripisovanje motivacije pogosto subjektivno, upoštevati pa je treba tudi, da je v označenem korpusu zbrana množica tvitov, ki so iztrgani iz konteksta, zato je še toliko težje presoditi, zakaj je nekdo uporabil tujejezično prvino.

2.3 Priprava vzorca

Na podlagi tvitov iz korpusa Janes 0.4 (Fišer et al. 2016) smo pripravili raziskovalni korpus tvitov, ki vsebujejo vsaj eno tujejezično prvino, in sicer v dveh korakih. Prvi vzorec (v nadaljevanju »splošni vzorec«) je vseboval 1600 naključnih tvitov z vsaj eno besedo, ki je bila računalniško označena kot tuja (z oznako Nj). Polovica tvitov je bila standardnih (oznaka L1), polovica nestandardnih (oznaki L2 in L3). Dodaten pogoj je bil, da gre za tvite uporabnikov, ki imajo vsaj 1000 tvitov, ter da je polovica avtorjev ženskih, polovica pa moških uporabnikov. V raziskavi smo upoštevali samo tvite zasebnih uporabnikov.

Med označevanjem splošnega vzorca se je izkazalo, da je v njem veliko lažno pozitivnih zadetkov oziroma tvitov, v katerih ni kodnega preklopa. Do tega prihaja zaradi šuma pri oblikoskladenjskem označevanju tvitov, ki so pogosto napisani v nestandardni slovenščini, zaradi česar je sicer slovenskim, a označevalniku neznanim nestandardnim besedam pogosto po pomoti pripisana oznaka Nj. Zato smo ustvarili še drugi vzorec (v nadaljevanju »fokusirani vzorec«), v katerega smo vključili 1600 tvitov, polovico ženskih in polovico moških avtorjev, vse z oznako nestandardno (L2, L3). Dodatno smo želeli rezultate izboljšati tako, da smo vanj vključili po 20 naključnih tvitov 80 avtorjev, ki imajo v korpusu Janes 0.4 največ tujih besed (Nj). Tako smo dobili vzorec tvitov oseb, za katere predvidevamo, da pogosto uporabljajo kodno preklapljanje.

3 KVANTITATIVNA ANALIZA

3.1 Splošne značilnosti preklapljanja

3.1.1 Pogostost preklopov

Najprej smo analizirali, kolikšen delež tvitov v izdelanem vzorcu vsebuje vsaj en kodni preklop. Ker smo že med označevanjem splošnega vzorca ugotovili, da je v raziskovalnem korpusu veliko šuma, smo prilagodili parametre za drugi, fokusirani vzorec, da bi dobili več pravih preklopov. Iz Tabele 1 je razvidno, da nam je to uspelo, saj se je delež relevantnih tvitov povečal s 26 % na 43 %. Kljub temu je zaradi različnih razlogov (v celoti tujejezični tviti, napake pri avtomatskem označevanju, neupoštevanje povezav na skladbe, članke ipd.) delež tvitov, ki smo jih morali izbrisati ročno, ostal dokaj velik. Ročno označeni korpus preklopov v tvitih tako vsebuje skupaj 1100 oz. približno tretjino izhodiščnih tvitov.

Tabela 1: Število in delež tvitov z vsaj enim kodnim preklopom.

	Splošni vzorec	Fokusirani vzorec	Skupaj
Število tvitov v raziskovalnem korpusu	1600	1600	3200
Število tvitov z vsaj enim preklopom	416	684	1100
Delež relevantnih tvitov	26 %	43 %	34 %

Nadalje nas je zanimalo, koliko je bilo v označenem vzorcu vseh kodnih preklopov. Rezultati so predstavljeni v Tabeli 2. Ne presenečа, da je preklopov več kot tvitov (skupaj 1381), saj lahko posamezen tvit vsebuje več kot en preklop. V povprečju posamezen tvit vsebuje 1,26 preklopa. V šestih tvitih so uporabljeni po štirje kodni preklopi, kar je tudi najvišje število preklopov v posameznem tvitu (primer [1]).

[1] uf, good luck. jst sm red velvet kapkejke jedla. biskvit, pofarban na rdeč. ne štekam hajpa razn tega da je pretty.

Tabela 2: Število in delež preklopov v korpusu preklopov.

	Splošni vzorec	Fokusirani vzorec	Skupaj
Skupno število preklopov	507	874	1381
Število (delež) medstavčnih preklopov	160 (32 %)	298 (34 %)	458 (33 %)
Število (delež) znotrajstavčnih preklopov	347 (68 %)	576 (66 %)	923 (67 %)
Povprečno število preklopov na tvit	1,22	1,28	1,26

Pri analizi splošnega vzorca, ki je vseboval 507 preklopov, smo ugotovili, da je v bolj standardnih besedilih delež kodnega preklapljanja nižji, saj se tri četrtine vseh preklopov pojavijo v tvitih z oznako standardnosti L2 (41 %) ali L3 (34 %), le četrtina pa jih je v tvitih z oznako L1. Standardnost besedil je bila označena avtomatsko, zato smo preverili, ali je morda prišlo do šuma zaradi nezanesljivosti avtomatskega označevanja. Izkazalo se je, da so deli tvitov, ki so zapisani v slovenščini, večinoma slovnično in pravopisno pravilni, tujejezični dodatek pa je (pogosto v obliki ključnika) dodan na konec slovenskega besedila (primer [2]):

[2] Ob navijanju komentatorjev za odpoved veleslaloma in druge serije poletov vsaj naši športniki dajejo vzgled. Bravo #orli #fairplay.

3.1.2 Jezik preklopov

V označenem korpusu preklopov smo identificirali 9 različnih jezikov, pri čemer ne razlikujemo med bosanščino, hrvaščino in srbščino. 90 % vseh preklopov je v angleščini, sledijo hrvaščina/srbščina/bosanščina (4,6 %) in nemščina (3,3 %), medtem ko so preklopi v ostale jezike zelo redki (skupaj 2,1 %, glej Tabelo 3).

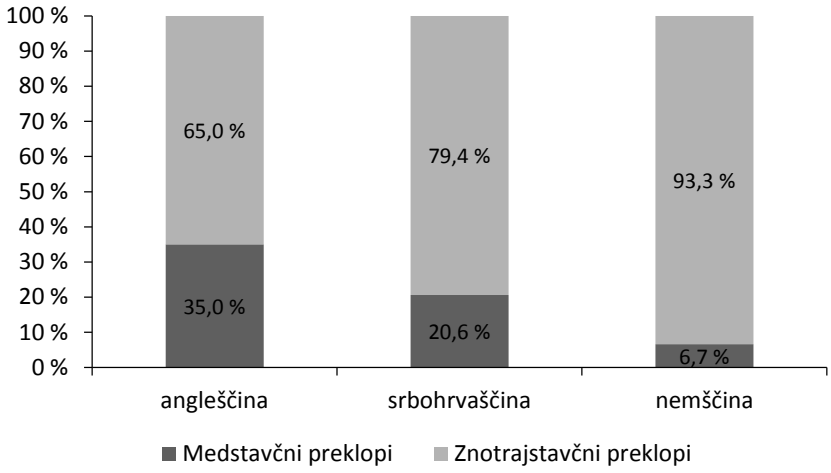
Tabela 3: Število in delež preklopov po jezikih v korpusu preklopov.

	Število preklopov	Delež
angleščina	1244	90,1 %
hrvaščina/srbščina/bosanščina	63	4,6 %
nemščina	45	3,3 %
latinščina	9	0,7 %
španščina	9	0,7 %
francoščina	6	0,4 %
italijanščina	3	0,2 %
arabščina	1	0,1 %
poljščina	1	0,1 %
skupaj	1381	100 %

Nato smo preverili, ali se razmerje med jeziki spremeni glede na vrsto preklopa. Razlike niso velike, vendar je delež angleščine med medstavčnimi preklopi še višji (95 %), manj je hrvaških/srbskih/bosanskih (2,8 %) in nemških preklopov (le 3 primeri), med ostalimi pa sta dva španska preklopa, dva latinska (oba »Mea culpa«) ter po eden iz italijanščine in arabščine.

Nasprotno je pri znotrajstavčnih preklopih razpršenost jezikov nekoliko večja: delež angleščine je nižji (87,5 %), več pa je hrvaščine/srbščine/bosanščine (5,4 %), nemščine (4,6 %) in ostalih jezikov (skupaj 23 primerov oziroma 2,5 %). Če se navežemo na kvalitativno analizo (glej razdelek 3.3), lahko po pregledu preklopov iz redkeje zastopanih jezikov ugotovimo, da gre predvsem za uveljavljene formulaične izraze (npr. »bon voyage«, »žnesekua«, »buenas noches«) ali kulturnospecifične pojme (npr. »prosecco«, »el clasico«, »bachata«).

Za najpogostejše tri jezike smo preverili, kakšno je razmerje med medstavčnimi in znotrajstavčnimi preklopi. Pri angleščini je razmerje približno 1 : 2, pri hrvaščini/srbščini/bosanščini približno 1 : 4, pri nemščini pa kar 1 : 14 v korist znotrajstavčnih preklopov, kar je razvidno tudi z Grafa 1.



Graf 1: Deleži medstavčnih in znotrajstavčnih preklopov v posameznem jeziku.

3.1.3 Dolžina preklopov

Že iz Tabele 2 je razvidno, da je znotrajstavčnih preklopov več kot medstavčnih, in sicer v razmerju 1 : 2. Razpon dolžine preklopov, merjene v besedah, znaša od 1 do 13, vendar je preklopov, daljših od 8 besed, zelo malo (vsega skupaj 24 primerov), najštevilčnejši pa so enobesedni preklopi, ki predstavljajo polovico vseh preklopov v označenem korpusu.

Ugotovitev, da govorci najpogosteje uporabljajo enobesedne tujejezične elemente, ni presenetljiva, saj lahko posamezne besede uporabijo tudi ljudje, ki tega jezika pravzaprav sploh ne obvladajo. Je pa ta rezultat pomemben, saj izpostavlja

slabost pristopov, ki vse enobesedne elemente štejejo za izposojenke, saj se na ta način izgubi pomemben del gradiva.

Povprečne dolžine znotraj- in medstavčnih preklopov, ki jih navajamo v Tabeli 4, kažejo, da so medstavčni preklopi v povprečju 70 % daljši od znotrajstavčnih. Glede na jezik preklopa izstopa nemščina, ki ima največji delež enobesednih preklopov (80 %) in najnižjo povprečno dolžino preklopa (1,40 besede).

Tabela 4: Povprečna dolžina preklopov glede na vrsto preklopa in jezik.

	Povprečno število besed
znotrajstavčni preklopi	1,73
medstavčni preklopi	2,94
vsi preklopi	2,13
angleščina	2,16
hrvaščina/srbščina/bosanščina	2,10
nemščina	1,40

Kot prikazuje Tabela 5, so glede na besedno vrsto preklopov v povprečju najdaljši »neuvrščeni« preklopi, kar ni presenetljivo, saj ta kategorija vsebuje daljše besedne zveze ali cele stavke, ki jim ni bilo mogoče določiti ene besedne vrste (primer [3]), dobrih 65 % preklopov, označenih s to kategorijo, pa so medstavčni preklopi (primer [4]):

[3] »a te je gasilc pošpricu v oči? :O ti mu kr povej, da po faci nima kej, he has to treat you with respect ;)

[4] i thought some Joe guy killed Bruce's parents? :) pa še to nismo zihr. Me zanima kako bojo v #Gotham to naredl :)

Na drugem mestu je kategorija predlogov, med katerimi sicer najdemo enobesedne (npr. »by« ali »via«), vendar prevladujejo predložne zveze (npr. »in the mood« ali »mit extra Creme«). Na tretjem mestu so samostalniki in samostalniške zveze, kar ne preseneča, saj pogosto vsebujejo prilastke (npr. »mission impossible« ali »Happy news«). Tudi v kategoriji prislovov najdemo večbesedne (npr. »all summer long« ali »Kind of«), vendar je tukaj delež enobesednih že več kot 60-odstoten.

V spodnji polovici Tabele 5 so kategorije, kjer je velika večina primerov (več kot 80 %) enobesednih. Med njimi je znaten delež dvobesednih preklopov še med pridevniki (14,5 %, npr. »Pretty cool« ali »custom made«), glagolih (13,7 %, npr. »come on« ali »se šunja«) in medmetih (9,1 %; »oh yeaaah« ali »damn right«). Za

veznike, členke, okrajšave in zaimke pa lahko rečemo, da so vedno enobesedni – do odstopanja v povprečni dolžini pri veznikih in členkih je prišlo zaradi treh preklopov, kjer gre za ponovitev ene besede (»and and AND«, »jp jp« in »jp, jp«).

Tabela 5: Povprečna dolžina preklopov glede na besedno vrsto.

	Povprečno število besed
neuvrščeno	3,7
predložna zveza	2,1
samostalniška zveza	1,6
prislovna zveza	1,5
veznik	1,2
pridevniška zveza	1,2
medmet	1,1
glagolska zveza	1,1
členek	1,0
okrajšava	1,0
zaimek	1,0

3.2 Slovnične značilnosti preklapljanja

Analiza preklopov glede na posamezno besedno vrsto pokaže, da kot preklopi najpogosteje nastopajo samostalniki oziroma samostalniške besedne zveze (33 %), kar je v skladu s sorodnimi raziskavami. Na drugem mestu so besedne zveze oz. stavki, ki jim ni bilo mogoče določiti besedne vrste (26 %), torej »otoki vrinjenega jezika«, ki so tvorjeni v skladu s slovnico tujega jezika. Na tretjem mestu so medmeti (15 %), ki so nas nekoliko presenetili, zato smo se odločili izdelati seznam uporabljenih medmetov. Izkaže se, da je velik delež medmetov posledica manjšega nabora besed, ki se velikokrat ponovijo. Med njimi izstopajo t. i. klepetalniške oz. spletne krajšave: »lol« (64 ponovitev oz. tretjina vseh medmetov), »btw« (11), »omg/omfg« (10), »wtf« (4), »imho« (2). Precej ponovitev imajo tudi izrazi »thanks« (14), »sorry« (9) in »please« (6) v različnih nestandardnih zapisih.

V nasprotju s pričakovanji je bil delež pridevniških besednih zvez (9 %) in glagolov (5 %) nekoliko nižji. Sklepamo lahko, da je v slovenščino zaradi pregibanja težje vriniti tuj glagol. Zato se večina tujih glagolov pojavlja v sklopu daljših fraz, ki so v našem korpusu označene kot neuvrščene.

Ti rezultati kažejo, da lahko kot kodni preklop nastopajo vse besedne vrste in da za preklapljanje v slovenščini omejitve za slovnične besedne vrste, kot jih je predvidel

Joshi (1982), ne veljajo. Podrobnejši pregled angleških slovničnih besed, ki so v vzorcu najpogostejše, pa pokaže, da je večina identificiranih členkov na različne načine zapisanih besed »yup« (jap, jep, jp, jup – skupaj 27 preklopov) in »nope« (nop, nope, noup – skupaj 14), poleg katerih se pojavijo le še »nah«, »yes«, »right« ter »the«. Med vezniki se trikrat pojavi »also«, dvakrat »and«, enkrat »or« in »aka«, med zaimki pa so naslednji: »everyone«, »he«, »she«, »what« in »why«. Na podlagi teh rezultatov lahko sklenemo, da je preklapljanje s slovničnimi besednimi vrstami sicer mogoče, a zelo redko in vezano na ozek nabor besedišča.

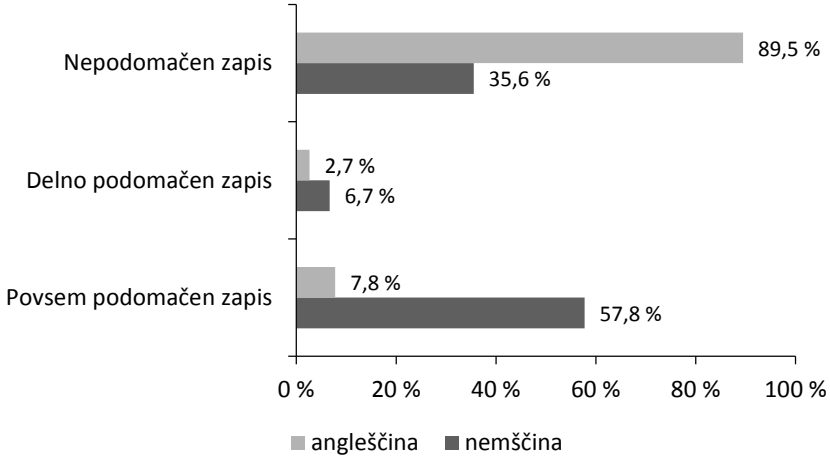
Na tem mestu velja omeniti še opažanje s seznama najpogostejših besed v preklopih, ki smo ga izdelali iz vseh besed, tudi tistih, ki so del besednih zvez in stavkov. Na njem imajo zelo visoko frekvenco prav slovnične besedne vrste, ki so običajno del »otokov vrinjenega jezika« (oziroma »neuvrščenih« preklopov po naši kategorizaciji besednih vrst), in sicer gre za zaimke (»I«, »you«, »my«, »it«), določne in nedoločne člene (»a«, »the«) ter predloge (»to«, »of«, »for«, »on«). Zaradi tega sklepamo, da je mogoče slovnične besede (oz. sistemske morfeme po Myers-Scotton) praviloma preklapljati le skupaj s celotno besedno zvezo, vendar pa bi to morali dodatno preveriti v korpusu, kjer bi bile slovnične značilnosti označene podrobneje (torej za vse besede, ki sestavljajo posamezen »neuvrščeni« preklop) oz. z dodatnimi kategorijami.

3.3 Stopnja integracije preklopov v slovenščino

3.3.1 Ortografska integracija

V analiziranem vzorcu močno prevladujejo nepodomačeni kodni preklopi (88 %), torej preklopi, v katerih so tujejezični segmenti navedeni citatno, v skladu z zapisom v izvornem jeziku (npr. »does a little happy dance«). Delno podomačenih (3 %, npr. »velkom back«) in povsem podomačenih (9 %, npr. »šejkam«) je relativno malo, kar bi lahko govorilo v prid teorijam, da je prav integriranost besede kriterij za ločevanje prevzetih besed od preklopov. Za cel razdelek 3.3.1 velja, da je na rezultate vsaj delno vplival tudi način vzorčenja, saj smo zajeli tvite, ki vsebujejo besede z oznako Nj, avtomatski označevalnik pa morda tuje besede s podomačenim zapisom zazna kot slovenske in jih ne označi z Nj.

Zaradi podobnosti slovenščine s hrvaščino/srbščino/bosanščino smo se pri podrobnejši analizi zapisovanja preklopov omejili na angleščino in nemščino. Kot je razvidno z Grafa 2, večjo stopnjo podomačenosti opazimo v nemških preklopih. Predpostavljamo, da je razlog v tem, da gre za besede, ki so jih govorci že prevzeli v slovenščino, čeprav jih ni v priročnikih, ki smo jih upoštevali kot merilo za ugotavljanje prevzetosti v slovenščino (glej 1.1).



Graf 2: Deleži angleških in nemških preklopov glede na ortografsko integracijo.

Glede na besedno vrsto preklopa so na nivoju zapisa najbolj podomačeni členki (40 % s povsem podomačenim zapisom), sledijo pa jim glagoli (30 %), pridevniki (21 %), prislovi (12 %), medmeti (8 %) in samostalniki (6 %). Za boljšo predstavo smo izluščili vse preklope, ki so na nivoju zapisa povsem podomačeni, in jih razvrstili po besedni vrsti in jeziku. Velik delež podomačenih členkov predstavljajo različni zapisi besede »yup« (npr. »jap, jep, jp, jup«). Bolj zanimiva kategorija so medmeti, kjer po številu izstopata »sori/sorči« (5x) in »tenks« (3x), med njimi pa so npr. tudi »dbest«, »fakof« in »džiizs«. Med pridevniki po ponovitvah izstopajo besede »kjut« (7x), hepi (3x) in »fakin(g)« (3x), med samostalniki pa »komp«/»kompjutr«.

Večjo podomačenost na nivoju zapisa smo pričakovali pri pregibnih besednih vrstah, saj smo predpostavljali, da si bodo uporabniki na ta način poenostavili tvorjenje besedila, a glede na nizek delež podomačenih zapisov teh besed sklepamo, da so preklopi med jeziki v tvitih večinoma popolni (da tviteraši iz enega jezikovnega koda popolnoma preidejo v drugega in ju ne zlivajo).

3.3.2 Morfološka integracija

Po analizi preklopov, pri katerih sta končnica in/ali obrazilo razvidna, smo ugotovili, da je delež neintegriranih preklopov (92 %) še višji kot pri ortografski integraciji (glej 3.1.1). Na visok delež vplivajo »neuvrščeni« preklopi, tvorjeni v skladu s slovnico vrinjenega jezika (primera [5] in [6]), nepregibne besedne vrste (primera [7] in [8]) in samostalniki z ničto končnico (primera [9] in [10]), ki smo

jim med označevanjem vzorca pripisali oznako »brez razvidne končnice«. Odprto ostaja vprašanje, ali bi uporabniki te besede v drugačnem kontekstu pregibali v skladu s slovensko slovnico ali pa obstaja večja verjetnost, da se preklopi pojavijo prav na stavčnih mestih, kjer pregibanje ni potrebno (primera [11] in [12]).

[5] @tviteraš kolegov mulc hoče met Aza's Arid pack, pa ga najdem samo za 1.6. Bo že moral potrpet, I guess :)

[6] @tviteraš vode dost. Migrena ni, tist ujamem dost zgodi. Tablet ne prime, feels weird 😐 kavo sm zmanjšala, ker mi v tej vročini ne paše.

[7] MK Založba na Hrvaškem v »knjigotrško megamrežo« http://t.co/eWNWSHWrC1 also, naj pripomnim, da sem prijetno presenečena nad min. Grilcem

[8] @tviteraš and and AND! Ladja potone! sicer pa men je Titanic kul, šla v kino zdej k je mel obletnico. #funtimes

[9] @tviteraš Zahteva mal več dela in organizacije kot metanje v smeti, feeling je pa waaaaaaay better. ;)

[10] Zrihtala touch screen, naštudirala bralnik črtnih kod und much more.

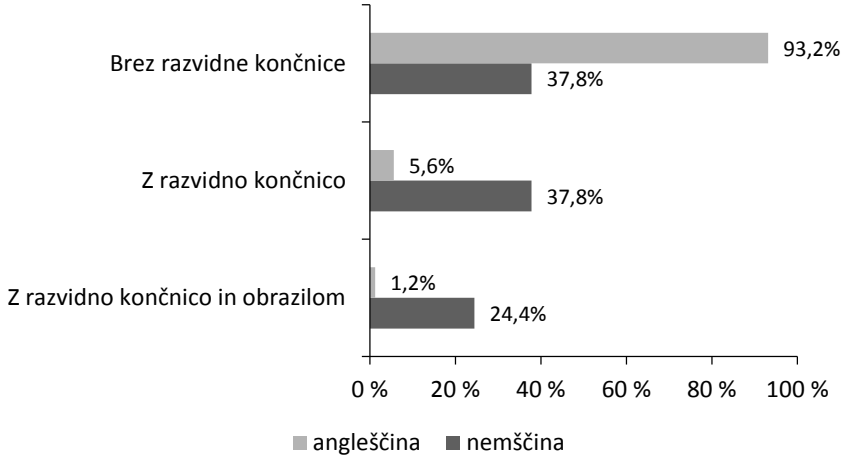
[11] @tviteraš sej pravm, francozi ownajo horrorland. Thanatomorphose. Prever trailer. Boom, tko se dela.

[12] @tviteraš uuu jaaaa, men se tut tej trailerji dopadejo... @ tviteraš

Pri primerjavi morfološke integriranosti glede na besedne vrste se je po pričakovanjih izkazalo, da so glagoli najbolj podvrženi dodajanju slovenskih končnic in obrazil (56 % glagolov je imelo razvidno končnico in/ali obrazilo). Na manjši delež pri samostalnikih (15 %) vplivajo že omenjene ničte končnice, rezultate za pridevniške zveze (6 %) pa je mogoče pripisati dejstvu, da so številni pridevniki del samostalniških zvez oziroma »otokov vrinjenega jezika«. Morda obstajajo določene slovnične omejitve pri preklapljanju pridevnikov na mestih, kjer bi v skladu s slovensko slovnico morali dobiti ustrezno končnico, a v naši raziskavi v večbesednih zvezah nismo označevali besednih vrst za vsako besedo, zato te hipoteze ne moremo empirično preveriti.

Podobno kot v 3.3.1 smo tudi tu izdelali seznam morfološko integriranih preklopov. Po številu ponovitev med glagoli izstopajo »laufati«, »šerati«, »filati«, »kenslati«, »lajkati«, »potegati«, »rentati«, med samostalniki pa se po dvakrat ponovijo le »follower«, »komp«, »taxi« (v različnih oblikah). Na splošno lahko ugotovimo, da do morfološke integracije prihaja pri pestrejšem naboru besed kot pri ortografski ter da tviteraši nimajo posebnih težav z dodajanjem slovenskih končnic tudi na nepodomačene osnove – samo v petih primerih so za to na primer uporabili vezaj.

Tudi pri morfološki integraciji smo podrobneje analizirali le preklope v angleščino in nemščino. Na Grafu 3 je očitna razlika med razporeditvijo deležev, in sicer je v nemščini precej več preklopov z razvidno končnico in/ali obrazilom, kar ustreza našim domnevam, da gre za starejše, bolj integrirane preklope, ki so na poti, da postanejo prevzete besede (ali pa glede na rabo to že so).



Graf 3: Deleži angleških in nemških preklopov glede na morfološko integracijo.

Na tem mestu je treba izpostaviti problematičnost izbranega kriterija za ločevanje prevzetih besed od kodnih preklopov. Med označevanjem se je izkazalo, da so številni germanizmi morda proti pričakovanjem že vključeni v priročnike, npr. »cajt«, »frišne« v SSKJ, »pucati«, »fuzbal«, »luft«, »šporhet« v Sloleksu ali »šmirglati« v SP, zato smo jih kljub morebitnim kvalifikatorjem šteli kot del slovenskega besedišča in jih nismo označevali kot kodne preklope. Razlog za večjo integracijo nemških preklopov se zagotovo skriva v tem, da je zgodovina nemško-slovenskih jezikovnih stikov v primerjavi s stiki z angleščino daljša, saj sega vse do druge polovice 8. stoletja, nemški kulturni vpliv na slovenščino pa je začel pojenjati po letu 1917 (Šekli 2015: 41–42).

3.3.3 Ortografska in morfološka integracija

Če iz korpusa preklopov izluščimo preklope, ki imajo hkrati oznako za (delno) podomačenost zapisa ter razvidno končnico in/ali obrazilo, dobimo 50 preklopov, med katerimi je 24 glagolskih (npr. »hajcajo«, »potegal«), 20 samostalnikih (npr. »frendi«, »kompu«) in šest pridevniških zvez (npr. »pofarban«). V korpusu preklopov to predstavlja 3,47 % vseh preklopov. Po kriteriju integracije bi lahko bile te besede kandidatke za prevzem v slovarje, pri čemer bi seveda morali upoštevati tudi pogostost njihove rabe in druge leksikografske kriterije za uslovarjenje besedišča.

3.4 Razlike glede na spol

V označenem korpusu je 730 preklopov v tvitih ženskih avtoric (53 %) in 651 moških (47 %),[31] vendar glede na majhno razliko in velikost vzorca ne moremo trditi, da ženske preklapljajo pogosteje kot moški. Moški tviti so v povprečju nekoliko daljši (2,4 besede pri moških in 2,14 pri ženskah), a imajo nekoliko večji delež znotrajstavčnih preklopov, ki so načeloma krajši (71 % znotrajstavčnih pri moških, 63 % pri ženskah). Pri primerjavi jezika preklopa glede na spol smo ugotovili, da je pri moških nekoliko višji delež hrvaških/srbskih/bosanskih preklopov (5,4 % v primerjavi s 3,8 % pri ženskah). Glede na rezultate lahko zaključimo, da med obema spoloma ni večjih razlik v kodnem preklapljanju na družbenem omrežju Twitter.

4 KVALITATIVNA ANALIZA

Na podlagi klasifikacij, predstavljenih v uvodu, smo sestavili nabor kategorij, ki smiselno ustreza opažanjem, zbranim med označevanjem in analizo kodnih preklopov. V nadaljevanju predstavljamo nekaj ilustrativnih primerov kodnih preklopov za vsako kategorijo skupaj s kontekstom, v katerem se pojavijo. Seveda bi bilo mogoče posamezne tvite uvrstiti tudi v več kategorij, saj se lahko funkcije in teme prekrivajo ali pa ker vsebujejo več različnih preklopov. V zvezi s slednjim smo se tudi odločili, da v vsakem primeru podčrtamo zgolj tisti preklop, ki je relevanten za obravnavano kategorijo.

Primere smo anonimizirali tako, da smo izbrisali oznake z imeni uporabnikov (@...), kadar to ne vpliva na skladenjsko celovitost tvita, v nasprotnem primeru pa smo uporabnika anonimizirali v @tviteraš, zaradi večje preglednosti pa smo spletne povezave okrajšali na http.

4.1 Diskurzne funkcije kodnega preklapljanja

4.1.1 Zgoščeno izražanje

Na Twitterju je zaradi prostorske omejenosti gospodarna raba jezika še pomembnejša kot sicer, zato ne preseneča, da si uporabniki v želji po vsebinsko zgoščenem izražanju včasih pomagajo tudi s tujimi jeziki. Angleščina je v tem pogledu zelo

31 V celotnem podkorpusu tvitov Janes sicer prevladujejo moški (53 % v primerjavi s 24 % žensk, ostali so označeni kot nevtralni; glej Erjavec et al. 2018).

priročna, saj omogoča nizanje samostalnikov v obliki levih prilastkov, kar v slovenščini načeloma ni mogoče. Primeri [13]–[15] prikazujejo, kako so se avtorji elegantno izognili daljšim pojasnjevalnim odvisnim stavkom.

[13] Če damo v enačbo še "priročnost" in "stroškovno učinkovitost", digital čisto povozi paper-based content rezultate. http

[14] ehm. cough, cough.. gag... joooj, neki se mi je zataknilo... valda ne bom zdaj še old ppl kašlja dobila?!?!?

[15] Pravkar sem odkril dismiss gumb. *does a little happy dance*

4.1.2 Ponovitve

Kljub kratkosti besedil na Twitterju avtorji včasih uporabijo dva izraza v različnih jezikih za isto stvar. Motivacije za to so lahko različne, npr. dodatna pojasnjevalna vrednost tujega termina primer [16], kulturna specifičnost pojma, ki potrebuje še slovensko razlago [18], nagovarjanje različnih občinstev [19] ali pa podkrepitev sporočila [20].

[16] Swine, swine ... morda tudi ptičja (avian flu) - skratka eno super zdravilo:) http

[17] Ti, kok se že imenuje tist čaj? "Barut, se mi zdi," odgovori misleč na bure baruta, oziroma sod smodnika. #gunpowder

[18] Nocoj v rojstni hiši Huga Wolfa v Slovenj Gradcu / Hoy en la casa natal de Hugo Wolf en Slovenj Gradec, Eslovenia http

[19] Dnz spim s slovarjem, da bova jutr na ti. :D Lahko noč oz. buenas noches! :) #španščina #matura

4.1.3 Navezovanje na sogovornika ali zunanje okoliščine

Uporabniki tujejezični del besedila včasih posebej zaznamujejo z navednicami, kar kaže, da ga občutijo kot tujek oziroma ga v svoj diskurz ne vključijo nezaznamovano. Razlog za to je lahko, da se navezujejo na tvit sogovornika [20] ali se odzivajo na druge zunanje okoliščine [21]. Kljub temu da gre za dobesedne navedke iz drugega konteksta, smo takšne primere označili na enak način kot ostale preklope, saj ustrezajo izhodiščni opredelitvi (kodno preklapljanje je uporaba več jezikov v istem stavku/diskurzu).

[20] al pa "oooor, you know" in naprej po slovensko :P

[21] Izraza "YOLO moment" še nisem slišala, mi je pa takoj prirastel k scru :)

4.1.4 Vljudnostni izrazi

Tudi izrazi iz te kategorije (primeri [22]–[25]) so na seznamu najpogostejših besed. Predvsem pri različnih okrajšanih zapisih besede »thanks« sklepamo, da k izbiri pripomore težnja k čim hitrejšemu in čim krajšemu sporazumevanju, kar je tudi na splošno značilnost računalniško posredovane komunikacije. Poleg tega lahko pogosto pojavljanje teh izrazov povežemo z večjo neformalnostjo komuniciranja na Twitterju.

[22] please, ne, ne more, niti 1 % šanse. On je v igri zaradi drugih interesov. Tudi R nominacije ne more dobiti, kaj šele volitve.

[23] lej sorry sej te nisem mislu tolk napadat kot je vceri izpadlo ampak trenutek za tvoj post ni bil na mestu ...

[24] aha tnx ... dvomim da bo z avtom tko da ni panike ...

[25] a, thank you, mogoče pridemo kej sparglje degustirat :)

4.1.5 Medmeti – klepetalniške krajšave

V kvantitativnem delu raziskave smo ugotovili, da so med preklopi pogosti angleški medmeti, predvsem klepetalniške krajšave. Na seznamu najpogostejših je na vrhu beseda »lol« (angl. *laugh out laughing*), med tistimi z veliko pojavitvami pa so še »omg« (angl. *oh my god*), »btw« (angl. *by the way*) in »wtf« (angl. *what the fuck*). V primerih [26]–[29] so najpogostejše klepetalniške krajšave prikazane v kontekstu celega tvita.

[26] Lol :D Sošolka si je vedno želela 3 otroke: Žak/Pak/Mak. No, Žaka že ima :)

[27] omg. pa kje ti ljudje rastejo? :D

[28] Na nebotičnik sva odšla... WTF, a Angleže hočemo prestrašit z baladami? :P

[29] ma deeej. Tam sm ze tut bla btw :p

4.1.6 Ekspresivni elementi

Tviteraši pogosto posežejo po angleščini, ko želijo še posebej poudariti svojo izjavo (primeri [30]–[33]). Morda je preklinjanje v tujem jeziku bolj sprejemljivo za ciljno občinstvo oziroma se zdi tudi uporabnikom manj vulgarno, po drugi strani pa morda slovenščina ne ponuja avtohtonih izrazov, ki bi ustrezno zajeli in prenesli želeno sporočilo.

[30] če je konc, je faking konc. Nič ni večno. Prosim, pejmo napreeej! ;)

[31] bemtiš če bi žvel nekje ob meji, madona da bi s principa tankal čez mejo. Fakof, res!

[32] damn right! Sem vedu da ti bo vsec ;)

[33] Bullshit! Makeup, cunje, razsvetljava, kamera, računalnik. Ozri se okoli sebe, princeska!

4.1.7 Ustaljene diskurzivne fraze in frazeološke enote

Med analizo smo naleteli na številne ustaljene diskurzivne fraze (primeri [34]–[38]) in frazeološke enote (primeri [39]–[42]), ki so del mentalnega leksikona in jih govorci, enako kot posamezne besede, prikličejo iz spomina kot celoto (Jakop 2006: 31). Ker oboje lahko opravljajo različne diskurzivne funkcije, jih skupaj obravnavamo kot nadkategorijo, ki bi si zaslužila samostojno raziskavo. Na tem mestu predstavljamo samo nekaj zanimivih primerov, ki odstirajo jezikovne izbire tviteraševv.

[34] hihi, my thoughts exactly. Ceprav razmisljam, da tko okorno hodi zato, da jo lazje fotkne. Recimo :D

[35] my bad, na drugo sem pozabla :(

[36] Ne, "for your information" ne morete kar dobesedno prevajati v "za vašo informacijo". #sampravm

[37] what's there to tell? Zvečer greš v savno, si tm do konca, pol pa domov:D

[38] si ti prebral lasten link in če si, what's your point? Ali pa si prebral samo naslov?

[39] zakon! :) (pa skromno pripominjam: great minds think alike ;))

[40] Torej lahko sedaj spet napišem za Majdo Potrata: Nomen est omen. Spet Potrata davkoplačevalskega denarja.

[41] u snooze u lose. Sem ful tulila, kdo bi.. udej je dobu

[42] Če ne še dlje. Sam ne zdržim jih dolg nosit. Sej več, less is more :)

Sem uvrščamo tudi znane citate (primeri [43]–[46]). Z dobesednim navajanjem v izvirniku je citat bolj zaznamovan, kot bi bil v prevodu, morda bi ga bilo težko prevesti ali pa bi se s prevodom pravzaprav izgubila aluzija, ki jo avtor poskuša doseči.

[43] Zakaj Nemčija(Merkel:Islam ist einTeil vonDeutschland)noče sprejeti tega dela,ki ji trka na vrata?Še več,podkupuje polit.&ga vsiljuje drugim

[44] Prijaznost pripelje le do neke tocke. Saj ves, good girls go to heaven, bad girls go everywhere else.

[45] Elementary, Dr. Watson: investicije se ne more krasti kar tako, kredite se raztala prijateljem kot karte.

[46] V #GOT niso imeli pojma, pravilno bi bilo "Spring is coming" :)

Med najpogostejše sestavine identificiranih ustaljenih diskurzivnih fraz sodita zaimka »I« (primeri [47]–[50]) in »you« (primeri [51]–[54]), ki se uvrščata v sam vrh najpogostejših besed v preklopih.

[47] i love you! Ful ss manj nesposobbo pocutim :)

[48] I see your nič nimam za oblečt and raise you nič nimam za obut. #endofseasonproblems

[49] Folk hodi na medene tedne. Midva bova kupila parking in kamin za v flat. And I couldnt be happier!

[50] Oh kako zelo odraslo od nje, res. Podpihovanje drugih proti meni, kot da smo se v os. Ljubica, I don't care, ti kar. :)

[51] Yes you are, grumpy. Sem ti odgovorila. Ti se pa zjasni, a iščeš fotko, ali izraz

[52] pejt v dnevno, daj zvočnike do konca na glas ... in pol pejt vn. Can't do it, can you? :P pridem popoldne :D

[53] u and me both...danes pogoltnem uspavalo pa bo :D kdaj se vračaš?

[54] ze doma??? Oh how i envy you!!!! :p

Med frazeološkimi enotami pa po pogostosti izstopajo angleški idiomi (primeri [55]–[58]) in frazni glagoli (primeri [59]–[62]).

[55] a si se odločil, da boš couch potato ratal? :)

[56] Dajmo definirat: trema pride, ko ti je nekaj še posebno pomembno in bi se rad še posebno izkazal. Give them a break. Saj bo :) #junaki

[57] Eh, pogojno tri leta. A slap on the wrist... RT @Delo: Gordani Kalan Živčec 10 mesecev zapora http

[58] On a side note: Bemti, nove baterije za UPS rabim ;) #prejsnjitvit

[59] Še malo pa bo ponedeljek..., so bring it on!!! http

[60] Tooo!!! A čutiš to, a? To je ta lahkota inbox zero! Keep it up!

[61] šment, sem imela v mislih stajling z balerinkami in enim retro kabrioletom :) oh well, dream on ....

[62] It goes on. Ustvaril jo je arhitekt Jeza, asistent pa je bil - Triler.

4.2 Tematska polja

Tematska polja smo tvitom pripisovali med ročnim označevanjem preklopov. Teme smo pripisali samo takrat, kadar je bila ta iz sobesedila jasno razvidna. Posameznemu tvitu smo pripisali le eno temo. V primerih, ko bi lahko tvit sodil v več tem, smo izbrali prevladujočo. V vzorcu smo identificirali sedem tematskih sklopov, pri čemer so tri domensko specifične (*pop kultura, računalništvo, šport*) in dve sporazumevalno specifični (*tviterščina, kulturnospecifični izrazi*).

4.2.1 Pop kultura

Zelo pogosta tema tvitov, ki vsebujejo kodno preklapljanje, so televizijske serije in filmi. V določenih primerih morda res ni na voljo primernih slovenskih ustreznic – vsaj ne tako kratkih in uveljavljenih oz. razumljivih (primera [64] in [65]).

[63] ane, čist underrated serija. poglej še Uk verzijo... Sam men niso bli ušeč :)

[64] no sej, spet ne vem ker dan je, navadn. sobota je drgač že. 😊 jst pa gledam zakaj ni streama nikjer, lol.

[65] jup. + Alison pa Felix rabta spinoff 😊

[66] mislm da smo tle kr vse weirdos. Dejmo se kregat okol fictional characters no :D

4.2.2 Računalništvo

Ne preseneča, da je veliko preklopov povezanih z izrazjem s področja novih tehnologij (primeri [67]–[70]). Tehnološkim navdušencem so angleški izrazi

najverjetneje bolj domači, nekateri pa imajo pred slovenskimi prednost tudi v tem, da so krajši, npr. »komp« za računalnik ali »app« za aplikacijo.

[67] Kako napisati scam e-mail? Slovensko - Nigerijski poslovni forum.

[68] Hmm gledam ja, samo če idejo pustimo na stran...kolko efekta pa to ima? Če je viral success nekaj tisoč views..moh.

[69] nisem mailchimpa nikol uporabljal. Ze tisti cartoon vmesnik me odbija :) a to za autoresponder rabis ali classic newsletter?

[70] resno, web player zalaufi, ne app na kompu.

4.2.3 Šport

Šport je na Twitterju poleg politike najpogostejša tema, zato ne preseneča, da jo je zaslediti tudi pri preklopih (primeri [71]–[74]). Do kodnega preklapljanja pride predvsem zaradi navijanja za tuje klube, kot preklop pa smo šteli tudi veznik »versus« (zapisan kot vs ali v), ki se pojavlja npr. pri navajanju tekem iz angleške nogometne ali ameriške košarkarske lige.

[71] Fino, Arsenal. Zdaj v pričakovanju el clasica. #nervous

[72] Marguč v igri za novica sezone v LP. Glasujmo, da postane Rookie of the season! :) http

[73] V soboto se oglasim s tekme Crystal Palace vs. Chelsea #CPLCHE #SelhurstPark Lani je ta obračun dobil Palace!

[74] Jutri bo zmagala, ker bo jezna! :) GO @TinaMaze! #Are

4.2.4 Kulturnospecifični pojmi

Različni kulturno pogojeni pojmi so pogosto trd oreh še za prevajalce, tako da ne preseneča, da je s tega področja precej kalkov ali citatno prevzetih besed. Med takimi, ki se v slovenščini še niso uveljavile do te mere, da bi bile vključene v priročnike, so tudi spodnji primeri preklopov ([75]–[78]).

[75] Od 09:30 v kuhinji...Cheesecake v hladilniku,predjed koncana,ostalo pa pripravljeno,da skoci v lonce....jupi;)

[76] Kdo ve kje kupiti 'plantain' banane? Ki niso banane... http

[77] lahko probaš tud sam nardit :) Horaa Osba'o

[78] Čemu je predsednica srbskega parlamenta na nedavnem obisku v Iranu nosila hijab?Jo je morda zeblo?

4.2.5 Tviterščina

Izmed vseh tematskih polj smo za tviterski žargon našli največ primerov ([79]–[82]), predvsem zaradi razširjene uporabe kratice RT, pojavita pa se tudi MT (angl. *modified tweet*) in FT (angl. *follow-up tweet*). Pogoste so besede z jedrom *follow* (slediti) in *handle* (angleški izraz za uporabniško ime tviteraša). V to kategorijo uvrščamo tudi tvite, v katerih se pojavi *selfie*.

[79] A notification si dobil za to? Al so to kakšni plačani followi, da se ti ne prikaže med notificationi?

[80] Hmm, če kdo pogleda tvoj tw account, bi se mu ziher zazdelo, da je imenovati te Haso še kar blaga oznaka za tako sarma yugoljublje.

[81] tvoj handle ti pristoji 😂

[82] Mi smo old-farts, ki smo se navadili na RT, še preden je TW vpeljal to "retweet" novotarijo :-)

5 SKLEP

V prispevku smo predstavili raziskavo kodnega preklapljanja v tvitih slovenskih uporabnikov, vzorčenih iz korpusa Janes. Za namene analize smo po izčrpnem pregledu sorodnih raziskav razvili označevalno shemo in smernice za označevanje ter vzorec ročno označili v eni od najbolj uveljavljenih anotacijskih platform WebAnno. Označevanje smo izvedli na petih ravneh: (1) jezik preklopa, (2) vrsta preklopa (medstavčni, znotrajstavčni), (3) stopnja integracije preklopa v slovenščino na nivoju zapisa (podomačen, delno podomačen ali nepodomačen zapis), (4) stopnja integracije preklopa v slovenščino na oblikoslovni ravni (z razvidno končnico in/ali obrazilom ali brez) ter (5) besedna vrsta preklopa. Označeni korpus preklopov smo nato analizirali kvantitativno in kvalitativno.

Rezultati so pokazali, da preklapljanje ni redek pojav, saj se je vsaj en preklop pojavil v več kot tretjini analiziranega vzorca. Kadar do preklapljanja pride, se velikokrat celo zgodi, da isti tvit vsebuje več kot en priklop, tako da v povprečju posamezen tvit vsebuje 1,26 preklopa, najvišje število identificiranih preklopov v istem tvitu pa je štiri, kar smo zaznali v šestih tvitih. Do znotrajstavčnega preklapljanja prihaja dvakrat pogosteje kot do medstavčnega. Približno polovica

preklopov je enobesednih, v povprečju pa je dolžina kodnega preklopa 1,26 besede. Rezultati so v skladu z ugotovitvami sorodnih raziskav, ki nakazujejo na to, da krajše preklope zlahka uporabijo tudi osebe, ki morda niso tako vešče tujega jezika, medtem ko je za prenos daljših, celo frazeoloških besednih zvez potrebnega več jezikovnega znanja.

Analiza preklopov glede na metapodatke, ki so pripisani vsem tvitom v korpusu Janes, kaže, da je delež preklopov občutno višji v nestandardnih tvitih, da pa ni bistvenih razlik v preklapljanju glede na spol tviteraša. S sociolingvističnim pristopom bi bilo zelo zanimivo raziskavo razširiti tako, da bi izvedli še analizo glede na druge pomembne sociodemografske podatke, kot so starost, stopnja izobrazbe in regionalna pripadnost uporabnikov družbenih omrežij. Poleg tega bi lahko preverili, kakšen je odnos avtorjev in bralcev do kodnega preklapljanja, ali se odnos spremeni glede na jezik preklopa oziroma ali obstajajo medgeneracijske razlike pri sprejemanju tujejezičnih prvin.

Analiza slovničnih značilnosti preklopov kaže, da je med preklopi največ samostalnikov in samostalniških zvez, ki jim sledijo medmeti, predvsem zaradi pogostih klepetalniških krajšav, medtem ko sta deleža pridevniških in glagolskih zvez nekoliko nižja od pričakovanj. Čeprav smo med preklopi identificirali tudi slovnične besedne vrste, zbrani podatki ne zadoščajo za sklepanje o tem, ali za kodno preklapljanje v slovenščini veljajo katere od slovničnih omejitev, ki so jih raziskovalci odkrili v drugih jezikih. Vsekakor bi bilo treba opraviti podrobnejše raziskave, da bi preverili, ali je preklapljanje res povsem prosto in odvisno zgolj od domišljije govorcev ali vendarle obstajajo določene konstrukcije iz slovenskih in tujejezičnih prvin, ki niso mogoče.

Glede stopnje prilagajanja preklopov v slovenščini smo ugotovili, da je zapis velike večine preklopov prevzet citatno ter da preklopi večinoma niso morfološko integrirani. Na ta rezultat gotovo vsaj v določeni meri vpliva uporabljeni način vzorčenja, ki temelji na avtomatsko pripisanih oblikoskladenjskih oznakah v korpusu za tujejezične besede (Nj). Zato bi bilo v prihodnje zanimivo ugotovitev preveriti na splošnejšem vzorcu.

V označenem korpusu preklopov smo identificirali 9 različnih jezikov, v katere preklapljajo uporabniki družbenega omrežja, med katerimi močno prevladuje angleščina (90 % vseh preklopov), ki ji sledi hrvaščina/srbščina/bosanščina (4,6 %) in nemščina (3,3 %), ostali jeziki so zelo redki. Večina kodnih preklopov na nivoju zapisa ali z obrazili ni prilagojenih slovenščini, temveč preklopi sledijo načelom vrinjenega jezika. Tu izstopajo preklopi iz nemščine, v katerih je stopnja integracije nekoliko višja, sklepamo, da zaradi tega, ker so uporabljeni germanizmi že dlje časa v rabi in imajo že status privzetih besed. V zvezi s tem bi bila zanimiva primerjava z drugimi (spletnimi) mediji, da bi ugotovili, ali je tako pogosto preklapljanje v

angleščino značilno za računalniško posredovano komunikacijo na splošno ali gre za specifično lastnost komuniciranja na Twitterju.

Kvalitativna analiza je pokazala, da so diskurzne funkcije preklopov zelo raznolike. Preklopi npr. pripomorejo k bolj zgoščenemu izražanju ter služijo kot dodatno pojasnilo, poudarek ali referenca na zunanje okoliščine. S preklopi tviteraši odgovarjajo sogovornikom, jih nagovarjajo, izražajo svoja čustva ali podkrepijo svoj odnos do napisanega. Posebno pozornost si zaslužijo ustaljene diskurzivne fraze in frazeološke enote, ki smo se jih v predstavljeni raziskavi zgolj dotaknili. V prihodnjih raziskavah bi bilo zanimivo raziskati interference, do katerih prihaja pri njihovi rabi v izvirniku in pri njihovem prenosu v slovenščino. Pregled zaznanih tematskih polj je pokazal, da se preklopi pogosto nanašajo na tviteraški žargon in informacijske tehnologije oziroma se pojavljajo v pogovorih o različnih prostočasnih dejavnostih, kot so šport, pop kultura ter hrana.

Sklenemo lahko, da kodni preklopi ne služijo zgolj zapolnjevanju leksikalnih vrzeli, temveč dajejo avtorjem širše možnosti za izražanje ter ustvarjanje svoje spletne identitete in sloga. V raziskavi se nismo ukvarjali z vrednostno presojo, ali tovrstna uporaba tujejezičnih prvin ogroža ali bogati slovenščino, vsekakor pa gre za pogost fenomen, ki si zasluži empirično obravnavo in dobro razumevanje. Mešanje jezikov je pomembno raziskovalno vprašanje tudi za jezikovne tehnologije, ki lahko dosegajo visoko stopnjo natančnosti le, če najprej ustrezno identificirajo jezik, iz katerega posamezna beseda oz. besedna zveza prihaja. To pa je toliko težje v nestandardnih, večjezičnih besedilih, ki so značilna za velik del računalniško posredovane komunikacije, zato je možnosti za izboljšave na tem področju še veliko, predstavljena raziskava pa nudi dragocen prvi uvid v problematiko tudi za te namene.

Zahvala

Avtorici se zahvaljujeta Tomažu Erjavcu za vso tehnično podporo pri raziskavi.

Literatura

Androutsopoulos, Jannis, 2013: Code-switching in computer-mediated communication. Herring, Susan C., Dieter Stein, Tuija Virtanen (ur.): *Handbook of the Pragmatics of CMC*. Berlin: Mouton de Gruyter. 667–694.

Appel, Rene in Pieter Muysken, 2005: *Language Contact and Bilingualism*. Amsterdam: Amsterdam University Press.

Clyne, Michael, 2009: Constraints on code switching: how universal are they? *Linguistics* 25/4. 739–764.

Deuchar, Margaret, 2006: Welsh-English code-switching and the Matrix Language Frame model. *Lingua* 116. 1986–2011.

Erjavec, Tomaž, Špela Arhar Holdt, Jaka Čibej, Kaja Dobrovoljc, Darja Fišer, Cyprian Laskowski in Katja Zupan, 2016: Annotating CLARIN.SI TEI corpora with WebAnno. *Proceedings of the CLARIN Annual Conference 2016*. Aix-en-Provence, Francija.

Gardner-Chloros, Penelope in Daniel Weston, 2015: Code-switching and multilingualism in literature. *Language and Literature* 24/3. 182–196.

Gardner-Chloros, Penelope, 2009: *Code-switching*. Cambridge, New York: Cambridge University Press.

Jakop, Nataša, 2006: *Pragmatična frazeologija*. Ljubljana: Založba ZRC.

Joshi, Aravind K., 1982: Processing Of Sentences With Intra-Sentential Code-Switching. *Proceedings of the 9th conference on Computational linguistics* – Volume 1. 145–150.

Myers-Scotton, Carol, 1997: *Duelling languages: Grammatical structure in code-switching*. Oxford: Clarendon Press.

Myers-Scotton, Carol, 2000: Code-switching as indexical of social negotiations. Wei, Li (ur.): *Bilingualism reader*. London; New York: Routledge. 127–153.

Myers-Scotton, Carol, 2002: *Contact Linguistics: Bilingual Encounters and Grammatical Outcomes*. Oxford; New York: Oxford University Press.

Poplack, Shana, 1980: Sometimes I'll start a sentence in Spanish y termina en español: toward a typology of code-switching. *Linguistics* 18(7/8): 581–618.

Poplack, Shana, 2015: Code Switching: Linguistic. *International Encyclopedia of the Social and Behavioral Sciences*, 2nd edition. Oxford: Elsevier Science Ltd. 918–925.

Reher, Špela, 2017: *Slovenščina na prepihu: kodno preklapljanje v objavah slovenskih uporabnikov Twitterja. Kvantitativna in kvalitativna analiza tvitov iz korpusa nestandardne slovenščine Janes*. Magistrsko delo. Ljubljana: Filozofska fakulteta v Ljubljani.

Rosenberg, Johanna, 2014: *"I chuckled. Rolig pun där." A grammatical investigation of Swedish-English code-switching in two web discussion forums*. Diplomsko delo. Lund University. https://lup.lub.lu.se/student-papers/search/publication/4253187

Sebba, Mark, 2012: Multilingualism in written discourse: An approach to the analysis of multilingual texts. *International Journal of Bilingualism* 17/1. 97–118.

Smernice za označevanje korpusa slovenskih tvitov JANES: tokenizacija, stavčna segmentacija, normalizacija, lematizacija in oblikoskladenjsko označevanje, 2016. http://nl.ijs.si/janes/wp-content/uploads/2014/09/JANES-jezikoslovne-smernice-v0.9.pdf/

Šekli, Matej, 2015: Pomenska polja nemških izposojenk v slovenščini. *Jezikoslovni zapiski* 21/2. 31–44.

Thomason, Sarah G., 2001: *Language contact: an introduction*. Edinburgh: Edinburgh University Press.

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

Spremembe pri pisemskem naslavljanju in poslavljanju v elektronski dobi

Helena Dobrovoljc

Izvleček

V poglavju je pozornost namenjena spremembam pri pozdravljanju, poslavljanju in naslavljanju v pisemskem sporazumevanju. To je nekoč potekalo na daljavo prek fizičnega, danes pa večinoma prek elektronskega prenosnika. Pokazati želimo na spontane in naučene spremembe pisemskega diskurza, ki prodirajo v vse tipe elektronskega pisemskega sporazumevanja in prikazujejo, kako so se v različnih obdobjih spreminjala družbena razmerja in piščev odnos do pisemskega sporazumevanja.

Ključne besede: pismo, elektronsko posredovana komunikacija, elektronska pošta, nagovor, pozdravljanje, poslavljanje, diskurzivne prvine, besedilo

1 UVOD

Pisma so ena od najbolj razširjenih in s formalnega vidika ustaljenih besedilnih zvrsti, zaradi česar jih je mogoče spremljati od obdobja epistolarnih stvaritev, pisem evangelistov in apostolov do poslovilnih, verižnih, ljubezenskih, javnih pisem in danes elektronskih pisemskih sporočil. Razmaha, kakršnega so doživela pisma v 21. stoletju, si pred desetletji nismo niti predstavljali, saj je v dobi računalniško posredovanega sporazumevanja in interneta možnosti za zasebno in uradno komuniciranje precej več kot v preteklosti, elektronska sporočila pa so doživela pravo ekspanzijo s kratkimi besedilnimi sporočili (SMS), z replikami v spletnih klepetalnicah in na forumih novičarskih portalov, objavami na Facebooku in Twitterju ipd., ki sicer služijo komunikaciji, a ne zadoščajo merilu pisemskega sporazumevanja v klasični obliki.

V prispevku se zato omejujemo zgolj na pisemsko sporazumevanje, ki je nekoč potekalo na daljavo prek fizičnega, danes pa večinoma prek elektronskega prenosnika, in ki še vedno ponuja mnogo raziskovalnih izzivov, saj že stoletja ohranja razmeroma tradicionalne konvencije in hkrati (predvsem v zasebni korespondenci) individualistični značaj: pisec in naslovnik sta zapletena v dialog, stopnjo formalnosti in odnosa med obema korespondentoma v pismu pa odražajo diskurzivne[1] prvine (naslavljanje in ogovarjanje ter pozdravljanje).

S primerjavo klasične in sodobne elektronske pisemske oblike v obdobju zadnjih desetih let bomo pokazali na porast novih nagovornih oblik, ki ustrezajo položaju polformalnega in se oddaljujejo od doslej favoriziranih oblik. Nove prvine, za katere ne moremo trditi, da jih pridobivamo pri institucionalnem učenju, prodirajo v vse tipe elektronskega pisemskega sporazumevanja in nakazujejo na piščev spremenjeni odnos do naslavljanja in posredno tudi do naslovnika.

2 METODA IN GRADIVO TREH OBDOBIJ

Za ponazoritev navedenih raziskovalnih hipotez smo izbrali jezikovne oz. stilistične priročnike, iz katerih je mogoče razbrati pomen in vlogo pisemskega sporazumevanja, kakor so ju interpretirali njihovi pisci v izbranih obdobjih zadnjih 150 let. Za ponazoritev pisemskih praks v dobi elektronskega sporočanja pa je bilo pregledano gradivo osebnega e-predala za obdobje dveh let, izvedena je bila tudi anketa o naslavljanju v elektronskih pismih. Za podkrepitev ugotovitev bomo mestoma, zgolj v utemeljevalni vlogi, uporabili tudi tretji gradivski vir – vprašanja jezikovnih uporabnikov v spletni jezikovni svetovalnici Inštituta za slovenski jezik Frana Ramovša pri ZRC SAZU.

1 *Diskurz*, kakor je interpretiran v okviru Foucaultovega pojmovanja družbene moči in praks: z njim označujemo *celoto verbalnih udejanjenj* (Foucault 2001: 117) v hierarhiji odnosov različnih obdobij.

Prispevek metodološko izrablja dva različna pristopa: pri preučevanju konvencij iz preteklih obdobij se poslužuje deduktivnega sklepanja o navadah pišočih, primerjavo s sodobnim gradivom pa nam omogočajo induktivne posplošitve le-tega, saj izhajamo iz pregleda posameznega pisemskega gradiva.

a) Pisemsko sporočanje v obdobju druge polovice 19. stoletja predstavlja anonimna – po navedbah strokovnjakov (Ahačič 2015) pa Končnikova – šolska slovnica, ki veliko pozornosti namenja besedilni pismenosti prav na primeru pisemskega sporazumevanja, kar priča o razširjenosti in pomembnosti te besedilne zvrsti v najzgodnejših obdobjih množične komunikacije v slovenščini.

b) Pisemsko sporočanje v obdobju druge polovice 20. stoletja:

b1) Za obdobje 1965–1975 sta bila izbrana stilistični priročnik Silve Trdina *Besedna umetnost* (1965) in Toporišičev šolski učbenik *Slovenski jezik* 1–4 (1965–1973) kot deli, na katere se pisci pisem, učitelji in predavatelji najpogosteje sklicujejo.

b2) Obdobje 1995–2005 predstavljata priročnik *Pišem, torej sem* (1994) publicista Draga Bajta in predlog za standardizacijo uradovalnega pisemskega sporočanja avtorice Monike Kalin Golob z naslovom *O dopisih* (2003).

c) Za prva desetletja 21. stoletja je bilo pregledano približno 500 pisem iz osebnega elektronskega poštnega[2] predala avtorice prispevka in razmeroma redko pisemsko gradivo besedilnega korpusa nestandardne slovenščine – *Janes* (Janes v0.4). Pri tem smo se oprli na že izvedeno raziskavo splošnih lastnosti pisemskega sporazumevanja v e-dobi (Dobrovoljc 2008) in upoštevali že ugotovljena dejstva ter predpostavke avtoritet na področju internetnega jezikoslovja[3] (Crystal 2011, Baron 2002a idr.), ki veljajo bodisi za vsa elektronska sporočila bodisi le za elektronsko pošto.

č) Interpretacija podatkov je potekala s pomočjo **ankete**[4] (v prilogi) na izbranem vzorcu uporabnikov e-pošte, starih od 63 do 20 let, ki so dnevni uporabniki e-poštnih storitev in ki v povprečju prejmejo 17 in

2 V nadaljevanju bo uporabljena krajša oblika, tj. *e-pošta*, *e-pismo*, *e-poštni predal*.

3 V akademskem svetu je pisemsko sporazumevanje deležno več pozornosti na interdisciplinarnem – socio- in psiholingvističnem področju. Zdi se, da se je prvotna ideja o olajšani in bolj učinkoviti komunikaciji na delovnem mestu že nekoliko umaknila študijam, ki poročajo o preobremenjenosti in nesorazmerni »zasutosti« s pošto, kakršne je deležen zaposleni posameznik v individualnem delovnem okolju (Derks, Bakker 2010). Med raziskovalnimi izzivi, ki bi bili zanimivi tudi za jezikoslovje in s katerimi bi se morali v opisanem kontekstu spopasti interdisciplinarno, so zagotovo (a) preučevanje nebesednih znamenj v uradni e-pošti; (b) razumevanje in razreševanje dvoumnih elektronskih sporočil; (c) primernost sporočanja t. i. »slabih novic« po e-pošti; (č) dinamika pošiljanja sporočil med sodelavci v isti organizaciji in njena funkcija; (d) vpliv e-sporazumevanja na vzdrževanje socialnih stikov ipd.

4 Vprašalnik o naslavljanju (2017) je predstavljen v obliki priloge.

pošljejo 11 e-pisem na dan. Odzvali so se 103 povabljeni anketiranci. Značaj vzorca torej ne zagotavlja možnosti statističnega posploševanja, rezultati ankete pa kljub temu pripomorejo k objektivizaciji preverbe predpostavk, ki jih prinaša primerjava gradiva različnih obdobij.

2.1 Šolska slovnica iz leta 1881

Že pred 130 leti je koroški šolnik Peter Končnik v *Slovenski slovnici z naukom, kako se pišejo pisma in opravilni sestavki* (1881), ki sicer ni podpisana (Ahačič 2015), pismo definiral podobno, kot ga opisuje sodobno jezikoslovje oz. besediloslovje:

> Pismo ali list je sploh pismeno poročilo, namenjeno neki osebi (včasi tudi več osebam). Pismo je namesto ustnega pogovora. Tudi pismene vloge na visoke osebe, da se spomnijo kake ustno izrečene prošnje (Promemoria, spomenica), in pa prošnje podane oblastvu (gosposki) so pisma glede na zapopadek, samo da po vnanji podobi uradno lice dobivajo. Pismo ali list obsega v sebi to, kar bi pisec temu, kteri ga prejme, z besedo povedal, ko bi ga imel pred seboj. Ako na pismo pride odpis ali ogovor, ali če si dve ali več oseb ena drugi pišejo, tu postane med njimi dopisovanje (korespondencija) (Slovenska slovnica 1881: 150).

Šolska slovnica (kasneje imenovana *Slovenska slovnica za obče ljudske šole*; dalje SSLŠ), kakor tudi nekaj njenih predhodnic in ponatisov, veliko pozornost namenja besedilni pismenosti prav na primeru pisemskega sporočanja, kar priča o pomembnosti in vsesplošni razširjenosti te besedilne zvrsti. Avtor slovnice v poglavju *Pisma različnega zapopadka*[5] (SSLŠ 1881: 134) podaja številne predloge za učinkovito pisemsko sporazumevanje v različnih komunikacijskih položajih, ob čemer predloge spreminja glede na namen posameznega pisemskega sporočila, in sicer razlikuje naslednje: *vprašanje*, *naznanilo*, *prosilni list* (kratka in daljša različica), *poročilo in prošnja*, *povpraševanje*, *voščilo ali čestitanje*, *voščilo za novo leto*, *prošnja za odpuščanje*, *izgovor*, *prijateljsko očitanje*, *povabilo*, *priporočilo*.

Subtilnost pri razlikovanju raznovrstnih položajev in diskurzivnih izhodišč pisemske besedilne predloge implicitno opozarja na pomen družbenih razmerij v tedaj izrazito socialno stratificirani jezikovni skupnosti. Avtor opozarja, da je upoštevanje etikete oz. zapisanih pravil predpogoj za »čislanost« in »omikanost« pišočega, saj se po zapisanem sodi »o pisavčevem umu, o njegovi omiki in njegovem značaju« (SSLŠ 1881: 151); pri tem mora pišoči paziti ne le na osebnostne posebnosti naslovnika (ali je »naš dobrotnik, naš prijatelj, naš višji ali podložnik, če zdrav ali bolan, če je mož sivih las ali nežen deček«, n. d.), ampak predvsem na socialni oz. stanovski položaj: »Višji stan zahteva globoko spoštovanje in včasih

5 Besedo *zapopadek* pisec uporablja v pomenu 'vsebina' (po SSKJ).

tudi veliko poniževanje, nižji stan pa vsaj tisto čislanje in ljubezen, ktero moramo sploh vsakemu človeku izkazovati« (n. d.).

Nadalje pisec predlóg opozori na pomen in povezanost ***pozdravljanja in naslavljanja*** ter ***vsebine*** pisma (»vvod pisma pripravlja na njegov zapopadek«, 1881: 145) ter ***sklepa*** pisma in tudi ***poslovilnega pozdrava***, kakor tudi ***nadpisa***, tj. napisa na pisemski ovojnici. Dodatek šolski slovnici torej ni le didaktični pripomoček z besedilnimi predlogami, temveč podaja priporočila slogovne in družbene ustreznosti, ki ponazarjajo pisemski diskurz in koncepte vljudnosti v drugi polovici 19. stoletja.

2.2 Pisanje pisem po Silvi Trdina in Jožetu Toporišiču

Medtem ko Končnik in avtorji pred njim največ pozornosti namenijo ogovoru naslovnika in razmerju pisec – naslovnik predvsem z vidika etikete, je stilistična pisemska teorija 20. stoletja pri opredelitvi vrste pisem in standardiziranju pisemskega sporočanja precej bolj uniformna, osredinjena na razliko formalno oz. uradno nasproti neformalnemu oz. neuradnemu, družbena razmerja oz. razlike med piscem in naslovnikom pa se umikajo v ozadje, kar je zagotovo posledica umika in kasneje tudi nezaželene hierarhije družbenih skupin oziroma težnje po uniformnosti in egalitarizmu v brezrazredni družbi.

Tradicionalno usmerjena Silva Trdina v svoji poljudno ubesedeni literarni teoriji *Besedna umetnost* (1965) nakazuje na estetsko komponento pisanja (oblika in čitljivost) in na pomembnost subjektivne obravnave vsakega naslovljenca, kar je za uradne napotke v tedanji družbi manj navadno:

> »Pismo odraža človekov značaj« (Trdina 1965: 289),
>
> »Redko kakšno delo razkriva avtorja v tolikšni iskrenosti, kakor store to pisma« (Trdina 1965: 290).

Pisma sicer deli glede na naslovnika, in to na (1) **osebna**, torej korespondenco osebnega značaja, ki je namenjena sorodnikom, prijateljem, znancem; (2) **poslovna**, ki »posredujejo poslovne zveze«; in (3) **uradna**, ki so administrativnega in političnega značaja, pošiljajo pa jih državni in drugi uradi (Trdina 1965: 288–289). A že za klasično pismo najbolj uradnega značaja, tj. poslovno pismo, Silva Trdina opozarja, da se »strogo uradni ton po daljšem poslovanju ali po osebnih stikih lahko omili in približa tonu osebnih pisem« (ibid. 1965: 290).

Precej bolj skopo predstavlja pisemsko sporočanje Toporišič: njegovi učbeniki za srednje šole *Slovenski jezik* 1–4 (1965–1973: 224–225), ki so izhajali sočasno z

delom Silve Trdina, pisma delijo glede na naslovnika kot (1) **privatna**, (2) **poslovna** (pisma, namenjena podjetjem, predstojnikom, lahko so tudi brez nagovora, saj je naslovnik zapisan v glavi dopisa), (3) **odprta** (pisma, namenjena več naslovnikom). Pri tem opozarja, da »pišoči naslovnika obravnava kot sogovornika« (1965: 224), le da je ta navzoč le v njegovih mislih.[6]

2.3 Pisma pri Dragu Bajtu

Tudi v »priročniku za pisanje« z naslovom *Pišem, torej sem*, v katerem se je pisemski korespondenci podrobneje posvetil Drago Bajt (1994: 94–96) in nadgradil dotedanje obravnave pisma, je opozorjeno na visoko stopnjo standardiziranosti oz. formaliziranosti pri poslovnih pismih ter na enostransko vodeno komunikacijo, ki se omejuje na nagovor hierarhično višje situiranega pisca in podrejenega naslovnika (npr. stranke). S tem je nakazana tudi protistavnost z **zasebno** korespondenco, saj Bajt trdi, da je zasebno pismo »najmanj standardizirano«[7] in se »podaja na področje govorne komunikacije z vsemi značilnostmi, ki jih knjižna norma prepoveduje, odsvetuje in preganja« (Bajt 1994: 95–96). Bajtovo razmišljanje o jeziku v zasebnih pismih je mogoče povezati z ugotovitvami sodobnih jezikoslovcev o naraščanju neformalnih in pogovornih prvin v jeziku novega medija (npr. Montego-Fleta et al. 2008; Baron 2002b). Zanimiv je tudi njegov dodatek – »Zasebno pismo je še danes napisano z roko in ne s pisalnim strojem ali računalnikom« (Bajt 1994: 96) –, ki priča o obravnavi zgolj klasične oblike pisemske besedilne vrste.

2.4 Predlog za standardizacijo dopisov Monike Kalin Golob

Desetletje za Bajtom se je dopisom, tj. pismom v uradovalni korespondenci, posvetila Monika Kalin Golob (2003) v brošuri *O dopisih*, in sicer z namenom njihove standardizacije. Avtorica navaja, da bo šele po odpravi pravopisnih pomanjkljivosti mogoče več pozornosti nameniti vsebini (in ne le obliki) dopisa (Kalin Golob 2003: 11), ki je »postal obrazec s predvidljivimi, če že ne obveznimi sestavinami« (ibid.: 16).

Dopisi so v delu Kalin Golobove obravnavani ne le kot besedilna zvrst, temveč kot torišče mnogih jezikoslovnih vprašanj in nasprotij – tako pri rabi ločil in posledično velike začetnice za nagovorom, kot tudi pri slogovnih posebnostih, npr.

6 Kasneje, v *Slovenski slovnici* (2000: 721), Toporišič pismo obravnava med tipičnimi besedilnimi vrstami praktičnega sporazumevanja, vendar ne obravnava njegove značilne členjenosti ali posebnosti glede na različnega naslovnika.

7 Kasneje je predlog za standardizacijo dopisa oz. pisma v uradovalnem jeziku objavila Monika Kalin Golob (2003).

vikanju in polvikanju –, ki jih je avtorica detektirala kot predavateljica jezikovnokulturne problematike v okviru GV. Tako natančno opredeli tudi dele dopisa in se ukvarja z jezikovnimi in pravopisnimi posebnostmi, na katere mora biti pišoči pozoren, kasneje (Kalin Golob 2015) pa ugotovljene posebnosti predlaga tudi v drugih standardizacijskih okvirih – pri prenovi pravopisnih pravil.

2.5 Elektronsko pismo

Pisemsko sporočanje je v elektronski dobi doživelo razcvet. Že po prvem desetletju 21. stoletja, ko je bilo naprednejšim akterjem družbenega dogajanja že jasno, da bodo informacijsko-komunikacijske tehnologije prevzele funkcijo klasičnih fizičnih prenosnikov in da v prihodnje svet ne bo potekal po vzporednih tirih analognega in elektronskega (Krek 2014), so statistiki napovedali, da se bo število e-poštnih računov na svetu povečalo z več kot 2,9 milijarde v letu 2010 na več kot 3,8 milijarde do leta 2014 (Radicati 2010), v letu 2015 pa 4,4 milijarde (Radicati 2017).

Primerjalno z drugimi besedilnimi sporočili v novem mediju (npr. besedili v spletnih klepetalnicah in kratkimi besedilnimi sporočili – SMS-ji), ki so doživela precej jezikoslovnih pretresov tudi v slovenščini, zlasti z vidika jezikovnih posebnosti elektronsko posredovanih sporočil (Jarnovič 2007; Jakop 2008; Dobrovoljc 2008; Erjavec in Fišer 2013; Michelizza 2014; Kalin Golob in Erjavec 2014), je bilo že ugotovljeno, da veljajo elektronska pisemska sporočila za prostorsko in tehnološko najmanj omejena in zato najbolj podobna klasičnim, neelektronskim »dopisovalnim« načinom (Crystal 2005, Baron 1998 et al.; za slovenščino Dobrovoljc 2008). Medtem ko pri drugih tipih elektronsko posredovanega sporazumevanja že sama storitev določa in tudi precizira okoliščine pisanja in besedilne zasnove (tviti so danes omejeni na 280 znakov, daljša besedila na Facebooku so v celoti dosegljiva le po kliku na povezavo »Prikaži več« ipd.), za e-pisma velja, da so (kljub možnosti skorajda sočasne izmenjave sporočil) od vseh »hitrih« oblik sodobnega dvosmernega sporočanja najbolj blizu klasičnim pisnim navadam, odvisnim zgolj od dopisovalcev. Ker za pošiljanje e-pisma ne potrebujemo posebnega fizičnega angažiranja, torej ne znamke, ovojnice in ne poštarja ali poštnega nabiralnika, in ker ga lahko shranimo ter beremo na več napravah hkrati, je razmah tega hitrega in gospodarnega načina sporazumevanja pričakovan in logičen.[8]

8 V poslovnem svetu elektronsko dopisovanje (zaradi dokumentiranosti, sledljivosti, preverljivosti, možnosti arhiviranja, hitrosti, udobnosti pri pošiljanju, hiperpovezav, možnosti sočasnega branja na več napravah, možnosti sodelovanja več oseb hkrati) v veliki meri že nadomešča telefonske pogovore – tako Baron (1998). Lastne izkušnje pa govorijo v prid pisanju pisem nasproti telefonu tudi zaradi obzirnosti in vljudnosti, saj naslovniku damo več možnosti, da se na odgovor pripravi. Podobno je pri telefonskem svetovanju o jezikovnih težavah, ki ga izvajamo na Inštitut za slovenski jezik Frana Ramovša ZRC SAZU že več desetletij: uporabnika, ki telefonsko zastavi zahtevno, nenavadno ali nerazumljivo vprašanje, zaprosimo, naj ga zapiše in pošlje po elektronski pošti, pri čemer sicer okrnimo njegovo spontanost in se odrečemo neposrednemu

Tudi zato se pričujoči pregled sporočevalnih navad osredinja prav na pisemsko sporočanje.

3 ZGRADBA PISMA

> »*Pismo ima predpisano zgradbo in tri vidne dele: uvod, jedro in zaključek.*« (Trdina 1965: 289)

Predpisana zgradba je ena od bistvenih določevalnih ali razlikovalnih lastnosti[9] stalnih oblik sporočanja oziroma besedilnih vrst (Toporišič 2000: 721 id.). Pismo v grobem določajo trije temeljni deli, ki jih različni obravnavani avtorji pojasnjujejo in poimenujejo razmeroma enotno kot trodelno strukturiran in diskurzno precej ustaljen besedilni vzorec:

- ***ogovor***, ***vvod***, ***sklep*** ali ***doveršek*** s ***podpisom*** (SSLŠ 1881: 152);
- ***uvod***, ***jedro*** in ***zaključek*** (Trdina 1965: 289), med neobveznimi sestavinami pa navaja tudi ***kraj*** in ***datum***; priporočljivo je pripisati tudi ***pošiljateljev naslov***; po mnenju avtorice je manj primerno dodati ***pripis*** (navadno uveden z okrajšavo *P. S.*), saj »pisma s pripisi niso lepa in očitujejo našo raztresenost« (Trdina 1965: 290);
- ***ogovor***, ***jedro***, ***pozdrav*** ali ***priporočilo*** (Toporišič 1965: 224–225);
- ***vljudnostni nagovor***, ***jedro*** in ***zaključek*** (Bajt 1994: 95); omenja tudi glavo s podatki o naslovniku in pošiljatelju, datum, zadevo ali predmet, jedro dopisa, sklepni pozdrav in podpis ter po izbiri tudi pripis.

Za poslovno pismo Kalin Golobova (2003) navaja naslednje zgradbene elemente:

- ***glava***, ***datum***, ***naslov*** (ki vključuje bodisi poimenovanje (npr. zadeva) ali splošno ime dokumenta bodisi nagovor naslovnika), ***jedro dopisa***, ***pozdravni del*** oz. zaključek dopisa in ***podpis***.

Kaj pa e-pismo? Teoretiki sodobnega internetnega jezikoslovja ugotavljajo, da pisna praksa pri e-pismih v veliki meri upošteva navade, pridobljene pri pisanju tradicionalnih pisem in naučene v klasični stilistični šoli. Tako na primer Crystal (2001: 104) opaža, da večina piscev e-pisem postavlja posamezne samostojne zgradbene elemente pisma v svojo vrstico. Sicer je pri e-pismih večina zgradbenih prvin zaradi zahtev e-poštnega programa določena samodejno: to je ***glava*** s

dialogu, a dobimo vprašanje, na katero je mogoče korektno odgovoriti. Za dopolnitev nabora prednosti e-sporočanja se zahvaljujem anonimni recenzentki ali recenzentu prispevka.

9 V *Slovenski slovnici* (2000: 722–723) Jožeta Toporišiča so med razlikovalnimi posebnostmi različnih besedil poleg zgradbe (tj. razčlenjenost besedila) tudi: funkcijska zvrstnost, prenosnik, dolžina besedila, koherenca, ton pisanja ipd.

podatki o naslovniku in pošiljatelju, lahko tudi ***podpis*** v obliki vizitke. Posebnost glede na klasična pisma je tudi, da e-pisma ne moremo poslati na napačno zapisan naslov: če se zmotimo pri zapisu, nas na napako opozori že e-poštni program sam, če zapišemo napačno ime, nas e-poštni strežnik v sprejemljivem času opozori, da naslovnik s tem naslovom ne obstaja. Program nas lahko celo opomni na pozabljene neobvezne sestavine (npr. ***zadeve*** oz. ***predmeta sporočila***), kar priča o že omenjeni visoki stopnji standardiziranosti in uniformnosti e-pisma kot besedilne vrste, ki najverjetneje izhaja iz tradicije klasičnih pisem.

3.1 Nagovor ali ogovor ter poslovilni pozdrav

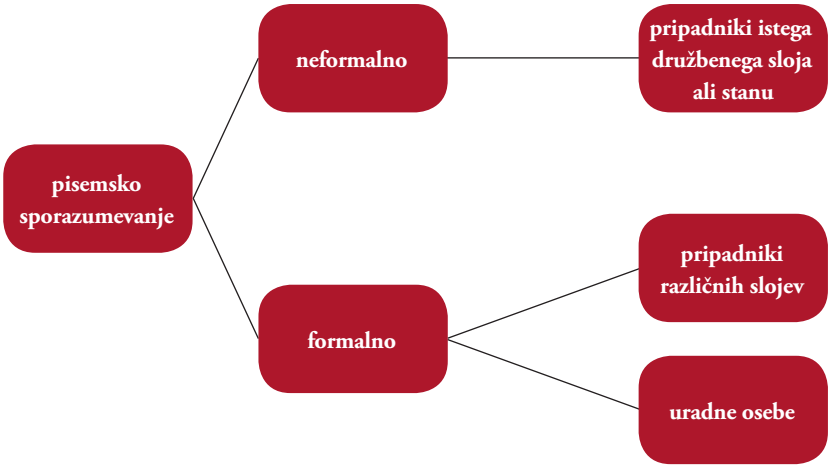
Nagovor ali ogovor imenujemo konvencionalizirani uvodni pisemski pozdrav. Pozdrav je sicer sporočilo, ki si ga posamezniki izmenjamo ob srečanju in slovesu in »ki izražajo potrditev obstoječih družbenih in medosebnostnih odnosov, spoštovanje, počastitev, naklonjenost, prisrčnost, dobrodošlost, dobronamernost, različna prepričanja in pripadnosti« (Makarovič 2013: 205). Gre torej za diskurzivno (in vedenjsko) prvino, ki nakazuje na socialne in medosebne okoliščine sporazumevanja, v pismih pa napoveduje tudi stopnjo formalnosti in način ubesedovanja v nadaljnji interakciji. Kot obvezni in ponavljajoči se sestavini nastopata pozdrav in slovo tudi kot formuli, ki olajšujeta ubesedovalni napor oziroma napor, ki bi ga imeli, če bi komunicirali iz oči v oči (Eckert in McConnell-Ginet 2013: 125), kar rezultira tudi v pomenski izpraznjenosti nagovornih in poslovilnih izrazov. Piščeva jezikovna izbira kaže na njegovo lastno družbeno identiteto in identitete naslovnikov, z njimi »kodira« družbene informacije (Pop 2012).

V pismih in elektronskih pismih se nagovor tesno povezuje tudi z ***zaključkom pisma*** ali poslovilnim pozdravom: od izbire nagovora in vrste nagovora je zato pogosto odvisna tudi primernost poslovilnega pozdrava. Redki avtorji opozarjajo tudi na pisemsko koherenco in prvine, ki to koherenco oz. sovisnost (Krstič 1993: 371, Toporišič 1992) zagotavljajo. Gre za ***ogovorne izraze*** v samem jedru pisma, tj. ***osebne in svojilne zaimke***, ki se – pisani z veliko oz. malo začetnico (*ti*/*Ti*, *vi*/*Vi*; *tvoj*/*Tvoj*, *vaš*/*Vaš*) – povezujejo s tikanjem ali vikanjem, v starejših obdobjih tudi z onikanjem (prim. Dobrovoljc 2011), torej z različnimi koncepti vljudnosti, in ki skupaj z nagovorom oz. ogovorom in podpisom tvorijo pisemski formalni obrazec.

Primerjava konvencij pisemskega oz. e-pisemskega sporazumevanja na časovni osi »19. stoletje–elektronska doba« nakazuje na spremembe, ki jih bomo v obravnavanih gradivskih virih prikazali ločeno za formalno in neformalno sporazumevanje.

3.1.1 Šolska slovnica iz leta 1881

Najobsežnejša navodila za pisemsko sporazumevanje je mogoče najti v šolski slovnici, ki (tudi z opozorilom na naslovnici) temu besedilnemu tipu namenja posebno mesto. Očitno je mogoče zaznati razliko med neformalnim sporazumevanjem, ki poteka med pripadniki ***istega družbenega sloja***, in neformalnim sporazumevanjem med pripadniki ***različnih slojev***, ki se na ravni rabe diskurzivnih prvin (nagovor, pozdrav) približuje formalnemu sporazumevanju z dogovorjenimi formulami in obrazci, vezanimi na protokol (Slika 1). Skrajno formalno pisemsko sporazumevanje avtor povezuje z vsakodnevnimi opravili poslovno-uradovalnega značaja: gre za pogosto brezosebne uradovalne pisemske obrazce.



Slika 1: Naslovniki formalnega in neformalnega pisemskega sporazumevanja v šolski slovnici iz leta 1881.

V šolski slovnici je mogoče najti cel nabor ***nagovorov***, ki jih priporoča za pisemsko naslavljanje prijateljev in znancev: *dragi* ali *dobri*, *ljubi prijatelj*, *preljubi*, *predragi*, *spoštovani*, *mnogocenjeni*, *preljubljeni prijatelj* (SSLŠ 1881: 152–153), še zlasti, če je namen pisanja povezan s posebnimi okoliščinami, zaradi katerih mu je pismo namenjeno (*moj srečni* ali *nesrečni*, *moj žalostni*, *ubogi*, *skerbni*, *prizanesljivi prijatelj* – str. 153). Opozarja tudi na možnost, da se formula nagovora vplete v prvo poved pisma (*Ne bi mi, dragi prijatelj, bil mogel bolj ustreči …* – str. 153) in ni izpostavljena na samem začetku pisma. Avtor opozori tudi na ***ogovorne zaimke*** v besedilu/jedru pisma, ki se povezujejo z nagovorom in pozdravom: »Tudi sredi lista, kjer v pismih na prijatle in znance stojí zaimek 'ti' ali 'vi,' rabijo

se včasi ogovori prejemnikovemu naslovu primerni, n. pr. 'Vaše Blagorodje, Vaša Milost' itd.« (1881: 153).

Pozornost je namenjena tudi ***sklepu*** vsebinskega dela pisma: »Kakor mora pismo imeti priličen vvod, tako mu se hoče tudi dobrega sklepa ali konca« (SS 1881: 154). Avtor opozarja, naj sklep izhaja iz vsebine pisma in se ravna »po višjem ali nižjem stanu tistega, komur pišemo« (str. 155), višji je namreč stan naslovnika (točka 3), več zagotovil vdanosti in spoštovanja naslovnik pričakuje od pisca pisma. Pismo stanovsko višjemu naslovniku torej ni oz. ne more biti več neformalno.

1. Pisma znancem in prijateljem nakazujejo največjo bližino med dopisovalcema, npr.: *Bodi zdrav!*, *Z Bogom!*, *Želeč vsega dobrega Tebi in Tvojim, sem in ostajem Tvoj –.*, *Ostani z Bogom in ljubi svojega zvesto udanega –. S celega serca sem Tvoj* ipd. (str. 155).

2. Podpisi v pismih manj znanim ljudem ali pripadnikom višjega stanu nakazujejo že formalno logiko obrazcev, ki oddaljeni od dejanskega pomena zapisanega: *Z globokim poštovanjem podpisuje se –*, *Sprejmite, Preblagorodni Gospod! zagotovilo globokega spoštovanja, s kterim se podpisuje* – (str. 155).

3. V pismih za najvišje osebe knežjega rodu je v ospredju protokolarna etiketa: *Z najglobokejšo spoštljivostjo sem (ostajem) –*; *Poln veselega upanja pričakujem, da bo milostivo (dobrotljivo) uslišana moja ponižna (preponižna) prošnja, ter ostajem z najglobokejšim spoštovanjem* (str. 156).

V predlogah za neuradno, neformalno dopisovanje oz. primerih »pisem različnega zapopadka« avtor uporabi naslednje nagovorne in poslovilne vzorce, ki so v Tabeli 1 prikazani v povezavi s pisemskim namenom (SSLŠ 1881: 125–148).[10] Ponavljanje tako nagovornih in poslovilnih pozdravov kaže, da namen pisma ne vpliva na izbiro nagovora ali poslovilnega pozdrava.

10 Dvojna poševnica označuje zapis v novi vrstici.

Tabela 1: Namen pisma ter namenu ustrezajoč nagovorni ter poslovilni vzorec po šolski slovnici iz leta 1881.

Namen pisma	Nagovor	Poslovilni pozdrav in podpis
vprašanje	*Ljubi Šimen!*	*Bodi zdrav! // Tvoj // Janez.*
naznanilo	*Draga sestra!*	*Pozdravljamo in poljubljamo Vas vse. // Tvoj // odkritoserčni brat // Jože.*
prosilni list	*Preljubi Blažek!*	*Tvoj // France.*
zahvalni list	*Spoštovani gospod stric!*	*Vaš // hvaležni stričnik Bogumil.*
naznanilo in prošnja	*Draga prijatlica!*	*Tvoja // odkritoserčna prijatlica // I. I.*
poročilo in prošnja	*Prespoštovani gospod doktor!*	*Vaš// ponižni služabnik // I. I.*
popraševanje	*Preljubi brat!*	*Serčno te pozdravljajo oče in mati, drugi bratje, sosebna, tvoj // zvesti brat // Tomaž.*
vošilo ali čestitanje	*Preljubi oče!*	*Poljubljaje Vam, dragi oče, roke sem in ostanem // Vaš pokorni sin // Štefan.*
vošilo za novo leto	*Predragi starši!*	*To vam danes obljubuje in bode tudi izpolnjeval // Vaš hvaležni sin // Janez.*
prošnja za odpuščanje	*Predragi stariši!*	*Naj nocoj zaspim v terdnem zaupanju, da se na me ne jezite, da mi zopet verjamete, in da še priserčno ljubite // Svojega // skesanega sina // Janeza.*
izgovor	*Ljubi prijatelj!*	*Serčno Te pozdravlja Tvoj // Bogoslav.*
prijateljsko očitanje	*Dragi prijatelj!*	*Prosim Vas torej dragi prijatelj, rešite me teh misli, ki mi nepokoj delajo, povejte mi, če tudi le z nekoliko versticami, da ste živi in zdravi, in da ste s svojo ljubeznijo vdani // svojemu vernemu// Janezu.*
povabilo	*Dragi Peter!*	*Priserčno Te pozdravlja // Tvoj zvesto vdani // France Podgornik.*

V posebnem dodatku (»Pridavek«) šolske slovnice (1881: 208–213) so v nekaj odstavkih podani tudi predlogi pisem, naslovljenih na visoke javne osebnosti (*cesarju in cesarici*; *nadvojvodam in nadvojvodinjam*; *kardinalu, ki je obenem knez nadškof*; *knezu nadškofu*; *knezu škofu*; *knezom in kneginjam*; *grofom in grofinjam*; *baronom in baronicam*; *vitezom in plemičem*) in na koncu še na »ljudi neplemenitega rodu, kteri se štejejo k omikanemu svetu pa niso duhovni« (SSŠL 1881: 209–212), ki so v tistem času odločali tudi o mnogih upravnih zadevah. Slovnica za vsako navedeno osebo prinaša podatek o najprimernejšem uvodnem nagovoru, nazivanju v pismu samem, podpisu pisca pisma in nadpisu, tj. napisu na ovojnici pisma. Navajamo primer za pismo baronom in baronicam (1881: 212):

Ogovor: *Preblagorodni gospod baron! Preblagorodna gospa baronica!*

V listu: *Vaše Preblagorodje, Vaša Milost*

Podpis: *Najponižniši služabnik, Najponižniški*

Nadpis: *Preblagorodnemu gospodu baronu (Preblagorodni gospé baronki)*

Gre torej za pisemske konvencije, ki jih Končnik predlaga za pisma, v katerih se zahvaljujemo, prosimo, vabimo ipd. in ki so protokolarno določene – najverjetneje povzete po nemškem vzoru. Še bolj formalno in brezosebno je pisemsko sporazumevanje, ki je namenjeno poslovanju in gre pravzaprav za uradna pisma (imenuje jih »listi«), pogosto naslovljena na pripadnike višjega stanu (Tabela 2). Danes bi jih poimenovali poslovna korespondenca.

Tabela 2: Vzorci opravilnih formalnih pisem v šolski slovnici iz leta 1881.

Vrsta pisma	Nagovor	Poslovilni pozdrav in podpis
naročilni list ali naročbenica	*Gospodu L. C. Hardmuth na Dunaju*	*S spoštovanjem Vaš vdani I. I. kupec*
	Gospodu Janezu Štepančiču v Ipavi	*S posebnim poštovanjem Vam vdani Pavel Poklukar // gostilničar*
ponudni list ali ponudnica	*Visokorodna Gospá! Milostiva Grofinja!*	*Blagovolite, Milostiva Gospa, sprejeti zagotovitev globokega spoštovanja, s kterim se podpisuje Vašega Visokorodja ves vdani služabnik I. I. // kipar*
	Poštovani Gospod!	*Vašega Blagorodja vdani služabnik I. I. pek*
izgovorni listi ali izgovornice	*Blagorodni Gospod!*	*S polnim spoštovanjem Vašega Blagorodja ponižni služabnik I. I. // krojač*
	Mnogo spoštovani gospod!	*Priporočaje se Vaši prijazni dobroti sem sem in ostanem s posebnim spoštovanjem Vašega Blagorodja vdani služabnik Matija Premerl, krojač*
opominjavni listi ali opomenice	*Blagorodni, // Prespoštovani gospod! Velecenjeni gospod!*	*S spoštovanjem Vašega Blagorodja vdani Nikolaj Terpin, knjigovez*

3.1.2 Silva Trdina, Jože Toporišič, Drago Bajt in Monika Kalin Golob

Glede na Končnikovo slovnico skromen nabor različnih nagovornih izrazov pri Toporišiču, Trdini in Bajtu priča o koreniti spremembi pisemskega vedenja in

družbenih konvencij. Od naštetih avtorjev še največ pozornosti pisemskemu »bontonu« nameni Silva Trdina, saj pravi: »naziv na naslovljenca […] mora biti prilagojen osebi, ki ji je pismo namenjeno« (Trdina 1965: 289) in navaja (ne glede na to, ali gre za zasebno ali uradno pisanje) naslednje pisemske ***nagovore*** oz. ogovore (str. 289):

Draga Nada.
Ljubi moj.
Predragi oče.
Preljuba mamica.
Spoštovani gospod profesor.

Za t. i. »zvalnikom« priporoča klicaj, enako kot šolska slovnica pa dopušča tudi možnost vejice, a le, če se nagovor vpne v prvo poved pisma, sicer ne. Podobno velja za sklep, ki ga je mogoče povezati z zadnjim stavkom, npr. *Vdano pozdravlja Vas vaš hvaležni dijak …* (str. 289). Silva Trdina ohranja nekatere vljudnostne navade, ki kažejo na večjo formalnost pri pozdravljanju, npr. izkazovanje vdanosti (očitno v pismu nadrejenemu ali učitelju), navade, ki – tudi glede na izbiro naslova *gospod* v pismu samem in *tovariš* na pisemski ovojnici – koreninijo še v obdobju pred drugo svetovno vojno.

Toporišič (1965: 224–225) je pri navajanju nagovorov manj stilno raznolik: navaja dve možnosti, in sicer neformalni nagovor *dragi* za prijatelja (*Dragi prijatelj*), formalni *spoštovani* za predsednika (*Spoštovani tov. predsednik*).[11] Besedilni zgled, s katerim ponazori razliko med formalnim in neformalnim, je Cankarjevo pismo materi (z nagovorom *Ljuba mati!* in pozdravom *Srečno! Ivan Cankar*) in uredniku Zvona, Antonu Aškercu (z nagovorom *Blagorodni gospod Aškerc!* in poslovilnim pozdravom *Z najodličnejšim spoštovanjem vdani Vam Ivan Cankar*) (str. 226). Izbrana zgleda tudi Toporišičev nauk (podobno kot teorijo Trdinove) za pisanje pisem postavljata v obdobje prve polovice 20. stoletja, ko je izkazovanje vdanosti obvezna sestavina pisemskega bontona.

Manj se s konkretnimi nagovori ukvarja Bajt (1994), vendar v drugem delu svojega priročnika *Pišem, torej sem*, ki ima naslov »Vzorci besednih vrst«, prikazuje značilne pisemske vzorce, v katerih se kažejo malenkostne razlike pri izboru nagovora v neformalnih[12] in formalnih položajih, predvsem v uporabi jedra samostalniške besedne zveze, ki sledi (*gospod*, *tovariš*). Še najbolj se je uveljavil nagovor *Spoštovani*, ki je pravzaprav pridevnik v samostalniški rabi in ustreza največ tipom naslovnika – ne glede na spol, stan, starost ipd.:

11 Raziskave o naslavljanju s *tovariš* in *gospod* oz. *tovarišica* in *gospa* … za slovensko pisemsko sporazumevanje v 20. stoletju še niso bile izvedene.

12 Zasebnih pisem Bajt ne prikazuje.

- pismo iz računovodstva (nagovor: *Spoštovani*, brez poslovilnega pozdrava),
- pismo iz uredništva revije (nagovor: *Spoštovani*, poslovilni pozdrav: *Lepo Vas pozdravljam*),
- uradno pismo iz društva (nagovor: *Spoštovani tovariš Bajt*, poslovilni pozdrav: *Lepo pozdravljeni*),
- osebno opravičilo razredničarki (nagovor: *Spoštovana gospa učiteljica*, poslovilni pozdrav: *S spoštovanjem // Irena Kastelic*).

Monika Kalin Golob (2003) sicer v okviru rabe nagovora pri uradnem dopisovanju opozarja na nevljudno krajšanje nazivov (npr. *ga./g.*) in na dejstvo, da imajo »slovesnejši dopisi« poleg priimka tudi rojstno ime. Prednost daje nagovoru *Spoštovani*, pri čemer opozarja, da je najprimernejši tudi z vidika neznanega (edninskega/množinskega) naslovnika, pri tem opozori tudi na doslednost ter pogosto polvikanje namesto vikanja pri ženskih naslovnicah (*Spoštovani* in ne *Spoštovana* …, če osebno ženskega spola vikamo) (str. 36).

Zanimivo razlikovanje uvede Kalin Golobova pri rabi ločila za uvodnim nagovorom, in sicer loči med uradnimi pismi, v katerih za nagovorom uporabimo klicaj, in zasebnimi pismi, ki končujejo nagovor z vejico, pismo pa nadaljuje zato jedro besedila z malo začetnico (str. 36–37). Posebna pozornost je v monografiji *O dopisih* namenjena vljudnosti in spoštljivosti (str. 44), kar je bilo v preostalih obravnavanih priročnikih iz druge polovice 20. stoletja umaknjeno v ozadje. Morda so tudi zato končni pozdravi predstavljeni kot avtomatizirani, a raznoliki zgradbeni elementi:

Posvetimo se najprej **zaključkom**. Imamo nekaj možnosti:

a) povsem avtomatizirani. S spoštovanjem; Lep pozdrav; Lepo Vas pozdravljam;

b) stavčni: V upanju na čimprejšnji odgovor Vas lepo pozdravljam;

c) priložnostni: Z odličnim spoštovanjem. (Kalin Golob 2003: 49)

V nadaljevanju navaja najbolj značilne tipe pisemskih zaključkov, ustrezne vsem trem zgradbenim različicam, in prilagojene funkciji posameznega dopisa.

3.1.3 Elektronsko pismo in pisemske navade v elektronski dobi

21. stoletje je v pisemsko sporazumevanje govork in govorcev slovenščine pripeljalo spremembe razmeroma ustaljenih vzorcev. Gradivo e-pisem, pregledanih v

osebnem e-predalu in v korpusu Janes[13] (predvsem pogovori med uporabniki in uredniki na Wikipediji), kvalitativno in kvantitativno dokazuje, da se formalno pisemsko sporazumevanje (tj. sporazumevanje med manj znanimi osebami ali v poslovnem svetu) pri naslavljanju in pozdravljanju ne omejuje zgolj na vzorce, ki so jih predstavljali priročniki v 20. stoletju, temveč udeleženci pisemskega sporazumevanja zavestno in načrtno uporabljajo nove izrazne načine, kar tudi utemeljujejo.

Pri interpretaciji novosti se opiramo na rezultate anketne raziskave (Vprašalnik o naslavljanju 2018), v kateri je na primer kar 80 odstotkov vprašanih prepričanih, da je izbira ogovora povezana z vljudnostjo oziroma da se pri dopisovanju skušajo ravnati po svojem dopisovalcu (33 %), kar priča o poudarjeni personalizaciji diskurza oz. prilagajanju vsakokratnemu naslovniku oz. dopisovalcu.

3.1.3.1 Pisma brez nagovora

Pri daljšem oz. kontinuiranem dopisovanju si vse pogosteje izmenjujemo sporočila brez nagovora, kar potrjujejo tudi izsledki v anketi. Večina vprašanih s svojimi odgovori pritrjuje trditvi antičnega misleca Teofrasta, da pismo brez pozdrava kaže na nevljudnost pisca (Teofrast; po Makarovič 2013: 222); kar 75 odstotkov vprašanih namreč meni, da je pismo brez nagovora neprimerno, da pismo brez nagovora priča o »pomanjkanju pisne in občevalne kulture«, »kulturni nerazvitosti« ali celo lenobi. Četrtina, ki ni takega mnenja, pojasnjuje opuščanje nagovora bodisi (1) s hitrim dopisovanjem, naglico in gospodarnostjo bodisi (2) s posebnim tipom komunikacije, pri katerem je glavna informacija pravzaprav v priponki.

Morda je dopisovanje brez nagovora posledica hitrega prevzemanja navad ob izmenjavi elektronskih sporočil drugega tipa, npr. spletnih klepetalnic ali veriženja SMS-jev, kjer sporazumevanje poteka skorajda sinhrono in vsaka nova interakcija zgolj nadaljuje prejšnjo, zato delujejo pozdravi ob vsakokratnem odzivu redundantno in jih dopisovalci tudi izpuščajo.

13 Gradivo referenčnega korpusa Kres prikazuje predvsem nagovore iz parlamentarnega življenja (*spoštovani poslanci*, *spoštovani gospod minister*, *predsednik*, *podpredsednik* ipd.), v korpusu Janes (Fišer et al. 2016) je pisemsko ogovarjanje prisotno predvsem v **pogovorih na Wikipediji**, ki so dovolj avtentični, da kažejo na vse bistvene elemente pisemskega sporazumevanja. V prispevku zato navajamo ponazoritve iz korpusa Janes.

Prvo vprašanje: *Pozdravljeni na vas se obracam glede izvida za avtoimunsko diagnostiko. […] Kaj lahko pricakujem v primeru ponovno pozitivnega izvida? Hvala*

Prvi odgovor: *Pozdravljeni, laboratorijski test še ne zadosti diagnostičnim kriterijem za antifosfolipidni sindrom. […] LP!*

Podvprašanje: *Hvala za odgovor. Toda ali je lahko to krivec? Lp*

Odgovor: *Ne morem s sigurnostjo reči, da ne. LP!*

Slika 2: Primer dopisovanja v spletni klepetalnici z opuščanjem nagovora v neprvih replikah (Vir: https://med.over.net/forum5/viewtopic.php?t=5700019).

Čeprav je pri formalnem dopisovanju izostanek nagovora pogosto prikrit z »zadevo«, v kateri zapišemo piščev namen, na kar opozarja tudi Kalin Golobova (2003: 35), po pregledu osebnega poštnega predala ugotavljam, da je pri uradnem e-pisemskem sporočanju takih primerov še vedno malo. Brez pozdrava oz. nagovora so le redka formalna pisma (14 od 141), čeprav so hkrati brez zadeve ali imena dokumenta, kot predlaga Kalin Golob (2003: 35), brez poslovilnega pozdrava pa še nekaj manj (2 od 141). Najpogosteje se poslovimo kar z *lep pozdrav* (Tabela 3) v različnih oblikah: ali z elipso glagolske in zaimkovne oblike *(želim vam) lep pozdrav* ali z izraženo prvoosebno glagolsko obliko (*lepo vas/te pozdravlja(m)*).

Tabela 3: Začetni in končni pozdrav pri formalnem e-pisemskem sporazumevanju.[14]

Nagovor	Pogostnost	Poslovilni pozdrav	Pogostnost
spoštovani (X)	52	*lep pozdrav*	79
spoštovana gospa (X)	46	*lepo Vas pozdravlja X*	19
spoštovana X	26	*LP*	14
pozdravljeni (X)	11	*lepo Vas pozdravljam*	7
dragi (X / gospod X)	4	*s spoštovanjem*	6
pozdravljena, X	4	*prisrčen pozdrav*	3
dober dan	2	*prijazen pozdrav*	2
lepo pozdravljeni	2	*lepo pozdravljeni*	2

Precej več e-pisem brez pozdrava je pisanih za neformalne položaje. Tako je v pregledanem gradivu brez nagovora oz. uvodnega pozdrava kar 167 od 330 pisem, brez poslovilnega pozdrava pa nekoliko manj (59 od 330), nekateri pisci izpuščajo

14 Upoštevane so tudi oblike z množinskim naslovnikom moškega spola (*spoštovani kolegi*) in množinskim naslovnikom, ki zajema osebe obeh spolov (*spoštovane kolegice in kolegi*).

pridevnik in nagovorijo kar z imenom (16 od 330), kar je pogosta, a po prepričanju četrtine anketiranih uporabnikov e-pošte (25 %) manj kultivirana možnost.

3.1.3.2 Nagovor v formalnih e-pismih

Pri formalnem e-pisemskem sporazumevanju se – sodeč po pregledanem gradivu in analizi vprašalnika (2017) – najpogosteje pojavljata nagovora *spoštovani* in *pozdravljeni* (v različnih oblikah), le redko pa si uradno dopisujemo tako, da uporabimo le ime, še redkeje uporabimo kar pozdrav, npr. *dober dan* ali *dobro jutro*. Anketiranci na vprašanje, ali tudi v elektronskem pisemskem sporazumevanju uporabljajo obliki, ki so se jih naučili v obdobju šolanja, navajajo, da jim te oblike »zvenijo starinsko, arhaično« in da klasičnih pisemskih nagovorov (*spoštovani*, *dragi*) ne uporabljajo, ker so preveč uniformirani ali tudi preveč »intimni« oziroma povzeti po angleščini (*dear* za *dragi*). Kljub navedenim dejstvom največ anketirancev (86 %) v uradni korespondenci uporablja obliko *spoštovani/spoštovana* – še zlasti ob prvem stiku z naslovnikom.

Spoštovani (spoštovana gospa, spoštovani gospod …)

V formalnih e-pismih je najpogosteje uporabljen nagovor *spoštovani* oz. nagovor s pridevnikom *spoštovani* (*spoštovana gospa*, *spoštovani bralci*, *spoštovani kolega* …), to dokazuje tako pregled rabe spletnih pisemskih besedil v korpusu Janes (Slika 3) kot tudi anketna raziskava, ki kaže, da uporabniki tudi najpogosteje prejemajo pisma s tem nagovorom. Ne le avtorji zgoraj opisanih priročnikov, tudi sestavljavci *Slovarja slovenskega knjižnega jezika* (dalje SSKJ) so nagovor *spoštovani* prepoznali kot »vljudnostni nagovor« (prim. geslo *spoštovati*). Pisemski pozdrav *spoštovani* se danes pojavlja v najrazličnejših skladenjskih položajih in je tudi besednovrstno konverzen:

a) Kot pridevnik v zvezi pridevnika in samostalnika, ki označuje naslov ali poklic naslovnika, ter desnega samostalniškega imenskega prilastka, ki ga sestavlja le priimek ali pa tudi osebno ime, npr. *spoštovana gospa/profesorica … (Mojca) Kalan*, *spoštovani gospod/doktor … (Andrej) Kalan*. Naslov in poklic ogovorjenega pogosto tudi okrajšamo (*g.*, *dr.*, *prof.* … in podobno), čeprav se v nekaterih okoljih raba okrajšav v nagovoru obravnava kot nevljudna (Kalin Golob 2003: 36).

b) V zvezi pridevnika in lastnega imena naslovnika, če gre za naslavljanje posameznika (*spoštovana Mojca*).

c) V zvezi pridevnika in občnoimenskega poimenovanja naslovnika (*oče*, *sestra*, *stric*, *teta* ipd.), ki razkriva medsebojni odnos med pošiljateljem

in naslovnikom (*spoštovani stric*, *spoštovani gospod*) oziroma skupinskim naslovnikom (*spoštovani kolegi in kolegice*, *spoštovani starši*).

č) Kot posamostaljeni nagovorni pridevnik, ki je rabljen v množinski (tudi zaradi vikanja) ali v edninski obliki (*spoštovani* mn., *spoštovana* ž).

Kot najprimernejša oz. najbolj univerzalna nagovorna različica se pojavlja *spoštovani* v primerih, ko nagovorimo skupinskega naslovnika ali neznano osebo. Kalin Golob (2003: 36) opozarja, da za »običajne dopise« zadošča le navedba naslova (*Spoštovana gospa Kalan*), »slovesnejšim« pa dodamo pred priimek še ime, lahko tudi funkcijo, akademske nazive, če jih poznamo ali vemo zanje (*Spoštovani gospod minister Kalan*) (str. 36).

Spoštovani gospod Jalen, menim, da če je vaš članek tako dober, sploh ne potrebuje samohvale!
Spoštovani g. Kocijančič,// se opravičujem za tako pozen odgovor na Vaše vprašanje,
Spoštovani gospod Ziga.
Spoštovani Klemen*! Obračam se osebne na Tebe kot glavnega administratorja.*
Spoštovani g. Yerpo *Ha? in ostali urejevalci!*
Spoštovani Klemen in Tone,
Spoštovani Zaplotnik!
Spoštovani R.P.,
Spoštovani Generalmajor, morda nekaj pojasnil o Wiki hierarhiji
Spoštovani IP 213.157.228.253.
Spoštovani wikipedist z nick "Žiga"!
Spoštovani gospod ali gospa oziroma gospodič ali gospodična, nažalost Vaš vzdevek ne razkrije spola zato se Vam že vnaprej opravičujem!
Spoštovani gost, prosim za prijavo.
Spoštovani uporabnik!
Spoštovani urejevalci, uredniki, administratorji in birokrati.
Spoštovani kolegi, // z vsem dolžnim spoštovanjem izjavljam, da jaz NISEM avtor komentarja o venetologih.
Spoštovani soavtorji strani o FERI-ju, posebej g. Majhenič.
Spoštovani Wikipedisti,
Spoštovani, nekdo, ki to zna na enostavnejši način, naj decimalne vejice zamenja s pikami, tako kot to pišemo v slovenščini.
Spoštovani! S kakšnim razlogom brišete vsebino (besedilo in slike), ki je bila predhodno že pregledana oz. odobrena s strani ostalih urednikov?

Slika 3: Korpus Janes: primeri rabe samostalniške in pridevniške oblike spoštovani iz pogovorov na Wikipediji (https://sl.wikipedia.org/wiki/).

Kljub skrajni formalizaciji te oblike pozdrava pa se v zvezi z nagovorom *spoštovani* pojavljajo tudi uporabniške dileme, ki jih izpričuje Jezikovna svetovalnica ZRC SAZU in ki so povezane z (1) vikanjem nagovorjenih oseb ženskega spola. Te je mogoče izvajati zgolj s samostalniško obliko (*spoštovani*), s pridevniško pa zaradi težnje po ujemanju s samostalniškim jedrom ne (*spoštovana gospa*) (prim. Jelovšek 2015). Druga dilema se navezuje na (2) ujemanje, ki je mogoče po bližini ali s celotno besedno zvezo, če je pridevnik *spoštovani* rabljen pred prirednim sestavljenim samostalniškim jedrom (*spoštovani bralci in bralke*, *spoštovani bralci*

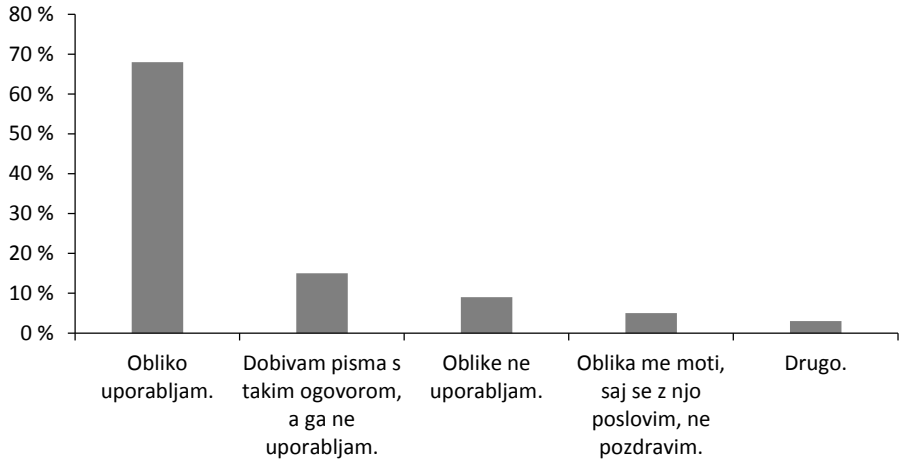
in spoštovane bralke, *spoštovane bralke in bralci*); prim. Dobrovoljc 2014). Če upoštevamo spolno vključujočo rabo jezika, je v evropskem prostoru mogoče opaziti težnje po ujemanju z bližnjim (Dobrovoljc 2018)

Pozdravljeni (pozdravljena, gospa; pozdravljeni, gospod …)

Dosedanja pisemska praksa in stilistika nista izpostavljali nagovora pri formalnem e-pisemskem sporazumevanju, ki ga vse pogosteje izkazuje raba in ga potrjujejo tudi anketni odgovori. Gre za nagovor *pozdravljeni*, ki ga srečujemo v različnih oblikah: *pozdravljena* (tudi *pozdravljena* + ime), *lepo pozdravljeni*, *pozdravljen* (tudi *pozdravljen* + ime) ali ime + *pozdravljeni*. Za tovrsten nagovor v nezasebnem pisemskem sporazumevanju se je odločilo kar 55 odstotkov vprašanih v izvedeni anketi, ki so svojo izbiro utemeljevali na različne načine, predvsem pa z dejstvom, da predstavlja neko ***manj formalno*** ali ***polformalno*** nagovorno možnost, npr.

- »*[S]poštovani/spoštovana* uporabljam v uradni korespondenci – običajno ob prvem »stiku«, kasneje jo zamenjam z obliko *pozdravljeni* oz. *pozdravljena*« (Komentar v Vprašalniku o naslavljanju, 2017).

Pomislek glede tega pisemskega nagovora, ki je sicer klasični pozdrav, ki izraža željo po zdravju naslovnika (Makarovič 2013), je povezan s **podvajanjem**, saj predstavljajo stalne zveze *lep pozdrav*, *z lepimi pozdravi*, *prisrčen pozdrav*, *lepo vas/te pozdravljam* – tudi v skrajšani obliki *lp* in podobno že razmeroma uveljavljeni poslovilni sodobnega klasičnega in elektronskega pisma (Dobrovoljc 2016). Zanimivo je, da se tega nagovora oprijema vse več pišočih, kar nas utrjuje v prepričanju, da pri sporazumevanju potrebujemo več ogovornih izrazov, ki bi (v registru) zajeli polformalne sporazumevalne položaje.

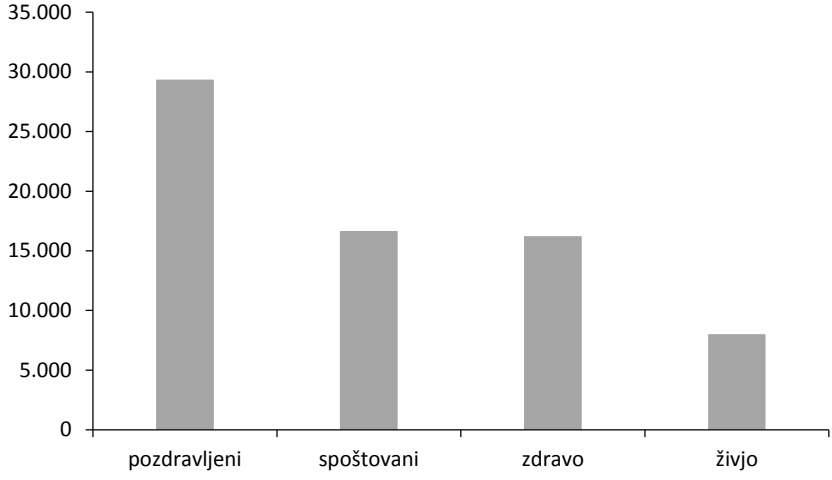


Slika 4: Odgovori anketirancev na vprašanje o rabi nagovora *pozdravljeni*.

Uporabniki, ki so v anketi odgovorili, da izraz uporabljajo, so svojo izbiro utemeljili z različnimi argumenti: včasih želijo doseči višjo stopnjo nevtralnosti oz. se izraziti v položajih med formalnim in neformalnim, gre torej večinoma za govorce, ki ne želijo uporabiti dveh preferiranih nagovornih izbir za formalno (*spoštovani*) in neformalno (*dragi*) – glej Trdina (1965) in Toporišič (1965), saj obliko *pozdravljeni* doživljajo kot najbolj nevtralno:

- »[O]blika je manj formalna kot *spoštovani* in bolj formalna kot npr. *živijo*/*živjo*, hkrati pa dovolj vljudna za različne sogovornike.«
- »Primerna je za različne naslovljence.«
- »Uporabljam jo le, kadar se želim prilagoditi svojim dopisovalcem.«
- »Pri rabi skrbim, da uporabim drugačen končni pozdrav.«

Navedeno potrjujejo tudi dopisovalci na uredniških straneh Wikipedije, ki sicer prikazujejo različna razmerja med piscem in naslovnikom, a bi njihovo korespondenco najlažje umestili med ***polformalna pisma***: nagovor *pozdravljeni* uporablja kar 42 odstotkov piscev pregledanih e-pisem.



Slika 5: Razmerja med nagovornimi izrazi v pogovornih oz. pisemskih besedilih na Wikipediji, kakor se izkazujejo v korpusu Janes.

Razmerja med nagovornimi izrazi v pogovornih oz. pisemskih besedilih na Wikipediji, kakor se izkazujejo v korpusu Janes, kažejo, da se nagovorni izrazi, pridobljeni skozi institucionalno ali neformalno izobraževanje v zadnjih desetletjih,

umikajo in da se elektronsko sporazumevanje nasploh približuje govoru. Gre za porast diskurzivnih pisemskih elementov, ki še najbolj ***posnemajo govorjeno*** sporazumevanje, npr. pozdravi *Dober dan!* ali *Pozdravljeni, Mojca* (vikanje) ali *Pozdravljena, Mojca* (tikanje).

3.1.3.3 Nagovor v neformalnih e-pismih

Poleg že omenjenega in uveljavljenega pozdrava (*draga/dragi (X)*), ki se sodobnim uporabnikom pogosto zdi – sodeč po odgovorih v anketi – preveč papirnat, oddaljen od dejanskih govornih situacij, hkrati pa je tudi preveč zaseben za rabo izven ozkega kroga najbližjih, je tudi v zasebnih in polzasebnih e-pismih vse pogosteje uporabljena oblika pozdrava *pozdravljeni.* Sicer je mogoče opaziti, da je zaradi dejstva, da je izbira pozdrava v zasebnih e-pismih vse bolj prepuščena piscu, nabor nagovorov v tem tipu e-pisem zelo pester. Najpogosteje je uporabljen pozdrav *živjo* (ki je reducirana oblika pozdrava *živijo*) ali nepodomačena različica *živio.* Nekaj pozdravov je starejših: npr. *zdravo*, ki je že v SSKJ opredeljen kot pozdrav med prijatelji in znanci, v obdobju socializma pa so ga uveljavljali v šolah. Nekateri pozdravi, težko bi jih imenovali nagovori, so v starejših slovarjih, pa tudi v SSKJ, opredeljeni kot medmeti za izražanje veselja ali razigranosti (*hojla*, *oj*, *ej*), tudi opozorila (*hej, ojla*). Pojavljajo se tudi pozdravi iz drugih jezikov: angleščine (*hello*), češčine (*ahoj*), tudi fonetično podomačeni (*haj*, *helou, halo*), ali npr. poslovilni pozdrav iz italijanščine (ciao): *čao* ali *čaw*[15] v funkciji uvodnega pozdrava. Nekateri pozdravi so individualno preoblikovani: okrajšani oz. poenobesedeni (*živ* in *ajd* iz *živijo* in srbskega *hajde*).

Podobno kot pri formalni e-pisemski komunikaciji tudi pri neformalni opažamo nabor različnih pozdravov, ki jih uporabljamo zelo individualizirano. Eden od anketirancev je opozoril tudi na razlike: »bližnji prijatelji, s katerimi smo izoblikovali nek čisto svoj jezik, tudi pri sorodnikih, pri katerih nekatere vikam« (glej Prilogo). Razlike med uveljavljenimi poslovilnimi pozdravi in novimi možnostmi so pri *neformalni* e-pošti zato precej bolj opazne. Stalnica oz. najpogosteje uporabljena oblika je *lep pozdrav* v različnih možnostih:

- z izraženo glagolsko (*pozdraviti*: *lepo te pozdravljam*) ali samostalniško obliko (*pozdrav*),
- v manjšalnicah (*pozdravček*) in
- z različnimi pridevniki ali brez njih: *LP*, *lep pozdrav*, *pozdrav*, *lepo te pozdravljam*, *pozdravček!*, *lep pozdravček*, *prisrčne pozdrave*, *sončen pozdrav*, *plp* ('prav lep pozdrav').

15 V Slovenskem pravopisu 2001 (www.fran.si) je izrecno poudarjena zgolj poslovilna funkcija pozdrava *čao*.

Približevanje govornim položajem je razvidno iz pozdravov *dijo*, *srečno*, *hajd*, *čao*, *bye*, *pa pa*, *se vidva*. Namesto pisemskega ali pogovornega poslovilnega pozdrava govorci uporabljajo tudi:

- zahvale (*najlepša hvala*, *thx*, *tenks*),
- želje za drugega (*drži se*, *dobro se drži*, *lep vikend*, *lep dan*, *uživaj*, *bodi v cvetju*, *floriraj*, *mej se, ol'd best*) in
- svoje želje (*piši mi*, *pokliči me*, *oglasi se*),
- opise telesnih stikov (*hug*, *cmok*, *poljub*, *lupčka*) – slednje pogosto nadomeščajo emodžiji, ikone in v funkciji emodžijev uporabljena ločila :) in ;).

4 SKLEP

Pregled predstavljenih pisemskih stilistik in učbenikov v prvem poglavju kaže, da so nagovori in poslovilni pozdravi v pismih precej ustaljeni pisemski vzorci, ki so se v 20. stoletju v primerjavi z 19. že opazno sprostili. Razlike med različnimi tipi nagovora in pripadajočega zaključka pisma so odvisne od vrste korespondence (poslovna, osebna, uradna …), stopnje formalnosti in neformalnosti med dopisovalcema/dopisovalci, trajanja korespondence in vloge/mesta posameznega pisma v njej ter vsebine poslanega.

Pregledano sodobno gradivo[16] potrjuje, da so ***nagovori*** pojmovani kot povsem samostojne diskurzivne enote, navadno sestavljene iz zveze pridevnika in različnih vrst samostalnikov, ki pogosto označujejo stan, družbeni položaj in ime naslovnika; na drugi strani pa ***poslovilni pozdravi*** odražajo naklonjenost do naslovnika ali mu sporočajo dobre želje. Danes se ta forma ohranja predvsem v formalnih pismih, v neformalnih pa se pojavlja cela paleta različnih možnosti, povzetih iz drugih jezikov in govorjenega jezika. Obe najbolj opazni prvini pisemskega diskurza, nagovor in poslovilni pozdrav, v neformalnem pismu vse bolj odražata osebno refleksijo pisca pisma in sta glede družbenih pričakovanj ter etikete vse bolj svobodna. Lahko ju razumemo tudi dobesedno, ne zgolj kot obrazec. Ta interpretacija temelji na odgovorih vprašanih v izvedeni anketi.

Glede odločitve o tem, ali je nagovor obrazec, ki nima pomenske vrednosti, se je enako število vprašanih opredelilo za pozitiven in negativen odgovor, saj je sedem odstotkov vprašanih odgovorilo, da je odgovor obrazec, zato o pomenu uporabljenih besed ne razmišljajo, po drugi strani pa je osem odstotkov

16 Prave stilne analize elektronskega pisemskega sporočanja še ni bilo opravljene. Normo pisemskega sporočanja smo želeli ugotoviti z opazovanjem in primerjavo posameznih vzorcev. Ker ob izvedeni anketi in pregledu e-predala avtorice prispevka ne moremo upravičiti reprezentativnosti vzorca za kako statistično posplošitev, se odločamo za induktivnost na osnovi homogenosti vzorca (elementi e-pisemskega sporočanja v e-pismih).

menilo, da nagovora, kot sta *dragi* in *spoštovani*, ne ustrezata njihovemu odnosu do naslovnika.

V kontrastivni primerjavi tradicionalnih in sodobnih načinov nagovora in poslovilnega pozdrava se je izkazalo, da sodobno ***formalno*** pisemsko in e-pisemsko sporazumevanje sicer kaže ohranjanje nekaterih navad, kakor se kažejo že v slovnici iz leta 1881 in v priročnikih druge polovice 20. stoletja, vendar primerjava opozarja na preureditev pojmovanj o vljudnosti pri medsebojnem naslavljanju. Vsekakor je najbolj opazno to, da se nagovori in pozdravi ***krajšajo***. V nagovorih se izgubljajo predvsem leva pridevniška določila, kot so *preljubi*, *predragi*, *velecenjeni*, *prespoštovani*, *blagorodni*, v poslovilnih pozdravih se opuščajo zagotovila vdanosti in ponižnosti (*vaš vdani*, *ponižni* …). Tako formalna kot ***neformalna*** pisemska etiketa se sprošča in pisma se personalizirajo, saj za večino piscev nagovor in pozdrav nista več pomensko izpraznjeni formuli, neodvisni od naslovnika. Pri poslovnih in uradnih e-pismih se pisemsko naslavljanje in pozdravljanje razvija v dve smeri. (1) V nekaterih okoljih se oba obrazca avtomatizirata – še zlasti če uporabljamo pisemske predloge v urejevalnikih besedil in jih uporabljajo samodejno, neodvisno od konkretne pisemske situacije. (2) V drugih okoljih pa tudi formalno sporazumevanje sledi neformalnemu in se pri nagovoru in poslovilnem pozdravljanju oddaljuje od togega obrazca in konvencije, ki sta personalizirana, a še vedno podrejena logiki klasičnega pisma. Narašča raba polformalnih oz. nenaučenih pisemskih oblik, npr. z uvajanjem novih oblik nagovora (*pozdravljeni*) ter sploh izpuščanjem nagovornih pisemskih oblik pri kontinuiranem dopisovanju.

V nadaljnjih raziskavah, ki jih spodbuja naraščajoče e-pisemsko komuniciranje, bo treba upoštevati tudi dejstvo, da po e-pošti komunicirajo pogosto sodelavci istega kolektiva, iste delovne skupine in celo ljudje, ki se nahajajo v istem prostoru, in da morda v te okviru ne bo več smiselno govoriti o pisemskem sporočanju, temveč zgolj o elektronski izmenjavi sporočil. Na dejstvo, da druge oblike elektronskih sporočil, tj. sporočila spremljevalne narave, ki nimajo značilne zgradbe, počasi izrivajo e-pisemsko komunikacijo, bi lahko pomislili tudi ob podatku, da se – sodeč po raziskavah skupine Radicati (2017) – e-pisemska izmenjava zmanjšuje, in sicer z 247 milijard e-pisem na dan v letu 2009 na 205 milijard e-pisem na dan v letu 2015, in to kljub večanju števila e-poštnih naslovov. Pisma so danes pogosto zgolj spremljevalna besedila ob dokumentih, ki jih ena oseba posreduje drugi v obliki priponke, zato je odsotnost spremljevalnih obrazcev toliko bolj razumljiva, saj prihaja do skrajne racionalizacije in zmanjšanja vseh vljudnostnih in družbenih pričakovanj. Po drugi strani psiholingvisti (Manger et al. 2003; Bertacco in Deponte, 2005 – po Dergs in Bakker 2010) opozarjajo, da vse bolj »plitvo« komuniciranje, kakršnega zasledimo pri neformaliziranih elektronskih sporočilih, ki izgubljajo pisemski značaj, zgolj na videz ustvarja vtis družbene bližine, ki spodbuja le površinsko obravnavo naslovnika. Povedano navaja k sklepanju, da bi morali diskurzivne prvine e-pisemskega sporočanja vzdrževati tudi v

e-dobi, saj imajo očitno posebno funkcijo pri izgradnji in ohranjanju medsebojnih odnosov.[17]

Literatura

Kozma Ahačič: Anonimni (Peter Končnik?). *Slovenska slovnica z naukom, kako se pišejo pisma in opravilni sestavki*. Ahačič, Kozma (ur.): *Slovenske slovnice in pravopisi: spletišče slovenskih slovnic in pravopisov od 1584 do danes*. http://www.fran.si/slovnice-in-pravopisi/30/18741870-anonimni.

Bajt, Drago, 1994: *Pišem, torej sem. Priročnik za pisanje*. Drugi natis. Maribor: Založba Obzorja.

Baron, Naomi S., 1998: Letters by Phone or Speech by Other Means: The Linguistics of Email. *Language & Communication* 2. 133–70.

Baron, Naomi S., 2002a: "Whatever.": A New Language Model?. Predavanje na srečanju Modern Language Association. 27.–30. december. New York.

Baron, Naomi S., 2002b: Who Sets Email Style? Prescriptivism, Capting strategies, and Democrating Communication Acces. *The Information Society* 18. Prispevek na konferenci Internet Research 1.0: The State of Discipline. Kansas. 14.–17. September 2000.

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

Crystal, David, 2005: *How language works. How babies babble, words change meaning and languages live or die*. New York, London: Penguin.

Crystal, David, 2011: *Internet linguistics. A student guide.* New York: Routledge.

Derks, Daantje in Arnold Bakker, 2010: The Impact of E-mail Communication on Organizational Life. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* 4/1. https://cyberpsychology.eu/article/view/4233/3277

Dobrovoljc, Helena, 2008: Jezik v e-poštnih sporočilih in vprašanja sodobne norma- tivistike. *Slovenščina med kulturami*. Košuta, Miran (ur.): Celovec: Slavistično društvo Slovenije. 197–210.

Dobrovoljc, Helena, 2011: Oblikovanje konceptov vljudnosti na izrazni ravni jezika v Japljevem prevodu posvetila cesarici Mariji Tereziji. Jesenšek, Marko (ur.). *Globinska moč besede*. Maribor: Mednarodna založba Oddelka za slovanske jezike in književnosti, Filozofska fakulteta. 197–210.

Dobrovoljc, Helena, 2014: Ujemanje pri nagovorih: spoštovani ... Jezikovna svetovalnica. http://svetovalnica.zrc-sazu.si/

Dobrovoljc, Helena, 2016: Nagovor oz. ogovorni izrazi pri pisanju pisem. Jezikovna svetovalnica. http://svetovalnica.zrc-sazu.si/

17 Na to opozarja pregled nagovorov v pisemskih pogovorih urednikov na Wikipediji, ki jih prinaša tudi korpus Janes, pri katerih ne gre za družabni klepet, temveč za stvarna opozorila s pogosto neprijetno vsebino za naslovnika, zato pisci večjo pozornost namenjajo spoštljivemu nagovoru, s katerim lažje ustvarijo distanco med piscem, naslovnikom in vsebino pisma.

Dobrovoljc, Helena, 2018: O rabi moških in ženskih oblik ter nezaznamovanosti moškega spola. Jezikovna svetovalnica. http://svetovalnica.zrc-sazu.si/

Eckert, Penelope in Sally McConnell-Ginet, 2003: *Language and Gender*. Cambridge: Cambridge University Press.

Erjavec, Tomaž in Darja Fišer, 2013: Jezik slovenskih tvitov: Korpusna raziskava. *Družbena funkcijskost jezika: Vidiki, merila, opredelitve*. Žele, Andreja (ur.): Ljubljana: ZIFF. 109–16.

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

Foucault, Michel, 2001: *Arheologija vednosti*. Ljubljana: Studia humanitatiS.

Jakop, Nataša, 2008: Pravopis in spletni forumi: kva dogaja? Košuta, Miran (ur.): *Slovenščina med kultura- mi*. Celovec: Slavistično društvo Slovenije. 210–19.

Jarnovič, Urška, 2007: Diskurzivne značilnosti sms-ov. *Jezik in slovstvo* 52/2. 61–79.

Jelovšek, Alenka, 2015: Vikanje pri nagovoru ženske osebe in Vikanje pri nagovoru dveh oseb. Jezikovna svetovalnica. http://svetovalnica.zrc-sazu.si/

Jezikovna svetovalnica Inštituta za slovenski jezik Frana Ramovša. https://svetovalnica.zrc-sazu.si

Kalin Golob, Monika, 2003: *O dopisih. (Kaj moram vedeti).* Ljubljana: GV Založba.

Kalin Golob, Monika, 2015: Pravopis v poslovno-uradovalnih besedilih. Dobrovoljc, Helena in Tina Verovnik (ur.): *Pravopisna razpotja: razprave o pravopisnih vprašanjih*. 1. izd. Ljubljana: Založba ZRC. 213–220.

Kalin Golob, Monika in Karmen Erjavec, 2014: Vrednotenje in pripadnost virtualni skupnosti slovenskega Twitterja. *Slavistična revija* 62/2. 217–233.

Končnik, Peter, 1881: *Slovenska slovnica z naukom, kako se pišejo pisma in opravilni sestavki*. https://www.dlib.si/details/URN:NBN:SI:DOC-IN88TM1E

Korpus Janes, Janes v0.4 Wikipedia. Jezikoslovna analiza nestandardne slovenščine. http://nl.ijs.si/noske/janes10.cgi/corp_info?corpname=janes_wiki

Krek, Simon, 2014: Papirnata slovenščina. *Delo*. 7. 2. 2014. http://www.delo.si/mnenja/gostujoce-pero/papirnata-slovenscina.html

Krstič, Adriana, 1993: Besedilnotipske značilnosti kratke zgodbe Vso pot do Pulsnitza Mihe Mazzinija. *Slavistična revija* 41/3. 359–380.

Makarovič, Gorazd, 2013: O pomenih besednih pozdravov. *Etnolog* 23. 205–236.

Michelizza, Mija, 2014: *Spletna besedila in jezik na spletu: primer blogov in Wikipedije*. Ljubljana: Založba ZRC.

Montego-Fleta, Begoña, Montesinos-López, Anna, Pérez-Sabater, Carmen, Turney, Ed, 2008: Computer mediated communication and informalization of discourse: The influence of culture and subject matter. *Journal of Pragmatics* 4. 770–779.

Pop, Anamaria Mirabela, 2012: *Stylistic Features of Business E-mails*. www.theroundtable.ro

Radicati, Sara, 2010, 2017: *Email Statistics Report*. Palo Alto: The Radicati Group. http://www.radicati.com/wp/wp-content/uploads/2010/04/Email-Statistics-Report-2010-2014-Executive-Summary2.pdf

Toporišič, Jože, 1965: *Slovenski jezik* 1–4 (1965). Maribor: Založba Obzorja.

Toporišič, Jože, 1992: *Enciklopedija slovenskega jezika*. Ljubljana: Cankarjeva založba.

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

Trdina, Silva, 1965: *Besedna umetnost*. Drugi del. Literarna teorija. Ljubljana: Mladinska knjiga.

Priloga: Vprašalnik o naslavljanju

Pri pisanju zasebnih in uradnih elektronskih pisem opažamo nove načine naslavljanja in pozdravljanja, ki jih v dosedanjih priročnikih za učenje sloga ni bilo mogoče zaslediti.

Pridevniške oblike *dragi X*, *draga* X, *spoštovana X, spoštovani X* namreč nadomeščajo drugi nagovorni načini, npr.

- *pozdravljeni X*,
- *X, pozdravljeni* (če osebo vikamo) in
- *X, pozdravljen/pozdravljena* (če osebo tikamo).

Pogosto uporabimo zgolj ime (*X*, …..) ali pa se pozdrav tudi opušča.

Pri neformalnih pismih uporabimo pogosto pogovorne izraze *Zdravo*, *Živijo*, *Oj*, *Hoj* … in podobno.

****

Da bi lažje ugotovili, kakšne so današnje navade pri naslavljanju v e-pismih, vas prosim, da si vzamete nekaj minut časa in odgovorite na naslednja vprašanja. Hvala!

Spol: **M** **Ž** (obkrožite)

Letnica rojstva:

Približno število poslanih e-pisem na dan:

Približno število prejetih e-pisem na dan:

1. Katere načine ogovora najpogosteje uporabljate pri **zasebni korespondenci**, torej če pišete prijateljem, sorodnikom …? Označite od 1 do 8.

ne uporabljam nagovora
uporabim le ime
dragi/draga X
spoštovani/spoštovana X
pozdravljeni/pozdravljen/pozdravljena X
živijo/živjo
zdravo
drugo

2. Katere načine ogovora najpogosteje uporabljate pri **nezasebni korespondenci** (služba, šola …)? Označite od 1 do 8.

ne uporabljam nagovora
uporabim le ime
dragi/draga X
spoštovani/spoštovana X
pozdravljeni/pozdravljen/pozdravljena X
živijo/živjo
zdravo
drugo _____

3. Katere načine ogovora najpogosteje uporabljajo vaši **zasebni** dopisovalci**?** Označite od 1 do 8.

dobivam večinoma pošto brez ogovora
uporabijo le ime
dragi/draga X
spoštovani/spoštovana X
pozdravljeni/pozdravljen/pozdravljena X
živijo/živjo
zdravo
drugo _____

4. Katere načine ogovora najpogosteje uporabljajo vaši nezasebni (poslovni) dopisovalci? Označite od 1 do 8.

dobivam večinoma pošto brez ogovora
uporabijo le ime
dragi/draga X
spoštovani/spoštovana X
pozdravljeni/pozdravljen/pozdravljena X
živijo/živjo
zdravo
drugo _____

5. Ali je nevljudno, če dobite pismo brez ogovora?

a) da

b) ne

6. Ali se vam zdi, da je za pisma brez ogovora kak poseben razlog? Prosim, navedite, če imate mnenje.

a) da

b) ne

7. Se vam zdi nevljudno, če prejmete pismo, kjer je namesto ogovora le vaše ime?

a) da

b) ne

8. Izberite trditve, ki so najbliže vašemu stališču,

a) izbira ogovora se mi ne zdi povezana z vljudnostjo

b) izbira ogovora je povezana z vljudnostjo

c) izbira ogovora zahteva preveč časa, zato ogovor izpuščam

d) noben od uradno priporočenih ogovorov (*dragi*, *spoštovani*) ne ustreza mojemu odnosu do naslovnika, zato ga kar izpuščam

e) uporabim tisti ogovor, ki ga navadno uporablja moj dopisovalec

f) uporabim le ime

g) ogovor pojmujem zgolj kot obrazec, ne razmišljam o pomenu besed

9. Pri učenju klasičnega pisemskega sporazumevanja učbeniki navajajo dve obliki: *dragi*/*draga* in *spoštovani*/*spoštovana*. Ali ju uporabljate tudi pri elektronski korespondenci?

a) da

b) ne

Če je odgovor DA: V katerih okoliščinah? ….

Če je odgovor NE: Zakaj ne? **…..**

10. Vse bolj se uveljavlja polformalna oblika ogovora – *pozdravljeni* (vikanje), *pozdravljen*, *pozdravljena*. Kaj menite o tej obliki glede na uveljavljeni *dragi/draga* in *spoštovani/spoštovana*? Izberite trditve ali dopišite svoje mnenje.

a) oblike ne uporabljam

b) oblika me moti, saj se z njo poslovim, ne pozdravim

c) oblike še nisem zasledil/zasledila

d) dobivam pisma s takim ogovorom, a ga ne uporabljam

e) obliko uporabljam, ker …

f) drugo: …

g) nimam mnenja

Napovedovanje spola slovenskih blogerk in blogerjev

Iza Škrjanec, Nada Lavrač, Senja Pollak

Izvleček

Napovedovanje spola avtorjev je zanimiv raziskovalni problem, izdelava napovednih modelov za razpoznavo spola pa je koristna za uporabo v različnih aplikacijah, npr. na področju trženja in analize kupcev. Cilj naše raziskave je razvoj in evalvacija napovednih modelov za avtomatsko razpoznavo spola avtorjev in avtoric slovenskih blogovskih zapisov. Za to nalogo uporabimo množico blogov 177 blogerjev in 96 blogerk, kot posamezno enoto klasifikacije pa upoštevamo vsa besedila posameznega avtorja v korpusu, združena v eno enoto. V prispevku primerjamo dva tipa modelov za napovedovanje spola avtorja besedil: model z ročno zgrajenimi pravili in model, zgrajen z metodami strojnega učenja. Modeli s pravili upoštevajo rabo slovničnega spola v delih besedila, v katerih se avtor nanaša nase. Za izgradnjo modelov s strojnim učenjem pa smo preizkusili več algoritmov in tipov značilk. Oba tipa modelov dosežeta klasifikacijsko točnost nad 85 %, najuspešnejši pa je model strojnega učenja, naučen na unigramih pojavnic s pomočjo metode podpornih vektorjev. Analiza najbolj informativnih značilk tega modela je pokazala, da se besedila blogerk in blogerjev razlikujejo na slovnični ravni (raba slovničnega spola in zaimkov), izbiri teme besedila (npr. večji poudarek na družini, ljubezni in spolnosti pri blogerkah) in slogovnih značilnostih (npr. raba kletvic v besedilih blogerjev).

Ključne besede: profiliranje avtorjev, spol, družbeni mediji, klasifikacija blogov

1 UVOD

Jezik je družbeni pojav in kot tak podvržen variaciji in spremembam. Med družbenimi dejavniki variacije sociolingvisti preučujejo tudi spol govorcev glede na jezikovne prakse, ki jih uporabljajo ženske in moški. Napredek v obdelavi naravnega jezika raziskovalcem omogoča, da modelirajo jezikovno variacijo v obsežnih besedilnih korpusih in z uporabo avtomatskih pristopov.

V prispevku primerjamo jezik moških in žensk s pristopom gradnje modelov za avtomatsko razpoznavo spola avtorjev besedil. Za to uporabimo podkorpus slovenskih blogovskih zapisov, ki so bili v okviru raziskovalnega projekta JANES zbrani in ročno označeni s podatkom o spolu avtorja.

Profiliranje avtorjev besedil glede na njihov spol je aktualen raziskovalni problem. Ena prvih odmevnih raziskav s tega področja je primerjala jezik govork in govorcev na podlagi Britanskega nacionalnega korpusa (BNC) in pokazala, da je v slogu pisanja žensk več poudarka na medosebni komunikaciji (npr. z rabo zaimkov), medtem ko je za moške med drugim bolj značilna raba členkov, njihov slog pisanja pa je informativne narave (Koppel et al. 2002). Prve raziskave avtomatske razpoznave spola avtorja so analizirale besedila v angleščini, kmalu pa so bili v področje raziskav vključeni tudi drugi jeziki, še posebej na podlagi virov iz družbenih medijev (Schler et al. 2006, Schwartz et al. 2013, Plank in Hovy: 2015, Peersman et al. 2011, Nguyen et al. 2013, Verhoeven et al. 2016, Ljubešić et al. 2017). Profiliranje avtorjev na podlagi spola je tudi ena od kategorij v sklopu vsakoletnega tekmovanja PAN (prim. Rangel et al. 2017).

Med jezike, ki jih pokrivajo raziskave iz računalniške stilometrije in profiliranja avtorjev, spada tudi slovenščina. Zwitter Vitez (2011, 2013) je predstavila prvo raziskavo o ugotavljanju avtorstva, in sicer je identificirala najverjetnejšega avtorja anonimnega besedila, ki je bilo objavljeno na uradni spletni strani ene od slovenskih parlamentarnih strank. Anonimno besedilo je primerjala z besedili potencialnih avtorjev s pomočjo vrste leksikalnih in berljivostnih značilk. Raziskovalni projekt JANES je prinesel nove možnosti za profiliranje avtorjev spletnih uporabniških vsebin, saj je korpus Janes opremljen z metapodatki o spolu in vrsti računa uporabnikov in uporabnic (Erjavec et al. 2018). V podkorpusu tvitov in blogov so ti metapodatki pripisani ročno. Verhoeven et al. (2017) so razvili model za identifikacijo spola avtorjev tvitov iz korpusa Janes in rezultate primerjali s podobnimi modeli, ki so bili naučeni na nemških, nizozemskih, italijanskih, španskih, francoskih in portugalskih tvitih (Verhoeven et al. 2016). Martinc et al. (2017) so razvili klasifikatorje za profiliranje avtorjev besedil v različnih jezikih, razviti modeli za določanje spola avtorjev tvitov pa so, skupaj z drugimi orodji za procesiranje nestandardne slovenščine, na voljo tudi v obliki spletnih delotokov

(Martinc et al. 2018). Z vidika spola avtorja pa sta bila podkorpusa tvitov in blogov analizirana v magistrskem delu Škrjanec (2017), na katerem je osnovano to poglavje.

V pričujočem poglavju uporabimo metode, s katerimi raziščemo, ali lahko na podlagi jezikovne variacije med spoloma avtomatsko razlikujemo med avtoricami in avtorji blogovskih zapisov. V poglavju predstavimo dva tipa napovednih modelov za določanje spola avtorja, in sicer model na podlagi ročno zgrajenih pravili in model, zgrajen z metodami strojnega učenja. Modeli s pravili upoštevajo rabo slovničnega spola v primerih avtorjevega samonanašanja v glagolskih zvezah, torej v sestavljenih glagolskih zvezah, pri katerih je pomožni glagol v prvi osebi ednine, deležnik na -l pa ima žensko ali moško obliko. Ti modeli predstavljajo osnovo za primerjavo s kompleksnejšimi in časovno bolj zahtevnimi modeli strojnega učenja, pri gradnji katerih smo eksperimentirali z različnimi značilkami in algoritmi.

Poglavje vsebuje sledeče razdelke. V drugem razdelku predstavimo korpus blogov, ki smo ga uporabili za analizo jezikovne variacije. Tretji razdelek opisuje metodologijo za gradnjo napovednih modelov. Četrti razdelek poda opis rezultatov klasifikacije, v petem pa se posvetimo napakam modela s pravili ter analiziramo tiste značilke izbranega modela strojnega učenja, ki imajo večjo težo pri klasifikaciji spola avtorja dokumenta. Šesti razdelek opiše sklepne ugotovitve in poda nekaj idej za prihodnje delo.

2 KORPUS BLOGOV

V raziskavi smo uporabili podkorpus blogov Janes-Blog (Erjavec et al. 2018), ki vsebuje bloge s portalov publishwall.si (18.515 besedil 615 uporabnikov) in rtvslo.si (23.515 blogov 243 uporabnikov), objavljene med oktobrom 2006 in januarjem 2016. Kot je podrobno opisano v Erjavec et al. (2018), je korpus bogato jezikoslovno označen, opremljen pa je tudi s številnimi dragocenimi metapodatki o besedilih (stopnja jezikovne in tehnične standardnosti, sentiment in jezik besedila) in njihovih avtorjih (tip uporabniškega računa, spol).

Glede na to, da smo želeli v eksperimentih izvesti binarno klasifikacijo avtorjev in avtoric, smo v analizo vključili le zasebne račune, ki imajo pripisano oznako ženskega ali moškega spola, torej smo izpustili korporativne račune in uporabnike z nedoločenim spolom. Poleg tega smo upoštevali le tiste uporabnike, ki so objavili vsaj 10 blogovskih zapisov v slovenščini. Končni podkorpus za eksperimente vsebuje 28.697 blogovskih zapisov skupno 273 avtorjev (od tega 64,84 % moških in 35,16 % žensk), kot prikazuje Tabela 1. Vsa slovenska besedila posameznega avtorja smo združili v en dokument, s čimer se uvrščamo med pristope profiliranja

na ravni uporabnika (Stamatatos 2009), ki ga uporabljajo tudi pri zasnovi tekmovanja PAN (Rangel et al. 2017); pri alternativnem pristopu pa se določa spol za vsako posamezno besedilo. Za eksperimente smo uporabili tako nelematizirana kot lematizirana besedila.

Tabela 1: Velikost učne množice blogov.

	Uporabniki	Blogovski zapisi	Pojavnice
Ženske	157	9.056	3.393.315
Moški	275	20.105	8.362.668
Ženske (10 ≥ besedil)	96	8.874	3.124.734
Moški (10 ≥ besedil)	177	19.823	6.968.164

3 METODOLOGIJA

Problem napovedovanja spola avtorja smo formulirali kot problem klasifikacije besedil, ki smo ga naslovili z gradnjo dveh tipov napovednih modelov: model napovedovanja s pravili in model z algoritmi strojnega učenja. Z vidika klasifikacije je naloga posameznemu avtorju pripisati razred, to je moški oz. ženski spol. V tem razdelku predstavimo njuno gradnjo in delovanje. Prvi pristop zahteva jezikovno znanje za določanje pravil, vendar pa je pristop z vidika razvoja in uporabe hitrejši. Za strojno učenje je potrebna velika množica ročno označenih dokumentov, ne potrebujemo pa nobenega jezikovnega znanja, saj se algoritmi značilnosti naučijo sami na podlagi primerov. V našem primeru modeli z ročno zgrajenimi pravili temeljijo na izražanju slovnično spola v samonanašanju, v modelih, naučenih s strojnim učenjem, pa so potencialno odločujoče vse uporabljene besede, model pa sam na podlagi učnih primerov pripiše težo posameznim značilkam (besedam, znakom oz. njihovi kombinaciji).

3.1 Modeli s pravili

Modeli s pravili uporabljajo ročno zgrajena klasifikacijska pravila za pripisovanje razreda avtorjem. Klasifikacijska pravila upoštevajo referenčni spol, o katerem sklepamo na podlagi rabe slovničnega spola v glagolskih zvezah. Referenčni spol je odvisen od tega, na koga se določena jezikovna oblika (npr. samostalnik ali zaimek) nanaša izven besedila (Motschenbacher 2010). Modeli s pravili so zgrajeni pod poenostavljeno predpostavko, da raba slovničnega spola odraža referenčni spol in da se slednji ujema s spolom avtorja. Model pri tem upošteva dele

besedila, v katerih se avtor nanaša sam nase. Tako raba ženskega spola v samonanašanju implicira, da gre za avtorico, medtem ko na podlagi moškega spola model predvideva, da gre za avtorja moškega spola.[1]

Modeli s pravili preverjajo rabo spola v samonanašalnih glagolskih zvezah. Upoštevajo rabo spola v glagolskih deležnikih na -l, in sicer pravila poiščejo pojavitve oblik pomožnega glagola: *sem*, *nisem*, *bom* in oblik z nestandardnim črkovanjem *sm* in *nism*. Razviti program v besedilu preveri, ali besede v okolici teh pomožnih glagolov nosijo oznako za spol, pri čemer pregleda okolico v velikosti dva (torej zadnji dve besedi tik pred pomožnim glagolom in prvi dve besedi po pojavitvi pomožnega glagola). Zanima nas končnica okoliških besed. Za vsakega avtorja izračunamo indikator za ženski ali moški spol na podlagi končnic okoliških besed: končnica *-la* signalizira žensko obliko deležnika na -l (npr. *naredi**la** sem*), medtem ko končnice *-al*, *-il* in *-el* nakazujejo moško obliko deležnikov (npr. *bom še vid**el***). Vsaka tovrstna pojavitev doda eno točko indikatorju za ženski oz. moški spol posameznega avtorja. Na koncu primerjamo vrednosti obeh indikatorjev: če večina najdenih indikatorjev (70 % ali več) spada k enemu razredu (ženski ali moški), pravila avtorju pripišejo ta razred. Poleg tega smo določili minimalno število indikatorjev na avtorja in pri tem primerjali uspešnost modela, če nastavimo minimalni prag na tri ali pet indikatorjev. V primeru, da besedila avtorja vsebujejo premalo indikatorjev ali pa je razmerje med obema razredoma preveč enakovredno, model avtorju pripiše razred *nedoločeno*.

Fišer et al. (2016) so opisali metodo, s katero so avtomatsko označili avtorje v korpusu Janes z oznako za spol. Tudi njihova metoda je osnovana na pravilih glede na rabo slovničnega spola, vendar so avtorja za razliko od našega pristopa, ki uporablja besedne oblike, uporabili oblikoskladenjske oznake, s katerimi so poiskali povedi, ki vsebujejo pomožni glagol in deležnik na -l. Podobno kot mi so tudi oni šteli indikatorje za ženski in moški spol. Če je bilo razmerje enih do drugih večje od 0,7, je model pripisal žensko oz. moško oznako, sicer je bil uporabnik uvrščen v razred nedoločenih. Poleg tega so dodali pogoj, da mora vsaj 1 % besedil uporabnika vsebovati indikatorje za spol, sicer je bil uporabnik uvrščen med nedoločene. S to metodo so uspešno označili 78 % blogerk in blogerjev. Na ta način so označili vse blogerje v korpusu, mi pa smo uporabili podmnožico blogerjev in blogerk (glej razdelek 2 in Tabelo 1). Njihov model prav tako ne vključuje pogoja glede števila indikatorjev v vseh besedilih. Zaradi teh razlik rezultati med našim modelom in modelom v Fišer et al. (2016) niso popolnoma primerljivi, vendar nas kljub temu zanima, kateri od obeh modelov z ročno grajenimi pravili je uspešnejši in primernejši za avtomatsko napovedovanje spola blogerjev.

1 Potrebno je poudariti, da za namene gradnje klasifikacijskih modelov predvidevamo, da je spol avtorja binaren razred (ženski ali moški). Vprašanje spolne identitete in samoidentifikacije posameznih avtorjev je relevantno na področju profiliranja avtorjev, vendar presega obseg tega poglavja.

3.2 Pristop z metodami strojnega učenja

V tem razdelku opišemo gradnjo klasifikacijskih modelov za določanje spola avtorja. Predstavimo pripravo besedil in eksperimente s tremi različnimi algoritmi strojnega učenja. Za gradnjo modelov in luščenje značilk smo uporabili knjižnico Scikit-learn (Pedregosa et al. 2011).

3.2.1 Priprava podatkov in učenje modelov

V enoto za klasifikacijo smo združili vse posamezne blogovske zapise enega avtorja v en skupni dokument, če so ti ustrezali kriterijem števila dokumentov na avtorja (glej razdelek 2). Vsako enoto smo predstavili v obliki vektorja značilk. Za model predstavitve smo izbrali vrečo besed (angl. *bag-of-words* ali BOW), v katerih enota besedila (npr. beseda, znak ali n-gram) predstavlja eno značilko. V eksperimentih smo kot značilke preizkusili posamezne besede ter besedne in znakovne n-grame (n-gram je enota n zaporednih besed ali znakov). Za primer vzemimo poved »*To pa je bila top sprostitev.*«, iz katere lahko zgradimo naslednje besedne n-grame dolžine ena (unigrame): »*To*«, »*pa*«, »*je*«, »*bila*«, »*top*«, »*sprostitev*«, ».«. Iz iste povedi dobimo naslednje besedne bigrame: »*To pa*«, »*pa je*«, »*je bila*«, »*bila top*«, »*top sprostitev*«, »*sprostitev.*«. Na podoben način, kot delimo poved na besedne n-grame, lahko besede razdelimo na znakovne n-grame. Iz besede »*bila*« tako lahko dobimo štiri znakovne unigrame (»*b*«, »*i*«, »*l*« in »*a*«) ali tri znakovne bigrame (»*bi*«, »*il*« in »*la*«). V svojih poskusih smo uporabili besedne uni- in bigrame ter znakovne tri- in tetragrame.

Za primerjavo lahko uporabimo različne metode za uteževanje značilk, npr. pogostost besede v posameznem dokumentu (angl. *term frequency*), binarno utež prisotnosti besede v besedilu ali pa mero TF-IDF (angl. *term frequency–inverse document frequency*). Mera TF-IDF upošteva pogostost besede v dokumentu in inverzno pogostost besede v celotnem korpusu, s čimer mera pripiše manjšo utež besedam, ki so pogoste v celotnem korpusu, ter večjo mero besedam, ki so bolj značilne za nekatere, ne pa vse dokumente (Kobayashi 2007). Z mero TF-IDF bi lahko uteži tako besedne kot tudi znakovne n-grame.

Za učenje modela, ki klasificira besedila glede na spol avtorja, smo testirali in med seboj primerjali tri algoritme strojnega učenja: naivni Bayesov klasifikator, logistično regresijo in metodo podpornih vektorjev (angl. *support vector machine*, v nadaljevanju SVM).

Vektorji značilk, ki predstavljajo besedila, so lahko precej veliki. Da bi zmanjšali velikost prostora značilk, prihranili čas za pripravo podatkov in morda izboljšali

točnost klasifikacijskega modela, pogosto uporabimo katero od metod za izbor značilk (Dhillon 2004). Pred učenjem modela smo število značilk najprej zmanjšali glede na število pojavitev in tako ohranili le tiste značilke, ki se pojavijo v besedilih najmanj 5 in največ 218 (80 %) blogerjev. Poleg tega smo uporabili metodo *SelectFromModel*[2] iz knjižnice Scikit-learn. Med učenjem algoritem vsaki značilki pripiše določen koeficient, metoda *SelectFromModel* pa iz prostora značilk odstrani tiste značilke, pri katerih je vrednost koeficienta nižja od določene vrednosti, pri čemer smo za učenje uporabili samodejne nastavitve praga. Pri modelih, naučenih z Bayesovim algoritmom, je ta koeficient logaritem verjetnosti značilke glede na razred. Modeli na podlagi logistične regresije in metode podpornih vektorjev pa koeficient pripišejo s pomočjo odločitvene funkcije.

Zgrajene klasifikacijske modele primerjamo glede na klasifikacijsko točnost. Ko gradimo napovedni model, nas zanima, kako uspešno lahko model napove razred na neznanih primerih, ki jih nismo uporabili v procesu učenja. Uspešnost modela ocenimo z uporabo prečnega preverjanja (Witten in Frank 2005). Pri tem postopku razdelimo učno množico na k delov, od katerih k-1 dele uporabimo za učenje, del, ki smo ga izpustili, pa uporabimo za testiranje. Proces ponovimo k-krat. V poskusih modele ocenimo in primerjamo z 10-kratnim prečnim preverjanjem in poročamo o aritmetični sredini in standardnem odklonu klasifikacijskih točnosti.

4 REZULTATI NAPOVEDOVANJA SPOLA BLOGERJEV

V tem razdelku predstavimo uspešnost za napovedovanje spola blogerjev, in sicer najprej poročamo o točnosti modelov s pravili, nato pa se osredotočimo na klasifikacijske modele strojnega učenja.

4.1 Modeli s pravili

Modeli s klasifikacijskimi pravili na podlagi spola deležnikov na -l klasificirajo blogerje v enega od treh razredov: *ženski*, *moški* ali *nedoločeno*. Tabela 2 predstavi klasifikacijsko točnost modela s pravili glede na to, katere oblike (standardne ali nestandardne skupaj z nestandardnimi) pomožnega glagola smo uporabili, da smo prepoznali glagolsko zvezo. Klasifikacijsko točnost smo izračunali tako, da smo preverili, koliko blogerjev in blogerk je model pravilno klasificiral glede na ročno oznako, avtorje, ki jih je model uvrstil v razred nedoločenih, pa razumemo kot nepravilno klasificirane primere. Tabela 2 vsebuje točnost glede na minimalno

2 Opis metode *SelectFromModel* je opisan na povezavi http://scikit-learn.org/stable/modules/feature_selection.html.

število indikatorjev. Kot lahko razberemo iz tabele, upoštevanje nestandardnega zapisa pomožnega glagola ne spremeni klasifikacijske točnosti. Najbolj točen je model, pri katerem mora imeti besedilo vsaj tri indikatorje za spol, točnost tega modela pa znaša 85,71 %. Za primerjavo tabela vključuje tudi rezultate za večinski klasifikator, torej rezultat, ki ga dobimo, če vsem dokumentom pripišemo večinski moški spol, ki ga najboljši pristop s pravili preseže za 21 %.

Tabela 2: Rezultati klasifikacije modela s pravili.

Minimalno št. indikatorjev	Pomožni glagol	Klasifikacijska točnost (%)
3	sem, nisem, bom	**85,71**
3	sem, nisem, bom, sm, nism	**85,71**
5	sem, nisem, bom	80,95
5	sem, nisem, bom, sm, nism	80,95
Večinski klasifikator		64,84

4.2 Modeli strojnega učenja

V tem razdelku predstavimo rezultate klasifikacije besedil z metodami strojnega učenja. Testirali smo tri različne algoritme (metoda podpornih vektorjev, logistična regresija in naivni Bayesov klasifikator). Preizkusili smo tudi več tipov značilk: besedne uni- in bigrame, znakovne bi-, tri- in tetragrame ter unijo teh značilk. V poskuse smo vključili tako nelematizirano kot lematizirano verzijo korpusa.

Tabela 3 prikazuje povprečje in standardni odklon točnosti pri 10-kratnem prečnem preverjanju. Tabela vključuje rezultate vseh treh algoritmov glede na različne značilke in obliko besedila (pojavnice pomenijo nelematizirano besedilo).

Tabela 3: Povprečna klasifikacijska točnost ± standardni odklon pri 10-kratnem prečnem preverjanju modela za napoved spola z uporabo različnih značilk, oblik besedila in treh algoritmov: metoda podpornih vektorjev (SVM), logistična regresija (LR) in naivnega Bayesov klasifikator (NB).

Značilke	Oblika	SVM (%)	LR (%)	NB (%)
besedni unigrami	pojavnica	**86,85 ± 6,00**	81,67 ± 5,89	64,89 ± 8,21
besedni uni- in birami	pojavnica	85,29 ± 8,34	79,81 ± 6,56	64,84 ± 6,29
znakovni (2-4)-grami	pojavnica	80,58 ± 8,17	78,76 ± 9,77	64,83 ± 6,25
besedni unigrami	lema	83,90 ± 7,82	80,57 ± 9,03	65,26 ± 10,89
besedni uni- in bigrami	lema	82,42 ± 7,12	78,02 ± 7,81	64,18 ± 1,65
Večinski klasifikator		64,85%		

Kot prikazuje Tabela 3, se je metoda podpornih vektorjev (SVM) odrezala bolje kot logistična regresija in Bayesov klasifikator. SVM doseže najvišjo klasifikacijsko točnost (86,85 % ± 6,00 %), ko uporabimo nelematizirane besedne unigrame kot značilke. SVM in logistična regresija presežeta večinski klasifikator v vseh preizkušenih kombinacijah z značilkami. V nasprotju pa Bayesov klasifikator preseže večinski klasifikator le v dveh poskusih, in sicer ko uporabimo nelematizirane ali lematizirane besedne unigrame kot značilke, vendar pa je razlika med rezultati minimalna, standardni odklon pa precej visok. Bayesov klasifikator se je torej v teh poskusih izkazal za manj uspešnega.

Če primerjamo uspešnost modelov glede na značilke, so se nelematizirani besedni n-grami izkazali kot bolj uporabni v primerjavi z lematiziranimi, kar velja tako za SVM kot za logistično regresijo. Oba omenjena algoritma dosežeta najvišjo točnost z nelematiziranimi besednimi unigrami (vrstica 1). Ko poleg besednih unigramov uporabimo še besedne bigrame, se točnost obeh modelov nekoliko zniža (za 1,56 % pri SVM-ju in za 1,86 % pri logistični regresiji), kar lahko pripišemo prevelikemu prileganju učni množici. SVM v poskusih z znakovnimi bi-, tri- in tetragrami doseže nižjo klasifikacijsko točnost kot v poskusih z besednimi n-grami. Znakovni n-grami se niso izkazali kot najboljši tudi za logistično regresijo, vendar pa so razlike v točnosti z besednimi n-grami manjše.

5 ANALIZA IN INTERPRETACIJA REZULTATOV KLASIFIKACIJE

V prejšnjem razdelku smo predstavili rezultate klasifikacije na podlagi pravil in klasifikacije z napovednimi modeli strojnega učenja glede na točnost napovedovanja spola. V razdelku 5.1 se najprej posvetimo modelu s pravili, in sicer analiziramo blogerje, ki jih je model klasificiral v napačni razred. Nato v razdelku 5.2 pregledamo značilke, ki so služile kot najbolj informativne za model strojnega učenja z najvišjo klasifikacijsko točnostjo, pri čemer nas zanimajo razlike in tudi podobnosti v besedilih blogerjev in blogerk.

5.1 Analiza napak modela s pravili

Model s pravili razvršča avtorje glede na rabo ženskega ali moškega spola v sestavljenih glagolskih zvezah v prvi osebi ednine. Kot prikazuje Tabela 2, dosežemo najvišjo točnost, ko je model manj strog in zahteva le tri indikatorje ženskega ali moškega spola v samonanašanju, da avtorja uvrsti v ženski ali moški razred (sicer

je bil avtor uvrščen v razred *nedoločeno*). Ta model uspešno klasificira 85,71 % avtorjev korpusa, Tabela 4 pa prikazuje razvrstitveno tabelo pravilnih in napačnih klasifikacij glede na ročne oznake. Na podlagi tabele lahko izračunamo, da je model pravilno klasificiral 79,17 % blogerk in 89,27 % blogerjev.

Tabela 4: Razvrstitvena tabela za najbolj uspešen model s pravili.

Napovedani razred	Dejanski razred	
	ženski	moški
ženski	76	2
moški	4	158
nedoločeno	16	17
skupaj	96	177

Najprej se osredotočimo na avtorje, ki jih je model uvrstil v razred nasprotnega spola. Kot vidimo v Tabeli 4, je model klasificiral štiri blogerke v moški razred. Pregledali smo njihova besedila; pri dveh od teh štirih blogerk je model našel zelo majhno število oznak za slovnični spol, vendar pa sta obe blogerki vseeno uporabili več moških kot ženskih oblik, saj sta objavili daljše zapise oz. citate v prvi osebi ednine moških avtorjev. Podobno velja za drugi dve blogerki, ki jih je model uvrstil v moški razred, saj je bila izmed več kot 40 oznak za spol večina moškega spola; tudi ti blogerki sta objavili več citatov in predvsem daljših leposlovnih zapisov v prvi osebi moškega spola. Izmed vseh blogerjev moškega spola so pravila razvrstila le dva blogerja v ženski razred. Po pregledu njunih besedil smo ugotovili, da je indikator za ženski spol večji od moškega predvsem zaradi pripovedi v prvi osebi ženskega spola in zaradi citatov ženskih oseb.

Tako za razred blogerjev kot blogerk velja, da je model avtorje največkrat napačno razvrstil v razred nedoločenih, saj je vanj uvrstil skoraj 17 % blogerk in skoraj 10 % blogerjev. Izmed 16 blogerk, ki jih je model klasificiral v razred nedoločeno, jih je šest vključevalo veliko število (nad 30) glagolskih zvez v prvi osebi ednine, vendar pa je število teh zvez v ženskem spolu skoraj izenačeno s številom zvez v moškem spolu. Po pregledu zapisov teh blogerk smo ugotovili, da so v besedila vključile veliko premega govora udeleženk in udeležencev dialoga. Nekateri izmed njihovih blogovskih zapisov pa so v celoti zapisani iz stališča pripovedovalca moškega spola.

Pri preostalih 10 blogerkah, ki jim spol ni bil pripisan (razred nedoločeno), model ni našel zadostnega števila (vsaj treh) indikatorjev za ženski ali moški spol v sestavljenih glagolskih zvezah v prvi osebi. To sicer še ne pomeni, da blogerke niso kako drugače izrazile svojega spola skozi besedne oblike. Tako lahko najdemo primere glagolskih zvez, ki jih naš klasifikator ni upošteval pri izračunu

indikatorja spola, saj med pomožnim glagolom in deležnikom leži več kot ena beseda, npr. v povedi: »Malo sem po naključju sledila.« Med ročnim pregledom smo našli tudi primere, v katerih je spol avtorice izražen v pridevnikih, npr.: »A v to sem prepričana.«

Pravila so v razred nedoločeno uvrstila 17 blogerjev. Štirje od njih so uporabili relativno veliko (40 ali več) tako ženskih kot moških oblik deležnikov na -l, in sicer predvsem v premem govoru. Klasifikator ostalim blogerjem spola ni pripisal in jih je uvrstil v razred nedoločenih, saj model ni našel več kot treh indikatorjev spola. Branje blogovskih zapisov teh avtorjev je pokazalo, da se v besedilih na splošno nase ne nanašajo pogosto, zato nismo našli prvoosebnih oblik niti v glagolih niti v pridevnikih.

5.2 Najbolj informativne značilke modela strojnega učenja

V tem razdelku analiziramo značilke, ki so bile uporabljene v klasifikacijskem modelu z najvišjo točnostjo. Kot smo pokazali v Tabeli 2, se je kot učni algoritem najbolje odrezal SVM z nelematiziranimi besednimi unigrami kot značilkami in tako dosegel klasifikacijsko točnost 86,85 %. Za interpretacijo vzamemo značilke, ki jim je klasifikator pripisal največje uteži in so bile torej najbolj informativne, da je klasifikator avtorju pripisal posamezni razred (spol). Iz klasifikacijskega modela jih izluščimo s funkcijo, ki vrača seznam značilk, ki imajo največje uteži za ženski oz. moški razred. Za ta prispevek smo s funkcijo izluščili 1.000 značilk za vsak razred, značilke smo na seznamu razvrstili po padajoči vrednosti uteži in vsakega od teh dveh seznamov nato analizirali ter primerjali med seboj.

Primerjanje vrednosti uteži nam lahko delno nudi vpogled v razlike v jezikovni rabi med ženskami in moškimi, saj lahko preverimo, katere značilke so bolj pomembne za blogerke in manj za blogerje ter obratno. Interpretacijo oz. pomembnost značilk za posamezni razred pa je treba vzeti z nekaj previdnosti, saj se moramo zavedati, da pri so pri algoritmu SVM značilke med seboj povezane in jih je težko interpretirati neodvisno od ostalih značilk, kar pa zaradi možnosti interpretacije v tem prispevku zanemarimo.

Na vrhu seznama najbolj informativnih značilk najdemo deležnike na -l v ženski (na seznamu ženskega razreda) oz. moški obliki (na seznamu moškega razreda). Za moški razred predstavljajo te oblike kar 15 % najbolj informativnih značilk, največje uteži pa imajo moške oblike bolj splošnih glagolov, npr. *šel, videl, dal*. Na seznamu ženskega razreda prav tako najdemo splošne glagole (npr. *imela, šla, vedela*, *dobila*), med 1.000 značilkami pa je deležnikov v ženski obliki 13 %. Na seznamih značilk se pojavljajo tudi druge besedne oblike, ki nakazujejo spol, in

sicer pridevniki, npr. *vesela* in *ponosna* na seznamu ženskega razreda ter *vesel* in *prepričan* na seznamu moškega razreda.

Poleg deležnikov in pridevnikov se na seznamih značilk pojavijo tudi druge besedne vrste. Pri obeh razredih imajo veliko utež različni prislovi, ki jih lahko razvrstimo med časovne, prostorske, številske ali modalne. V splošnem je na seznamu ženskega razreda več časovnih prislovov, še posebej takih, ki izražajo pogostost (npr. *znova, včasih, pogosto, nikdar*). Na seznamu moškega razreda najdemo več časovnih prislovov, ki so vezani na eno točko v času (npr. *nocoj, sinoči, včeraj*). Zanimive razlike med najbolj informativnimi značilkami obeh razredov se pojavijo v zaimkih. Med značilkami, informativnimi za ženski razred, so predvsem osebni in svojilni zaimki za prvo osebo ednine (npr. *moja/mojega/mojih, mene/zame/menoj*) in dvojine (npr. *naju/nama, najin*). Na seznamu moškega razreda je veliko manj zaimkov, najdemo pa osebne zaimke za prvo osebo množine (*naši/naše*), tretjo osebo ednine (npr. *njej, njegovega/njegov*) ali množine (*njihovi*).

Med prvimi stotimi značilkami z največjimi utežmi so predvsem deležniki ženskega oz. moškega spola, nato pa se na obeh seznamih začnejo pojavljati tudi samostalniki in lastna imena. Čeprav so na seznamu besedni unigrami, torej le posamezne besede brez konteksta, lahko na podlagi seznama sklepamo o temah, ki so bolj značilne za enega od razredov in manj za drugega. Med značilkami ženskega razreda so pogoste besede, ki zaznamujejo družino in družinske člane (npr. *otroci/otroka/otroke*, *mama/mami, očka, starši, sestro, družina, otroštvo*). Poleg tega na tem seznamu najdemo več besed, ki se nanašajo na romantične zveze in spolnost (npr. *spolnost, ljubček, zaljubljenost, seks*). Med značilkami, ki so bolj tipične za ženski razred, je besedišče, povezano s čustvi in občutki, ki so lahko pozitivni (*ljubezen/ljubezni, strasti, nasmeh*), najde pa se več primerov negativnih čustev (*otožnost, jokala, solze, samota, zavist, žalostna, sram, sramota*).

Poleg družinske in ljubezenske tematike lahko s seznama visoko uteženih značilk ženskega razreda sklepamo še o eni temi, ki je bolj priljubljena pri blogerkah kot blogerjih. Na seznamu ženskega razreda namreč najdemo več besed, povezanih s prehrano, kar nakazuje na to, da gre v besedilih za recepte ali nasvete glede prehranjevanja, npr. *cvetače, zelenjave, maslo, kokosovo, testo, penino, cimet*. Med manj vidnimi, vendar še vedno razlikovalnimi temami je tudi zdravje, saj na seznamu ženskega razreda najdemo več besed, ki se nanašajo na zdravstvene zadeve (npr. *zdravljenje, kemoterapija/kemoterapije, rakom/rak, zboleti, cepiv*).

Besedišče, ki je povezano s političnim dogajanjem, se pojavlja med najbolj informativnimi značilkami obeh razredov (npr. na seznamu ženskega razreda najdemo primere *demokratske, socialisti, isis*), vendar pa ta tema veliko bolj opredeljuje seznam značilk moškega razreda in je prevladujoča glede na ostale. Na seznamu moškega razreda najdemo besede, ki se nanašajo na državo, državne organe in

mehanizme (*država, sodišča, volitvah, vlade, referendum*), politične funkcije (*predsednik, državljanov*), Cerkev in različne politične in ekonomske ureditve (*demokracija, kapitalizem*). Seznam vključuje tudi besede, ki so povezane z Jugoslavijo (*udba*, *komunisti*, *jla*) in drugo svetovno vojno (*nob, belogardisti*). Na seznamu moškega razreda se v različnih oblikah pojavita tudi besedi *politika* in *politično* (*politično/političnega/politične/političnega*). Poleg obsežnega besedišča s področja politike je za moški razred zelo značilna športna tematika, saj se to besedišče pojavlja le pri moškem razredu (*ligi, žogo, prvenstvo, tekmo/tekem*).

Razlike med blogerkami in blogerji, o katerih sklepamo na podlagi seznamov najbolj informativnih značilk, se vezane tudi na slog in register pisanja. Na seznamu moškega razreda najdemo denimo primere vulgarizmov (*hudiča, jebo/jebe, scat*) in slabšalnih poimenovanj za manjšinsko skupino (*cigan/ciganov*).

Kot je bilo izpostavljeno, se moramo zavedati, da smo pri interpretaciji rezultatov zanemarili dejstvo, da so pri modelih dejansko značilke med seboj povezane. Tovrstna interpretacija nam omogoča delno razumevanje razlik med besedili, na podlagi besed, ki so bile za klasifikacijski model zelo pomembne. Za preverjanje relevantnosti razlik pa bi bili potrebni dodatni statistični testi.

6 SKLEP

V tem prispevku smo se lotili vprašanja avtomatske napovedi spola avtorja na podlagi besedil, s ciljem napovedati spol slovenskih blogerk in blogerjev, katerih zapisi so zbrani v korpusu spletnih uporabniških vsebin Janes. V prispevku predstavimo dva tipa napovednih modelov: model s pravili (ročno zgrajeni klasifikacijski modeli) in modele zgrajene z metodami strojnega učenja (avtomatsko zgrajeni klasifikacijski modeli).

Klasifikator na podlagi pravil je zasnovan tako, da spol avtorjem pripišemo glede na izražanje slovničnega spola z deležnikom v glagolskih zvezah v prvi osebi ednine. Klasifikator pripiše avtorjem ženski ali moški spol glede na število glagolskih zvez, ki izražajo določen spol oz. jih uvrsti v razred nedoločeno, če je premalo indikatorjev za katerega koli od spolov. V poskusih smo z najboljšim modelom s pravili uspešno klasificiral 85,71 % avtorjev v učni množici (glej Tabelo 2), pri čemer je presegel večinski klasifikator za okoli 21 %. Zanimivo je, da upoštevanje nestandardnih zapisov pomožnega glagola (*sm*, *nism*) ni spremenilo uspešnosti modela. Iz rezultatov lahko sklepamo, da je raba spola v glagolskih zvezah v primerih samonanašanja precej predvidljiva tudi v blogih, saj velika večina avtorjev uporablja dovolj prvoosebnih glagolskih oblik, pri katerih sta pomožni glagol in deležnik relativno blizu, pomožni glagol pa je zapisan na standardni način. Kljub

temu da je napoved spola avtorjev v delu Fišer et al. (2016) zastavljena nekoliko drugače, lahko ugotovimo, da klasifikacijska točnost našega modela s pravili ta model preseže za skoraj 8 %.

V razdelku 4.2 poročamo o uspešnosti modelov za napoved spola, ki smo jih zgradili s tremi različnimi algoritmi strojnega učenja (SVM, logistična regresija in naivni Bayesov klasifikator) ter njihove rezultate primerjali glede na uporabljene značilke. Modele smo testirali z 10-kratnim prečnim preverjanjem. Rezultati so pokazali, da se je najbolje odrezal SVM z uporabo besednih unigramov nelematiziranega besedila, saj je dosegel 86,85-% klasifikacijsko točnost (Tabela 3). Tako SVM preseže večinski klasifikator za 22 %, medtem ko je od modela s pravili boljši le za 1 %.

Avtomatsko zgrajeni model z najvišjo klasifikacijsko točnostjo smo ovrednotili tudi kvalitativno. Analizirali smo tiste značilke, ki jim je model pripisal največje uteži pri odločitvi za uvrstitev v ženski ali moški razred. Analiza je pokazala, da je med najbolj informativnimi značilkami veliko število takih, ki vsebujejo informacijo o spolu (deležniki na -l in pridevniki). Zanimive pa so razlike glede na druge besedne vrste, in sicer je med značilkami ženskega razreda več zaimkov. Podoben trend v besedilih avtoric je bil opažen tudi na primeru angleških sporočil na družbenem omrežju Facebook (Schwartz et al. 2013), angleških blogovskh zapisih (Schler et al. 2006), angleškega pisanega in govorjenega jezika (Newman et al. 2008), kot tudi na pilotni analizi slovenskih tvitov (Verhoeven et al. 2017).

Razlike v najbolj informativnih značilkah so vezane tudi na različne teme, ki so bolj značilne za avtorje enega spola in manj za drug spol. O osredotočanju na teme, kot sta družina in ljubezenski odnosi, torej na teme, ki zaznamujejo socialne procese, v povezavi z besedili avtoric poročajo tudi Schler et al. (2006), Newman et al. (2008) in Schwartz et al. (2013). Schler et al. (2006) in Schmid (2003) so podobno kot mi opazili, da moški v primerjavi z ženskami večkrat pišejo ali govorijo o politiki in športu. Blogovske zapise v korpusu Janes smo preučevali tudi v Škrjanec in Pollak (2016), kjer smo s pomočjo metode gručenja podatkov izdelali ontologije tematik in s tem identificirali prevladujoče teme, o katerih pišejo slovenski blogerji. Rezultati so pokazali, da tako blogerke kot blogerji pišejo o politiki, družini, romantičnih odnosih, okolju in prehrani; blogerji se s primerjavi z blogerkami več posvečajo temam o športu, glasbi, literaturi, Cerkvi, begunski krizi in temam o naravi. V zapisih blogerk pa je več poudarka na religiji, socialni politiki in čustvih.

Raba vulgarizmov in kletvic naj bi bila bolj značilna za jezik moških (Bamman et al. 2014, Newman et al. 2008, Schwartz et al. 2013), kar se je pokazalo tudi na našem seznamu značilk, vendar bi bila potrebna tudi podrobnejša analiza teh besed v kontekstu.

Glede na našo analizo in primerjavo dveh tipov napovednih modelov lahko zaključimo, da je za slovenske bloge pristop s pravili o rabi spola v glagolskih zvezah lahko enako uspešen kot bolj kompleksni modeli, zgrajeni z algoritmi strojnega učenja. Model s pravili ima to prednost, da ne potrebuje označenih podatkov za učenje, vendar pa se je treba zavedati, da lahko avtorji zavestno manipulirajo z rabo slovničnega spola v samonanašanju, da prikrijejo svoj spol. V tem pogledu bi bilo zanimivo in koristno preveriti, kako uspešni so modeli, iz katerih bi izključili besedne značilke, ki imajo referenčni spol (deležniki in pridevniki). Če stopimo korak dalje, Daelemans (2013) trdi, da bi morali biti modeli za profiliranje avtorjev neodvisni od žanra in teme besedila, za kar bi bilo potrebno ohraniti le slogovne značilke, odstraniti pa tiste, ki se nanašajo na razlikovalne teme.

Svojo raziskavo o avtomatski napovedi nameravamo razširiti na več načinov. Model s pravili je možno izboljšati tako, da poleg upoštevanja spola v glagolskih zvezah vključimo tudi pridevnike. Kompleksnejša naloga pa bi bila ugotavljanje citiranih delov besedila, ki so bili, kot je pokazala analiza napak, pogosto razlog za napačno klasifikacijo. Za boljšo uspešnost z metodami strojnega učenja pa bi potrebovali večjo učno množico, v nadaljnjih eksperimentih pa bomo upoštevali nove kombinacije besednih in znakovnih n-gramov ter alternativne metode rangiranja značilk (Guyon et al. 2002). Posebno pozornost bomo posvetili nadaljnji interpretaciji razlik med avtorji blogov glede na spol, kjer bomo izsledke predstavljene raziskave preverili s statističnimi testi, kar smo delno že obravnavali (Škrjanec 2017). Interpretacija informativnih značilk modela, naučenega na lematiziranih besedilih, nam bo omogočila, da več pozornosti namenimo značilkam, ki so neodvisne od izražanja slovničnega spola. Nenazadnje bomo modele preizkusili tudi na drugih žanrih spletnih uporabniških vsebin, ki jih vsebuje korpus Janes (forumi, komentarji novic in uporabniške in pogovorne strani na Wikipediji).

Literatura

Bamman, David, Jacob Eisenstein in Tyler Schnoebelen, 2014: Gender identity and lexical variation in social media. *Journal of Sociolinguistics* 1/2. 135–160.

Daelemans, Walter, 2013: Explanation in computational stylometry. *Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing* (CICLing'13). Vol 2. Berlin, Heidelberg: Springer. 451–462.

Dhillon, Inderjit, Kogan, Jacob in Nicholas, Charles, 2004: Feature selection and document clustering. Berry, Michael (ur.): *A Comprehensive Survey of Text Mining*. New York: Springer. 73–100.

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

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

Guyon, Isabelle, James Weston, Stephen Barnhill in Vladimir Vapnik, 2002: Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning* 46/1. 389–422.

Jurafsky, Dan in James H.Martin, 2009: *Speech and Language Processing, second edition*. Upper Saddle River, New Jersey: Prentice-Hall, Inc.

Kobayashi, Mei in Aono, Masaki, 2007: Vector space models for search and cluster mining. Berry, Michael W. in Malu Castellanos (ur.): *Survey of Text Mining: Clustering, Classification and Retrieval.* Springer. 109–127.

Koppel, Moshe, Shlomo Argamon in Anat R.Shimoni, 2002: Automatically categorizing written texts by author gender. *Literary and Linguistic Computing* 17/4. 401–412.

Ljubešić, Nikola in Tomaž Erjavec, 2016: Corpus vs. lexicon supervision in morphosyntactic tagging: The case of Slovene. *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016). Portorož, Slovenia: European Language Resources Association (ELRA). 1527–1531.

Ljubešić, Nikola, Fišer, Darja in Tomaž Erjavec, 2017: Language-independent gender prediction on Twitter. *Proceedings of NLP+CSS: Second Workshop on Natural Language Processing and Computational Social Science*. Vancouver, Kanada: ACL. 1–6.

Martinc, Matej, Iza Škrjanec, Katja Zupan in Senja Pollak, 2017: PAN 2017: author profiling - gender and language variety prediction. Cappellato, Linda, Nicola Ferro, Lorraine Goeuriot in Thomas Mandl (ur.): *Working notes papers of CLEF 2017 Conference and Labs of the Evaluation Forum*. CEUR Workshop Proceedings.

Martinc, Matej, Senja Pollak in Ana Zwitter Vitez, 2018: Delotoki za nadaljnje analize nestandardne slovenščine. Fišer, Darja (ur.): *Viri, orodja in metode za analizo spletne slovenščine*. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani. 16–43.

Motschenbacher, Heiko, 2010: *Language, Gender and Sexual Identity: Poststructuralist Perspectives*. Amsterdam: John Benjamins.

Nguyen, Dong, Rilana Gravel, Dolf Trieschnigg in Theo Meder, 2013: TweetGenie: Automatic age prediction from tweets. *ACM SIGWEB Newsletter* 4/4. 1–6.

Osrajnik, Eneja, Darja Fišer in Damjan Popič, , 2015: Primerjava rabe ekspresivnih ločil v tvitih slovenskih uporabnikov in uporabnic. *Zbornik konference Slovenščina na spletu in v novih medijih*. Ljubljana: Znanstvena založba Filozofske fakultete. 50–74.

Peersman, Claudia, Daelemans, Walter in Van Vaerenbergh, Leona, 2011: Predicting age and gender in online social networks. *Proceedings of the Third International Workshop on Search and Mining User-generated Contents*, ACM. 37–44.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay, 2011: Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12. 2825–2830.

Plank, Barbara and Hovy, Dirk, 2015: Personality Traits on Twitter or How to Get 1,500 Personality Tests in a Week. *Proceedings of the Sixth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. 92–98.

Rangel, Francisco, Paolo Rosso, Martin Potthast in Benno Stein, 2017: Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. Cappellato, Linda, Nicola Ferro, Lorraine Goeuriot in Thomas Mandl (ur.): *Working notes papers of CLEF 2017 Conference and Labs of the Evaluation Forum*. CEUR Workshop Proceedings.

Schler, Jonathan, Koppel, Moshe, Argamon, Shlomo, in Pennebaker, James, 2006: Effects of age and gender on blogging. *Proceedings of AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs* 6. 199–205.

Schwartz, H. Andrew, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, Martin E. P. Seligman in Lyle H. Ungar, 2013: Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one* 8(9).

Stamatatos, Efstathios, 2009: A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology* 60/3. 538–556.

Škrjanec, Iza, 2017: *Gender-based analysis of Slovene user-generated content*. Magistrsko delo. Ljubljana: Mednarodna podiplomska šola Jožefa Stefana.

Škrjanec, Iza in Senja Pollak, 2016: Topic ontologies of the Slovene blogosphere: A gender perspective. Fišer, Darja in Michael Beißwenger (ur.): *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities*, 27-28 September 2016, Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia. 62–65.

Verhoeven, Ben, Daelemans, Walter in Plank, Barbara, 2016: TwiSty: A multilingual Twitter stylometry corpus for gender and personality profiling. V *Proceedings of the 10th Language Resources and Evaluation Conference (LREC 2016)*. ELRA, Portorož, Slovenia.

Verhoeven, Ben, Iza Škrjanec in Senja Pollak, 2017: Gender Profiling for Slovene Twitter Communication: The Influence of Gender Marking, Content and Style. *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*. 119–125.

Zwitter Vitez, Ana, 2011: Povej mi karkoli in povem ti, kdo si: Ugotavljanje avtorstva besedil. Kranjc, Simona (ur.): *Obdobja 30: Meddisciplinarnost v slovenistiki*. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete. 565–570.
Zwitter Vitez, Ana, 2013: Le décryptage de l'auteur anonyme : l'affaire des électeurs en survêtements. *Linguistica* 53/1. 91–101.
Witten, Ian H. in Eibe Frank, 2005: *Data Mining: Practical Machine Learning Tools and Techniques.* San Francisco: Morgan Kaufmann.

Zahvala

Avtorice se zahvaljujemo recenzentom za koristne komentarje in predloge.

O avtorjih

Špela Arhar Holdt je znanstvena sodelavka na Centru za jezikovne tehnologije Univerze v Ljubljani (Filozofska fakulteta, Fakulteta za računalništvo in informatiko). Ukvarja se s korpusnim jezikoslovjem in uporabniškimi študijami. Sodelovala je pri pripravi vrste korpusnih virov za slovenščino, označevalnih sistemov in korpusnih vmesnikov. V raziskavah razvija in evalvira postopke za analizo slovenskih korpusov, zlasti za uporabo korpusnih podatkov na področjih (k uporabnikom usmerjenega) slovarskega ter slovničnega opisa, normativistike in jezikovne didaktike.

Jaka Čibej je doktorski študent na Oddelku za prevajalstvo Filozofske fakultete Univerze v Ljubljani. Kot raziskovalec se ukvarja s slovensko računalniško posredovano komunikacijo (zlasti s proučevanjem regionalnih jezikovnih različic v spletni slovenščini in razlik med računalniško posredovano komunikacijo ter pisno standardno slovenščino), z gradnjo korpusnih in drugih jezikovnih virov za slovenščino ter z uporabo množičenja in vključevanjem jezikovnih uporabnikov v gradnjo jezikovnih virov. Od leta 2018 kot projektni sodelavec deluje tudi na Institutu »Jožef Stefan« in na Centru za jezikovne vire in tehnologije Univerze v Ljubljani.

Helena Dobrovoljc je višja znanstvena sodelavka na Inštitutu za slovenski jezik Frana Ramovša pri Znanstvenoraziskovalnem centru SAZU in predavateljica sodobnega slovenskega knjižnega jezika na Fakulteti za humanistiko Univerze v Novi Gorici. Akademski naziv magisterij je pridobila na Oddelku za slovenistiko Filozofske fakultete Univerze v Ljubljani s pravopisno problematiko, doktorski naslov s področja teorije jezikovne naravnosti pa na Oddelku za splošno in primerjalno jezikoslovje iste fakultete. Kot raziskovalka se ukvarja s sodobno knjižno slovenščino in njeno normo, pravopisom, sociolingvistiko in slovaropisjem. Posveča se tudi normativnim in slovničnim zadregam jezikovnih uporabnikov in od leta 2012 moderira Jezikovno svetovalnico, ki deluje na Inštititu za slovenski jezik. Od leta 2012 je vodja skupine za pripravo novega

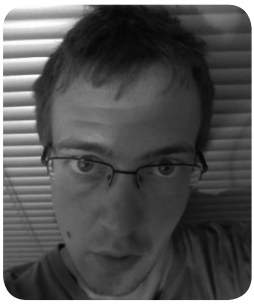
pravopisnega slovarja (ePravopis), od leta 2013 pa Predsednica pravopisne komisije pri SAZU in ZRC SAZU, ki pripravlja nova pravopisna pravila. Je urednica zbirke znanstvenih monografij z naslovom Lingua Slovenica, ki izhaja pri Založbi ZRC, in članica Komisije za standardizacijo zemljepisnih imen Vlade Republike Slovenije.

Tomaž Erjavec je zaposlen kot svetnik na Odseku za tehnologije znanja na Institutu »Jožef Stefan«, zaposlen pa je bil tudi na Univerzi v Edinburgu in Tokijski univerzi. Področja njegovega raziskovanja so jezikovne tehnologije in digitalna humanistika s poudarkom na izdelavi in označevanju jezikovnih virov slovenskega jezika. Na področjih jezikovnih tehnologij in korpusnega jezikoslovja je poučeval na Univerzi v Novi Gorici, na Univerzi v Gradcu in na Mednarodni podiplomski šoli Jožefa Stefana. Je ustanovni predsednik slovenskega Društva za jezikovne tehnologije, sodeluje pri izdelavi standardov za zapis jezikovnih virov pri SIST in ISO TC 37 in je nacionalni koordinator raziskovalne infrastrukture za jezikovne vire in tehnologije CLARIN.SI.

Darja Fišer je docentka in predstojnica katedre za leksikologijo, terminologijo in jezikovne tehnologije na Oddelku za prevajalstvo Filozofske fakultete Univerze v Ljubljani ter znanstvena sodelavka na Odseku za tehnologije znanja na Institutu »Jožef Stefan« v Ljubljani. Poučuje predmete, vezane na korpusno jezikoslovje in jezikovne tehnologije. Raziskovalno se trenutno ukvarja predvsem z računalniško posredovano komunikacijo in leksikalno semantiko z uporabo korpusnega jezikoslovja in procesiranja naravnega jezika. Je predsednica Slovenskega društva za jezikovne tehnologije, vodja usmerjevalnega odbora FoLLI za največjo evropsko poletno šolo jezika, logike in računalništva ESSLLI in direktorica za področje uporabnikov pri evropski raziskovalni infrastrukturi za jezikovne vire in tehnologije CLARIN.

Nada Lavrač je vodja Odseka za tehnologije znanja na Institutu »Jožef Stefan« v Ljubljani in profesorica na Univerzi v Novi Gorici ter na Mednarodni podiplomski šoli Jožefa Stefana, kjer je tudi predstojnica programa Informacijske in komunikacijske tehnologije. Osrednja tematika njenega raziskovalnega in pedagoškega dela je podatkovno rudarjenje in odkrivanje zakonitosti v podatkih, s poudarkom na relacijskem in semantičnem podatkovnem rudarjenju. Predmet njenih raziskav sta tudi analiza besedilnih podatkov in računalniška ustvarjalnost. Poleg vodenja programske skupine Tehnologije znanja je vodila vrsto slovenskih raziskovalnih projektov s področja podatkovnega rudarjenja ter sodelovala v vrsti evropskih projektov, vključno z več projekti s področja podatkovnega rudarjenja, računalniške ustvarjalnosti in analize medicinskih podatkov.

Nikola Ljubešić je docent na Oddelku za informacijsko in komunikacijsko znanost na Univerzi v Zagrebu in znanstveni sodelavec na Oddelku za tehnologije znanja na Institutu »Jožef Stefan« v Ljubljani. Raziskovalno se ukvarja predvsem s strojnim učenjem za leksikalno semantiko in analizo družabnih medijev, zaznavanjem semantičnih premikov, napovedovanjem medjezikovnih leksikalnih lastnosti, jezikovno analizo nestandardnih besedil, normalizacijo nestandardnih besedil, profiliranjem uporabnikov in zaznavanjem neprimernih vsebin na družabnih medijih. Poučuje obdelavo naravnega jezika in strojno učenje in je član Združenja za računalniško jezikoslovje (ACL), Slovenskega društva za jezikovne tehnologije in Hrvaškega društva za jezikovne tehnologije.



Matej Martinc je pridobil naziv univerzitetni diplomirani filozof in sociolog kulture leta 2011 in naziv diplomirani inženir računalništva in informatike leta 2015 na Univerzi v Ljubljani. Trenutno je doktorski študent na Mednarodni podiplomski šoli Jožefa Stefana, kjer se ukvarja predvsem z raziskavami na področjih računalniške obdelave naravnega jezika, profiliranjem avtorjev besedil in računalniške kreativnosti .

Mija Michelizza je znanstvena sodelavka na Inštitutu za slovenski jezik Frana Ramovša ZRC SAZU. Je soavtorica tretje izdaje Slovarja slovenskega knjižnega jezika (eSSKJ), sodelovala je tudi pri drugi izdaji Slovarja slovenskega knjižnega jezika (SSKJ^2) in pri Slovarju novejšega besedja slovenskega jezika (SNB). Doktorirala je na temo spletnih besedil in jezika na spletu, ukvarja se z leksikologijo in leksikografijo, tudi v povezavi z novostmi, ki jih prinašajo splet in drugi elektronski mediji.

Maja Miličević Petrović je izredna profesorica na Oddelku za splošno jezikoslovje Fakultete za filologijo v Beogradu. Poučuje predmete, vezane na poučevanje tujih jezikov, psiholingvistiko, korpusno jezikoslovje in kvantitativne metode v jezikoslovju. Raziskovalno se ukvarja predvsem z vlogo transferja pri učenju tujega jezika, jezikovnimi lastnostmi prevodnih besedil in z računalniško posredovano komunikacijo. V največji meri se posveča srbščini, italijanščini in angleščini, zlasti pa jo zanima raziskovalna metodologija. Izvedla je številne seminarje in spletne tečaje, namenjene statistični analizi in splošnim metodološkim vprašanjem pri proučevanju jezika.

Senja Pollak je podoktorska sodelavka na Odseku za tehnologije znanja Instituta »Jožef Stefan«. Magisterij iz računalniškega jezikoslovja je pridobila na Univerzi v Antwerpnu, doktorirala pa je na Oddelku za prevajalstvo Filozofske fakultete Univerze v Ljubljani. Področja njenega raziskovanja so rudarjenje besedil, korpusno jezikoslovje in računalniška kreativnost. V okviru nacionalnih in evropskih projektov se je v zadnjem času posvečala analizi jezikovnih značilnosti računalniško posredovane komunikacije in finančnega poročanja, razvoju računalniških metod za potrebe prevajalske industrije ter metodam računalniške kreativnosti za odkrivanje novega znanja. V okviru evropskega projekta SAAM (2017–2020) pa predmet svojih raziskav širi na področje avtomatskega razpoznavanja upada kognitivnih sposobnosti na podlagi analize besedil in govora, kar je tudi tema njenega podoktorskega usposabljanja na Institutu Usher Univerze v Edinburgu.

Damjan Popič je zaposlen na Oddelku za prevajalstvo Filozofske fakultete v Ljubljani, kjer poučuje predmete, vezane na sistem slovenskega jezika, metodologijo in tvorjenje (znanstvenih) besedil ter uporabo informacijskih tehnologij pri prevajalskem in raziskovalnem delu. Raziskovalno se ukvarja predvsem s vprašanji sociolingvistike in korpusnega jezikoslovja.

Špela Reher je z odliko magistrirala iz prevajanja na Filozofski fakulteti v Ljubljani pri mentorici doc. dr. Darji Fišer. V magistrski nalogi je raziskovala uporabo kodnega preklapljanja v tvitih slovenskih uporabnikov, ki so vključeni v korpus Janes. Trenutno ob študiju prava na Pravni fakulteti v Ljubljani deluje na področju prevajalstva kot samostojna podjetnica.

Iza Škrjanec je magistrska študentka Univerze v Saarlandu. Diplomirala je iz medjezikovnega posredovanja na Filozofski fakulteti Univerze v Ljubljani. Nato je vpisala magistrski študij Informacijskih in komunikacijskih tehnologij na Mednarodni podiplomski šoli Jožefa Stefana, kjer je septembra 2017 zagovorila magistrsko delo »Analiza slovenskih spletnih uporabniških vsebin z vidika spola« pod mentorstvom prof. dr. Nade Lavrač in dr. Senje Pollak. Trenutno svoje znanje izpopolnjuje na magistrskem študiju računalniškega jezikoslovja na Univerzi v Saarlandu v Nemčiji.

Ana Zwitter Vitez je docentka za francoski jezik na Fakulteti za humanistične študije in vodja Centra za jezike Univerze na Primorskem. Pedagoško pokriva področja modernega francoskega jezika, prevajanja in jezikovnih tehnologij. Raziskovalno je dejavna pri preučevanju računalniško posredovane komunikacije, govorjenega diskurza in avtorstva anonimnih besedil.

Abstracts

The Janes corpus of Slovene user-generated content
Tomaž Erjavec, Nikola Ljubešić, Darja Fišer

The chapter presents the first extensive and richly annotated corpus of user-generated content for Slovene, which contains tweets, forum posts, news comments, user and talk pages from Wikipedia, and blogs and blog comments. First, we describe the harvesting procedure for each data source and provide a quantitative analysis of the corpus. Next, we present automatic and manual procedures for enriching the corpus with metadata, such as user type and gender, and text sentiment and standardness level. Finally, we present the encoding of the corpus and the procedure of making the publicly available version of the corpus, and its availability.

Keywords: corpus construction, computer-mediated communication, user-generated content, Internet Slovene, non-standard Slovene

Manually annotated Janes corpora for linguistic research and training language technology tools
Jaka Čibej, Špela Arhar Holdt, Tomaž Erjavec, Darja Fišer

In this chapter, we first present the general procedure and workflow of corpus compilation (from data preparation, guidelines, annotation platform and annotation campaign to final data conversion, publication and distribution), with particular emphasis on the largest of the corpora: Janes-Norm (approximately 185,000 tokens) and Janes-Tag (approximately 75,000 tokens), the main purpose of which is to improve language technology tools for tokenization, sentence segmentation, normalization, lemmatization and morphosyntactic tagging. The second part of the chapter consists of an overview of all manually annotated Janes corpora: in addition to the already mentioned Janes-Norm and Janes-Tag, it describes Janes-Syn (syntax in CMC), Janes-Kratko (shortening phenomena in CMC), Janes-Vejica (comma use in CMC), Janes-Preklop (code switching in CMC), and Janes-Geo (use of non-standard linguistic elements in CMC depending on the users' regional origin). The overview provides short descriptions of the content, structure and purpose of each corpus.

Keywords: Slovene, computer-mediated communication, lemmatization, normalization, morphosyntactic tagging, open data, Text Encoding Initiative, CLARIN.SI

Tools for processing non-standard Slovene
Nikola Ljubešić, Tomaž Erjavec, Darja Fišer

This chapter discusses problems associated with automatically processing non-standard language and methods we developed to resolve these problems. We consider the tasks of predicting text standardness, text segmentation, text normalisation, text rediacritisation, morphosyntactic tagging and named entity recognition. We show that the error of language

tools trained on standard language, when applied to non-standard language, increases drastically but that prior text normalisation or tool adaptation can effectively deal with non-standard input if a reasonable amount of annotated non-standard text is manually annotated and supervised machine learning models are either trained or updated on it.

Keywords: language technology, non-standard language, text normalisation, morphosyntactic tagging, named entity recognition.

Workflows for analysing non-standard Slovene
Matej Martinc, Senja Pollak, Ana Zwitter Vitez

In recent years, much effort has gone into the development of infrastructures that would simplify scientific research, increase interdisciplinary cooperation and provide easier access to research methods and data to a wide range of users. This chapter describes the implementation of a set of tools (widgets) for natural language processing into the visual programming platform ClowdFlows, which will allow linguists and other potential users to perform easier and faster text analysis. The new tools can be used for a range of different natural language processing tasks and will support corpus management and visualization of different corpus statistics. The use of the developed tools is explained and presented in two implemented workflows. The first workflow shows how the described tools can be connected into a system for building new corpora of tweets with the help of a TweetCat streaming tool. The second workflow deals with the analysis of the existing corpus of Eurovision comments, where we focus on the lexical and morphosyntactic differences between positive and negative comments. We conclude with a claim that the implemented tools enable the development of new and the analysis of existing corpora, expand the possibilities for quantitative analysis and in general reduce the complexity of natural language processing.

Keywords: natural language processing, corpus analysis tools, visual programming, ClowdFlows, non-standard Slovene

Spelling practices in Internet Slovene
Darja Fišer, Maja Miličević Petrović

This chapter presents a quantitative analysis of instances of non-standard spelling found on Slovene Twitter. The analysis is based on a manually normalized, lemmatized and part-of-speech tagged tweet sample. The focus is on transformations identified in non-standard forms (compared to the standard ones), on their distribution by part of speech and lemma, as well as the distribution of three different transformation types: deletions, insertions and replacements. The results show that a higher percentage of all transformations is covered by lexical words, but looking within PoS classes, function words are transformed to a greater extent. Deletions constitute the most common transformation

type; given that they mostly take the form of vowel deletions at word end, they point to a similarity between Slovene Twitterese and spoken language.

Keywords: non-standard spelling, computer-mediated communication, Twitter, Slovene

(Non-)standardness in Slovene CMC: the case of the comma
Damjan Popič, Darja Fišer

This chapter deals with comma placement in Slovenian tweets. We investigate to what extent comma placement in Slovenian CMC is used in compliance with standard Slovene, and in what circumstances comma placement deviates the most from the standard language. We aim to enhance previous research into the most common faults in comma placement and try to provide a more comprehensive representation of the use of the comma in Slovenian CMC, all the while making comparisons to the most recent findings in studies dealing with standard language. The results show that the standard use of the comma in the Slovenian computer-mediated communication is more common than the non-standard one. However, we can say that, in a significant portion of the dataset, the comma is omitted on purpose, in keeping with the informality of this type of communication.

Keywords: the comma, computer-mediated communication, Slovene

Regional language variants in Slovene computer-mediated communication: A corpus-based approach with the manually annotated Janes-Geo corpus
Jaka Čibej

In this chapter, we present the compilation and analysis of the manually annotated Janes-Geo corpus, which represents the first step in corpus-based studies of regional language variants in Internet Slovene. The Janes-Geo corpus contains approximately 64,000 tokens written by approximately 270 Twitter users classified into one of nine Slovene regions based on automatically generated metadata on the user's regional origin. The corpus was manually annotated with non-standard language elements according to a bottom-up typology. The purpose of the Janes-Geo corpus is two-fold: to discover the most frequent forms of linguistic non-standardness in Internet Slovene, and to compare the differences in the use of non-standard language elements between users from different regions. In addition to the method of automatically coding metadata on the regional origin of Twitter users, the chapter also describes the annotation process, the structure of the corpus, and some of the main differences between its regional subcorpora, e.g. the frequency of vowel or consonant omissions, various non-standard morphological elements, the most frequent non-standard vocabulary, and the most frequent grapheme transformations.

Keywords: regional language variants, Slovene, tweets, geolocation, computer-mediated communication

Tweets as a lexicographic resource for the analysis of semantic shifts in Slovene
Darja Fišer, Nikola Ljubešić

In this chapter we show the potential of Twitter to monitor lexicographic novelties, focusing on changes in the use of established vocabulary. The approach is based on a comparison of the target word's semantic profiles from a reference corpus and a corpus of tweets with the method of distributional modeling of words. We also propose a typology of semantically identified semantic shifts. We evaluate the results of the approach with a corpus-based manual lexicographic analysis. In addition to easily recognizable noise due to pre-processing errors in both corpora, we distinguish between, the presented approach yields valuable candidates for semantic shifts, especially those that were triggered by daily events and informal communication circumstances.

Keywords: semantic shifts, distributional semantics, corpus-based lexicography, social media

A corpus approach to syntax of computer-mediated Slovene
Špela Arhar Holdt

This chapter presents the activities that were focused on the syntax of computer-mediated Slovene. In the first part of the chapter, we describe the preparation of the Janes-Syn training corpus, a sampled corpus of 200 tweets, which were manually syntactically annotated with the JOS dependency system. We present the adaptations of the annotation system to address the following features of computer-mediated Slovene: genre-specific elements (emoticons, emojis, references to websites, user names and hashtags); the use of foreign language within the Slovene tweets; syntactical fragmentality; and nonstandard use of punctuation. The second part of the chapter presents a linguistic analysis of word order in Janes-Syn. For the study, three linguists independently annotated segments of tweets with presumably marked word order. The study examined their agreement rate, the typology of the annotated word-order problems, and the correspondence of the identified problems to the automatically assigned tags about language standardness of a specific tweet. The analysis revealed important inconsistencies in the linguistic perception of word-order markedness, while the comparison with the automatically assigned tags highlighted the need to better define the concepts of markedness and (non)standardness at the word-order level. These questions should be further addressed with the inclusion of spoken-language data. The typology of annotated problems underlined a number of previously non-examined syntactical features of computer-mediated Slovene and provided guidelines for future corpus-based research of word order in Slovene.

Keywords: Computer-mediated Slovene, corpus linguistics, syntactic annotation, tweets, word order.

Spoken elements in non-standard internet Slovene
Ana Zwitter Vitez, Darja Fišer

Communication in on-line forums, social media and news portals is frequently seen as a hybrid between spoken and written discourse. In order to examine the stereotype, we compare the features of written, spoken, and CMC discourse through an analysis of keyword forms in the corpora Kres, Gos, and Janes. The results show that at the PoS level, CMC is closer to standard written texts than spoken discourse with forum posts being the closest and tweets and news comments with a positive sentiment deviating the most from the written standard. At the lexical level, inter-speaker interactive elements that are typical of spoken discourse are most frequent in tweets and news comments. The results of the analysis could play a role in the rethinking of Slovene register variation and could also be included in the production of new language resources.

Keywords: spoken discourse, computer-mediated communication, informal communication, corpus analysis, interactive elements

The use of hashtags in Slovenian tweets
Mija Michelizza

This chapter deals with hashtags usage according to the role of hashtags in Slovenian tweets. Hashtags as a type of metadata serve as a means of categorization, but they can also perform various communication roles. Hashtags were arranged according to their role in tweets into eight categories which have been shown as relevant already in the Wikström's study (2014): topic tags, hashtag games, meta-comments, parenthetical explanations and additions, emotive usage, emphatic usage, humorous and playful usage, popular culture and tradition. It should be taken into account that categories in this categorization are not mutually exclusive; the same hashtag can be used in categorization as well as appear in any of the communication roles. In the latter, we notice that hashtags are more commonly syntactically integrated, but further research on this topic is needed. Hashtags often show the connection between tweeting while following other media as a backchannel. Some hashtags can be very long as they contain phrases or whole sentences; they rarely contain non-letter symbols other than the hash. Although hashtags represent a newer linguistic element that stands out in computer-mediated communication, they mostly remain within their roles and do not appear in most tweets from a random sample of the analyzed corpus.

Keywords: Hashtag, hash, Twitter, computer-mediated communication, Slovenian

Code-switching in Slovene tweets
Špela Reher, Darja Fišer

This chapter introduces the quantitative and qualitative analysis of code-switching (CS) in Slovenian tweets. The analysis was carried out on a sample of tweets from the Janes corpus that were manually annotated by using our own 5-level annotation scheme, which included language and type of CS, orthographic and morphologic assimilation of code-switches to the Slovenian language, as well as the part of speech. The quantitative analysis showed that CS is not a rare phenomenon, that intrasentential CS is more common than intersentential, that there were about 50% of single-word switches and that closed-class words also appear as code-switches. Among the languages used in CS, English clearly dominates, and most code-switches keep the orthographic and morphologic features of the source language. The qualitative analysis showed that CS fulfils various discourse functions, such as referential, expressive, or phatic. In terms of the semantic fields, CS was often related to popular culture, in particular TV shows, sport, food and Twitter. We also found many idiomatic expressions, phrasal verbs, collocations and even some proverbs among the code-switches.

Keywords: code-switching, borrowing, computer mediated communication, corpus linguistics

New conventions in opening and closing phrases of letters in the electronic age
Helena Dobrovoljc

The chapter looks at the old and new conventions of letter-writing, especially the opening and closing phrases. The main medium of letter communication in the past was of a physical nature, whereas today the medium is electronic. We introduce the spontaneous and learned changes in the letter discourse that are making their way into all types of electronic written communication and are an indication of how the social relationships and the writer's attitude towards letter communication have changed throughout the last century.

Key words: letter-writing, computer-mediated communication, electronic mail, opening and closing phrases, discourse elements, text

Predicting gender of Slovene bloggers
Iza Škrjanec, Nada Lavrač, Senja Pollak

Predicting the gender of text authors presents and interesting research problem; moreover, gender prediction models can be of use in various applications, such as marketing and user profiling. Our study aims to build and evaluate models for the automated gender prediction of Slovenian bloggers. For this task, we use a dataset of blog entries by 177 male and 96 female Slovenian bloggers. All blog entries by an individual user are merged and considered

a single classification instance. We compare two types of gender prediction models: a rule-based and a statistical classifier. The rule-based classification model takes into account the use of referential gender in self-referencing contexts. When building statistical models, we experiment with different features and learning algorithms. Both types of models perform with a classification accuracy over 85%. The most successful model is a token unigram learned with support vector machines. The analysis of the most informative features of this model has shown that the blogs by female and male authors display variation in terms of grammar (the use the grammatical gender and pronouns), topic (e.g. more pronounced topics of family, love, and sexuality in entries by female bloggers) and style (e.g. the use of profane language in entries by male authors).

Keywords: author profiling, gender, social media, blog classification

Imensko kazalo