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.

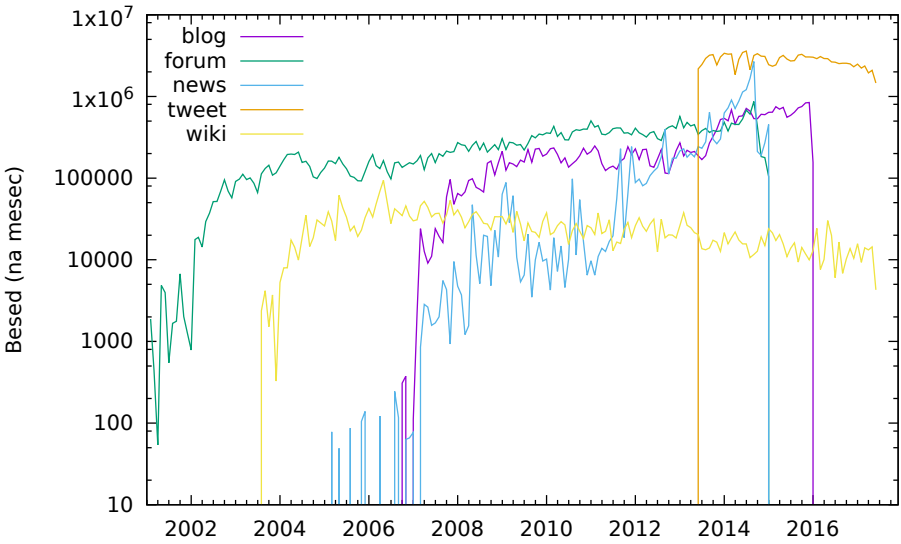
Poleg metapodatkov, ki jih je bilo možno zajeti iz besedila, smo celoten korpus oz. posamezne podkorpuse dodatno obogatili z metapodatki, ki so bili dodani bodisi avtomatsko bodisi ročno. V nadaljevanju razdelka predstavimo statistike za najpomembnejše metapodatke, podrobneje pa razložimo tudi bolj zanimive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

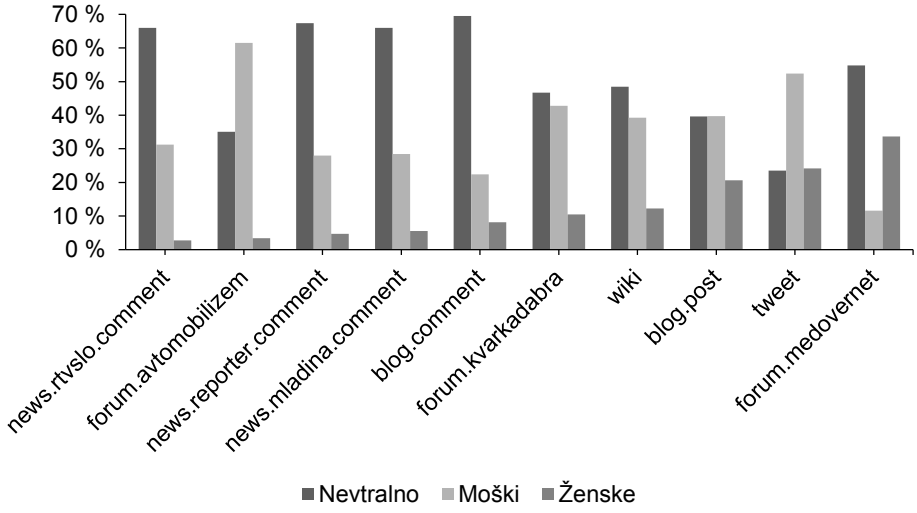
4.3 Spol avtorja

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

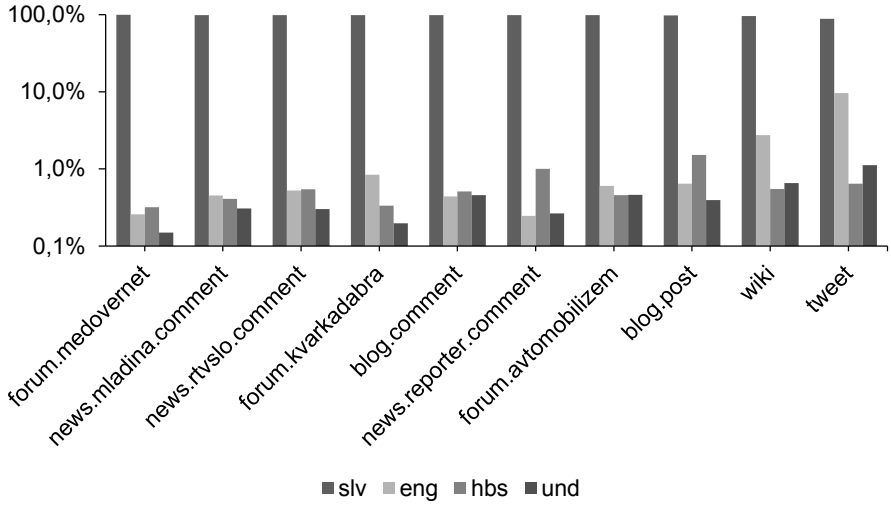
Glede na to, da namen sporočanja močno opredeljuje izbiro jezikovnih sredstev, smo nekatere podkorpuse opremili tudi s podatkom o tipu avtorja, pri čemer ločujemo med osebnimi računi posameznikov, ki uporabniške spletne vsebine objavljajo v svojem imenu kot obliko preživljanja prostega časa, in uradnimi 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

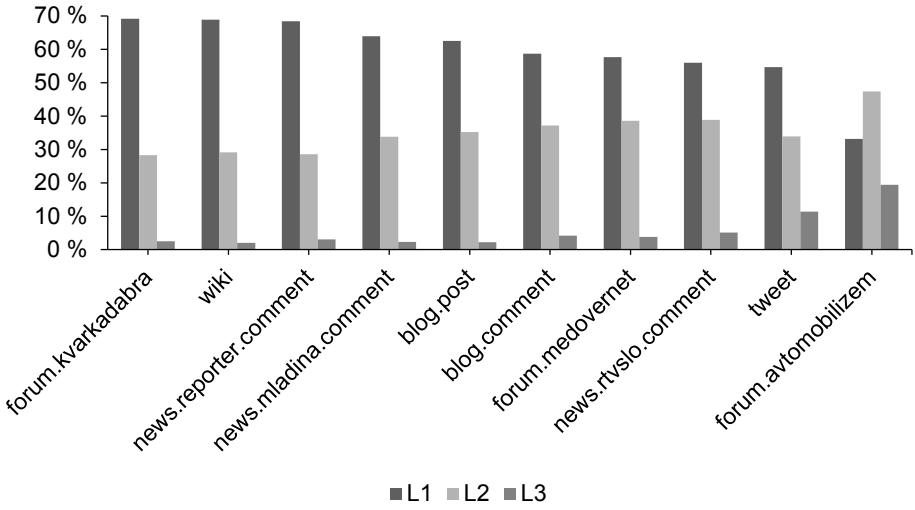
Ker so že prve analize pokazale, da zgrajeni korpus vsebuje številna besedila podjetij (novice, oglasi) in javnih ustanov (obvestila), ki tako po komunikacijskem namenu kot jezikovni podobi v ničemer ne odstopajo od klasičnih besedil na njihovih spletnih straneh, smo se odločili razviti postopek, ki 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.

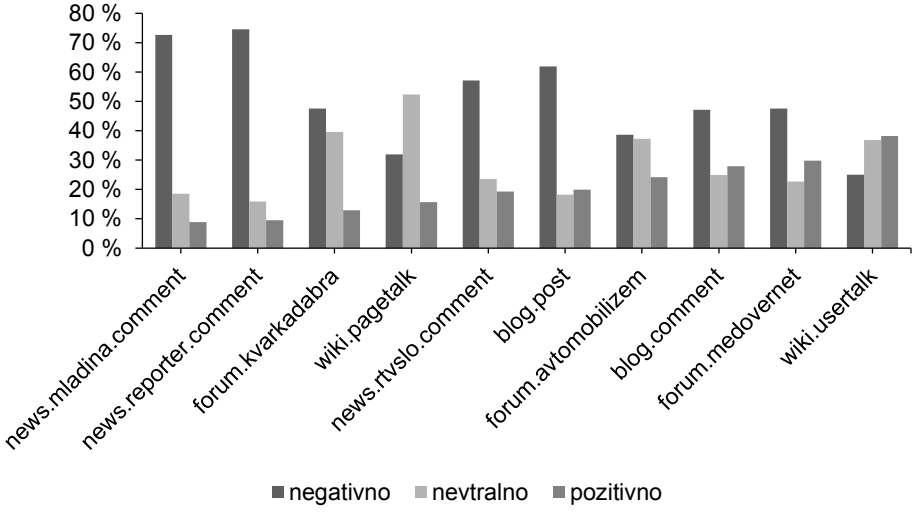
Ob zavedanju, da avtomatska kategorizacija ni zelo zanesljiva, Slika 5 vizualizira razporeditev sentimenta v posameznih virih v korpusu, ki so urejeni naraščajoče glede na pozitivni sentiment. V večini virov (komentarji na novice, forumi in blogi) prevladuje negativni sentiment, najizraziteje na portalih *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.