JANES V0.4: KORPUS SLOVENSKIH SPLETNIH UPORABNIŠKIH VSEBIN

Darja FIŠER
Filozofska fakulteta Univerze v Ljubljani, Inštitut »Jožef Stefan«

Tomaž ERJAVEC
Inštitut »Jožef Stefan«

Nikola LJUBEŠIĆ
Inštitut »Jožef Stefan«, Filozofska fakulteta Univerze v Zagrebu

Fišer, D., Erjavec, T., Ljubešić, N. (2016): JANES v0.4: Korpus slovenskih spletnih uporabniških vsebin. Slovenščina 2.0, 4 (2): 67–100.
URL: http://www.trojina.org/slovenscina2.0/arhiv/2016/2/Slo2.0_2016_2_04.pdf.

V prispevku predstavimo najnovejšo različico korpusa spletne slovenščine Janes, ki vsebuje tvite, spletne forume, novice in uporabniške komentarje nanje, blogovske zapise in komentarje nanje ter uporabniške in pogovorne strani na Wikipediji. 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, spol in regija avtorja ter sentiment in stopnja tehnične in jezikovne standardnosti posameznega besedila. Prispevek sklenemo z opisom delotoka za jezikoslovno označevanje korpusa, ki vključuje tokenizacijo, stavčno segmentacijo, rediakritizacijo, normalizacijo, oblikoskladenjsko označevanje in lematizacijo.

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

1 UVOD

Kljub zgledni podprtosti slovenščine z referenčnimi in specializiranimi korpusi nobeden od njih ne vsebuje besedil, ki jih na spletu ustvarjajo uporabniki sami.

Ker njihov pomen za jezikoslovje, tehnologije, pa tudi družbo nasploh z množično razširjenostjo družbenih omrežij (Statistični urad RS 2015) strmo narašča in ker številne tuje (Crystal 2011, Baron 2008, Beißwenger 2013) ter prve domače jezikoslovne raziskave (Dobrovoljc 2012, Erjavec in Fišer 2015, 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. Poleg najrazličnejših jezikoslovnih raziskav je korpus namenjen tudi razvoju jezikovnotehnoloških orodij, ki se bodo s šumnimi spletnimi besedili, ki vsebujejo fonetiziran zapis besed, dialektalne, pogovorne, slengovske in tujejezične izraze, tipkarske napake, pogosto pa so zapisani tudi brez šumnikov, uspešneje spopadala, kot to uspeva obstoječim, ki so bila naučena na standardni slovenščini (Ljubešić idr. 2014).

Prispevek ima naslednjo strukturo. V drugem razdelku opišemo zvrstnost korpusa, načela vključevanja virov in postopek zbiranja besedil ter korpus kvantificiramo. V tretjem razdelku predstavimo načine pridobivanja metapodatkov, s katerimi so opremljena besedila v korpusu in omogočajo širok nabor natančnejših in primerjalnih jezikoslovnih analiz, podamo pa tudi analize korpusa po posameznih metapodatkih. Četrti razdelek prinaša opis postopkov in orodij za jezikoslovno označevanje korpusa, peti razdelek pa na kratko opiše zapis korpusa, čemur sledijo sklepne ugotovitve in načrti za nadaljnji razvoj korpusa.

2 GRADNJA IN ZVRSTNOST KORPUSA

V najnovejšo različico korpusa Janes je vključenih pet zvrsti javno objavljenih uporabniških spletnih vsebin, in sicer tviti, forumi, novice in komentarji nanje, blogi in komentarji nanje ter uporabniške in pogovorne strani na Wikipediji. Med slovenskimi uporabniki popularna družbena omrežja, kot so Facebook, WhatsApp in Snapchat, vsebujejo večinoma zasebno komunikacijo med

uporabniki, zato jih v korpus nismo vključili. Zajem tvitov in uporabniških ter pogovornih strani na Wikipediji je celovit v smislu, da smo v korpus vključili vse uporabnike in njihove objave s teh platform, ki smo jih identificirali. Za razliko od njih pa smo zaradi časovnih in finančnih omejitev za zajem forumskih sporočil, komentarjev na novice in blogov izbrali zgolj manjši nabor virov, ki so v slovenskem spletnem prostoru najbolj priljubljeni, ponujajo največ jezikovne produkcije in/ali predstavljajo pomemben del slovenskega spletnega prostora. Čeprav se zavedamo, da s tem nismo zajeli vseh tem, s katerimi se spletne uporabniške vsebine ukvarjajo, in besedišča, ki je na njih uporabljeno, predvidevamo, da smo kljub vsemu zajeli zadovoljiv vzorec jezikovne rabe, ki je za ta način komunikacije med govorci slovenščine značilna. Zbiranje besedil je bilo oportunistično, kar pomeni, da smo z izbranih virov v korpus zajeli vso gradivo, ki smo ga lahko. Tako zgrajen korpus sicer ni uravnotežen, vendar lahko služi kot osnova za izdelavo uravnoteženega podkorpusa, izdelavo katerega načrtujemo v sklepni fazi projekta, saj so načela za uravnoteženje spletne besedilne produkcije vse prej kot rešen raziskovalni problem in jih je zato potrebno predhodno skrbno preučiti in oblikovati.

V nadaljevanju razdelka opišemo vire in metode, ki smo jih uporabili za zajem posameznih zvrsti besedil, ki so zajeta v korpusu.

2.1 Zajem besedil

2.1.1 TVITI

Tvite smo zajeli z namenskim orodjem TweetCat (Ljubešić et al. 2014), ki je bilo izdelano prav za gradnjo korpusov tvitov manjših jezikov. Orodje uporablja Twitter Search API, da najde uporabnike, ki tvitajo v ciljnem jeziku (v primeru korpusa Janes je to slovenščina). Orodje 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. Za slovenščino smo uporabili *še, kaj, že, če, ampak, mogoče, jutri, zdaj, vendar, kje, oziroma, tudi, sploh, spet, všeč, ravnokar, končno,*

kdaj, preveč in *očitno*. Ko orodje identificira uporabnike, ki potencialno tvitajo v ciljnem jeziku, izvede natančnejšo identifikacijo jezika uporabnika na njegovi časovnici (največ zadnjih 200 tvitov), saj je točnost določanja jezika močno odvisna od količine besedila na razpolago. Avtorji pretežno ciljnega jezika so dodani v indeks uporabnikov, ki jim nato orodje ves čas sledi in sproti shranjuje njihove objavljene tvite. V množico potencialno zanimivih uporabnikov so zajeti tudi vsi uporabniki, ki jim sledijo že identificirani tviteraši, 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. Ker je slovenščina na družbenem omrežju Twitter redek jezik, je to zaporedje filtrov potrebno, da bi res zajeli čim več slovenskih in čim manj tujejezičnih tviteraševob zavedanju, da je identifikacija jezika težak problem, toliko bolj za besedila na tviterju, ki so zelo kratka, niso napisana v standardnem jeziku in vsebujejo veliko tujejezičnih prvin. Vsa filtriranja so narejena na uporabnikih in ne na tvitih: ti so za preverjeno slovenske uporabnike vsi vključeni v korpus, ne glede na to, kateri jezik jim je pripisan.

2.1.2 FORUMI

V korpus smo vključili zdravstvene posvetovalnice s foruma *med.over.net* ter specializirana foruma s področja avtomobilizma in znanosti *avtomobilizem.com* in *kvarkadabra.net*, s čimer smo želeli zajeti najaktivnejše forume, pokriti raznovrstnen nabor tem in zaobjeti raznolike segmente jezikovne rabe v slovenskih forumih. To smo ocenili na podlagi števila registriranih uporabnikov posameznega foruma, števila in dinamike objavljenih sporočil ter nabora aktivnih tem na forumih. 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,[1] kar je bilo ozko grlo pri nadaljnjem širjenju virov besedil. S pomočjo ekstraktorja besedila smo iz zajetega materiala 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, kar olajšuje ciljno uporabo korpusa v nadaljnjih raziskavah.

2.1.3 KOMENTARJI NA NOVICE

Z novičarskih portalov smo zajeli osrednji nacionalni javni medij *rtvslo.si* ter dva ožje usmerjena politična tednika, levi politični opciji naklonjena *mladina.si*[2] in desno usmerjeni *reporter.si*. Za vključitev vira v korpus je bila ključna politika novičarskih portalov, saj številni portali dostop do novic zaračunavajo (npr. *vecer.com*), 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. Tudi zajem komentarjev je 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.

[1] Za luščenje besedil iz zajetih spletnih strani smo uporabili Pythonovo knjižnico Beautiful Soup: https://www.crummy.com/software/BeautifulSoup/.
[2] 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.

2.1.4 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 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 npr. *blog.siol.net* in *ednevnik.si*.

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

2.1.5 UPORABNIŠKE IN POGOVORNE STRANI NA WIKIPEDIJI

Zajem pogovornih strani z Wikipedije smo opravili z lastnim orodjem, ki obdela izvoz Wikipedije. 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 Wikipedija strani (*pagetalk*), smo za omogočanje natančnejših analiz v korpusu eksplicitno ločili od komentarjev na uporabniških straneh posameznih avtorjev slovenske Wikipedije (*usertalk*).

2.2 Postprocesiranje

Zajete podatke vseh petih podkorpusov smo še 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 unikod znak, pri ostalih identificiranih napakah pa smo izbrisali bodisi nelegalne znake bodisi celotno besedilo. V tej fazi smo poskrbeli tudi, da podkorpus ne vsebuje praznih in podvojenih besedil in da se

zapiše kot veljaven dokument XML.

2.3 Velikost korpusa

Korpus Janes v0.4 vsebuje nekaj več kot 9 milijonov besedil, v katerih je dobrih 200 milijonov pojavnic oz. 175 milijonov besed. Zgrajeni korpus je zelo heterogen, tako glede na količino, dolžino in starost vključenih besedil kot tudi glede na avtorstvo, kar prikažemo s kvantitativno analizo korpusa v nadaljevanju razdelka.

(Pod)korpus in vir	Št. besedil	Št. besed	Št. pojavnic	Št. besed/ besedilo
tweet[3]	**7.503.199**	**90.180.337**	**107.053.232**	**12,0**
forum	**772.953**	**39.760.357**	**47.067.665**	**51,4**
avtomobilizem	569.594	21.920.804	25.630.690	38,5
medovernet	122.613	11.614.852	13.798.691	94,7
kvarkadabra	80.746	6.224.701	7.638.284	77,1
blog	**404.281**	**28.827.667**	**34.535.479**	**71,3**
rtvslo.comment	324.586	11.627.874	14.070.544	35,8
rtvslo.post	23.515	8.082.593	9.622.171	343,7
publishwall.post	18.515	7.294.410	8.634.578	394,0
publishwall.comment	37.665	1.822.790	2.208.186	48,4
news[4]	**299.219**	**12.521.553**	**14.838.652**	**41,8**
rtvslo.comment	267.909	10.343.112	12.240.183	38,6
mladina.comment	26.011	1.890.119	2.253.572	72,7
reporter.comment	5.299	288.322	344.897	54,4
wikipedia	**75.699**	**3.844.631**	**4.766.697**	**50,8**
usertalk	50.510	2.635.840	3.266.873	52,2

[3] Imena podkorpusov in njihovih delov so v angleščini, da ustrezajo tipologiji v samem korpusu. V splošnem je bil pri kodiranju korpusa princip, da so metapodatki (imena elementov in atributov) v angleškem jeziku, saj s tem jasno ločimo metajezik od objektnega jezika.

[4] Samih spletnih novic (torej *news.post*) v tabelo in statistiko nismo vključili, saj ne spadajo med uporabniško generirane vsebine.

pagetalk	25.189	1.208.791	1.499.824	48,0
Janes 0.4	**9.055.351**	**175.134.545**	**208.261.725**	**19,3**

Tabela 1: Velikost podkorpusov Janes 0.4 po vrsti besedila in posameznih virih.

Kot prikazuje Tabela 1, je v korpusu Janes v0.4 največji podkorpus tvitov s preko 100 milijoni pojavnic, s čimer predstavlja več kot polovico 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*), medtem ko spletnih novic (torej *news.post)* v tabelo in statistiko nismo vključili, saj ne spadajo med uporabniško generirane vsebine. 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 še za polovico manjši. 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.

V Tabeli 1 je podana tudi povprečna dolžina besedil v besedah. Besedila v korpusu so tipično zelo kratka, saj v povprečju vsebujejo manj kot 20 besed, kar sledi iz narave zajetih besedilnih zvrsti. Po pričakovanju so najdaljši blogovski zapisi s skoraj 400 besedami na besedilo na portalu *publishwall*, najkrajši pa tviti, dolžina katerih je zaradi odločitve ponudnika platforme omejena na največ 140 znakov. Zanimivo je, da je dolžina ostalih besedil precej bolj primerljiva, od malo pod 36 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.

3 KORPUSNI METAPODATKI

Pomembna odlika korpusa Janes je bogastvo metapodatkov o posameznih besedilih ali skupinah besedil, kar nam omogoča bistveno bogatejše jezikoslovne analize. Nekateri metapodatki so bili zajeti neposredno, predvsem URL izvornega besedila, pri tvitih pa preko Twitter API-ja še uporabniško ime avtorja, datum in čas pošiljanja, število posredovanj (*retweets*) in všečkov (*favourites*) in, kjer je uporabnik to funkcijo vklopil, tudi geolokacija.

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,[5] naslov in datum objave besedila, za forume pa tudi naslov posameznega podforuma in teme.

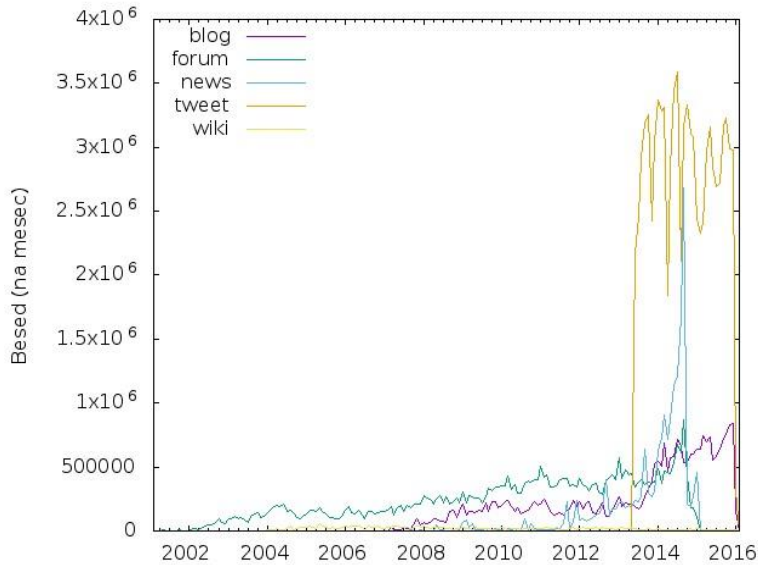
Poleg metapodatkov, ki jih je bilo možno zajeti iz samega besedila, smo celoten korpus oz. posamezne podkorpuse dodatno obogatili z metapodatki, ki so bili dodani bodisi avtomatsko bodisi ročno. V nadaljevanju razdelka predstavimo statistike po nekaterih bolj zanimivih metapodatkih, podrobneje pa razložimo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

3.1 Starost besedil

Za večino zvrsti besedil smo zajem izvedli samo enkrat, in sicer februarja 2015 za (komentarje na) novice in za forume, ter januarja 2016 za (komentarje na) bloge in komentarje na Wikipediji. 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 Twitter API samo

[5] Izjema so novice, ki velikokrat niso podpisane ali imajo v najboljšem primeru samo okrajšavo imena avtorja, zato *news.post* ne vsebuje avtorjevega imena.

zadnjih 500 tvitov uporabnika, zato je pomembno, da se tviti zbirajo sproti. TweetCat skoraj neprekinjeno obratuje od začetka zbiranja do danes, pri čemer smo v Janes 0.4 vključili samo tvite od junija 2013, ko se je zajem tvitov začel, pa do januarja 2016. Ob začetku zbiranja smo pridobili tudi manjše število starejših tvitov, ki jih v korpus nismo vključili, ker bi v kakršnikoli diahroni raziskavi povzročili več škode kot koristi, saj so starejši tviti na voljo samo pri izrazito atipičnih uporabnikih, ki tako tvitajo tako malo, da se tako stari tviti vsa ta leta ostanejo v kvoti 500 tvitov, kot jih Twitter API omogoča zajeti.



Slika 1: Starost besedil v podkorpusih.

Kot prikazuje Slika 1, so bila besedila, vključena v korpus, objavljena v obdobju 2001–2015. 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). Najstarejši komentarji na novice so sicer iz leta 2005, vendar je teh zelo malo. Velika večina jih je iz 2014, kar je posledica tehničnih rešitev novičarskih portalov. Kot rečeno, je najmlajši vir besedil družbeno omrežje Twitter, pri čemer nihanja

niso posledica začasne neuporabe Twitterja, temveč kažejo na obdobja, ko zaradi težav s strežnikom zbiranje tvitov ni delovalo.

3.2 Avtorstvo besedil

Besedila v korpusu je napisalo več kot 96.000 avtorjev, kjer kot enega avtorja štejemo eno uporabniško ime znotraj enega vira. Potrebno je poudariti, da je tako definirano število avtorjev zgolj ocena, saj lahko ista oseba uporablja različna uporabniška imena znotraj enega vira ali enako ime v različnih virih, zgodi pa se lahko tudi, da več oseb uporablja isto uporabniško ime v istem viru.

(Pod)korpus	Št. uporabnikov	Št. besed/ uporabnika	Št. besedil/ uporabnika
tweet	**8.749**	**10.307,5**	**857,6**
forum	**64.489**	**616,5**	**12,0**
avtomobilizem	12.793	1.713,5	44,5
medovernet	49.484	234,7	2,5
kvarkadabra	2.212	2.814,1	36,5
blog	**6.591**	**4.373,8**	**61,3**
rtvslo.comment	3.138	3.705,5	103,4
rtvslo.post	243	33.261,7	96,8
publishwall.post	615	11.860,8	30,1
publishwall.comment	3.040	599,6	12,4
news	**14.430**	**867,7**	**20,7**
rtvslo.comment	12.921	800,5	20,7
mladina.comment	1.273	1.484,8	20,4
reporter.comment	236	1.221,7	22,5
wikipedia	**2.389**	**1.609,3**	**31,7**
usertalk	1.493	1.765,5	33,8
pagetalk	896	1.349,1	28,1
Janes 0.4	**96.648**	**1.812,1**	**93,7**

Tabela 2. Avtorstvo besedil v korpusu Janes v0.4.

Kot kaže Tabela 2, je posamezni avtor v povprečju napisal nekaj čez 1.800 besed oz. skoraj 94 besedil, pri čemer se tudi tu številke zelo razlikujejo glede na podkorpus in vir. Kar devetkrat več besedil, ki obenem vsebujejo petkrat več besed od povprečja, objavljajo uporabniki omrežja Twitter. Ne glede na spletni portal komentatorji sestavijo za slabo polovico besedil glede na povprečje, pri čemer največ besed posamezni komentator prispeva na portalu *mladina*, najmanj pa na portalu *rtvslo*. Največ nihanja 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, in sicer več kot šestkrat manj od povprečja, medtem ko posamezni avtor na forumu *kvarkadabra* objavi skoraj dvakrat več besed od povprečja, s čimer po številu prispevanih besed na avtorja zaseda drugo mesto, tik za uporabniki omrežja Twitter.

3.3 Spol avtorja

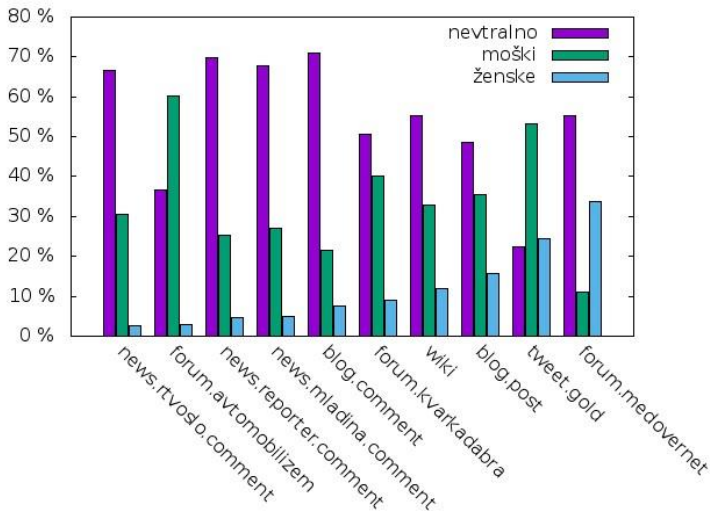
Eden najpomembnejših sociodemografskih podatkov v sociolingvističnih in drugih raziskavah je spol avtorja. Glede na to, da je v slovenščini spol v glagolskih oblikah v pretekliku in prihodnjiku eksplicitno izražen, smo ga vsem avtorjem v korpusu na podlagi prevladujoče oblike v njihovih besedilih pripisali avtomatsko. Za določanje spola avtorjev smo uporabili oblikoskladenjsko označeni korpus (glej razdelek 3.4), 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 prevladujoč spol, sicer smo mu pripisali nevtralnega. Ta hevristika je približna, saj bi za bolj natančno opredelitev spola potrebovali skladenjsko razčlenjen korpus, ker lahko samo tako določimo

celoten povedek v pretekliku ali prihodnjiku, pa še tu ostaja problem z navedki objav drugih uporabnikov.

Natančnost metode smo evalvirali s pomočjo ročno pregledanega seznama oznak za spol za vseh 8.749 avtorjev 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 %, v vseh ostalih primerih smo uporabnikom neupravičeno pripisali nevtralni spol. Z drugimi besedami, metoda je konzervativna in avtorju raje pripiše nevtralni spol, kot da bi se motila pri pripisovanju dejanskega spola, to pa v sociolingvistične in jezikovnotehnološke raziskave, ki v ospredje postavljajo značilnosti izražanja moških in žensk, ne vnaša šuma, temveč zgolj zmanjšuje vzorec za analizo.

Slika 2[6] poda razporeditev spolov po podkorpusih in posameznih virih, urejena pa je po naraščajočem deležu ženskih avtorjev. V vseh virih močno prevladuje nevtralni spol, še najbolj v komentarjih na bloge (71 %) in spletne novice (67 %–70 %), razen v tvitih, ki so bili edini označeni ročno, in na forumu *avtomobilizem*, kjer prevladujejo moški. Moških je tudi sicer 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 Twitter (53 %) in na forumu *kvarkadabra* (40 %). Najmanj žensk sodeluje v komentarjih na *rtvslo* (23 %) in na forumu *avtomobilizem* (3 %), največ pa na že omenjenem forumu *medovernet* (34 %), na Twitterju (24 %) in na blogovskih portalih.

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



Slika 2: Spol avtorjev besedil v podkorpusih.

3.4 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 izobraženi 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 tip avtorja pripisali le avtorjem v podkorpusu tvitov , kjer so poleg individualnih uporabnikov zelo aktivne tudi medijske hiše, javne ustanove in zasebna podjetja. V nadaljevanju projekta podobno načrtujemo tudi s podkorpusom blogov, v katerem so prav tako zajeti nekateri profesionalni pisci, katerih način in teme pisanja se najverjetneje razlikujejo od ljubiteljskih blogerjev. Čeprav smo 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 podkorpusa tvitov je pokazala, da 76 % uporabnikov, zajetih v podkorpusu tvitov, tvita v osebnem imenu, medtem ko je korporativnih računov oz. računov javnih ustanov 24 %. Zanimiva je tudi primerjava tipa uporabnika z njegovim spolom, saj bi pričakovali, da so tviti ustanov vedno po spolu avtorja deloločljivi. To sicer večinoma drži, ne pa vedno, saj je za 16 % institucionalnih uporabniških računov spol mogoče določiti; ta je v 13 % moški, v 3 % pa ženski.

3.5 Regija avtorja

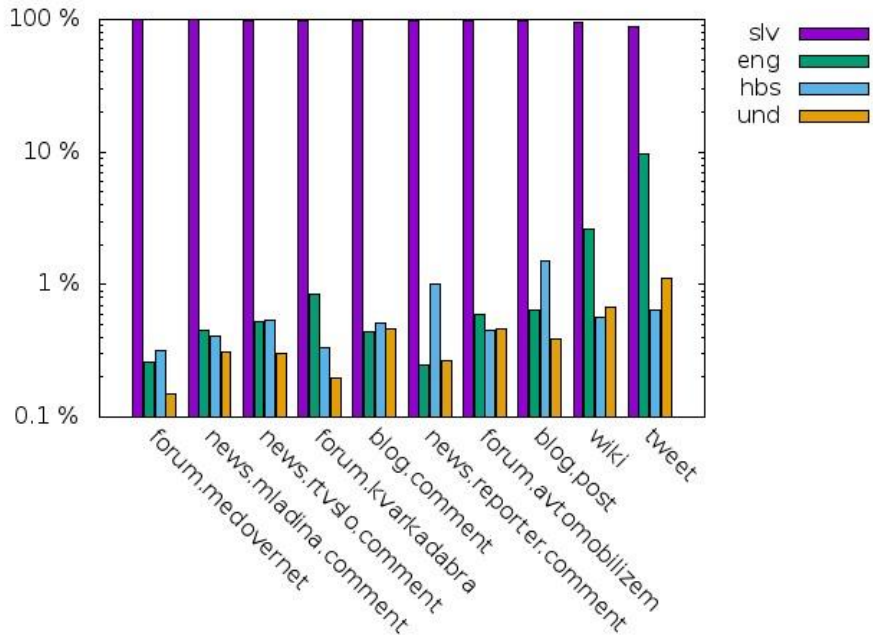
Ker orodje TweetCat, ki ga uporabljamo za zajem tvitov v korpus, omogoča tudi zajem podatkov o geolokaciji, tj. s koordinatami kraja, iz katerega je bil tvit poslan, smo se za omogočanje raziskav regionalnih značilnosti računalniško posredovane komunikacije odločili dodati tudi podatek o regionalni pripadnosti avtorjev v podkorpusu tvitov (Čibej in Ljubešić 2015). S pomočjo orodja Google Maps API v3 Tool6 smo Slovenijo razdelili na 9 koordinatnih poligonov, ki predstavljajo 7 narečnih skupin (gorenjsko, dolenjsko, štajersko, panonsko, koroško, rovtarsko in primorsko) ter Ljubljano in Maribor. Za vsak geolociran tvit v korpusu smo nato preverili, iz katere regije je bil poslan. Uporabnikom, ki so več kot 90 % tvitov z geolokacijo poslali iz ene same regije in so obenem poslali vsaj 3 tvite, smo pripisali metapodatek o regionalni pripadnosti.

Ker je geolociranje funkcionalnost, ki jo mora uporabnik eksplicitno vklopiti, je geolociranih besedil v korpusu razmeroma malo: uporabnikov je sicer res 1.901, kar je 21,7 % vseh, vendar je njihovih geolociranih tvitov samo 160.888, kar predstavlja 2,1 % celotnega podkorpusa. Ker zajem tvitov poteka kontinuirano, označevanje regije avtorja tudi redno osvežujemo.

3.6 Jezik besedil

Kjub temu da smo besedila za korpus zbirali iz slovenskih virov oz. pri tvitih od 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*,[7] 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 šumnikov) in vsebujejo mešanico jezikov. Neposredno označevanje korpusa z *langpy* zato vrne veliko število jezikov (92), od katerih je večina uporabljena zelo malokrat in so tipično tudi napačno identificirani. V nadaljevanju raziskav načrtujemo izboljšati prepoznavanje jezika, ki bo posebej prilagojeno posebnostim spletnih uporabniških vsebin, med katerimi igrata najpomembnejšo vlogo nestandardna ortografija in izrazito kratka dolžina besedil. Zaenkrat pa smo rezultate *langpy* označevanja 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).

[7] Dostopno kot del distribucije Pythona.



Slika 3: Zastopanost jezikov po podkorpusih.

Kot vidimo na sliki 5, 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 (< 1 %), z izjemo komentarjev na Wikipediji in tvitov. Pri *wiki* je 2,6 % besedil identificiranih kot angleških, medtem ko je pri *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 velikokrat tvitajo tudi svojim tujejezičnim sledilcem.

Glede na razmeroma majhno količino neslovenskih besedil so vse analize v nadaljevanju razdelka opravljene na celotnih podkorpusih.

3.6 Standardnost besedila

Ker so 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.

Razvili smo avtomatsko metodo (Ljubešić et al. 2015), ki besedilo opredeli glede na njegovo stopnjo standardnosti, pri čemer se izkaže, da je koristno ločiti med dvema vrstama (ne)standardnosti, in sicer tehnično in jezikovno. Tehnična (ne)standardnost (T) je ožje pravopisna in v je veliki meri motivirana z naravo medija (kratkost in instantnost sporočil), naprav, s pomočjo katerih uporabniki komunicirajo (neergonomske tipkovnice na pametnih telefonih in tabličnih računalnikih) in okoliščin komuniciranja (na avtobusu, med hojo, na koncertu) ter se izraža predvsem v (ne)uporabi velikih začetnic, ločil, presledkov in prisotnosti tipkarskih napak. Jezikovna (ne)standardnost (L) pa je bolj leksikalne in skladenjske narave ter upošteva izbor in zapis besed, njihove oblikoslovne lastnosti ter besedni red. Za obe meri uporabljamo lestvico od 1 (povsem standardno) do 3 (zelo nestandardno). Za vtis, kakšne lastnosti besedil smo upoštevali, v Tabeli 3 za obe vrsti (ne)standardnosti podamo po en primer za vsak podkorpus.

Podkorpus	**T=1 / L=3**	**T=3 / L=1**
tweet	*A nis bla včer na Bledu?*	*komunistična ideologija ubijaj,kradi laži…..zelo primerna za aktualno vlado,,,,,*
forum	*/.../ Z postovanjem vas vse pozdravljam in vam zelim da odprite ocesa, kot sam jaz to naredila, in vam zelim veliko zdravlja v sem skupaj. Se nekaj, gospo Barko pozdravljam i zelo cenim njen trud. /.../*	*/.../ Težave ,ki jih jaz opazim so ,hoja po prstih,težave pri likovnem ,nenatančnost pri geometriji in pri izdelovanju raznih izdelkov(letos so pri naravoslovju izdelovalo hišico iz kartona). Zavedam se ,da se bo do konca šole pokazalo še kaj,a kdo izdela vso šolo ,da se kdaj ne bi kje zataknilo. /.../*

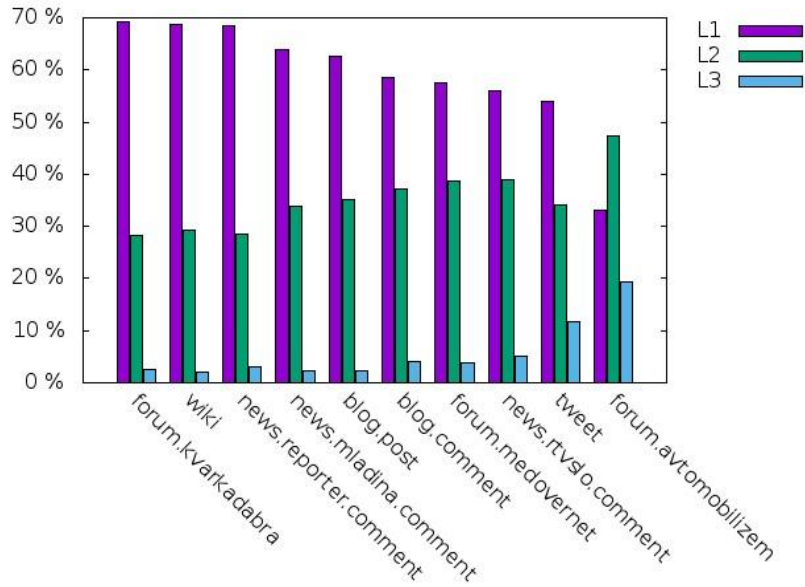
news.comment	*Men so drugač vsi ful lepi, ampak zver je pa ekstra kjut. Pa ful lep nasmešek ma. Pa obrvi..*	*/.../ Zadeva je nerodna in zgled zelo slab ,kar se tiče ostalih članic ,ki prav tako visij (m)o na nitki ! Morda pa za to potrebujemo Patrije in za 11 milijonov streliva ? /.../*
blog.post	*/.../ Ankat sva z bratam šla n hliv pbrat ječke ud kur . Mama nama ih je dala, vsakm nekaj. Jest sm ih mila u varžetih, u bertahu, u usakm varžetu anga, ke sn se bala, de se drgač ječk zdrbi. /.../*	*/.../ PROTI SVETLOBNEMU ONESNAŽEVANJU V TNP ¶ AAG SODELUJE V FEBRUARJU V AKCIJI "OBJEM TOPLINE"–PROSIMO VAS GLASUJTE ZA AAG – NAJDETE NAS POD ŠT.6 – GLASIJETE LAHKO ENKRAT NA DAN! /.../*
blog.comment	*Metek v čelo prvome ka prijde, če glij ka nikaj nej krijv.*	*/.../ Baje so 3-krat v nekajletnem presledku glasovali o tem in celo 2-krat so poslanci izglasovali PROTI , tretjič (? leta 2011 so pa izglasovali , da je omejitev na 5 let nazaj !! /.../*
wiki	*/.../ Klemen: te matra branje? Sm reku da kar je blo RESNEJŠE od spuščene zračnice, in ne spuščena zračnica. Če te matra naslov, ga pa spremeni. Mene zanima seznam mrtvih nemcev v sloveniji. Par zadev sm zbral js. Spodaj jasno piše da ni popoln, če maš vir do česa manjkajočega ga pa dodaj. Najprej se fehta folk k sodelovanju, ko pa neki dodaš, te napadejo oldboysi, kako je zanič. /.../*	*lp zanima me kdo v slo izdeluje karte, rabim ponudbo na xxxxxxx.*

Tabela 3: Primeri tehnično in jezikovno (ne)standardnih besedil v korpusu Janes v.04.

Postopek temelji na regresijskem strojnem učenju, zato smo najprej ročno označili 1.200 tvitov, komentarjev in forumskih sporočil s celoštevilčno mero 1-3 tako za T kot za L. Tu se je potrebno zavedati, da kljub razmeroma natančnim navodilom označevalcem ocena stopnje nestandardnosti vseeno vsaj do neke mere ostaja subjektivna odločitev.

Regresorju smo nato definirali značilke, ki bi lahko bile pomembne za določanje stopnje standardnosti. Glede na izbrane značilke in s pomočjo učne množice se

je program naučil pripisati vsakemu besedilu zvezno vrednost od 1 do 3 za obe meri standardnosti. S tem programom smo nato določili obe stopnji standardnosti vsem besedilom v korpusu. Evalvacija je pokazala, da je povprečna absolutna napaka metode 0,38 za določanje tehnične in 0,42 za določanje jezikovne standardnosti. Ti rezultati torej niso dovolj dobri, da bi npr. v L1 lahko pričakovali izključno jezikovno standardna besedila, vendar je tudi ta približna ocena za filtriranje poizvedb v korpusu v praksi vseeno že zelo koristna, saj uporabniku omogoča, da analizo omeji na tisti segment korpusa, ki ga v določeni raziskavi zanima.



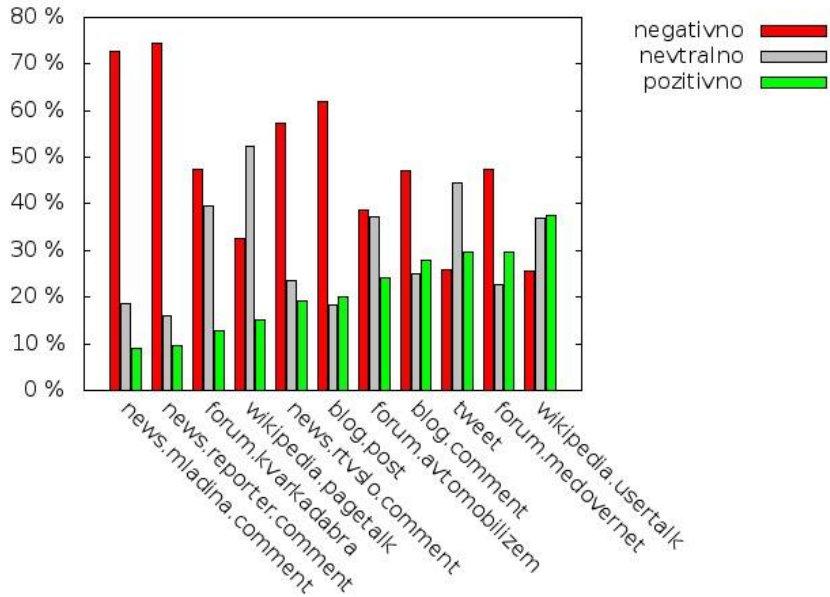
Slika 4: L-standardnost podkorpusov.

Na sliki 3 podamo podatke o razmerju stopenj lingvistične standardnosti besedil po posameznih podkorpusih in nekaterih bolj zanimivih virih, pri čemer so stolpci urejeni padajoče glede na L3. Gledano v celoti so besedila v korpusu precej bolj standardna, kot bi morda pričakovali, saj je standardnih več kot polovica besedil v vseh virih, razen v forumu *avtomobilizem*. Poleg njega, kjer

je zelo nestandardnih petina vseh besedil, po nestandardnosti izstopajo še tviti, ki vsebujejo 12 % takšnih besedil, 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 %.

3.7 Sentiment besedila

Označevanje sentimenta na področju uporabniško ustvarjenih vsebin postaja vse popularnejše (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 je na negativen, pozitiven in nevtralen, pri čemer se kot nevtralna kategorizira tudi besedila, kjer je sentiment mešan.



Slika 5: Sentiment podkorpusov.

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 2014), ki pa zaradi projektnih okoliščin žal niso dostopni za neposredno uporabo v našem korpusu.

Natančnost določanja sentimenta smo evalvirali na vzorcu 600 besedil (Fišer et al. 2016), v katerega smo vključili po 120 besedil iz vsake vrste besedil v korpusu, razen komentarjev na bloge, za katere smo ugotovili, da se obnašajo zelo podobno kot komentarji na novice in jih zato v nadaljevanju raziskave nismo posebej obravnavali. Vzorec smo uravnotežili še z enakim številom besedil po virih, npr. po 40 komentarjev na novice za vsakega od vključenih treh spletnih portalov. S tem smo zagotovili večjo raznolikost vzorca, saj bi sicer v njem prevladali viri, ki so v korpus prispevali največ besedil (npr. reporter.si v korpus prispeva v primerjavi z rtvslo.si zgolj 5 % besedil). Vsakemu besedilu v vzorcu so trije anotatorji ročno pripisali sentiment, pri čemer so imeli tudi možnost, da besedilo označijo kot nerelevantno, ker je npr. napisan v tujem jeziku, avtomatsko generiran ipd. Po tem izločanju je končni vzorec vseboval 555 besedil.

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 ujemanje, 0 pa naključno ujemanje med označevalci. Za naloge, kot je bila naša, velja, da je ujemanje še sprejemljivo, kadar je koeficient alfa vsaj 0,4. (Mozetič et al., 2016). Kot je razvidno iz Tabele 3, je določanje sentimenta precej subjektivna naloga in težak problem za računalnike. Vsi rezultati ujemanja so pod 0.6, kar je sicer sprejemljivo, a daleč od popolnega ujemanja. Avtomatsko pripisovanje sentimenta je, pričakovano, slabše od ujemanja med označevalci. Čeprav je skupni rezultat nad pragom sprejemljivosti, ta za tri od petih tipov besedil ni 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 ti med seboj niso strinjali. S tem smo sistem

kaznovali tudi v primerih, ko se je morda ujemal z enim od označevalcev.

	skupaj	Wiki	News	Blog	Forum	Tweet
Anotatorji	0,563	0,464	0,513	**0,594**	0,464	0,547
Sistem	0,432	0,402	0,394	**0,446**	0,245	0,372
Št. besedil	555	107	115	115	119	99

Tabela 4. Ujemanje, izraženo s koeficientom Krippendorffove alfe za različne vrste besedil v vzorcu.

Glede na rezultate lahko ugotovimo, da je določanje sentimenta tako za označevalce kot za avtomatski sistem najlažje za bloge. To je verjetno posledica dejstva, da so blogi v povprečju najdaljša besedila v korpusu, kar bralcu oz. sistemu omogoča, da bolje razpozna avtorjev sentiment. Označevalcem so bili drugi najlažji tviti, medtem ko je avtomatski sistem na njih dosegel drugi najslabši rezultat. To je zanimivo, saj je bil sistem naučen ravno na tvitih in bi tako pričakovali, da bo na njih dosegal najboljše rezultate. Podrobna analiza problematičnih tvitov (Fišer in Erjavec, 2016) je pokazala, da so zelo kratki in fragmentarni ter ne vsebujejo dovolj konteksta za določanje sentimenta, prav tako pa mnenja v njih niso eksplicitno izražena, so pogosto sarkastični, ironični ali cinični in jim je brez poznavanja širšega konteksta sentiment težko določiti.

Ob zavedanju, da avtomatska kategorizacija ni zelo zanesljiva, Slika 4 vizualizira razporeditev sentimenta v posameznih virih v korpusu, ki so urejeni naraščajoče glede na pozitivni sentiment. Ni presenetljivo, da v večini virov, predvsem pa v komentarjih na novice, blogih in forumih, prevladuje negativni sentiment, najizraziteje na portalih *reporter* in *mladina*, kjer je negativnih kar tričetrt komentarjev. Nevtralni sentiment prevladuje na pogovornih straneh Wikipedije in v tvitih, ki vsebujeta 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.

3 JEZIKOSLOVNO OZNAČEVANJE KORPUSA

V pričujočem razdelku predstavimo razvite avtomatske metode jezikoslovnega označevanja besedil v korpusu, pri čemer je bila veriga označevalnih orodij posebej prilagojena za označevanje nestandardnih besedil.

3.1 Stavčna segmentacija in tokenizacija

Korpus Janes je tokeniziran in stavčno segmentiran z novim orodjem (Ljubešić in Erjavec 2016), ki pokriva slovenščino, hrvaščino in srbščino. Tako kot klasični tokenizatorji tudi naš dela na osnovi pravil, ki so implementirana kot regularni izrazi, njegova novost pa je v tem, da opcijsko podpira tudi procesiranje nestandardnega jezika, kjer uporablja bolj ohlapna klasična kot tudi dodatna pravila. Primer prvega je, da lahko pika konča poved, čeprav se naslednja beseda ne začenja z veliko začetnico ali ji celo ne sledi presledek. Pri tem pa še vedno drži, da pojavnice, ki se končajo s piko in so na seznamu okrajšav, ki ne končajo povedi, kot npr. *prof.*, ne končujejo povedi. Eno izmed dodatnih pravil je posvečeno identifikaciji emotikonov, kot npr. *:-]*, *:-PPPP*, ^_^ itd.

V okviru projekta smo ročno popravili stavčno segmentacijo in tokenizacijo za 4.000 tvitov, kar je okoli 100.000 pojavnic (Čibej et al. 2016). Evalvacija orodja na tej podatkovni množici je pokazala, da bi bilo stavčno segmentacijo tvitov mogoče še precej izboljšati (86,3-% natančnost), medtem ko je tokenizacija zadovoljiva (99,2-% natančnost) ob upoštevanju, da sta obe nalogi za nestandardni jezik tvitov zelo zahtevni in verjetno ni orodja, ki bi lahko doseglo 100-% natančnost. V nadaljevanju nameravamo evalvacijo razširiti tudi na druge vrste besedil, vključenih v korpus, s čimer želimo preveriti stabilnost rezultatov segmentacije in tokenizacije za različne fenomene računalniško posredovane komunikacije, ki se pojavljajo v različnih vrstah spletnih uporabniških vsebin.

3.2 Rediakritizacija

Veliko besedil, zajetih v korpusu Janes, je napisanih brez strešic na šumnikih, kar zelo otežuje nadaljnje postopke obdelave kot tudi siceršnjo uporabnost korpusa. Zato smo besede v korpusu najprej rediakritizirali z namensko razvitim orodjem (Ljubešić in dr. 2016). Orodje se modela rediakritizacije nauči na običajnih besedilih s šumniki in njihovih avtomatsko generiranih različicah, v katerih so strešice na šumnikih odstranjene. Orodje kombinira verjetnost prevoda besede (torej verjetnost, da se beseda brez strešic prevede v neko drugo besedo, ki morda vsebuje strešice) in kontekstualno verjetnost (verjetnost, da je na nekem sičniku v besedi strešica ali ne), ki je ocenjena na podlagi velikega jezikovnega modela. Naši eksperimenti so pokazali, da najboljše rezultate tako na standardnih kot nestandardnih besedilih dosežemo, če se program nauči modelov iz velikih količin tako standardnih kot nestandardnih besedil. Za slovenščino sta bila modela naučena na slovenskih besedilih, zajetih iz Wikipedije, tvitov in spletnih besedil. Evalvacija je pokazala, da metoda na pojavnicah standardnega besedila (Wikipedia) doseže natančnost 99,62 %, na nestandardnih besedilih (tviti) pa 99,12 %.

3.3 Normalizacija

V naslednjem koraku smo rediakritizirane besedne pojavnice normalizirali z metodo, ki temelji na statističnem strojnem prevajanju črk. Cilj normalizacije je, da besedam, ki v zapisu odstopajo od standarda (npr. *jest, jst, jas, js*), pripiše njihovo standardno ustreznico (*jaz)*. Prevodni model je bil naučen na ročno normaliziranem vzorcu 4.000 tvitov, kar je približno 100.000 besed, medtem ko smo za model ciljnega (torej standardnega) jezika uporabili kombinacijo modelov, ki so bili naučeni na korpusu Kres in velikem vzorcu tvitov, ocenjenih kot zapisanih v pretežno standardni slovenščini. Čeprav smo ugotovili, da lahko dosežemo nekaj odstotnih točk boljše rezultate, če normaliziramo celotne povedi, smo zaradi velikosti korpusa in počasnosti metode raje izbrali pristop, v katerem normaliziramo posamezne besede. S tem sicer izgubimo kontekst

besede, a zato normaliziramo le besedišče korpusa, kar pomeni več velikostnih razredov hitrejše procesiranje.

3.4 Oblikoskladenjsko označevanje in lematizacija

Kot zadnji korak jezikoslovnega označevanja smo rediakritizirane in normalizirane pojavnice označili z njihovo oblikoskladenjsko oznako in lemo, za kar smo uporabili orodje, ki je bilo razvito za slovenščino, hrvaščino in srbščino (Ljubešić in Erjavec 2016). Orodje uporablja strojno učenje na osnovi pogojnih naključnih polj in za razliko od klasičnih označevalnikov uporablja leksikon samo posredno, v obliki značilk. Za slovenščino je bilo orodje izšolano na ročno označenem korpusu ssj500k 1.3 (Krek et al. 2013) in oblikoskladenjskem leksikonu Sloleks 1.2 (Dobrovoljc et al. 2015). Novi označevalnik zmanjša relativno napako za skoraj 25 % glede na prejšnje rezultate pri označevanju slovenščine in doseže 94,3 % natančnost na testni množici, ki zajema zadnjo desetino ssj500.

Po vzoru Bartz et al. (2014) smo pri označevanju korpusa Janes vpeljali tudi nove oblikoskladenjske oznake, namenjene boljšemu označevanju spletnega jezika, in sicer *Nw* (e-poštni ali URL naslovi), *Ne* (emotikoni ali emojiji, npr. *:-)* oz. ☺), *Nh* (ključniki, npr. *#kajdogaja*) in *Na* (sklici, npr. *@dfiser3*).

Lematizacija, ki je ravno tako del orodja, upošteva predvideno oblikoskladenjsko oznako in dostopen oblikoskladenjski leksikon, strojno naučen model pa se uporabi samo v primerih, ko para *oblikoskladenjska oznaka : besedna oblika* ni v leksikonu.

5 ZAPIS KORPUSA

Korpus Janes je zapisan v jeziku XML, ki omogoča strukturiranje korpusa, zapis metapodatkov in jezikoslovnih oznak ter strojno preverljivost pravilnosti zapisa. Trenutno je vsak podkorpus shranjen v svoji datoteki in kodiran po lastni shemi XML, ki čim bolje izraža strukturo podkorpusa in njegovih metapodatkov. Tako je npr. vsak tvit svoje besedilo, medtem ko so novice

strukturirane kot novica in njeno besedilo, ki mu sledijo komentarji, forumi pa so urejeni na posamezne podforume, teme in besedila. V nadaljevanju projekta bomo podkorpuse zapisali v enotnem formatu Iniciative za kodiranje besedil TEI (TEI 2016), pri čemer bomo za zapis uporabili razširitev sheme TEI, kot jo predlaga interesna skupina TEI za zapis uporabniško generiranih vsebin (Beißwenger et al. 2012).

Kot trenutno zadnjo stopnjo obdelave smo podkorpuse pretvorili v vertikalni format, primeren za uvoz v konkordančnik (no)SketchEngine (Rychlý 2007), posamezne podkorpuse v vertikalnem formatu pa združili še v celotni korpus Janes 0.4, pri čemer celotni korpus vsebuje samo strukture in metapodatke, ki so skupni vsem podkorpusom. Posamezni podkorpusi in celoten korpus Janes so bili nato uvoženi v lokalni instalaciji konkordančnikov noSketchEngine in SketchEngine. Dostop imajo zaenkrat samo projekti partnerji, saj je pred javno objavo potrebno korpuse še do te mere osiromašiti, da ne bodo kršili avtorskih pravic, zaščite osebnih podatkov ali pogojev uporabe spletnih platform, s katerih so bile vsebine zajete (glej Erjavec et al., 2015).

6 ZAKLJUČEK

V prispevku smo predstavili gradnjo, opremljanje z metapodatki in jezikoslovno označevanje trenutne različice korpusa spletne slovenščine Janes v0.4 ter podali statistike po korpusnih (meta)podatkih. V primerjavi s tipičnimi spletnimi korpusi se predstavljeni razlikuje po tem, da smo posvetili veliko naporov ohranitvi strukture izvornih virov in zajemu čim več metapodatkov. Posebej smo se posvetili tudi vidiku nestandardnosti jezika v korpusih, kjer smo pred oblikoskladenjskimi označevanjem in lematizacijo besedila tokenizirali s posebej za nestandardni jezik prilagojenim tokenizatorjem, zapis besed rediakritizirali in standardizirali, besedilom v korpusih pa smo dodali tudi oznako za stopnjo standardnosti na tehnični in jezikovni ravni. Prestavljena in evalvirana na novo razvita orodja so skupaj z razvitimi modeli tudi odprtokodno dostopna na Githubu.

V prihodnjem delu načrtujemo končno različico 1.0 korpusa Janes, ki se od predstavljene verjetno ne bo bistveno razlikoval po vsebovanih besedilih, z izjemo na novo zbranih tvitov in komentarjev na Wikipediji, imel pa bo boljše jezikoslovno označevanje, kar bomo dosegli s pomočjo obsežnih ročno označenih učnih množic in na njih temelječimi eksperimenti. Posvetili se bomo tudi enotnemu in standardiziranemu zapisu korpusa in njegovemu filtriranju, da ga bomo lahko javno objavili za raziskovanje prek spletnih konkordančnikov, pa tudi za prevzem prek raziskovalne infrastrukture CLARIN.SI. Javno bomo objavili tudi vse ročno označene učne množice, predstavljene v prispevku in zgrajene v preostanku trajanja projekta.

ZAHVALA

Avtorji se zahvaljujejo anonimnima recenzentoma za koristne pripombe in Jasmini Smailović za označevanje sentimenta v korpusu Janes v0.4. Raziskava, opisana v prispevku, je bila opravljena v okviru nacionalnega temeljnega projekta »Viri, orodja in metode za raziskovanje nestandardne spletne slovenščine« (J6-6842, 2014–2017), ki ga financira ARRS.

LITERATURA

Baron, N. (2008): Always On: Language in an Online and Mobile World. Oxford University Press.

Bartz, T.; Beißwenger, M.; Storrer, A. (2014): Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommunikation: Phänomene, Herausforderungen, Erweiterungsvorschläge. Journal for Language Technology and Computational Linguistics 28 (1): 157–198.

Beißwenger, M. (2013): Raumorientierung in der Netzkommunikation. Korpusgestu¨tzte Untersuchungen zur lokalen Deixis in Chats. Die Dynamik sozialer und sprachlicher Netzwerke, 207–258. Springer.

Beißwenger, M.; Ermakova, M.; Geyken, A.; Lemnitzer, L.; Storrer, A. (2012): A TEI Schema for the Representation of Computer-mediated Communication. Journal of the Text Encoding Initiative 3 (2012).

Čibej, J.; Fišer, D.; Erjavec, T. (2016): Normalisation, Tokenisation and Sentence Segmentation of Slovene Tweets. Proceedings of the workshop Normalisation and Analysis of Social Media Texts at LREC'16. Portorož, Slovenia, May 28 2016.

Čibej, J.; Ljubešić, N. (2015): »S kje pa si?« – Metapodatki o regionalni pripadnosti uporabnikov družbenega omrežja Twitter. Zbornik konference Slovenščina na spletu in v novih medijih. Ljubljana: Znanstvena založba Filozofske fakultete, 10–14.

Crystal, D. (2011): *Internet Linguistics*: *A Student Guide*. Routledge, New York.

Dobrovoljc, H.; Jakop, N. (2012). *Sodobni pravopisni priročnik med normo in predpisom*. Založba ZRC.

Dobrovoljc, K., Krek, S., Holozan, P., Erjavec, T., Romih, M. (2015):

Morphological lexicon Sloleks 1.2. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1039.

Erjavec, T. Fišer, D. (2013): Jezik slovenskih tvitov: korpusna raziskava. Družbena funkcijskost jezika: vidiki, merila, opredelitve, 109–116. Znanstvena založba Filozofske fakultete.

Erjavec, T.; Čibej, J.; Fišer, D. (2015): Pravna podlaga za zagotavljanje prostega dostopa korpusov spletnih besedil. Smolej, M. (ur.). OBDOBJA 34: Slovnica in slovar – aktualni jezikovni opis. Ljubljana: Znanstvena založba Filozofske fakultete, 193–199.

Fišer, D.; Erjavec, T. (2016): 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, D.; Smailović, J.; Erjavec, T.; Mozetič, I.; Grčar, M. (2016): Sentiment Annotation of the Janes Corpus of Slovene User-Generated Content. Proceedings of the 10th Languate Technologies and Digital Humanities Conference, 29.9.-1.10. 2016, Ljubljana: Filozofska fakulteta.

Krek, S., Erjavec, T., Dobrovoljc, K., Može, S., Ledinek, N., Holz, N. (2013): *Training corpus ssj500k 1.3,* Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1029.

Krippendorff, K. (2012). Content Analysis, An Introduction to Its Methodology. Sage Publications, Thousand Oaks, CA, 3rd edition.

Lebar, L.; Petrovčič, A.; Petrič, G. (2012): Analiza slovenskih spletnih forumov. Poročilo. http://www.nebojse.si/portal/Dokumenti/Analiza_slovenskih_spletnih_forumov.pdf

Liu, B. (2015): Sentiment analysis. Mining opinions, sentiments, and emotions. Cambridge University Press.

Ljubešić, N.; Erjavec, T. (2016): Corpus vs. Lexicon Supervision in Morphosyntactic Tagging: The Case of Slovene. Proceedings of LREC'16 Conference, Portorož, Slovenija.

Ljubes˘ić, N.; Erjavec, T. Fišer, D. (2014a): Standardizing Tweets with Character-Level Machine Translation. Lecture notes in computer science, 164–75. Springer.

Ljubešić, N.; Erjavec, T. in Fišer D. (2016): Corpus-Based Diacritic Restoration for South Slavic Languages. Proceedings of LREC'16 Conference, Portorož, Slovenija.

Ljubešić, N.; Fišer, D.; Erjavec, T. (2014): TweetCaT: a tool for building Twitter corpora of smaller languages. Proceedings of LREC'14 Conference, Reykjavik, Islandija.

Ljubešić, N.; Fišer, D.; Erjavec, T.; Čibej, J.; Marko, D.; Pollak, S.; Škrjanec, I. (2015): Predicting the level of text standardness in user-generated content. Proceedings of RANLP'15 Conference, 7-9 September 2015, Hissar, Bulgaria. Hissar: 371–378.

Michelizza, M. (2015): Spletna besedila in jezik na spletu. Primer blogov in Wikipedije v slovenščini. Lingua Slovenica 6. ZRC.

Mozetič, I.; Grčar, M.; Smailović, J. (2016). Multilingual Twitter sentiment classification: The role of human annotators. PLoS ONE, 11(5):e0155036.

Rychlý, P. (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, J.; Grčar, M.; Lavrač, N.; Žnidars˘ič, M. (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/.

JANES V0.4: CORPUS OF SLOVENE USER-GENERATED CONTENT

The paper presents the current version of the Slovene netspeak corpus Janes, which contains tweets, forum posts, news comments, blogs and blog comments, and user and talk pages from Wikipedia. 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, gender and region, and text sentiment and text standardness level. Finally, we give a detailed account of the linguistic annotation workflow which includes tokenization, sentence segmentation, rediacritisation, normalization, morphosyntactic tagging and lemmatization.

Keywords: corpus construction, computer-mediated communication, user-generated content, Internet Slovene, non-standard Slovene