GRADNJA IN ANALIZA KORPUSA SPLETNE SLOVENŠČINE JANES

[1]Darja Fišer, [2]Tomaž Erjavec, [1]Jaka Čibej, [2,3]Nikola Ljubešić
[1]Filozofska fakulteta, Ljubljana; [2]Institut »Jožef Stefan«, Ljubljana;
[3]Filozofski fakultet, Zagreb

UDK 811.163.6'27'322:004.738.52

Spletna besedila tako po svetu kot v Sloveniji predstavljajo vse večji delež jezikovne produkcije, uporabniške spletne vsebine pa postajajo vse pomembnejši vir znanja in vplivajo tudi na nadaljnji razvoj jezika. Če želimo ta potencial izkoristiti, moramo temeljito proučiti spletni segment jezikovne rabe, ki se razlikuje od klasične jezikovne produkcije. Prvi korak v to smer je izgradnja korpusa spletne slovenščine Janes, ki ga predstavljamo in analiziramo v pričujočem prispevku.

gradnja korpusa, spletna slovenščina, računalniško posredovana komunikacija, uporabniške spletne vsebine, jezik družbenih omrežij

Web texts represent an increasing segment of language production both worldwide and in Slovenia. User-generated content is thus becoming an increasingly important source of knowledge and affects future language development. In order to harness this potential, it is necessary to conduct a thorough analysis of internet language use, which differs from traditional language production. The first step in this direction is the construction and analysis of the Janes corpus of internet Slovene, which is presented in this paper.

corpus building, internet Slovene, computer-mediated communication, user-generated content, language of social media

1 Uvod

Jezikovna produkcija se v vse večji meri udejanja na spletu, in to v lokalnih jezikih, kar potrjujejo številni tuji in domači statistični podatki o spletnih uporabnikih in njihovem udejstvovanju na družbenih medijih. Po podatkih Internet World Stats[1] iz leta 2013 je angleščina glede na število uporabnikov na spletu sicer še vedno na prvem mestu, a je število uporabnikov jezikov, ki so se na lestvici uvrstili pod deseto mesto, že preseglo 440 milijonov in še narašča. Podatki Statističnega urada RS[2] kažejo, da je v prvem četrtletju leta 2014 internet uporabljalo 72 % oseb med 16. in 74. letom starosti, kar 81 % med njimi celo vsak dan ali skoraj vsak dan. Obenem jih je 58 % sodelovalo na družbenih omrežjih, kar pomeni, da več kot polovica slovenskih uporabnikov spleta ustvarja uporabniške spletne vsebine.

1 http://www.internetworldstats.com
2 http://www.stat.si/StatWeb/glavnanavigacija/podatki/prikazistaronovico?IdNovice=6560

Jezik uporabnikov na spletu se precej razlikuje od tistega v klasični jezikovni produkciji, kar kažejo številne tuje (Crystal 2011; Baron 2010; Beißwenger idr. 2013), pa tudi prve domače jezikoslovne raziskave (Michelizza 200; Dobrovoljc 2008; Erjavec, Fišer 2014). Proučevanje novomedijskega jezika je pomembno za sodoben jezikoslovni opis slovenščine, za izdelavo ustreznih leksikografskih, normativnih in pedagoških priročnikov ter za razvoj jezikovnotehnoloških orodij, za katera je značilno, da so bila naučena na standardni slovenščini, zato se z uporabniškimi spletnimi vsebinami bistveno težje spopadajo (Ljubešić idr. 2014b). Da bi omogočili celovit in podroben vpogled v slovenščino novih medijev ter razvoj jezikovnih tehnologij zanjo, smo zgradili korpus spletne slovenščine Janes, ki ga predstavljamo v prispevku.

2 Gradnja korpusa

2.1 Izbira in zajem besedil

V trenutno različico[3] korpusa Janes smo vključili štiri zvrsti uporabniških spletnih vsebin, in sicer tvite, forumska sporočila, komentarje na spletne novice in blogovske zapise. V nadaljevanju razdelka opišemo vire in metode, ki smo jih uporabili za zajem.

Tviti so bili zajeti z namenskim orodjem TweetCat (Ljubešić idr. 2014a), ki je bilo izdelano 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, zajem tvitov pa je nato potekal skoraj dve leti. S tem smo dobili podkorpus tvitov, ki poleg besedila posameznega tvita vsebuje tudi njegove metapodatke, in sicer uporabniško ime avtorja, datum in čas pošiljanja ter število njegovih posredovanj (ang. *retweets*) in všečkov (ang. *favourites*).

Za zajem forumskih sporočil in novičarskih portalov, s katerih smo zajeli komentarje uporabnikov, smo izbrali po nekaj virov, ki so v slovenskem spletnem prostoru najbolj priljubljeni, ponujajo največ jezikovne produkcije in/ali predstavljajo pomemben del slovenskega spletnega prostora. Forumska sporočila smo zajeli z domen med.over.net, avtomobilizem.com in kvarkadabra.net, komentarje na novice pa s spletnih portalov rtvslo.si, mladina.si in reporter.si. Ker se spletna mesta po sestavi med seboj razlikujejo, smo za vsak vir posebej napisali ekstraktor besedila. 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 in identifikacijska številka pripadajočega članka, datum objave komentarja, uporabniško ime avtorja ter identifikacijska (zaporedna) številka komentarja. Vsi komentarji so z identifikacijskimi številkami razvrščeni glede na

3 Trenutna različica korpusa JANES (0.3) je bila zgrajena 5. marca 2015.

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 lahko z iskanjem po identifikacijskih številkah tem ali podforumov opazujemo tudi značilnosti izbranih vsebinskih podsegmentov forumov.

Za gradnjo podkorpusa blogovskih zapisov smo uporabili kar deduplicirano različico splošnega korpusa slovenskega spleta slWaC 2.0 (Erjavec, Ljubešić 2014), iz katerega smo zajeli vsa besedila, pri katerih se v domeni 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 podkorpusov forumov in komentarjev zanje 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.

Vsi našteti podkorpusi so bili nato združeni v korpus Janes, 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 in oblikovanje korpusa

Zajete vire smo najprej očistili podvojenih in praznih besedil, nato pa smo jih jezikoslovno označili. Prvi korak označevanja sta bili tokenizacija in stavčna segmentacija, za kar smo uporabili standardno knjižnico mlToken za slovenski jezik, ki je del programa ToTaLe (Erjavec idr. 2005). 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 idr. 2014b). Z orodji za standardno slovenščino programa ToTaLe smo nato normalizirane besede še oblikoskladenjsko označili in lematizirali.

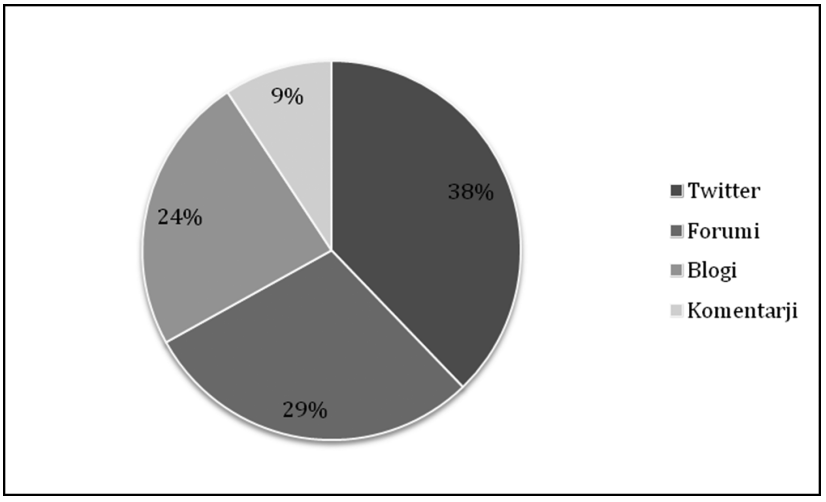
V zadnjem koraku smo posamezne podkorpuse in celoten korpus uvozili v spletni konkordančnik NoSketch Engine (Erjavec 2013). Dostop do njih je omejen na sodelavce projekta, ob zaključku projekta pa načrtujemo tudi splošno (tako prosto kot odprto dostopno) različico korpusov, ki pa bo morala upoštevati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

3 Analiza korpusa

3.1 Sestava korpusa

Korpus Janes v0.3 vsebuje dobrih 161 milijonov pojavnic. Kot prikazuje Slika 1, predstavljajo največji delež v korpusu tviti, sledijo jim forumska sporočila, blogovski zapisi in komentarji spletnih novic. V podkorpusu forumskih sporočil dobro polovico

predstavljajo sporočila s foruma Avtomobilizem (54 %), sledijo sporočila s foruma Medover.net (29 %), najmanj pa jih je s foruma Kvarkadabra (16 %). V podkorpusu komentarjev na novice jih je velika večina s portala RTV Slovenija (82 %), sledijo komentarji s spletnih portalov Mladine (15 %) in Reporterja (2 %).



Slika 1: Sestava korpusa Janes v0.3

Besedila, vključena v korpus, so bila objavljena v obdobju 2001–2015, a jih je skoraj polovica (49 %) iz leta 2014 in pri vseh podkorpusih, razen pri forumskih sporočilih, predstavljajo največji delež zajetega gradiva.

Korpus vsebuje nekaj nad 4,8 milijona besedil s skupno dolžino dobrih 134,5 milijona besed. Napisalo jih je okoli 85.000 različnih avtorjev (brez upoštevanja blogov, za katere nimamo podatkov). Največje število avtorjev imajo forumska sporočila (73 %), najmanjše pa tviti (9 %). Število besedil po posameznih podkorpusih niha v obratnem vrstnem redu: največ (skoraj 3,7 milijona) jih je namreč v podkorpusu tvitov. To je razumljivo, saj jih zaradi omejenega števila 140 znakov na sporočilo pogosto imenujemo tudi mikroblogi. Ta besedila so zato tudi najkrajša v korpusu in znašajo v povprečju 13 besed na besedilo, prav tako jih je posamezni avtor v povprečju napisal največ (skoraj 490), s čimer je v korpus prispeval tudi največ besed (dobrih 6.600).

Avtorji v podkorpusih forumskih sporočil in komentarjev na novice so kljub veliki razliki v številu (več kot 63.500 oz. slabih 14.300) prispevali primerljivo število besedil (skoraj 777.000 forumskih sporočil oz. 300.000 komentarjev, kar znaša v povprečju dobrih 620 besedil na avtorja forumskih sporočil oz. slabih 880 na avtorja komentarjev). Primerljiva je tudi njihova povprečna dolžina (51 oz. 42 besed na besedilo). Na drugem koncu spektra so blogi, ki jih je v korpusu najmanj (skoraj 62.300) in so v povprečju tudi najdaljši (518 besed na besedilo), kar je 14-kratna dolžina povprečnega tvita, 12-kratna dolžina komentarja in 10-kratna dolžina forum-

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

3.2 Oblikoskladenjska analiza

Poleg tega, da je podkorpus tvitov najobsežnejši po številu vsebovanih pojavnic (61 mio.) in številu stavkov (7,3 mio.), vsebuje tudi daleč največ različnih besednih oblik (2,7 mio.) in lem (2 mio.) ter praktično celoten nabor oblikoskladenjskih oznak, ki jih vsebuje korpus Janes (99 %). Čeprav bi lahko k višjemu številu lem in oznak delno prispevale tudi napake avtomatskega označevanja, lahko z gotovostjo trdimo, da podkorpus tvitov vsebuje najbogatejše in najbolj raznoliko besedišče v korpusu Janes.

Glede na obseg podkorpusov forumskih sporočil in blogov, ki znaša okoli tri četrtine (oz. dve tretjini) podkorpusa tvitov, je namreč število besednih oblik približno za dve tretjini (oz. tri četrtine) nižje v podkorpusu forumov (oz. blogov), medtem ko je različnih lem v forumih in blogih v primerjavi s tviti le še za slabo četrtino (oz. petino). Najmanj stavkov (šestkrat manj kot pri tvitih) vsebuje podkorpus komentarjev, ki v primerjavi z drugimi vsebuje tudi najmanjši nabor besednih oblik in lem – prvih je več kot petkrat, drugih pa kar desetkrat manj kot v korpusu tvitov.

V primerjavi z uravnoteženim korpusom Kres vsebuje korpus Janes več kot sedemkrat več medmetov in 1,5-krat več okrajšav. Po drugi strani pa je v korpusu Janes število številnikov praktično razpolovljeno in se v njem pojavlja za slabo tretjino manj pridevnikov. Analiza ključnih oblikoskladenjskih oznak v korpusu Janes glede na korpus Kres pokaže, da je za komunikacijo na družbenih omrežjih značilna raba glavnih in pomožnih glagolov prve in druge osebe v sedanjiku, osebni, svojilni, kazalni in nedoločni zaimki v prvi osebi ednine ter samostalniki oz lastna imena moškega spola v ednini, kar daje prvi vpogled v vsebino in način izražanja v uporabniških vsebinah (sporočanje osebnih mnenj, občutij, dejavnosti).

3.3 Leksikalna analiza

Pojave na leksikalni ravni smo analizirali s pomočjo seznamov ključnih besednih oblik glede na korpus Kres, ki smo jih za celoten korpus in vse podkorpuse izračunali s pomočjo logaritma verjetnosti. Z vsakega seznama smo jih analizirali prvih sto, leksikalne prvine pa razvrstili v eno od devetih kategorij: nevtralno besedišče (*ampak*), pogovorni/narečni/slengovski izrazi (*fajn*), govorne prvine (*pač*), novomedijsko besedišče (*#junaki*), krajšanje (*lp*), emotikoni (*:d*), tematsko obarvano besedišče (*volitve*), nestandardni zapis šumnikov (*vec*), nestandardno pisanje skupaj/narazen (*nevem*).

Med najbolj ključnimi sto besednimi oblikami za celotni korpus Janes je nenevtralnega besedišča 60 %. Največ med njimi je pogovornih, narečnih in slengovskih izrazov (23 %), govornih prvin (13 %) in besedišča, značilnega za družbene medije (10 %). V podkorpusu tvitov je nenevtralnega besedišča kar 81 %. V forumskih sporočilih ga je slaba polovica (46 %), komentarji na novice ga vsebujejo tretjino (33 %), najmanj, dobro četrtino, pa ga najdemo v blogih (29 %), kar je glede na značilnosti analiziranih besedilnih zvrsti tudi v skladu s pričakovanji.

Novomedijsko izrazje močno prednjači v podkorpusu tvitov (41 %). Največ pogovornih, narečnih in slengovskih izrazov najdemo v forumskih sporočilih (25 %). S tematsko obarvanim besediščem so izrazito bogati komentarji na novice (23 %). Zanimivo je, da smo največ govornih prvin (18 %) našteli v blogih, saj se ti zdijo še najbolj oddaljeni od govorjenega diskurza. Domnevamo, da gre za šum, saj v podkorpusu blogov zaenkrat ne ločujemo besedil blogovskih zapisov in komentarjev nanje. Strategije krajšanja sporočil so razmeroma enakomerno zastopane v vseh podkorpusih, medtem ko so emotikoni in izpuščanje šumnikov izrazito značilni za interaktivne besedilne zvrsti, kot so tviti, komentarji in forumi, medtem ko nestandardno pisanje skupaj/narazen ni posebej značilno za nobenega od podkorpusov.

Za natančnejši vpogled v leksiko, posebej specifično za novomedijsko slovenščino, smo analizirali seznam tistih besed v korpusu, ki tako odstopajo od učnega leksikona in korpusa standardne slovenščine, da jih lematizator ni znal lematizirati. Pregledali in kategorizirali smo 100 najpogostejših besed, ki skupaj predstavljajo skoraj 30.000 pojavitev v korpusu. Polovica med njimi so različni emotikoni, ki bi jih bilo treba dodati v leksikon. Sledijo besede, zapisane brez šumnikov (12 %), nestandardno zapisane besede (11 %) in več besed, zapisanih kot ena (6 %), ki se jih moramo naučiti ustrezno standardizirati še pred lematizacijo. V manjšini primerov naletimo na težave zaradi neznanih okrajšav, neznanih besed, ki so bile prevzete iz tujih jezikov, neznanih lastnih imen, zapisa ulomka in zatipkanih besed (skupaj 11 %).

4 Zaključek

V prispevku smo predstavili prve rezultate gradnje in analize korpusa Janes, ki vsebuje tvite, forumska sporočila, blogovske zapise in komentarje na spletne novice. Ta predstavlja prvi korak k proučevanju značilnosti novomedijske slovenščine, kot jo uporabljajo ne zgolj avtorji klasičnih (tiskanih) besedil, temveč vsi, ki ustvarjajo uporabniške spletne vsebine. Predstavljeni korpus prvič pri nas omogoča celovite, poglobljene in ponovljive jezikoslovne raziskave slovenske računalniško posredovane komunikacije. Temeljita analiza tega segmenta spletne komunikacije bo pripomogla tudi k razvoju novih jezikovnotehnoloških orodij za slovenščino (npr. izboljšanih označevalnikov ter aplikacij za zaznavanje žaljivega govora in rudarjenje podatkov). Ker v nadaljevanju projekta načrtujemo tako prosto kot odprto dostopno različico korpusa, bodo podatke pri svojem delu lahko uporabljali tako raziskovalci kot zainteresirana javnost.

Zahvala

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

Literatura

BARON, Naomi S., 2010: *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. Barbara Frank-Job, Alexander Mehler, Tilmann Sutter (ur.): *Die Dynamik sozialer und sprachlicher Netzwerke. Konzepte, Methoden und empirische Untersuchungen an Beispielen des WWW*. Wiesbaden: VS Verlag für Sozialwissenschaften. 207–258.

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

DOBROVOLJC, Helena, 2008: Jezik v e-poštnih sporočilih in vprašanja sodobne normativistike. Miran Košuta (ur.): *Slovenščina med kulturami, Zbornik Slavističnega društva Slovenije 19*. Celovec, Ljubljana: Slavistično društvo Slovenije. [295]–314.

ERJAVEC, Tomaž, 2013: Korpusi in konkordančniki na strežniku nl.ijs.si. *Slovenščina 2.0* 1. 24–49. http://www.trojina.org/slovenscina2.0/arhiv/2013/1/Slo2.0_2013_1_03.pdf

ERJAVEC, Tomaž, FIŠER, Darja, 2013: Jezik slovenskih tvitov: korpusna raziskava. Andreja Žele (ur.): *Družbena funkcijskost jezika (vidiki, merila, opredelitve), Obdobja 32*. Ljubljana: Znanstvena založba Filozofske fakultete. 109–116.
http://www.centerslo.net/files/file/simpozij/simp32/zbornik/Erjavec.pdf

ERJAVEC, Tomaž, IGNAT, Camelia, POULIQUEN, Bruno, STEINBERGER, Ralf, 2005: Massive multi-lingual corpus compilation: Acquis Communautaire and ToTaLe. *Archives of Control Sciences* 15. 529–540.

ERJAVEC, Tomaž, LJUBEŠIĆ, Nikola, 2014: The slWaC 2.0 corpus of the Slovene web. *Jezikovne tehnologije: zbornik 17. mednarodne multikonference Informacijska družba – IS 2014, 9.–10. oktober 2014, [Ljubljana, Slovenia]: zvezek G*. Ljubljana: Institut »Jožef Stefan«. 50–55.
http://is.ijs.si/zborniki/2014_IS_CP_Volume-G_%28LT%29.pdf

LJUBEŠIĆ, Nikola, ERJAVEC, Tomaž, FIŠER, Darja, 2014a: Standardizing tweets with character-level machine translation. *Computational linguistics and intelligent text processing: 15th International Conference, CICLing 2014, Kathmandu, Nepal, April 6-12, 2014: proceedings: part II*. Heidelberg [etc.]: Springer. 164–175.

LJUBEŠIĆ, Nikola, FIŠER, Darja, ERJAVEC, Tomaž, 2014b: TweetCaT: a tool for building Twitter corpora of smaller language. *Zbornik konference Ninth International Conference on Language Resources and Evaluation, May 26–31, 2014, Reykjavik, Iceland. LREC 2014: proceedings*. [S. l.]: ELRA. 2279–2283. http://www.lrec-conf.org/proceedings/lrec2014/pdf/834_Paper.pdf

MICHELIZZA, Mija, 2008: Jezik SMS-jev in SMS-komunikacija. *Jezikoslovni zapiski* 14/1. 151–166.