Korpus tvitov slovenskih politikov Janes TwePo

Urška Bratoš*

* Maribor

Povzetek

V prispevku predstavljamo izdelavo korpusa tvitov slovenskih politikov Janes TwePo. Pojasnjujemo izbiro predmeta raziskovanja, proces pridobivanja virov in oblikovanje končnega seznama ter osnovne metapodatke. S pomočjo konkordančnika Sketch Engine smo nato na podlagi korpusnih metapodatkov izvedli še kratko analizo tviterskega jezika politikov. Analiza teh podatkov je med drugim pokazala, da politiki na Twitterju uporabljajo standardno slovenščino, vendar manj kot pričakovano.

1. Uvod

Družbeno omrežje Twitter je bilo ustanovljeno leta 2006 in je danes s 335 milijonov mesečno aktivnih uporabnikov (statista.com) eno izmed priljubnejših spletnih platform. Med njegovimi uporabniki so tudi politiki, ki Twitter uporabljajo predvsem za širjenje političnih sporočil in samopromocijo v političnih kampanjah (Golbeck et al., 2010). Ker predstavljajo institucionalni glas ljudstva in zasedajo najpomembnejše funkcije v državi, se od politikov pričakuje določena mera profesionalnosti tudi na družbenih omrežjih, katerih prvotni namen je komuniciranje zasebne narave. Eden izmed temeljnih pokazateljev profesionalnosti politikov je uporaba jezika. Dosedanje raziskave kažejo, da se jezik na družbenih omrežjih, kot že prej npr. v SMS-ih (Kalin Golob, 2009), v marsičem razlikuje od pisnega standarda (Erjavec in drugi 2015). Še pred pojavom družbenih omrežij je internetni jezik analiziral Crystal (2001) in ugotovil, da ima značilnosti tako govorjenega kot zapisanega jezika. V njem so pogoste krajšave, uporaba emocionalne ikonografije, spletni neologizmi, neobičajna raba ločil in simbolov ter igriva ponavljanja in grafološke inovacije. Stavki jezika družbenih omrežij so enostavni in zgoščeni, pojavlja se slogovna heterogenost in jezikovna inovativnost (Strehovec, 2003). Za jezik družbeno izpostavljenih posameznikov, ki so izvoljeni predstavniki ljudstva, se zdi, da tovrstnih odstopanj ni pričakovati, vendar poglobljene analize jezika javnih osebnosti ali politikov na tem omrežju na slovenskem gradivu še niso bile narejene.

V prispevku predstavljamo izdelavo korpusa tvitov slovenskih politikov Janes TwePo. Kot kaže že ime, smo besedila pridobili iz gradiva korpusa Janes (Fišer et al., 2016). S pomočjo konkordančnika Sketch Engine (Kilgarriff et al., 2004) smo nato izvedli še korpusno analizo, ki jo omogočajo metapodatki, in sicer z naslednjim ciljem: ugotoviti standardnost, sentiment in ključne besede tviterskega jezika politikov.

2. Gradnja korpusa

V korpus Janes TwePo smo iz korpusa Janes (Fišer et al., 2016) zajeli vse tvite slovenskih politikov, parlamentarnih strank, ministrstev in vlade, napisane v obdobju od 1. junija 2013 do 5. januarja 2016, nato pa zaradi obsega raziskavo omejili le na analizo tvitov posameznih politikov. Ustvarili smo podkorpus Janes TwePo-Posamezniki in vanj vključili le tiste politično delujoče fizične osebe, ki so v izbranem časovnem okvirju opravljale funkcijo predsednika države, predsednika vlade, ministra, poslanca, evroposlanca ali župana. Tvite pravnih oseb, torej institucij, kot so vlada, parlamentarne stranke in ministrstva smo iz analize izključili.

Podatki o predsedniku države, predsednikih vlade, ministrih, poslancih, evropskih poslancih in evropskih komisarjih so dostopni na uradnih spletnih straneh. Pri oblikovanju seznama poslancev in strank smo si pomagali tudi s poročilom o delu državnega zbora v dveh obdobjih (2011–2014 in 2014–2018, dz-rs.si), medtem ko smo podatke o županih pridobili pri pristojni osebi z Ministrstva za javno upravo RS.

Ko smo glede na časovni okvir in politično funkcijo pripravili abecedno urejen seznam politikov, nas je zanimalo dvoje: ali ima politik na Twitterju ustvarjen uporabniški račun in ali je ta profil zaveden v korpusu Janes. Sledil je ročni vnos podatkov za vsakega politika posebej. Velja še omeniti, da smo v korpus vključili tudi politike, katerih računi niso več aktivni, so pa zajeti v korpus Janes, medtem ko v korpus nismo vključili računov s potencialno relevantnimi uporabniškimi imeni, ki pa jih nismo mogli zanesljivo identificirati (so brez slike in profila), prav tako nismo vključili tvitov politikov, ki sicer imajo račun na Twitterju, vendar so objavili manj kot 50 tvitov.

Metapodatki, ki spremljajo vsak tvit v Janes TwePo, so naslednji:

- **uporabniško ime na Twitterju**
- **ime in priimek politika**
- **spol politika**
- **raven politične funkcije:** lokalna, državna ali evropska
- **politična funkcija:** predsednik države, predsednik vlade, minister, poslanec, evropski poslanec, župan
- **politična stranka, ki ji politik pripada** (pri županih stranka, ki ga podpira)

V kategorijah raven *politične funkcije, politična funkcija* in *stranka* je bilo lahko izbranih več vrednosti (npr. v primerih, ko so poslanci Državnega zbora RS postali evropski poslanci, ali ko so politiki z daljšim stažem opravljali več kot eno funkcijo in ko so politiki zamenjali stranko).

Da bi zmanjšali možne napake, ki se ob ročnem preverjanju lahko zgodijo, smo postopek razvrščanja in preverjanja vseh podatkov ponovili 5-krat. Nato smo iz korpusa Janes zajeli vso relevantno gradivo in izdelali korpus. Pri tem smo prevzeli vse obstoječe Janesove metapodatke in oblikoskladenjske oznake ter pripisali še nove metapodatke, naštete zgoraj. Korpus smo naložili na konkordančnik Sketch Engine, po predstavitvi prispevka na konferenci pa ga bomo objavili tudi v repozitoriju CLARIN.SI.[1]

3. Zgradba korpusa in prve analize

3.1. Velikost korpusa

Celoten korpus Janes TwePo vsebuje nekaj več kot 104 tisoč tvitov, ki obsegajo skoraj 1,8 milijona pojavnic (dobrih 1,3 milijone besed) in 75 tisoč lem (Tabela 1). Podkorpus Janes TwePo-Posamezniki vsebuje (ki ne vključuje vlade, ministrstev in parlamentarnih strank), pa vsebuje nekaj več kot 77 tisoč tvitov, ki obsegajo približno milijon besed in 65 tisoč lem, kar predstavlja slabe tri četrtine celotnega korpusa tvitov politikov.

	TwePo	TwePo-Posamezniki
Št. tvitov	104.369	77.643
Št. pojavnic	1.791.166	1.284.212
Št. besed	1.354.241	1.056.858
Št. lem	74.251	64.845

Tabela 1: Velikost korpusa Janes TwePo in Janes TwePo-Posamezniki.

3.2. Spol avtorja tvita

Podkorpus TwePo-Posamezniki vsebuje tvite skupno 78 slovenskih politikov[3] od tega 50 (64 %) politikov moškega spola in 28 (36 %) političark, pri čemer so od skupno 77.643 tvitov, ki so zajeti v naš podkorpus, 62.745 tvitov (81 %) napisali moški, 14.898 (19 %) pa ženske (Tabela 2). V povprečju to pomeni 1255 tvitov na posameznega politika in 532 tvitov na posamezno političarko – povedano drugače: v obravnavanem obdobju je bilo na Twitterju aktivnih za tretjino več politikov kot političark, ki so v skupni obseg tvitov prispevali štirikrat toliko sporočil kot političarke.

Spol	Št. tvitov	Delež v %
Moški	62.745	81
Ženski	14.898	19
Skupaj	**77.643**	**100**

Tabela 2: Število tvitov in delež v korpusu Janes TwePo glede na spol njihovih avtorjev oz. avtoric.

3.3. Raven politične funkcije avtorja tvita

Podatki kažejo, da so največji odstotek tvitov objavili politiki, aktivni na državni ravni, sledijo tviti evropski politikov, nato pa tviti lokalnih politikov (Tabela 3). To seveda ni presenetljivo, saj je tudi število politikov na evropski in lokalni ravni znatno manjše kot na državni. Ustreznejšo primerjavo zato kaže relativni obseg: relativno gledano, "vodijo" župani, ki so objavili 1,6-krat toliko sporočil kot državni poslanci.

Raven politične funkcije	Frekvenca	Delež v %
Slovenska	61.989	78
Evropska	12.055	15
Lokalna	5.094	7
Skupaj	**79.138**	**100**

Tabela 3: Število in delež tvitov v korpusu Janes TwePo glede na raven politične funkcije njihovih avtorjev oz. avtoric.

3.4. Politična funkcija avtorja tvita

Največ tvitov so napisali poslanci Državnega zbora RS, in sicer 54.950, sledijo evropski poslanci z 12.055 tviti. Ministri so objavili 5.686 tvitov, 5.094 tvitov so napisali župani, predsednik države je napisal 3.113 tvitov,[2] 1.465 sta jih napisala premiera (Tabela 4).

Politična funkcija	Št. tvitov	Delež v %
Poslanec	54.950	67
Evropski poslanec	12.055	15
Minister	5.686	7
Župan	5.094	6
Predsednik države	3.113	4
Predsednik vlade	1.465	2
Skupaj	**82.363**	**100**

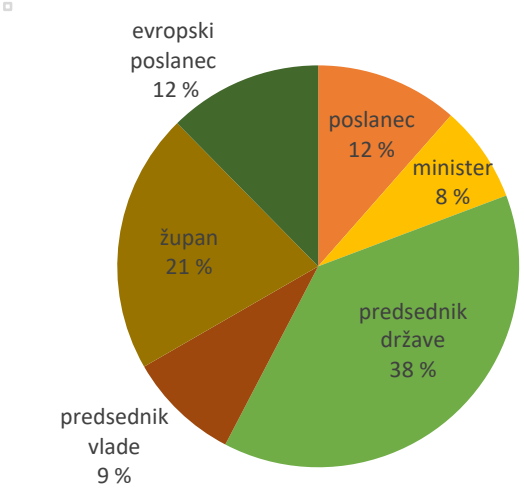
Tabela 4: Število in delež tvitov v korpusu Janes TwePo glede na politično funkcijo njihovih avtorjev oz. avtoric.

Če podatke delimo s številom politikov, ki opravljajo določeno funkcijo, dobimo podatek o tvitersko najaktivnejši politični funkciji: največ tvitov je napisal predsednik države, sledijo župani, predsednika vlade, poslanci in evropski poslanci ter ministri (Slika 1).

[1] http://www.clarin.si/info/o-projektu/

[2] Na profilu predsednika države tvite objavi predsednik sam ali njegova ekipa (označeno s PRS). V analizi so zajeti vsi tviti tega profila.

[3] Celoten korpus Janes TwePo zajema tvite skupno 78 politikov, 9 parlamentarnih strank (SDS, NS, SLS, SD, PS, ZL, ZAB, SMC, DL), 3 ministrstev (Ministrstvo za izobraževanje, znanost in šport RS, Ministrstvo za kulturo RS in Ministrstvo za obrambo RS) in 1 vlade.



Slika 1: Delež tvitov v korpusu Janes TwePo glede na politično funkcijo njihovih avtorjev oz. avtoric.

3.5. Politična stranka, ki ji avtor tvita pripada

V podkorpusu analizirani politiki pripadajo skupno 23 različnim strankam. Podatki v Tabeli 5 kažejo, da so na Twitterju najaktivnejši člani stranke Slovenska demokratska stranka (SDS) z napisanimi nekaj manj kot 28 tisoč tviti. S skoraj polovico manj tvitov, 16 tisoč, sledi Stranka modernega centra (SMC, prej Stranka Mira Cerarja), s skoraj 10 tisoč tviti je na tretjem mestu Državljanska Lista (DL), na četrtem pa so Socialni demokrati (SD) z dobrimi 9 tisoč tviti. Najmanj aktivni so bili v obdobju, ki ga zajema naš podkorpus, nestrankarski kandidati za župane, Združena levica (ZL), stranka Zares in Demokratična stranka upokojencev (Desus).

Politična stranka, ki ji politik pripada	Št. tvitov	Delež v %
SDS	27.604	28
SMC	16.027	16
DL	9.932	10
SD	9.269	9
SLS	4.890	5
PS	4.496	5
Zveza za primorsko ZZP	3.461	4
SMS STRANKA MLADIH ZELENI SLOVENIJE	3.461	4
MAŠA KLAVORA IN SKUPINA VOLIVCEV	3.461	4
NSI IN SLS	2.450	2
NP	2.156	2
SD IN STRANKA EVROPSKIH SOCIALISTOV	2.098	2
ZAAB	1.853	2
NSI	1.781	2
ZBOR ČLANIC IN ČLANOV ZAAB KOMEN	1.594	2
LDS	1.398	1
LISTA DR. IGORJA ŠOLTESA	941	1
DESUS	598	1
ZARES	565	1
ZL	177	0
LEANA TOMIČ IN SKUPINA VOLIVCEV	33	0
TJAŠA MÖDERNDORFER S SKUPINO VOLILCEV	6	0
JOŽE MERMAL IN SKUPINA VOLIVCEV	6	0
Skupaj	**98.257**	**100**

Tabela 5: Število in delež tvitov v korpusu Janes TwePo glede na strankarsko pripadnost njihovih avtorjev oz. avtoric.

3.6. Število tvitov po posameznih avtorjih

Podatki iz podkorpusa kažejo, da je vseh 78 analiziranih politikov objavilo skupno 77.643 tvitov, kar v povprečju pomeni 995,4 tvita na posameznega avtorja.

Največ tvitov je v analiziranem obdobju objavil Marko Pavlišič (poslanec stranke DL, obdobje 3 let in 6 mesecev), in sicer 9.115 (Tabela 6), kar pomeni, da je v povprečju dnevno objavil nekaj več kot 7 tvitov. Drugi na Twitterju najaktivnejši politik je bil Kamal Shaker (poslanec SMC), tretji pa Bojan Krajnc (poslanec SMC). Na petem mestu je bil poslanec SDS Tomaž Lisec, na šestem pa predsednik države Borut Pahor. Opaziti je, da so na tviterju od vseh političnih funkcij najaktivnejši poslanci državnega zbora, in to ne glede na starostno generacijo, in da ne gre za predsednike ali predsednice strank. Izstopa tudi podatek, da so največ tvitov napisali člani stranke SDS, vendar pa so se prav med najaktivnejše politike na Twitterju uvrstili le trije člani te stranke, kar pomeni, da je tvitanje članov SDS "razpršeno". Janez Janša je glede na število objavljenih tvitov zasedel 12. mesto.

Politik	Št. tvitov	Delež v %
Marko Pavlišič (DL)	9.115	23
Kamal Shaker (SMC)	6.355	16
Bojan Krajnc (SMC)	5.956	15
Uroš Brežan (SLS)	3.461	9
Tomaž Lisec (SDS)	3.297	8
Borut Pahor (SD)	3.113	8
Jožef Jerovšek (SDS)	2.596	6
Roman Jakič (PS)	2.340	6
Andrej Čuš (SDS)	2.106	5
Tanja Fajon (SD)	2.098	5
Skupaj	**40.437**	**100**

Tabela 6: Politiki, ki so v korpus Janes TwePo prispevali največ tvitov (prvih 10 mest), ter število in delež njihovih tvitov.

Najmanj tvitov je v analiziranem obdobju objavila ekipa Zorana Jankovića (poslanec PS). Druge gl. v Tabeli 7.

Politik	Št. tvitov	Delež v %
Zoran Janković	6	2
Ljubo Žnidar	17	6
Dragutin Mate	18	7
Danijel Krivec	21	8
Zvonko Lah	26	10
Aljoša Jerič	31	12
Peter Vilfan	33	12
Jasna Gabrič	33	12
Polonca Komar	39	15
Irena Tavčar	42	16
Skupaj	**266**	**100**

Tabela 7: Politiki, ki so v korpus Janes TwePo prispevali najmanj tvitov (zadnjih 10 mest), ter š tevilo in delež njihovih tvitov.

3.7. Tviti politikov: sentiment in standardnost

Iz metapodatkov, vključenih v korpus Janes TwePo (in pridobljenih že iz korpusa Janes, Fišer in Erjavec, 2016; Ljubešić et. al., 2015), je bilo mogoče avtomatsko pridobiti tudi podatke o tem, v katerem jeziku politiki pišejo tvite, kateri sentiment v njih prevladuje in kakšna je raven njihove skladnosti s standardno slovenščino.

V korpusu Janes TwePo je večina tvitov napisana v slovenščini (93 %), manj kot 5 % tvitov je napisanih v angleščini. Slednjo uporabljajo tako poslanci Državnega zbora RS kot evropski poslanci, saj občasno nagovarjajo tuje naslovnike. Ostala 2 % tvitov sta zapisana v nemškem, italijanskem, hrvaškem ali drugem jeziku.
Nadalje nas je zanimala analiza sentimenta besedila, ki pokaže, ali je avtor oz. avtorica tvita temi, o kateri tvita, naklonjena ali ne. Sentiment je lahko označen kot negativen, pozitiven ali nevtralen (Fišer in drugi, 2016). Slaba polovica tvitov, ki so jih objavili politiki, ima v našem korpusu po metapodatkih nevtralen sentiment (45 %), 30 % jih ima negativen sentiment, 25 % pa pozitivnega (Tabela 8).

Sentiment	Št. tvitov	Delež v %
Nevtralen	35.348	45,5
Negativen	23.066	29,7
Pozitiven	19.229	24,8
Skupaj	**77.643**	**100**

Tabela 8: Število in delež tvitov v korpusu Janes TwePo glede na njihov sentiment.

Sentiment smo nato povezali s podatki o spolu avtorjev tvita in ugotovili, da so političarke napisale približno enako število tvitov z negativnim in pozitivnim sentimentom, moški politiki pa so napisali več tvitov z negativnim odnosom do vsebine (Tabela 9).

Sentiment	Spol	Št. tvitov	Delež v %
Negativen	Moški	18.861	82
	Ženski	4.202	18
Skupaj		**23.063**	**100**
Pozitiven	Moški	14.711	77
	Ženski	4.518	23
Skupaj		**19.229**	**100**

Tabela 9: Število in delež tvitov v korpusu Janes TwePo glede na sentiment in spol avtorja oz. avtorice.

O standardnosti jezika tvitov je mogoče ugotoviti naslednje: politiki so v svojih tvitih uporabljali pretežno standardno slovenščino (68 % tvitov, Tabela 10), zelo nestandardnih zapisov je bilo malo (6 %), npr.:

- a) zamenjava šumnikov s sičniki:
  Dragi zupani! Ce ne ze prej, ste danes spoznali ...
  (Anja Bah Žibert, 23. 10. 2013)

- b) nepravilna raba ločil, velikih in malih začetnic:
- *juhuhu se bomo tu... tudi v letu 2015:)*
  (Anja Bah Žibert, 31. 12. 2014)

- c) napačno pisanja skupaj in narazen ali tipkarske napake:
  osebnovkljucit humanitarno, se vam ne zdi???
  (Igor Šoltes, 11. 9. 2015)

- č) uporaba pogovornega jezika:
  Buh jim je zaplosku. #snežet #potres
  (Uroš Brežan, 29. 8. 2015)

Standardnost	Št. tvitov	Delež v %
L1	52.427	68
L2	20.419	26
L3	4.797	6
Skupaj	**77.643**	**100**

Tabela 10: Število in delež tvitov v korpusu Janes TwePo glede na njihovo standardnost.

4. Zaključek

Twitter je v zadnjih letih postal pomemben kanal za politično komuniciranje, zato smo se odločili, da bomo za analizo jezika politikov zgradili korpus. Pri tem nam je bil v pomoč že pripravljeni korpus spletne slovenščine Janes, katerega kar obsežen del so tudi tviti. Kot podkorpus "izločeni" Janes TwePo, zajet v tukajšnjo analizo, tako vsebuje kar 1,2 milijona pojavnic in je tako relevantno dobra osnova za različne jezikovne analize.
V nadaljevanju želimo raziskavo nadgraditi z natančnejšo analizo sentimenta, določiti, kako je sentiment izražen z jezikovnimi sredstvi, in analizirati ključne besede ter podrobneje proučiti na eni strani pravopisna in slovnična, na drugi pa besedna odstopanja od normativnih pravil in nevtralnega stila.

5. 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. Za tehnično pomoč pri izdelavi, označevanju in objavi korpusa se zahvaljujem prof. dr. Tomažu Erjavcu. Prispevek je del magistrske naloge z naslovom *Jezik politikov na družbenem omrežju Twitter in tviti politikov kot vir novinarskega sporočanja*, ki nastaja na Fakulteti za družbene vede pod mentorstvom doc. dr. Nataše Logar in doc. dr. Darje Fišer. Mentoricama se lepo zahvaljujem za branje in popravljanje prispevka.

6. Literatura

David Crystal. 2011. *Internet Linguistic*. London in New York: Routledge.

Darja Fišer in Tomaž Erjavec. Analysis of sentiment labeling of Slovene user-generated content. V: Darja Fišer (ur.), in Michael Beisswenger (ur.). *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities, 27-28 September 2016, Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia*. 1st ed. Ljubljana: Znanstvena založba Filozofske fakultete. 2016, str. 22-25, ilustr. http://nl.ijs.si/janes/wp-content/uploads/2016/09/CMC-2016_Fiser_Erjavec_Analysis-of-Sentiment-Labeling.pdf.

Darja Fišer, Jennifer Golbeck, Justin M. Grimes in Anthony Rogers. 2010. Twitter Use by the U.S. Congress. *Journal of the American Society for Information Science and Technology* 61, 8, 1612–1621.

Monika Kalin Golob. 2009. Med pisnim in govornim ali zgolj po svoje: SMS-sporočila. V: Tanja Oblak Črnič in Breda Luthar (ur.): *Mobilni telefon in transformacija vsakdana.* Ljubljana: FDV. 81–95.

Adam Kilgarriff, Pavel Rychly, Pavel Smrz in David Tugwell. 2004. The Sketch Engine.V: *Proceedings of EURALEX 2004*, str.105–116, Lorient.

Nikola Ljubešič, Darja Fišer, Tomaž Erjavec, Jaka Čibej, Dafne Marko, Senja Pollak, Iza Škrjanec. Predicting the level of text standardness in user-generated content. V: *Proceedings*, International conference Recent Advances in Natural Language Processing, Hissar, Bulgaria, 7-9 September, 2015. Hissar: [s.n.]. 2015, str. 371-378, ilustr. http://lml.bas.bg/ranlp2015/docs/RANLP_main.pdf.

Number of monthly active Twitter users worldwide from 1st quarter 2010 to 2nd quarter 2018 (in millions). *https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/* (2. september 2018).

Janez Strehovec. 2003. *Umetnost interneta. Umetniško delo in besedilo v času medmrežja.* Ljubljana: Študentska založba.

Nada Šabec. 2014. Raba slovenščine in angleščine v fizičnem in virtualnem prostoru. *Slavistična revija*, letnik 62/2014.