Slovar tviterščine

Polona Gantar,[*] Iza Škrjanec,[‡] Darja Fišer, [*†] Tomaž Erjavec[†]

* Oddelek za prevajalstvo, Univerza v Ljubljani, Aškerčeva 2, 1000 Ljubljana
apolonija.gantar@ff.uni-lj.si,
darja.fiser@ff.uni-lj.si
‡ Ljubljana
skrjanec.iza@gmail.com
† Odsek za tehnologije znanja, Institut »Jožef Stefan«, Jamova cesta 39, 1000 Ljubljana
tomaz.erjavec@ijs.si

Povzetek

V prispevku opišemo postopek izdelave Slovarja tviterščine, ki je osnovan na korpusu uporabniško generirane slovenščine Janes. Najprej opišemo postopek luščenja korpusnih podatkov ter izdelavo geslovnika in se osredotočimo na kategorizacijo tviterske leksike z vidika standardizacije in stopnje podomačenosti. Nato predstavimo zgradbo slovarskega gesla in tip v slovar vključenih podatkov ter način urejanja v spletnem orodju za izdelavo slovarjev Lexonomy. Prispevek zaključimo z idejami za nadaljnje leksikalne analize tviterske leksike.

Dictionary of Slovene Twitterese

The paper describes the creation of the Slovene Twitterese Dictionary which is based on the corpus of Slovene user-generated texts Janes. First, the procedure of extracting corpus data and headword list creation are described, focusing on the categorisation of the Twitterese vocabulary with respect to standard Slovene and the levels of lexical adoption. Then, the structure of the dictionary entry and the types of lexicographic information included in the dictionary are described, along with the dictionary writing, editing and browsing platform Lexonomy. The paper concludes with plans for future lexical analysis of Slovene Twitterese.

1 Uvod

V prispevku predstavimo proces izdelave slovarja tviterščine na podlagi korpusa Janes Tviti v0.3.4 (Erjavec et al., 2015), ki poteka v okviru projekta »Viri, orodja in metode za raziskovanje nestandardne spletne slovenščine«.[1] Slovarček, ki naj bi ob koncu projekta vseboval pribl. 800 gesel, bo predstavljal besedišče, ki se v okviru širše pojmovane računalniško posredovane komunikacije pojavlja v tvitih, tj. kratkih, na 140 znakov omejenih objavah (s fotografijami, videi ali brez) na družbenem omrežju Twitter.

Kot bomo podrobneje opisali v nadaljevanju prispevka, je besedišče, ki smo ga vključili v slovar, vezano na slovenske tvite, in sicer take, ki po avtomatski oceni (Ljubešić et al., 2015) vsebujejo pretežno nestandardno besedilo. V prispevku podrobneje pojasnimo kategorizacijo tviterske leksike, trenutno stanje, v katerem jo je mogoče pregledovati v spletnem slovarčku, ter orodje za izdelavo spletnega slovarčka.

2 Tviterščina v okviru računalniško posredovane komunikacije

V širšem kontekstu pojmovanja računalniško posredovane komunikacije velja, da je za leksiko na družbenih omrežjih značilen nestandarden, pogosto fonetiziran zapis besed, številne specifične okrajšave in veliko tujejezičnih elementov (Crystal, 2001; Baron, 2003).

Za angleščino so leksiko računalniško posredovane komunikacije začeli zbirati že zelo zgodaj, med najvidnejšimi pionirskimi zbirkami je *The Internet Dictionary* (Crumlish et al., 1995), eden najobsežnejših pa *A glossary of netspeak and textspeak* (Crystal, 2004). Sorodna zbirka je tudi glosar tipičnih okrajšav iz SMS-sporočil (Anjaneyulu, 2013). Medtem ko za spletno slovenščino nekaj parcialnih raziskav leksike že obstaja (Kalin Golob, 2008; Šabec, 2009; Erjavec in Fišer, 2013; Michelizza, 2015), celovitih popisov ali slovarskih zbirk zanjo še nimamo, kot tudi ne opisov leksike, specifične za družbena omrežja, kot sta denimo Twitter in Facebook.

Jezikoslovne analize različnih jezikov na družbenem omrežju Twitter od angleščine (Giai, 2013) in španščine (Álvarez et al., 2012), pa vse do indonezijščine (Bruggman in Conners, 2016) in malezijščine (Isa, 2014) sistematično kažejo, da je jezik v tvitih izrazito dinamičen, se hitro spreminja in prilagaja razvoju platforme ter je močno odvisen od namena in okoliščin komuniciranja in tako včasih bolj podoben javnemu pisnemu komuniciranju v klasičnih medijih, kot so novice ali blogi, drugič pa bolj zasebnim SMS-sporočilom oziroma pogovorom v spletnih klepetalnicah. Hu et al. (2013) so pokazali, da je jezik tvitov bistveno bolj konservativen in manj neformalen kot v SMS-sporočilih in spletnih klepetalnicah. Facchinetti (2015) ugotavlja, da v tvitih zaradi omejitve dolžine posameznega tvita na 140 znakov ne izstopa nobena jezikovna značilnost, kot so npr. strategije krajšanja sporočil, saj je povprečna dolžina tvitov v njenem korpusu zgolj polovična dovoljene, krajšave, nadomeščanje besed z nealfabetičnimi simboli ali črkami pa redke. Pogosto prisoten fonetiziran in prozodičen zapis besed ter fragmentirana skladnja sta značilna tudi za ostale oblike računalniško posredovane komunikacije (Herring, 2012), prve raziskave slovenskih tvitov v primerjavi s sorodnimi jeziki, kot sta srbščina in hrvaščina, pa poleg izjemno pogostega fonetiziranega zapisa kažejo na izrazito veliko prisotnost tujejezičnih prvin in različnih stopenj prevzetosti besed (Fišer et al., 2015).

[1] Spletna stran projekta: http://nl.ijs.si/janes/.

3 Namen in metoda izdelave slovarja

Namen slovarčka je izdelati nabor za slovenske tvite najbolj značilne leksike, pri čemer smo »značilnost« določali na podlagi različnih statističnih parametrov in jezikoslovnih kategorij. Posledično bodo informacije in način njihovega prikaza v slovarju namenjene seznanjanju potencialnih uporabnikov s tipično tvitersko leksiko, možnostmi zapisa posamezne besede in načinom vključevanja v slovenski jezik tako z vidika stopnje jezikovne podomačenosti kot z vidika pomena in rabe – zadnje predvsem v smislu izbire registra in govornega položaja (nestandardna beseda; kletvica, žaljivka ipd.).

Slovar oz. slovarska baza bo namenjena tudi leksikalnim analizam tviterskega žanra v širšem kontekstu leksike spletnih uporabniških vsebin za slovenščino, analizam jezikovnega (ne)standarda, načina podomačevanja tujih besed ter možnostim vključevanja v bodoče slovarske priročnike za slovenščino.

Pri oblikovanju geslovnika smo izhajali iz korpusa Janes Tviti v0.3.4 (Erjavec et al., 2015) in podatke o pogostnosti primerjali s korpusom Gigafida v1.0 ter s preostalim delom korpusa Janes, torej s forumi, komentarji in blogi (prav tam). Na ta način smo želeli izluščiti leksiko, ki je za tvite (a) bodisi bolj specifična in se v drugih žanrih računalniško posredovane komunikacije ne pojavlja oz. se pojavlja razmeroma redko (b) bodisi se v tvitih pojavlja opazno pogosteje kot v preostalih žanrih računalniško posredovane komunikacije. Pričakovali smo tudi, da je leksika, značilna za računalniško posredovano komunikacijo, prisotna tudi zunaj specializiranih spletnih žanrov. V ta namen smo frekvenčne podatke primerjali s korpusom Gigafida, hkrati pa upoštevali še vključenost in opis tviterske leksike v obstoječih enojezičnih slovarjih za slovenščino (Slovar slovenskega knjižnega jezika 2, Slovar novejšega besedja slovenskega jezika) ter novosti v pomenu in rabi.

Pri luščenju podatkov smo poleg frekvence upoštevali še razmerje med pojavnicami in lemami ter oblikoskladenjske oznake v korpusu. Na ta način je bilo mogoče prepoznati variantnost posameznih zapisov ter na podlagi ugotovitev analizirati proces vključevanja tujejezične leksike v slovenski jezik. Variantnost smo v slovarčku prikazali z nizom konkurenčnih variantnih zapisov, stopnjo podomačenosti pa z navedbo leksikalne kategorije (nova beseda) in stopnje podomačenosti (tuje podomačeno ali nepodomačeno).

3.1 Luščenje podatkov

Kot rečeno, je bil osnova za izdelavo geslovnika korpus tvitov, ki je bil zajet z namenskim orodjem TweetCat (Ljubešić et al., 2014b), s katerim smo identificirali uporabnike, ki tvitajo pretežno v slovenščini, in zajeli njihove tvite. Poleg besedila smo zajeli tudi metapodatke, kot so uporabniško ime avtorja, datum in čas pošiljanja ter število posredovanj (ang. *retweets*) in všečkov (ang. *favourites*) zajetega tvita. Korpus Janes Tviti v0.3.4 vsebuje več kot 56 milijonov besed oz. 4 milijone tvitov, ki jih je napisalo okoli 7.600 avtorjev. Metapodatke smo obogatili še z ročno določenimi podatki o lastnostih avtorja (tip računa, spol) ter avtomatsko pripisanimi podatki o sentimentu tvita.

Razvili smo tudi metodo, ki vsakemu tvitu v korpusu avtomatsko pripiše stopnjo jezikovne (zapis besed, raba slengizmov, narečnih in tujejezičnih besed, besedni red, slovnica) in tehnične standardnosti (raba ločil, presledkov, velikih/malih tiskanih črk). Ti dve meri (imenovani L in T) sta pripisani vsakemu tvitu v korpusu, in sicer z ocenami 1 (zelo standardno) – 3 (zelo nestandardno) (Ljubešić et al., 2015). Za namene slovarja smo upoštevali le tvite, ki so zapisani v jezikovno nestandardni slovenščini (L2 in L3) ter tako dobili podkorpus, ki vsebuje okoli četrtino celotnega korpusa, kar je milijon tvitov oz. 14 milijonov besed.

Na podlagi korpusa smo nato naredili frekvenčna leksikona lem in pojavnic. Prvi seznam je za izdelavo geslovnika sicer bolj uporaben, vendar so avtomatsko pripisane leme, posebej za besede v nestandardnem zapisu, velikokrat napačne, zato smo v oporo izdelali še drugega, kar nam je omogočilo tudi identifikacijo variantnih zapisov posamezne leme. Oba seznama smo z metodo frekvenčnega profila (Rayson in Garside, 2000) primerjali z lemami oz. pojavnicami celotnega korpusa Janes v0.3 ter s korpusom Gigafida. Tako smo dobili seznama ključnih besed in besednih oblik tviterščine v obeh virih, kjer vsak izpostavi njene specifične lastnosti, prvi glede na druge spletne uporabniške vsebine, drugi pa glede na splošnejše besedišče. Seznama obeh virov smo združili, v informacijo pa označili tudi tiste leme in oblike, ki se pojavljajo v Slovarju slovenskega knjižnega jezika (SSKJ). Vsaka vrstica v izdelanih seznamih tako vsebuje lemo oz. obliko, pogostost pojavitve na tisoč besed v korpusu Janes Tviti v0.3.4, frekvenco in ključnost glede na korpus Janes v0.3 in Gigafido ter informacijo, ali je lema oz. oblika zastopana v SSKJ.

3.2 Kategorizacija in izbor leksike

Seznama ključnih lem in besednih oblik smo prekrižali tako, da smo identificirali medsebojno povezane oblike, kar nam je predstavljalo izhodišče, t. i. širši geslovnik za jezikoslovno kategorizacijo. S posameznimi leksikalnimi kategorijami, ki jih prikazujemo tudi uporabniku, smo želeli opredeliti način vključevanja prevzete leksike v slovenski oblikoskladenjski sistem (krajšava; nova beseda; stopnja podomačenosti) in izpostaviti pomenske lastnosti, zlasti pomenske premike, ter značilnosti rabe, kot je npr. izbira registra (kletvica, žaljivka) in nestandardnosti. V nadaljevanju opišemo merila za določitev posamezne kategorije in podkategorije ter prenos informacije v spletni slovarček.

3.2.1 Nestandardni zapis in nestandardne besede

Z izrazoma nestandardna beseda in nestandardni zapis besede razumemo besede in zapise besed, ki jih ni mogoče pričakovati v besedilih, ki predstavljajo gradivno osnovo za standardizacijski opis jezika. Gre za besedila, ki sodijo v sfero javne pisne rabe, zlasti s področja javne uprave, izobraževanja in komunikacijskih medijev, ter znanstvena besedila (Skubic, 2005; Frawley, 2003). Za razumevanje pojma nestandardna leksika v slovarju tviterščine je zato treba najprej izpostaviti dejstvo, da je bil izbor tvitov v prvi vrsti vezan na pripis tehnične in jezikovne (ne)standardnosti (gl. poglavje o luščenju podatkov ter Ljubešić et al., 2015), kjer smo se namenoma odločili le za tvite z oznako L2 in L3, tj. za zgolj nestandardne tvite. Prvi kriterij nestandardnosti je torej statistični in ustreza predpostavki, da je določitev jezikovnega standarda lahko vezana le na razmeroma ozek segment jezika, kjer je kodifikacija zaželena in konsenzualno sprejeta (Krek,

2015).[2] Z namenom, da bi besedišče, ki se postopno vključuje v slovenščino zlasti prek angleščine (npr. *frendica*, *čekirati*, *luzer*, *lider*, *biznis*, *kul*) ločevali od že relativno ustaljenih prevzetih besed (npr. *cajteng*, *fršlok*, *kofe*), smo ta segment tviterske leksike dodatno opredelili kot nestandarden, seveda ob zavedanju, da je večina novejše prevzetih besed, ki v slovarju nimajo te opredelitve, prav tako nestandardnih v že zgoraj opredeljenem pomenu te besede.

Pri izdelavi ožjega geslovnika smo iz širšega geslovnika najprej izločili besede, ki so se znašle na seznamu ključnih lem in pojavnic zaradi nestandardnega, pogosto govornega zapisa besede, kar je glede na upoštevano stopnjo nestandardnosti že na ravni korpusa pričakovano. Te besede oz. oblike same na sebi niso posebej značilne za tvite, saj enakovredno pripadajo splošnemu besedišču, npr. *saj, jaz, kar*, *sem* ipd. Razlog, da so bile v postopku luščenja zajete v širši geslovnik, je njihov specifičen zapis, npr. *sj*, *js*, *kr*, *sm* ipd. Te besede nas z vidika pomena in rabe v slovarju tviterščine niso zanimale.

Nestandardne zapise besed smo ločevali od t. i. nestandardnih besed. S to kategorijo smo označevali (a) besede, ki imajo v jeziku prepoznavno standardno različico, npr. *bajta – hiša, crkavati – umirati, izležavati – lenariti*; tudi na ravni izbire registra, npr. *govno – sranje*, in so navadno prevzete iz tujega jezika, zlasti iz srbohrvaščine ali nemščine, npr. *švoh*, *cajteng*, *kao*, *čelav*. Novejših besed, prevzetih iz angleščine in nemščine, načeloma nismo opredeljevali s kategorijo standardnosti, čeprav imajo nekatere prav tako standardno ustreznico, npr. *hengati – družiti se; invajtati – povabiti*, ampak zgolj glede na stopnjo podomačenosti, o čemer več v razdelku 3.2.2. V nekaterih primerih smo kot nestandardne označili tudi (b) besede, ki nimajo ustrezne standardne različice, posledično pa so lahko prisotne tudi v standardnih besedilih, npr. *afnati se*, *pofočkati*, *štopati*, kar je povezano z njihovimi specifičnimi pomenskimi lastnostmi in izbiro registra, kot se kaže npr. v visoki stopnji pozitivnega ali negativnega vrednotenja. Kot nestandardne smo določili tudi (c) besede ki imajo ob standardni ustreznici tudi nestandardno obliko, ki je nastala bodisi kot posledica krnjenja, npr. *depresija → depra*; združevanja, npr. iz besedne zveze: *pornografski film → pornič*, ali izbire obrazila, npr. *penzionist → penzič*, *profesionalec → profič*.

Kot dodatno merilo smo za prepoznavanje nestandardnosti upoštevali lastnosti rabe, kot je denimo (d) izbira registra, npr. *govno, jeben, komunajzar*, *nadrkan*, in (e) pomenske lastnosti besede, npr. *hud* v pomenu '*lep, kakovosten*', *pičiti* v pomenu '*oditi, hitro iti*'. Zadnjo kategorijo bi bilo zato morda ustrezneje poimenovati »nestandardni pomen«. Kot zanimivost je mogoče dodati, da imajo besede, ki smo jih v širšem geslovniku označili s kategorijo »nestandardna«, če so vključene v SSKJ, v njem navadno kvalifikator *pogovorno* in *zlasti v sproščenem ožjem krogu* ter *nižje pogovorno*, *nizko*, *vulgarno*, *slabšalno*, *ekspresivno*, v nekaterih primerih tudi *zastarelo* (npr. *šetati*, *pušiti*, kjer gre tudi za pomenski premik) ali *starinsko* (npr. *glupost*).

3.2.2 Nove besede

Kategorija ***nova beseda*** je besedam v geslovniku pripisana na podlagi podkategorij, ki določajo stopnjo podomačenosti, zato oznake *nova beseda* slovarskim uporabnikom nismo prikazovali (zastopana je v slovarski bazi), je pa na dejstvo, da gre za besedo, ki se postopoma integrira v slovenski jezik, mogoče sklepati iz njene stopnje podomačenosti. Čeprav gre, kot rečeno, v večini primerov za nestandardne besede, ki imajo bodisi standardne ustreznice (npr. *browser – iskalnik*, *comp – računalnik, čekirati – preveriti; prijaviti se*) bodisi se zunaj korpusa tvitov skoraj ne pojavljajo (npr. *dejtati*, *folovati*, *ritvitati*) oz. se pojavljajo zelo redko (npr. *guglati, dron, logirati se*), jim oznake nestandardnosti v slovarju nismo eksplicitno pripisovali. Deloma zato, ker je njihova nestandardnost določena že z izborom besedil, deloma pa zato, ker zaradi relativno kratke prisotnosti v slovenščini njihovega standardizacijskega statusa še ni mogoče predvideti (prim. zlasti besede, kot so *bizarka*, *internetiti*, *odslediti*, *virtualka*, *tiskovka* ipd.).

Kategorijo ***tuje podomačeno*** smo pripisovali besedam, ki poleg enega ali več podomačenih zapisov ohranjajo tudi zapis v izvirniku, npr. *follower – folover*, in glede na to, ali je zapis oz. kateri od variantnih zapisov pisno in/ali glasovno podomačen, npr. *happy – hepi, cute – kjut*. Upoštevali smo tudi pregibanje po slovenskem oblikoslovnem vzorcu, kjer smo bili pozorni na prekrivnost osnovne oblike, kjer glasovna in pisna podomačitev ni potrebna, se pa beseda pregiba po slovenskem sistemu v neimenovalniških sklonih, npr. *link – linka; bed – biti v bedu*, včasih tudi prek postopne pisne in glasovne podomačitve, npr. *junk, džank – junka, džanka*. Kot stopnjo podomačenosti smo upoštevali tudi sposobnost tvorbe novih podomačenih oblik, ki se lahko uveljavljajo postopoma prek pisnega in glasovnega podomačevanja, npr. *follower*, *follover, folower, folover – followat* (najpogosteje), *follovat, folovat – pofollowat* (najpogosteje), *pofollovat pofolovat*. Pri izboru za ožji geslovnik smo upoštevali tudi, ali je katera od variant tipična predvsem za tvite in ali je že opisana v obstoječih slovarjih.

V slovarčku so posamezne variante, če so v korpusu tvitov izkazane vsaj trikrat, predstavljene znotraj variantnega niza (gl. sliko 4) in hkrati kot iztočnice. Na ta način je vsaki varianti na ravni iztočnice pripisana ustrezna kategorija podomačenosti, npr. *kjut* – tuje podomačeno; *cute* – tuje nepodomačeno, hkrati pa so na vsako varianto vezani tudi drugi podatki v slovarskem geslu, npr. potencialne kolokacije in korpusni zgledi. Oznaka ***tuje nepodomačeno*** je tako rezervirana za tiste variantne oblike, ki so glede na podomačeno obliko dovolj pogosto zastopane ali celo prevladujejo, npr. *annoying – anojing, deal – dil*, bodisi podomačene oblike (še) niso razvile, se pa v korpusu pojavljajo razmeroma pogosto, in sicer v slovenskem kontekstu, npr. *hardcore, multitasking* ipd. V to skupino sodijo tudi frazeološke enote, npr. *pitaj boga, lagano sportski, kein problem*.

3.2.3 Novi pomeni

Kategorijo ***nov pomen*** smo uporabljali za označevanje besed, ki so prišle v ožji geslovnik zaradi izkazanega pomenskega premika glede na obstoječi opis v SSKJ ali SNB, npr. *štekati*, *sledilec*, *koma*, *pičiti*. Te informacije v

[2] Tvite je pri izbiri standardizacijsko primernih besedil sicer smiselno upoštevati glede na dejstvo, da se je z razmahom spleta in s prehodom s papirja na zaslon možnost javne objave in dostopa do besedil bistveno povečala ter da je veliko še do nedavnega zasebnih žanrov prešlo v javno sfero (forumi, klepetalnice, družbena omrežja itn.) (Gorjanc et al., 2015a).

slovarskem geslu ne prikazujemo eksplicitno v obliki leksikalne kategorije, pač pa na ravni pomenskega opisa in kolokacij, če so izkazane, ter s tipičnimi korpusnimi zgledi, kar ilustrira slika 1.

štekati
SSKJ: *pogovorno*: razumeti, razumeti se
Novo: občasno prenehati delovati za krajši čas, predvsem v zvezi z elektronskimi napravami in programsko opremo (*a še komu Firefox zadnje čase šteka za popizdit*).

Slika 1: Primer obravnave novega pomena obstoječe besede.

3.2.4 Kratice, krajšave in alfanumerični znaki

Med kraticami in krajšavami ter alfanumeričnimi znaki,[3] kot npr. *gr8, ju3* ipd., smo v ožji geslovnik sprejeli zgolj tiste, ki se ne nanašajo na lastna imena, npr. izdelkov, in sicer ne glede na stopnjo podomačenosti: *iPhone, ajfon, ajfoun*, politična telesa, družbe in podjetja: *DZ, RKC, nyt* (New York Times), na zapise datotečnih formatov ter splošno rabljene kratice in krajšave, npr. *mr., cca, ipd.* itd. Enotna kategorija a ***krajšava***,[4] je tako v slovarskem geslu pripisana tistim občnoimenskim kraticam, okrajšanim besedam in zvezam, ki so za tviterski žanr tipične, npr. *app/ep; omg/omb/omajgad; tnx/thanks/tenks/thnx; bd/bday/rd*, in drugim. Tujejezične okrajšave s podomačenim zapisom imajo v slovarskem geslu v pomenskem razdelku vedno naveden tujejezični ustreznik in slovenski prevod, npr. *gr8/grejt* – krasno (great); *bdw/btW* – mimogrede (by the way).

4 Zgradba gesla in vrsta slovarskih podatkov

Tip slovarske informacije in zgradba gesla so navadno pogojene s preučitvijo uporabnikov in njihovih slovarskih potreb, vendar pa razmisleki v zvezi z naborom slovarskih informacij v slovarju tviterščine zaradi omejenosti projektne aktivnosti ne temeljijo na konkretnih uporabniških izkušnjah ali empiričnih raziskavah. Ob popisu tipične leksike, variant in pomenskih lastnosti smo si prizadevali izpostaviti predvsem tiste lastnosti tviterske leksike, ki jih je bilo mogoče v čim večji meri formalno in objektivno prepoznati v korpusu. Sem sodi zlasti variantnost oblik in stopnja podomačenosti v pisni in glasovni podobi. Dodano vrednost slovarja tako predstavlja kategorizacija na ravni standardnosti (nestandardna beseda) ter stopnje podomačenosti (tuje podomačeno/nepodomačeno), poleg tega pa še opis pomena in rabe, navadno v obliki kratkega pojasnila, ki je pri tujejezičnih besedah največkrat slovenski (*tenks – hvala*), pri nestandardnih pa standard(izira)ni sinonim (*komad – pesem*), ter navedba potencialnih kolokacij, zapis v izvornem jeziku in korpusni zgledi. Za pridobitev najmanj treh korpusnih zgledov smo uporabili aplikacijo GDEX v orodju Sketch Engine, ki je že bila preizkušena pri avtomatskem luščenju leksikografskih podatkov iz korpusa Gigafida (Kosem et al., 2013). Elemente, ki jih predvideva polni geselski članek v slovarju tviterščine so podani v sliki 2.

5 Orodje za izdelavo spletnega slovarja in objava slovarja na spletu

Pri izbiri slovarskega vmesnika smo upoštevali prosto dostopnost, čim večjo neodvisnost pri vnosu podatkov v shemo XML in možnost prenosa slovarskih gesel na lastni strežnik. Ključna je bila tudi fleksibilna nastavitev elementov v zgradbi slovarskega gesla, ki mora omogočati hierarhično ureditev in prilagoditev vizualizacije, saj se zastopanost elementov geselske zgradbe med posameznimi gesli razlikuje. Med možnimi platformami (Termania, Razvezani jezik, Wiktionary, DEBWrite in Lexonomy) smo se odločili za Lexonomy (Měchura, 2012), ki ga podrobneje opišemo v nadaljevanju, hkrati pa smo podatke prenesli tudi v program za izdelavo slovarjev iLex (Erlandsen, 2004), kjer bomo urejali slovarsko bazo.

iztočnica	osnovna oblika
kategorija	krajšava/nestandardna beseda/nova beseda
podkategorija	tuje podomačeno/tuje nepodomačeno
variantni zapisi	osnovne oblike var. zapisov
pomen	
slovenski del	
pomenski opis	standardni sinonim; kratek pomenski opis ali opis rabe; slovenski prevod
kolokacije	
tujejezični del	
izvirni zapis	zapis v jeziku izvirnika
zgledi	
zgled	korpusni zgled

Slika 2: Elementi predvideni v geselskem članku.

Lexonomy [5] je prosto dostopno spletno orodje za izdelavo slovarjev in njihovo neposredno objavo na spletu. Slovarski vmesnik deluje v spletnem okolju, zato namestitev programa na lastni računalnik ni potrebna. Uporabniki lahko pregledujejo objavljene slovarje v iskalniku ali pa si ustvarijo lastni račun in kreirajo svojo bazo podatkov (slika 3). Program omogoča preprosto izdelavo zgradbe slovarske baze, ki jo je mogoče v procesu izdelave slovarja enostavno spreminjati, in podpira skupinsko delo, saj lahko isti slovar ureja več uporabnikov.

Uporabnik lahko izbere že izdelano predlogo za kreiranje preprostega enojezičnega slovarja, mogoče pa si je ustvariti lastno predlogo in podatke kadarkoli objaviti na spletu. Podatke je v formatu XML mogoče izvoziti ali uvoziti, kar omogoča obdelavo in nadgradnjo v drugih programskih orodjih ter združevanje z drugimi podatkovnimi bazami.

Slika 4 prikazuje slovarsko geslo za iztočnico *dafaq*, ki vključuje opredelitev leksikalne kategorije in stopnjo podomačenosti. Sledi niz variantnih zapisov ter pomeni. Vsak registriran pomen lahko vsebuje pomenski opis,

[3] Za zadnje se uveljavlja tudi izraz *kratkopisne kratice* (Logar, 2004).

[4] Za nadpomenko krajšava smo se odločili zato, ker se formalne lastnosti, kot je npr. pika pri okrajšavah, in zapis s samimi velikimi črkami pri kraticah bodisi ne uporablja ali pa se uporablja nerazlikovalno.

[5] Lexonomy: http://www.lexonomy.eu/_en/.

kolokacije ter slovenski prevod ali pa zapis v izvirnem jeziku. Kot rečeno, so korpusni zgledi iz korpusa izluščeni s pomočjo aplikacije GDEX v orodju Sketch Engine. Namen korpusnega zgleda je potrditi registriran pomen in prikazati njegovo rabo ter tipično besedilno okolje na čim bolj avtentičen način.

Trenutni slovarček [6] vsebuje 21 testnih gesel, v nadaljevanju pa nameravamo v program uvoziti približno 1000 iztočnic, skupaj s pripisano leksikalno kategorijo in stopnjo podomačenosti. V program bomo avtomatsko uvozili tudi variantne zapise, pri čemer bo vsak variantni zapis, če je v korpusu dovolj pogost in je njegova raba razpršena med različnimi uporabniki Twitterja, v slovarju predstavljen kot samostojno geslo s podatki, vezanimi zgolj na konkretno varianto.

Slika 3: Izdelava slovarske baze v programu Lexonomy.

Slika 4: Prikaz gesla v Slovarju tviterščine na spletu v programu Lexonomy.

[6] Slovarček je prosto dostopen na http://lexonomy.cjvt.si/slovar-tviterscine/.

Čeprav je ena od osnovnih prednosti programa Lexonomy možnost neposredne objave slovarskega gesla na spletu, je njegova pomanjkljivost predvsem v tem, da ni mogoče vzdrževati razlike med slovarsko bazo, kamor želimo vključiti tudi korpusne metapodatke, kot so npr. tip uporabnika, spol, sentiment, standardnost, regija ter različne statistične vrednosti za posamezno lemo, vendar jih hkrati (še) ne želimo prikazovati navzven oz. jih želimo uporabnikom prikazovati na načine, ki jih v trenutni obliki Lexonomy ne omogoča, npr. v obliki grafov, preglednic ipd. Prav tako je naš namen čim več podatkov v slovarju neposredno povezati s korpusnimi viri in obstoječimi slovarji, ki so dostopni na spletu. Zaradi tega smo se odločili, da bomo slovar tviterščine kot bazo hranili tudi v programu iLex, kamor je mogoče za potrebe združevanja in medsebojnega povezovanja leksikalnih baz vključevati tudi podatke, ki jih ne želimo prikazovati navzven, so pa za leksikalne analize in nadaljnje nadgradnje koristni.

6 Zaključek in prihodnje delo

V nadaljevanju bomo slovarsko bazo nadgradili s podatki slovarskega tipa, kamor sodi oblikovanje pomenskih opisov, izbor relevantnih kolokacij, dodajanje tujejezičnih elementov pri razvezavi kratic in okrajšav ter izbor dobrih zgledov. Hkrati razmišljamo tudi o vključitvi podatkov leksikonskega in slovničnega tipa ter o izboljšavah avtomatskega luščenja podatkov iz korpusa. Pri gradnji slovarske baze tviterske leksike imamo ves čas v mislih možnost integracije v druge slovarske baze, npr. v slovarsko bazo za izdelavo Slovarja sodobne slovenščine (Gorjanc et al., 2015).

Obstoječo slovarsko bazo bomo v prihodnje izkoristili za nadaljnje raziskave tviterske leksike, zlasti za ugotavljanje načina integracije tujejezičnih elementov v slovenski jezik, kjer se uveljavljajo različne možnosti tako na ravni zapisa in morfologije kot tudi na ravni besedotvorja in skladnje. V nadaljnje analize želimo vključiti tudi podatke o tipu uporabnika in njegovi regijski pripadnosti in nenazadnje tudi podatke o analizi sentimenta in druge korpusne metapodatke. Predvidevamo, da bo na tej podlagi mogoče spremljati trend podomačevanja in določiti leksiko, ki se postopno vklaplja v slovenski leksikalni fond.

7 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.

8 Literatura

Irina Álvarez Argüelles in Alfonso Muñoz Muñoz. 2012. An insight into Twitter: a corpus based contrastive study in English and Spanish. *Revista de Lingüística y Lenguas Aplicadas*, 7.1: 37–50.

Thotapally Anjaneyulu. 2013. A glossary: usage abbreviations of mobile phone SMS. *et Cetera*, 70.2: 141.

Naomi S. Baron. 2003. Language of the Internet. Ali Farghali (ur.): *The Stanford Handbook for Language Engineers*. Stanford: CSLI Publications. 59–127.

Claudia Brugman in Thomas Conners. 2016. Comparative study of register specific properties of Indonesian SMS

and Twitter: implications for NLP. *Winter Storm*. College Park, Maryland.

Christian Crumlish et al. 1995. *The Internet Dictionary: The Essential Guide to Netspeak*. SYBEX Inc.

David Crystal. 2001. *Language and the Internet*. Cambridge: University Press.

David Crystal. 2004. *A glossary of netspeak and textspeak*. Capstone.

Tomaž Erjavec in Darja Fišer. 2013. Jezik slovenskih tvitov: korpusna raziskava. *Družbena funkcijskost jezika: (vidiki, merila, opredelitve), Obdobja 32*. Ljubljana: Znanstvena založba Filozofske fakultete, 109–116.

Tomaž Erjavec, Darja Fišer in Nikola Ljubešić. 2015. Razvoj korpusa slovenskih spletnih uporabniških vsebin Janes. *Zbornik konference Slovenščina na spletu in v novih medijih*, Ljubljana, 25.–27. november 2015. Ljubljana: Znanstvena založba Filozofske fakultete, 20–26, http://nl.ijs.si/janes/wpcontent/uploads/2015/11/Konferenca2015.pdf.

Jens Erlandsen. 2004. iLex – new DWS. *Third International Workshop on Dictionary Writing systems (DWS 2004)*. Brno, 6. – 7. September 2004.

Roberta Facchinetti. 2015. English in social media: A linguistic analysis of tweets. *XIV Simposio Internacional de Comunicación Social. Santiago de Cuba*. 19-23.

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

William J. Frawley (ur.). 2003. *International Encyclopedia of Linguistics.* Oxford: Oxford University.

Enrico Giai. (2013). *Twenglish: A New Variety of English? A quantitative analysis of a Twitter based corpus,* http://www.tesionline.com/intl/thesis.jsp?idt=48368.

Vojko Gorjanc, Polona Gantar, Iztok Kosem in Simon Krek (ur.). 2015. *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete.

Vojko Gorjanc, Simon Krek in Damjan Popič. 2015a. *Med ideologijo knjižnega in standardnega jezika*. Ljubljana: Znanstvena založba Filozofske fakultete. 32–48.

Yuheng Hu, Kartik Talamadupula in Subbarao Kambhampati. 2013. Dude, srsly?: The Surprisingly Formal Nature of Twitter's Language. *Zbornik ICWSM* 2013.

Monika Kalin Golob. 2008. SMS-sporočila treh generacij. Miran Košuta (ur.): *Slovenščina med kulturami, Zbornik slavističnega društva Slovenije 19*. Celovec, Ljubljana: Slavistično društvo Slovenije. 283–294.

Iztok Kosem, Polona Gantar in Simon Krek. 2013. Avtomatizacija leksikografskih postopkov. *Slovenščina 2.0*, 1(2): 139–164. http://slovenscina2.0.trojina.si/arhiv/2013/2/Slo2.0_2013_2_07.pdf.

Simon Krek. 2015. Standardni in knjižni jezik – drugi poskus. Smolej, M. (ur.). *Obdobja 34: Slovnica in slovar – aktualni jezikovni opis*. Ljubljana: Znanstvena založba Filozofske fakultete, 401–407.

Nikola Ljubešić, Darja Fišer, Tomaž Erjavec, Jaka Čibej, Dafne Marko, Senja Pollak in Iza Škrjanec. 2015. Predicting the level of text standardness in usergenerated content. *10th International Conference on Recent Advances in Natural Language Processing: Proceedings of RANLP 2015 Conference*, 7–9. September 2015, Hissar, Bulgaria. Hissar: 371–378.

Nataša Logar. 2004. Nove tehnologije in nekateri nesistemski besedotvorni postopki. Kržišnik, E. (ur.) *Obdobja 22: Aktualizacija jezikovnozvrstne teorije na Slovenskem – členitev jezikovne resničnosti.* Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete, 121-132.

Michal Boleslav Měchura. 2012. Léacslann: A platform for building dictionary writing systems. V Ruth Vatvedt Fjeld and Julie Matilde Torjusen (ur.). *Proceedings of the 15th EURALEX International Congress. 7-11 August 2012.* 855-861. Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo.

Mija Michelizza. 2015. Spletna besedila in jezik na spletu: primer blogov in Wikipedije v slovenščini. *Zbirka Lingua Slovenica, 6*. Ljubljana: Založba ZRC, ZRC SAZU.

Isa Na. 2014. *Language Use On Twitter Among Malaysian L2 Speakers*. Doktorska disertacija, University of Malaya Kuala Lumpur.

Paul Rayson in Roger Garside. 2000. Comparing Corpora Using Frequency Profiling. *Zbornik ACL Workshop on Comparing Corpora*. Hong Kong. 1–6. .

Andrej E. Skubic. 2005. *Obrazi jezika*. Ljubljana: Študentska založba.

Nada Šabec. 2011. The Globalizing Effect of English on the Language of the Slovene Media. V Vukanovic, Marija Brala; Krstanovic, Irena Vodopija (ur.). *The Global and Local Dimensions of English: Exploring Issues of Language and Culture*. Berlin: Dr. W. Hopf, 133–126.