Razvoj korpusa slovenskih spletnih uporabniških vsebin Janes

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

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

Povzetek

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

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

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

1 Uvod

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

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

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

2 Korpus Janes

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

2.1 Izbor in zajem besedil

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

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

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

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

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

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

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

2.2 Jezikoslovno označevanje

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

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

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

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

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

Tabela 1: Velikost korpusa Janes v0.3.

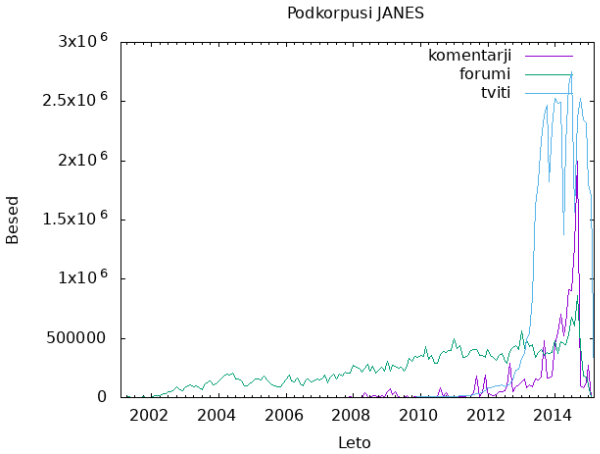
2.3 Sestava

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

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

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

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



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

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

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

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

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

3 Korpus Janes Tviti

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

3.1 Označevanje tipa in spola uporabnikov

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

3.2 Označevanje sentimenta

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

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

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

Tabela 2: Ujemanje ročno in avtomatsko pripisanih oznak sentimenta.

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

Tabela 3: Razporeditev ročno in avtomatsko pripisanih oznak sentimenta.

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

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

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

3.3 Sestava

Tabela 4 poda velikost korpusa Janes Tviti v0.3.4 v celoti in glede na posamezne označene metapodatke. Korpus vsebuje več kot 56 milijonov besed oz. 4 milijone tvitov, ki jih je napisalo okoli 7.600 avtorjev. Med njimi prevladujejo moški (53 %), ki so v korpus prispevali tudi največji delež tvitov in enak delež pojavnic (56 %). Žensk je približno pol manj, in sicer slabih 25 %. Objavile so 27 % tvitov in enak delež pojavnic v korpusu. V podobnem deležu se pojavljajo uporabniki, ki jim ni bilo mogoče pripisati spola (22 %), za katere je zanimivo, da so v korpus prispevali najmanjši delež besedil, in sicer nekaj nad 17 % tvitov in le za odstotek večji delež pojavnic. Ti podatki kažejo, da moški in ženske tvitajo približno enako pogosto in v podobni dolžini, medtem ko uporabniki, ki jim spola nismo mogli določiti, objavijo precej manj sporočil, ki pa so nekoliko daljša.

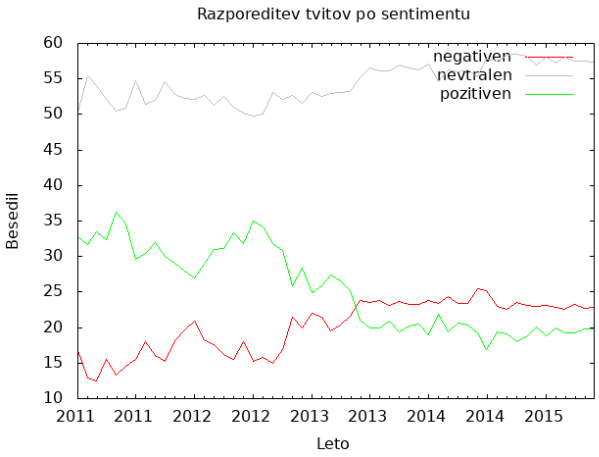
Dobre tri četrtine uporabnikov, zajetih v korpusu, tvita v osebnem imenu (76 %), medtem ko je slaba četrtina korporativnih računov oz. računov javnih ustanov (24 %). Zasebni uporabniki so v korpus prispevali 79 % tvitov oz. dobrih 80 % pojavnic, institucijski pa le 21 % tvitov oz. za odstotek manj pojavnic.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

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

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

4 Označevanje nestandardnosti besedil

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



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

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

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

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

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

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

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

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

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

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

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

5 Zaključek

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

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

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

Zahvala

Avtorji se zahvaljujejo Jasmini Smailović za označevanje sentimenta v korpusu Tvitov, Sašu Rutarju pa za izdelavo programske kode za to nalogo. Hvala tudi Jasmini, Marku Robniku Šikonji, Katji Zupan in anonimnima recenzentoma za koristne pripombe. Za vse preostale napake avtorji krivijo drug drugega. Raziskava, opisana v prispevku, je bila opravljena v okviru nacionalnega temeljnega projekta "Viri, orodja in metode za raziskovanje nestandardne spletne slovenščine" (J6-6842, 2014–2017), ki ga financira ARRS.

6 Literatura

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

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

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

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

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

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

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

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

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

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

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

Nikola Ljubešić, Darja Fišer, Tomaž Erjavec, Jaka Čibej, Dafne Marko, Senja Pollak in Iza Škrjanec. 2015. Predicting the level of text standardness in user-generated content. V: *RANLP - Recent Advances in Natural Language Processing*.

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

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

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