JEZIK SLOVENSKIH TVITOV: KORPUSNA RAZISKAVA

Tomaž Erjavec
Institut »Jožef Stefan«, Ljubljana

Darja Fišer
Filozofska fakulteta, Ljubljana

UDK 004.773:811.163.6'276'354

V prispevku predstavimo korpus in analizo nestandardne slovenščine z družbenega omrežja Twitter. Korpus, ki vključuje tvite iz prvih štirih let obstoja omrežja, vsebuje 360.000 tvitov oz. pet milijonov besed. Slovenščina, uporabljena v njih, je zelo bogata in se precej razlikuje od slovenščine, uporabljene v slovenskem uravnoteženem korpusu ccKRES, in sicer predvsem po pogovorni, bolj fonetični ortografiji, rabi prvin, ki so bolj značilne za govorjeni jezik, ter po pogosti rabi tujejezičnih besed.

tviti, nestandardna slovenščina, korpusna analiza, ortografija, besedišče

This paper presents a corpus of Slovene tweets and the analysis of non-standard Slovene as used on the Twitter social network. The corpus, which comprises tweets from the first four years of Twitter's existence, contains 360,000 tweets or 5 million tokens. The Slovene used in the analysed tweets is substantially different from the balanced corpus of standard Slovene ccKRES. The distinguishing features of »Twitter Slovene« are a more colloquial, phonetic orthography, frequent use of spoken language elements and an abundance of foreign words.

Tweets, non-standard Slovene, corpus analysis, orthography, vocabulary

1 Uvod

Družbena omrežja postajajo vse popularnejša tudi med slovenskimi uporabniki. Zaradi tehničnih značilnosti medija in okoliščin, v katerih tovrstna komunikacija poteka, se jezik, ki se uporablja za komunikacijo na družbenih omrežjih, precej razlikuje od standardne slovenščine. Proučevanje rabe slovenščine na spletu je relevantno s sociolingvističnega vidika, analiza nestandardne slovenščine pa je potrebna tudi zaradi zagotavljanja učinkovitih računalniških orodij za procesiranje jezika, saj lahko le tako govorcem omogočimo polno funkcionalnost spletnih orodij, kot so pametni brskalniki, orodja za povzemanje besedil, sintetizatorji govora, strojni prevajalniki ipd.

Jezik, ki ga uporabljamo v komunikaciji preko računalnika, je pod drobnogled vzel Crystal in ugotovil, da t. i. internetni jezik ni ne pisni ne govorjeni, temveč vsebuje značilnosti obeh (Crystal 2001: 47). Čeprav Crystal ni analiziral jezika spletnih družbenih omrežij (Facebook se je namreč pojavil šele leta 2004, Twitter pa dve leti kasneje), bi ga lahko uvrstili v njegovo kategorijo jezika spletnih klepetalnic (prav tam: 129), kamor je sicer uvrstil jezik, ki ga uporabljamo na forumih in dopisnih seznamih. Za tovrstno komunikacijo je značilno, da poteka v realnem času ali z zamikom in pri tem uporablja nestandardno ortografijo (npr. izključno male tiskane črke, opuščanje večine ločil in večkratno ponavljanje črk za čustveno poudarjanje zapisane izjave), nestandarden zapis besed in pogoste specifične okrajšave.

Da je angleški klepetalniški jezik resnično bogat z okrajšavami, akronimi in emotikoni, je pokazala raziskava N. Baron (2003), do

podobnih zaključkov pa sta prišla tudi Kaalep in K. Muischnek (2011), ki sta analizirala rabo internetne estonščine. Rabo nestandardne slovenščine je že raziskovala M. Kalin Golob (2008), ki je analizirala SMS-sporočila, ki so zelo podobna tvitom, in v njih identificirala številne podobne pojave, ki glede na starost govorca in komunikacijsko funkcijo prav tako nihajo od povsem knjižnih do povsem pogovornih ter vsebujejo neknjižno rabo ločil, fonetično pisavo in krajšave, opuščanje pomožnega glagola oz. nedoločnika in emotikone. Da je izsledke analiz različnih besedilnih zvrsti, v katerih je uporabljena nestandardna slovenščina, mogoče posplošiti, dokazujejo tudi rezultati raziskav o rabi slovenščine v elektronskih sporočilih (Dobrovoljc 2008) in na spletnih forumih (Jakop 2008). Ti so precej podobni in v veliki meri veljajo tudi za tvite, čeprav imajo tviti to posebnost, da je njihova dolžina veliko bolj omejena, zato je v njih še toliko bolj poudarjena ekonomična raba jezika, vsem pa je skupno to, da v njih izstopa vpliv narečnih glasovnih posebnosti, nepravopisna raba ločil ter velikih in malih tiskanih črk, prevzete besede iz tujih jezikov, čustveno zaznamovane besede, mašila in krajšave.

V pričujočem prispevku analiziramo jezik slovenskih tvitov oz. kratkih sporočil na enem najpopularnejših družbenih omrežij Twitter, ki jih objavljajo uporabniki omrežja, njihovi sledilci pa jih lahko berejo, nanje odgovarjajo, posredujejo svojim sledilcem, jih citirajo in všečkajo. Uporabniki Twitterja so raznoliki in tvitajo iz zelo različnih razlogov. Na omrežju Twitter imajo račune javne ustanove in osebnosti ter zasebna podjetja, ki s tviti javnost obveščajo o svojih dejavnostih, zasebni uporabniki pa omrežje uporabljajo predvsem za komentiranje dnevnega dogajanja in za zabavo. Zato lahko pričakujemo, da se bodo tviti tako raznolikih uporabnikov med seboj razlikovali tudi po jezikovni plati, ki niha med dvema ekstremoma, od standardne pisne slovenščine do zelo pogovornega, narečno obarvanega slenga. Za Twitter je značilno tudi, da je jezik zaradi tehničnih okoliščin izrazito ekonomičen, saj je največja dolžina tvita omejena na 140 znakov, uporabniki pa pogosto tvitajo tudi preko mobilnih naprav in v situacijah, ki otežujejo dolgovezno in natančno tipkanje (npr. na avtobusu, stoje, med koncertom). Zato pričakujemo, da bo tudi v slovenskih tvitih veliko okrajšav in neupoštevanja pravopisnih pravil, predvsem glede rabe vejice in velike začetnice, prav tako pa tudi izpuščanja šumnikov in samoglasnikov ter zatipkanih besed.

2 Korpus slovenskih tvitov Tweet-sl

Zaradi priljubljenosti tvitov so se pojavili agregatorji, ki jih zbirajo in analizirajo ter ponujajo tiste, ki naj bi bili najbolj zanimivi za določen profil uporabnikov. Z agregatorja sitweet.com so nam prijazno odstopili bazo slovenskih tvitov, ki so nastali med 1. 12. 2007 in 20. 2. 2011, tako da vsebuje obdobje približno štirih let.

Osnovna baza vsebuje veliko tvitov, ki za jezikoslovne analize slovenskih besedil niso zanimivi in predstavljajo šum v podatkih. Zato smo dobljeno bazo najprej prečistili, tako da smo odstranili tujejezične tvite in tvite, ki samo povzemajo vsebino spletnih strani. Da bi iz korpusa izločili neslovenske tvite, smo iz njega izbrisali sporočila, ki ne vsebujejo črk č, š ali ž. S tem smo sicer izgubili nekatera slovenska sporočila, saj smo odstranili tudi taka, v katerih avtorji šumnike iz tehničnih razlogov (angleška tipkovnica na računalniku, počasnejši dostop do šumnikov na mobilnih telefonih) nadomeščajo s črkami c, s in z. Z drugim filtrom smo odstranili sporočila, ki vsebujejo spletne naslove, s čimer smo želeli izločiti vsiljene tvite oz. posredovana reklamna sporočila, ki za našo analizo niso relevantna, saj ne odslikavajo jezika slovenskih tvitov. Iz baze smo nato odstranili tudi imena pošiljateljev in prejemnikov sporočil, s čimer smo sicer izgubili nekaj potencialno zanimivih informacij, a ker je bil naš cilj narediti prosto dostopen korpus, nismo želeli

načenjati problema varovanja osebnih podatkov.

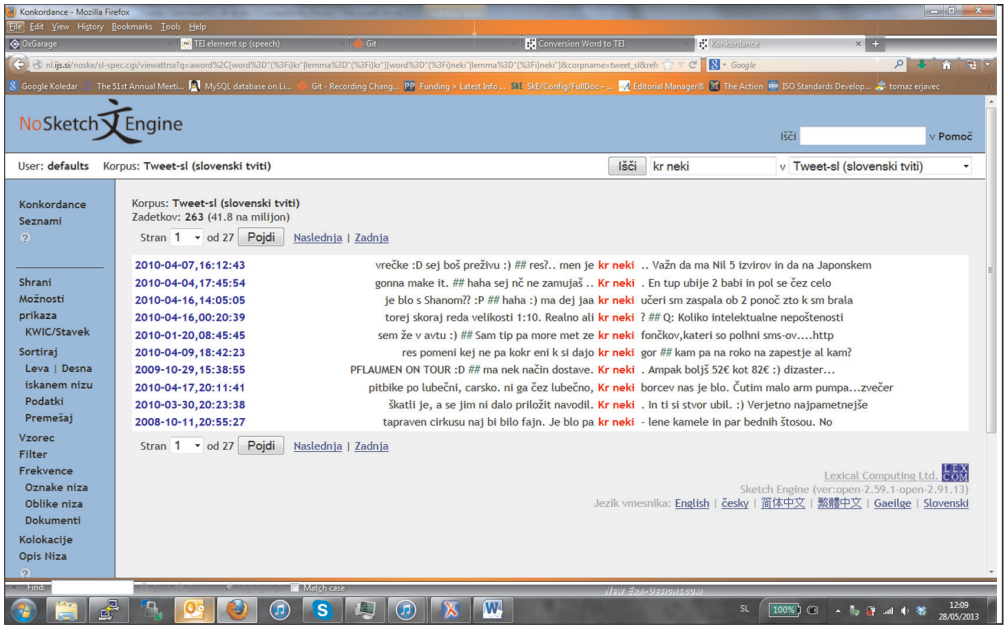
Korpus smo nato avtomatsko jezikoslovno označili, za kar smo uporabili program ToTaLe (Erjavec idr. 2005), ki opravi tokenizacijo, oblikoskladenjsko označevanje (tagiranje) in lematizacijo. Tu je treba poudariti, da je bil model slovenščine, ki ga uporablja ToTaLe, naučen na standardni slovenščini, zato je natančnost pripisanih jezikoslovnih oznak manjša, še posebej pri najbolj zanimivih besedah za našo analizo, torej tistih, ki jih ne najdemo v standardni slovenščini. Pripisane oblikoskladenjske oznake sledijo priporočilom za oblikoskladenjsko označevanje JOS[1] (Erjavec, Krek 2008) in so v korpusu zapisane v angleškem jeziku (npr. *Xf* 'Residual, foreign word'), vendar jih je s pomočjo tabele za konverzijo enostavno prevesti v slovenske (npr. *Nj* 'Ostalo, tujejezična beseda').

Korpus Tweet-sl vsebuje 367.510 tvitov in z upoštevanjem dejstva, da pri tokenizaciji prihaja tudi do napak, 6.405.594 pojavnic, od tega 5.021.853 besednih in 1.383.741 ločil. Besednih različnic je v korpusu 369.983, različnih lem pa 214.887. Korpus je dostopen za pregledovanje in raziskovanje preko spletnih konkordančnikov CUWI in noSketchEngine (Rychlý 2007),[2] pa tudi za prenos.[3] Slika 1 ilustrira osnovno uporabo konkordančnika noSketchEngine skozi iskanje besedne zveze *kr neki*. Oba konkordančnika sta podrobneje opisana v Erjavec (2013) in omogočata bogat nabor iskalnih in prikazovalnih funkcij, npr. iskanje po oznakah in z regularnimi izrazi, izpis frekvenčnih seznamov ali kolokacij ter shranjevanje rezultatov.

3 Korpusna analiza slovenskih tvitov

3.1 Besednovrstne značilnosti

Besednovrstne značilnosti korpusa tvitov smo analizirali tako, da smo pogostost besednovrstnih oznak v njem primerjali s tistimi iz uravnoteženega korpusa ccKRES



Slika 1: Primer konkordanc korpusa Tweet-sl s konkordančnikom noSketchEngine

1 Oblikoskladenjske specifikacije JOS so dostopne na http://nl.ijs.si/jos/msd.
2 Oba konkordančnika sta dostopna preko http://nl.ijs.si.
3 Za dostop do celotnega korpusa kontaktirajte enega od avtorjev.

(Logar Berginc idr. 2012), ki je prostodostopni del uravnoteženega korpusa sodobnega slovenskega jezika KRES. Korpus ccKRES ima približno 10 milijonov besednih pojavnic in je v istem velikostnem razredu kot Tweet-sl, prav tako pa je bil tudi enako avtomatsko jezikoslovno označen.

Analiza pokaže, da je v tvitih bistveno več pojavnic, ki jim je označevalnik pripisal oznako neuvrščeno, pri čemer so na prvem mestu tujejezične pojavnice, sledijo programske napake, izstopajo pa tudi tipkarske napake, ki so na osmem mestu. Medtem ko je v korpusu ccKRES skoraj 1,5-krat več občnih samostalnikov kot v tvitih, v tvitih najdemo veliko več lastnih imen. Jezik tvitov zaznamuje predvsem občutno pogostejša raba medmetov, členkov, prislovov in okrajšav, ki so sicer značilne za govorjeno slovenščino, v uravnoteženem korpusu pa je nekoliko več pridevnikov.

Emotikoni so eden od najbolj značilnih pojavov v komunikaciji na družbenih omrežjih ter so zanimivi, ker pokažejo na polariteto sentimenta posameznega tvita in se tudi velikokrat uporabljajo za avtomatsko zaznavanje sentimenta v besedilih (Smailović idr. 2011). V korpusu smo najpogosteje našli smeške. Ti so bodisi veseli, od katerih so najbolj pogosti :), :D, ;), :P, :)), :-) in =), ali pa žalostni, najpogosteje :(in :-(. Identificirali smo 92 različnih smeškov, pri čemer je pri večini razlika v številu oklepajev (npr. :((, ;))), pa vse do več kot 20 oklepajev). Smeški so uporabljeni v kar 132.641 tvitih, kar predstavlja 36 % vseh tvitov v korpusu. Po drugi strani so negativni smeški veliko redkejši, saj jih vsebuje samo 4.125 tvitov oz. 3,1 % vseh tvitov v korpusu. Poleg smeškov so zelo pogosti še:

- emotikoni za izražanje presenečenja, nejevolje, jeze: :/, :| , :o,
- emotikoni za izražanje naklonjenosti, odobravanja: <3, \o/, \o, xoxo in +1.

Pogosti so tudi simboli, ki imajo v tvitih poseben pomen, kot je na primer # na začetku besede, ki označuje oznako teme sporočila oz. hashtag (npr. *#odmevi*) in @ na začetku besede, ki označuje ime uporabnika (npr. *@huferka*), pisano nestično pa isti simbol pomeni »v«, ki mu sledi navedba lokacije (npr. @ *Tivoli*).

3.2 Ortografske značilnosti

Ortografske značilnosti tvitov smo analizirali na podlagi seznama besednih oblik v korpusu, ki jih lematizator ni znal lematizirati in torej močno odstopajo od besed v učnem leksikonu, torej od standardne slovenščine. Skupaj jih je 10.332; 82 % besednih oblik na tem seznamu je enopojavnic, več kot desetkrat pa se v korpusu pojavi 249 različnih besednih oblik, med njimi je najpogostejših pet: *nism* (1988 pojavitev), *morm* (1388), *rečt* (527), *vidmo* (304) in *vidm* (303).

Med ortografskimi odstopanji je najpogostejši nestandardni (pogovorni oz. narečni) zapis besed, za katerega je značilna bolj fonetična pisava (npr. *dubu*, *bote*, *nouga*, *priem*, *lhka*) in izpuščanje nenaglašenih samoglasnikov (npr. *spomnm*, *vrjamm*, *kšnga*, *vedt*, *smešn*). Fonetičnost je izrazito poudarjena pri nekaterih pogovornih in narečnih besedah, ki so zapisane z neslovenskimi znaki (npr. *ful*, *kul*, *itaq*, *qrac*, *weš*, *prawš*, *al'*, *mal'*). Številne tuje, predvsem angleške besede so zapisane v skladu s slovensko izgovarjavo (npr. *lejm*, *ekšn*, *fensi*, *safr*, *informejšn*), veliko pa jih je zapisanih citatno (npr. *trećі*, *randesvous*, *geeky*). Za čustveno zaznamovane tvite je značilno tudi podvajanje črk, predvsem samoglasnikov, največkrat zadnjih (npr. *jeeeeeee*, *prosiiim*, *zeeeebbbeee*, *čokolaaadaaa*, *počitniceeeeee*). Pri nestandardni ortografiji v korpusu za številne besede najdemo več variant, ki v različni meri odstopajo od norme (npr. *lahko*, *lahk*, *lohk*, *lahka*, *lohka*, *lohko*, *lhk*, *lhko*, *lhku*, *lahku*, *lhka*, *lohku*, *lehko*, *lhke*, *lejko*, *lejku*). Za lažje procesiranje in uporabo korpusa bi morali vse te variante normalizirati tako, da bi jim pripisali isto standardno lemo.

Na seznamu besednih oblik, ki jih lematizator ni znal lematizirati, najdemo tudi

precej okrajšav (npr. *rd*, *ln*, *nm*), med katerimi so nekatere sestavljene tudi iz številčno-črkovnih kombinacij (npr. *mi2*, *5ra*, *3p*), druge pa so prevzete iz angleščine (npr. *lol*, *wth*, *imho*). Številne okrajšave so specifične za Twitter (npr. *RT*, *MT*, *FF*, *DM*) oz. navajajo druga popularna družbena omrežja (npr. *fb*, *fsq*, *flickr*, *tumblr*).

Zelo pogosto je pisanje skupaj, kar se v skladu s pravopisom piše narazen, predvsem prislov *ne* pred osebno glagolsko obliko (npr. *nemorš*, *neboš*, *neveš*, *navš*, *naujo*) in člen *ta* pred posamostaljenim pridevnikom (npr. *tanajlažji*, *tamalim*, *unadva*), kar otežuje tokenizacijo, posledično pa tudi avtomatsko oblikoskladenjsko označevanje in lematizacijo z modeli standardne slovenščine. Pisanje skupaj je v tvitih pogosto tudi takrat, kadar uporabnik preseže maksimalno število dovoljenih znakov 140 in tvit skrajša z brisanjem nekaterih presledkov v sporočilu, največkrat za sicer nestičnimi ločili, včasih pa tudi med besedami, kar lahko oteži razumevanje sporočila, močno pa je prizadeto tudi avtomatsko procesiranje tvitov. Zaradi okoliščin, v katerih tviti nastajajo, so v njih pogoste tipkarske napake, ki prav tako negativno vplivajo na označevanje korpusa, in nestandardna raba malih in velikih začetnic, kar otežuje predvsem avtomatsko označevanje lastnih imen.

3.3 Leksikalne značilnosti

Za analizo leksikalnih značilnosti slovenskih tvitov smo izdelani korpus primerjali s korpusom ccKRES. Za primerjavo smo uporabili metodo frekvenčnega profila

Tabela 1: Prvih 20 lem s primerjalnega seznama korpusa Tweet-sl s ccKRES glede na LL; lema je specifična za tistega od korpusov, ki ima večjo številko v svojem stolpcu (natisnjena krepko)

Lema	LL	Tweet-sl.pm	ccKRES.pm	Tweet-sl	ccKRES
d	61.274	**6,6**	0,15	33.121	1.533
in	35.662	13,38	**28,44**	67.216	284.460
pa	34.472	**20,16**	8,47	101.247	84.747
jaz	24.154	**12,14**	4,68	60.978	46.826
ki	23.344	3,18	**9,98**	15.959	99.813
ja	22.761	**3,31**	0,28	16.603	2.763
p	21.672	**2,49**	0,09	12.499	868
ne	17.611	**14,19**	6,95	71.238	69.470
še	17.360	**10,06**	4,21	50.542	42.105
iti	13.398	**4,68**	1,39	23.490	13.939
a	13.191	**4,72**	1,44	23.709	14.358
danes	11.636	**2,69**	0,53	13.513	5.253
xd	11.220	**1,06**	0	5.308	37
kaj	11.170	**5,3**	1,97	26.637	19.747
v	11.009	16,4	**24,69**	82.362	246.897
že	10.316	**6,18**	2,63	31.036	26.311
če	10.190	**6,62**	2,94	33.263	29.399
jst	10.003	**0,98**	0,01	4.940	89
no	9.795	**1,78**	0,24	8.945	2.423
rt	9.769	**0,94**	0,01	4.741	63
sm	9.074	**0,96**	0,02	4.842	203

(angl. frequency profiling), ki sta jo vpeljala Rayson in Garside (2000), ter z njo poiskali besedišče, ki je najbolj specifično posameznemu korpusu. Najprej smo izdelali frekvenčni seznam lem za vsakega od obeh korpusov, nato pa za vsako lemo izračunali njeno logaritemsko verjetnost (angl. log-likelihood, LL). LL upošteva tako frekvenci elementa kot tudi velikosti obeh korpusov, ki ju primerjamo; večji, kot je, bolj je element značilen za enega od njiju. V tabeli 1 prikazujemo prvih 20 lem z največjimi vrednostmi LL.

Podrobno smo analizirali prvih 500 uvrščenih lem na seznamu in ugotovili, da jih je 66 % značilnejših za tvite. V korpusu ccKRES prednjačijo prvine pisnega standardnega jezika, za katerega so značilni daljši, kompleksnejši stavki (npr. vezniki *ter*, *kateri*, *vendar*, *toda*, *kajti*, *temveč* in prislovi *nato*, *predvsem*, *zlasti*, *namreč*, *tedaj*), pa tudi samostalniki in pridevniki, ki razkrivajo besednovrstno sestavo korpusa ccKRES (npr. pravni jezik: *člen*, *odstavek*, *postopek*, *dejavnost*, *primer*; *evropski*, *državen*, *strokoven*, *javen*, *slovenski*).

Za tvite so značilne predvsem prvine govorjenega jezika, kot so številni pogovorni prislovi (npr. *ful*, *kul*, *fajn*, *super*, *itak*, *jasno*, *baje*, *glih*, *sploh*, *ziher*), medmeti (npr. *hehe*, *ej*, *eh*, *uf*, *jao*, *pač*, *fak*, *mah*, *jes*, *šit*) in prislovi (*pa*, *če*, *da*, *ker*, *ampak*). Samostalnike in glagole, ki izstopajo v tvitih, bi lahko razdelili na tri skupine:

- tematsko obarvani: *vikend*, *petek*, *kava*, *pivo*, *vreme*, *dež*, *sneg*, *sonček*, *karta*, *reklama*; *delati*, *gledati*, *čakati*, *poslušati*, *spati*, *učiti*, *brati*, *priporočati*, *deževati*, *snežiti*;
- pogovorni, narečni in slengovski: *fotka*, *folk*, *komad*, *cajt*, *fora*, *fuzbal*, *šiht*, *žur*, *bus*, *faks*; *rabiti*, *probati*, *ratati*, *pasati*, *štekati*, *zrihtati*, *jebati*, *hrkati*, *laufati*, *furati*;
- žanrskospecifični, med katerimi so številni v angleščini (za številne obstajajo in se uporabljajo tudi slovenske ustreznice, vendar angleški izrazi po izračunu LL izstopajo glede na korpus ccKRES): *tvit*, *link*, *blog*, *email*, *verzija*, *aplikacija*, *profil*, *hashtag*, *update*, *account*; *tvitati*, *klikniti*, *blokirati*, *followati*, *googlati*, *lajkati*, *inštalirati*, *čivkati*, *resetirati*.

Korpusa se močno razlikujeta tudi po rabi okrajšav, saj so za ccKRES značilne: *str.*, *dr.*, *št.*, *g.*, *RS*, v tvitih pa izstopajo: *rt*, *lj*, *lol*, *mb*, *fb*, *cc*, *btw*, *wtf*, *slo*, *tv*, *ng*, *omg*, *ftw*, *ju3*, *tnx*, *ce*, *svn*, *o.o.* Da je jezik tvitov res zelo drugačen od standardne slovenščine, pa ne nazadnje kaže tudi dejstvo, da je delež napak v tokenizaciji, tegiranju in lematizaciji tvitov občutno višji, saj jih je med pregledanimi 500 lemami 22 %, v korpusu ccKRES pa samo 1 %. Največ težav je z lematizacijo pogovorno zapisanih besed (npr. *jst*, *js*, *sm*, *nism kr*, *maš*, *u*, *tut*, *men*, *morm*), pri tokenizaciji največ težav povzročajo emotikoni, ki jih program obravnava kot ločene pojavnice, oblikoskladenjsko označevanje pa je problematično za dvoumne besede, kot so pridevniki in prislovi oz. zaimki in prislovi (npr. *oblačno*, *ta*).

4 Zaključek

V prispevku smo predstavili prvi slovenski korpus tvitov in analizirali njegovo besedišče. Socialna omrežja prinašajo v jezik veliko novosti, ki so zanimive tako jezikoslovno kot tudi s stališča razvoja jezikovnih tehnologij. Korpus Tweet-sl je namenjen ravno takim raziskavam in je dostopen tako preko konkordančnikov kot tudi za prenos.

Pri analizi korpusa smo se osredotočili predvsem na uporabljeno besedišče. V korpusu slovenskih tvitov izstopa raba medmetov, členkov, prislovov in okrajšav ter lastnih imen, s čimer so tviti bolj podobni govorjeni kot pisni slovenščini. Kot je za družbena omrežja značilno, je tudi v analiziranih tvitih zelo pogosta uporaba emotikonov in čustveno zaznamovanega zapisa besed in ločil, pri katerem so iste črke oz. ločila večkrat ponovljeni. Kljub temu da v korpusu najdemo tudi tvite, ki so napisani v povsem standardni slovenščini, večina tviteraševbesede zapisuje pogovorno, uporablja slengizme in narečne izraze, precej pa je tudi tujejezičnih besed, ki

so velikokrat zapisane v skladu s slovensko fonetiko.

Korpus Tweet-sl vsebuje besedila do začetka leta 2011, ki so glede na mladost in hiter razvoj tega medija zdaj že zastarela. Zato si želimo vzpostaviti zajem tvitov in sproti posodabljati korpus. S stališča avtomatskega označevanja so besedila tvitov problematična, saj so trenutni modeli za slovenski jezik naučeni na standardni slovenščini, zaradi česar je avtomatsko označevanje tvitov slabše. Za izboljšanje stanja bi bilo treba ročno normalizirati najpogostejše nestandardne besedne oblike, za druge pa bi lahko uporabili metode normalizacije, kot smo jih za starejšo slovenščino (Erjavec 2011). Tudi analiza besedil bi bila lahko še bolj poglobljena, npr. s primerjavo skladnje s korpusom standardne slovenščine ali primerjava besedišča tvitov z besediščem v korpusu slovenskega govorjenega jezika GOS (Verdonik, Zwitter Vitez 2011).

Zahvala

Avtorja se zahvaljujeta Matiji Rijavcu, ki je zagotovil bazo tvitov s strežnika http://sitweet.com, ter anonimnima recenzentoma za koristne pripombe in nasvete.

Literatura

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

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

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. 197–210.

ERJAVEC, Tomaž, IGNAT, Camelia, POULIQUEN, Bruno, STEINBERGER, Ralf, 2005: Massive Multi-Lingual Corpus Compilation: Acquis Communautaire and ToTaLe. *Proceedings of the 2nd Language & Technology Conference*. Poznan. 32–36.

ERJAVEC, Tomaž, KREK, Simon, 2008: Oblikoskladenjske specifikacije in označeni korpusi JOS. Tomaž Erjavec, Jerneja Žganec Gros (ur.): *Zbornik Šeste konference Jezikovne tehnologije*. Ljubljana: Institut »Jožef Stefan«. 49–53.

ERJAVEC, Tomaž, 2011: Automatic linguistic annotation of historical language: ToTrTaLe and XIX century Slovene. *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. ACL.

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

GRČAR, Miha, KREK, Simon, DOBROVOLJC, Kaja, 2012: Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Institut »Jožef Stefan«. 89–94.

JAKOP, Nataša, 2008: Pravopis in spletni forumi – kva dogaja? Miran Košuta (ur.): *Slovenščina med kulturami, Zbornik slavističnega društva Slovenije 19*. Celovec, Ljubljana: Slavistično društvo Slovenije. 210–219.

KAALEP, Heiki-Jaan, MUISCHNEK, Kadri, 2011: Morphological analysis of a non-standard language variety. *Proceedings of the 18th Nordic Conference of Computational Linguistics NODALIDA*. NEALT: Northern European Association for Language Technology.

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

LOGAR BERGINC, Nataša idr., 2012: *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko, Fakulteta za družbene vede.

RAYSON, Paul, GARSIDE, Roger, 2000: Comparing corpora using frequency profiling. *Proceedings of the ACL Workshop on Comparing Corpora*. Hong Kong. 1–6.

RYCHLÝ, Pavel, 2007: Manatee/Bonito – A Modular Corpus Manager. *1st Workshop on Recent Advances in Slavonic Natural Language Processing*. Brno: Masaryk University. 65–70.

SMAILOVIĆ, Jasmina, ŽNIDARŠIČ, Martin, GRČAR, Miha, 2011: Web-based experimental platform for sentiment analysis. *Proceedings of the 3rd International Conference on Information Society and Information Technologies – ISIT 2011*. Dolenjske Toplice.

VERDONIK, Darinka, ZWITTER VITEZ, Ana, 2011: *Slovenski govorni korpus Gos*. Ljubljana: Trojina, zavod za uporabno slovenistiko.