Tviti kot leksikografski vir za analizo pomenskih premikov v slovenščini

Darja Fišer, Nikola Ljubešić

Izvleček

V poglavju predstavimo potencial družbenega omrežja Twitter za spremljanje leksikalnih novosti s poudarkom na proučevanju sprememb v rabi že ustaljenega besedišča. Pristop temelji na primerjavi semantičnih profilov besed, zgrajenih iz referenčnega korpusa in iz korpusa tvitov, z metodo distribucijskega modeliranja besed. Rezultate pristopa ovrednotimo z ročno, korpusno podprto leksikografsko analizo, ki jo nadgradimo s predlogom tipologije za samodejno zaznane pomenske premike. Poleg šuma, do katerega prihaja zaradi napak pri predprocesiranju uporabljenih korpusov, predstavljeni pristop izpostavi velik delež dragocenih kandidatov za pomenske premike, med katerimi prednjačijo tisti, ki se v tvitih pojavijo zaradi dnevnih dogodkov in neformalnih sporočanjskih okoliščin.

Ključne besede: leksikalna semantika, distribucijsko modeliranje besed, pomenski premiki leksikografije, družbena omrežja

1 UVOD

Besede skozi rabo ves čas dobivajo nove pomene ali pomenske odtenke ali pa izgubljajo tiste, ki niso več v rabi (Mitra et al. 2015). Pomen besed se tipično spreminja sistematično (Campbell 2004) in se po navadi širi ali oža (Sagi idr. 2009), velikokrat pa besede dobivajo tudi nove pozitivne ali negativne konotacije, procesa, ki ju v leksikalni semantiki imenujemo amelioracija in pejoracija (Cook in Stevenson 2009). Klasičen primer širjenja pomena v slovenščini predstavlja *miška*, ki je tradicionalno pomenila poljsko žival, danes pa z njo poimenujemo tudi računalniško miško. Nasprotni proces se je zgodil s pomenom besede *gas*, ki se je v 19. stoletju uporabljal tudi za plinske ulične svetilke, danes pa ga v nestandardnem jeziku uporabljamo le še za plin. Amelioracijo lahko npr. opazimo pri prislovu *hudo*, ki je v standardni slovenščini negativno konotirana, medtem ko ima v nestandardni rabi izrazito pozitiven naboj. Obratni primer semantičnega premika, pejoracije, opazimo pri samostalniku *blondinka*, ki je v standardni slovenščini nevtralen, v nestandardni pa rabljen skoraj izključno slabšalno.

Detekcija novih pomenov je velik, a pomemben izziv za leksikografijo, ki je nujen za posodabljanje slovarskih gesel glede na sodobno rabo, uporabniške spletne vsebine in družbena omrežja pa so zaradi množične priljubljenosti in živahne jezikovne rabe idealen vir za tovrstne informacije. Aktualen popis semantičnega inventarja potrebujejo tudi različne jezikovnotehnološke aplikacije, kot sta npr. odgovarjanje na vprašanja in strojno prevajanje.

Zaradi obsežnosti jezikovnega gradiva, ki jih je za to nalogo potrebno analizirati, so lahko učinkoviti le avtomatizirani pristopi. V tem prispevku predstavljamo pristop za samodejno detekcijo pomenskih premikov, ki smo ga prvič predstavili že v Fišer in Ljubešić (2016), kjer smo ga preizkusili na manjšem eksperimentu. Ker so bili prvi rezultati spodbudni, smo za namene pričujoče publikacije bistveno nadgradili jezikoslovno analizo kandidatov, ki je sedaj opravljena na obsežnejšem korpusnem gradivu, poenotenih protokolih gradnje besednih skic v obeh korpusih ter z izboljšano tipologijo in smernicami za razvrščanje kandidatov. Zaradi boljšega razumevanja razvite metode in prizadevanj za njene nadaljnje izboljšave smo za to poglavje nekoliko spremenili tipologijo pomenskih premikov. V kategorijo kandidatov, pri katerih z ročno analizo ni bilo zaznanega pravega pomenskega premika, smo uvedli podkategoriji za kandidate, ki so se na vrh seznama uvrstili zaradi napak pri predprocesiranju korpusa, in dejanske lažne kandidate (glej razdelek 4.3). Zato se rezultati analize, predstavljene v tem poglavju, razlikujejo od tistih, ki smo jih objavili v predhodnih publikacijah, kar se kaže predvsem v zmanjšanem deležu identificiranih večjih premikov proti povečanemu deležu identificiranega šuma v predprocesiranju korpusov.

2 SORODNE RAZISKAVE

Medtem ko so samodejno prepoznavanje pomenov besed temeljito proučevali že številni raziskovalci (Sparck Jones, 1986; Ide in Véronis 1998, Schütze 1998, Navigli 2009), so avtomatski pristopi k detekciji pomenskih premikov še vedno razmeroma slabo raziskani, čeprav so zelo pomembno teoretično-analitično vprašanje v leksikalni semantiki pa tudi dragocen aplikativen izziv v sodobni leksikografiji, ki bi lahko prispeval k lažjemu in hitrejšemu posodabljanju slovarskih gesel, kar je cilj, ki z vse večjo dostopnostjo diahronih, tematsko in žanrsko specifičnih korpusov postaja vse bolj dosegljiv.

Večina raziskav na področju samodejnega prepoznavanja pomenskih premikov se osredotoča na diahrono sledenje sprememb v rabi in pomenu besed na podlagi zelo obsežnih zgodovinskih korpusov, ki zajemajo besedila izpred nekaj desetletij ali celo stoletij (Mitra et al. 2015; Tahmasebi et al. 2011; Hamilton et al. 2016). Drugi priljubljeni pristop je primerjava besednih pomenov v dveh ali več korpusih, ki vsebujejo besedila iz različnih časovnih obdobij ali žanrov. Cook et al. (2013), na primer, nove besedne pomene identificirajo s primerjavo t. i. *ciljnega korpusa* z *referenčnim korpusom*, za kar uporabijo metode tematskega modeliranja za indukcijo besednih pomenov. Preprostejši in potencialno robustnejši pristopi ne zahtevajo predhodnega razlikovanja med specifičnimi pomeni, temveč merijo kontekstualno razliko leksema v dveh ali več korpusih. Gulordava in Baroni (2011), na primer, pomenske premike merita s pomočjo distribucijske podobnosti med besednimi vektorji, zgrajenimi iz dveh različnih korpusov. Podoben pristop uporabimo tudi v tem poglavju, v katerem uporabimo distribucijsko modeliranje za prepoznavanje novih pomenov v jeziku slovenskih tvitov.

3 METODA

Pristop, uporabljen v pričujočem poglavju, temelji na distribucijskem modeliranju pomena besed. Za zaznavanje pomenskih premikov je ključna gradnja in primerjava dveh distribucijskih semantičnih modelov za vsako iztočnico iz dveh korpusov: prvega za rabo iztočnice v splošnem jeziku slovenščini, za kar smo uporabili referenčni korpus Gigafida,[1] drugega pa za iztočnico, kot se uporablja v spletni slovenščini, za kar smo uporabili korpus Janes-Tweet (glej Erjavec et al. 2018).

1 http://www.gigafida.net

Za gradnjo in primerjavo distribucijskih modelov smo uporabili orodje *word2vecf* (Levy in Goldberg 2014b).[2] Kot kontekstne značilke smo upoštevali površinske oblike in se tako izognili pogostemu šumu, do katerega prihaja pri označevanju in lematizaciji nestandardnih besedil. Značilke smo zajeli iz kontekstnega okna dveh besed na vsaki strani iztočnice, pri čemer ločil nismo upoštevali. Relativnega položaja iztočnic nismo kodirali.

Na ta način smo dobili vektorske predstavitve 200 dimenzij za vsako od 5425 lem, ki se v korpusu Janes-Tweet pojavijo vsaj 500-krat. Z znižanjem tega precej strogega frekvenčnega praga bi sicer zlahka pridobili večji nabor besed, a smo ta kriterij uporabili, ker se v pričujoči raziskavi osredotočamo na splošno besedišče, ki je pogosto v različnih žanrih.

Za izračun pomenskih premikov smo uporabili kosinusno podobnost, pretvorjeno v mero razdalje (kosinusno podobnost odštejemo od 1) med vektorjema za iztočnico, zgrajenima iz standardnega in nestandardnega korpusa. Pri tem smo izhajali iz predpostavke, da je razdalja med vektorjema iztočnice, ki se v obeh korpusih uporablja v istem pomenu (npr. *banana*), mnogo manjša med vektorjema iztočnice, ki se pojavlja v različnih pomenih (npr. *miška*).

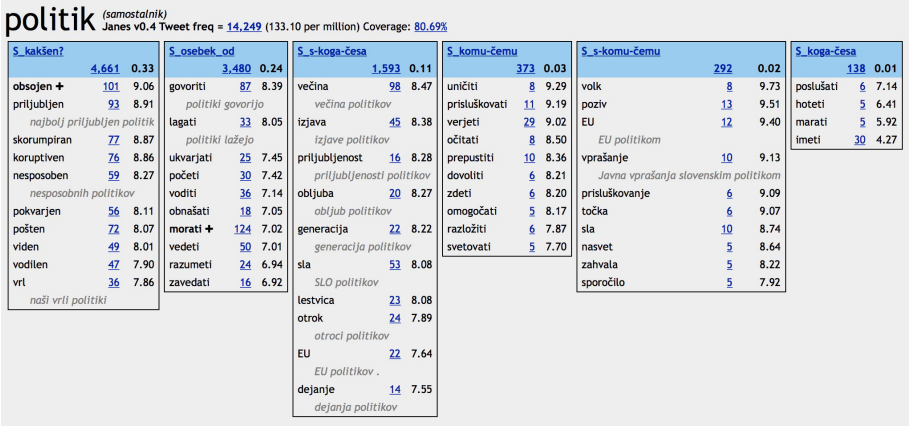
Predstavljena metoda je razmeroma preprosta in ne upošteva dejstva, da je večina besed v korpusu uporabljenih v številnih pomenih, prav tako pa tudi ne ločuje med različnimi vrstami pomenskih premikov. Vendar je za alternativni pristop potrebno predhodno razdvoumljanje večpomenskih besed, ki je že sama na sebi zelo zahtevna naloga, tako da bi v naš postopek v podatke vnašala precej šuma, še posebej, ker delamo z nestandardnim jezikom. Dodatna omejitev razdvoumljanja je, da ne zmore prepoznati novih pomenov, ki so eden naših glavnih ciljev. Ne glede na to smo prepričani, da je predlagani preprost pristop lahko neposredno uporaben za leksikografsko delo, saj izpostavi lekseme, ki so bodisi (1) uporabljeni v različnih pomenih bodisi (2) imajo drugačno frekvenčno distribucijo pomenov v obeh korpusih, kar je oboje pomembno za opis rabe besed. Tako preprost in robusten pristop bi bilo enostavno integrirati v leksikografski delotok (Gantar et al. 2015).

4 ANALIZA

Metodo smo preizkusili z ročno analizo 200 lem, katerih konteksti se v referenčnem korpusu in korpusu tvitov najbolj razlikujejo. Za to smo uporabili primerjavo besednih skic iste leme v obeh korpusih v orodju Sketch Engine (Kilgarriff et. al. 2014), ki temeljijo na slovničnih vzorcih za slovenščino, ki sta jih razvila

2 https://bitbucket.org/yoavgo/word2vecf/

Krek in Kilgarriff (2006). Besedne skice so povzetki slovničnega in kolokacijskega vedenja iztočnice. Prikazujejo kolokatorje iztočnice in so razvrščene glede na slovnična razmerja, na primer na besede, ki so predmet glagolu, besede, ki služijo kot osebek glagola itd., kot ponazarja Slika 1.



Slika 1: Besedna skica za besedo »politik« v korpusu Janes-Tweet. Oznake na vrhu vsakega stolpca so imena slovničnih relacij, npr. S_kakšen? (pridevnik + samostalnik). Sivo obarvane fraze prikazujejo, kako se iztočnica povezuje s svojimi kolokatorji, npr. »najbolj priljubljen politik«. Kolokacije v krepkem tisku in znakom + ponujajo nadaljnje besedne skice za večbesedno zvezo, npr. »pravnomočno obsojen politik«. Število pojavitev pri vsaki kolokaciji vsebuje povezave na konkordance.

S primerjavo besednih skic v korpusih Janes-Tweet in Gigafida smo izvedli analizo pomenskih premikov, kot ponazarja Tabela 1. V obeh korpusih smo izračunali besedne skice za besedo »pirat«, pri čemer smo v obeh korpusih upoštevali le tiste kolokatorje, ki se z iztočnico pojavijo vsaj petkrat, in število kolokatorjev v posamezni semantični relaciji omejili na 25. Za razvrščanje kolokatorjev smo uporabili asociativno mero logDice. Čeprav zaradi prostorskih omejitev v tabeli navajamo samo pet najmočnejših kolokatorjev treh najbolj produktivnih besednih skic za iztočnico, smo v analizo zajeli vse slovnične relacije in kolokatorje v obeh korpusih. Po potrebi smo analizirali tudi konkordance kolokatorjev.

Tabela 1: Najpogostejših pet kolokatorjev za tri najproduktivnejše slovnične relacije besede »pirat« v korpusih Janes in Gigafida. Besede v krepkem tisku zaznamujejo nov pomen v korpusu Janes, ki v korpusu Gigafida ni izkazan.

Janes			**Gigafida**		
iztočnica: pirat pogostost: 1.034 (9,65 na milijon) pokritost leme:[3] 69,05 %			iztočnica: pirat pogostost: 9.941 (7,05 na milijon) pokritost leme: 86,37 %		
Relacija	Kolokator	Frekv. / logDice	Razmerje	Kolokator	Frekv. / logDice
Pridevnik + iztočnica	somalski	7 / 10,48	Pridevnik + iztočnica	somalijski	203 / 11,19
	somalijski	5 / 9,61		somalski	48 / 9,10
	islandski	**6 / 8,80**		zdelan	28 / 8,29
	vesoljski	6 / 7,70		karibski	60 / 8,18
	spleten	8 / 3,97		novodoben	38 / 6,71
Glagol + iztočnica v rodilniku	**voliti**	**6 / 7,45**	Glagol + iztočnica v rodilniku	preganjati	20 / 6,65
	podpreti	**5 / 6,11**		kaznovati	6 / 5,12
	/	/		snemati	15 / 5,07
	/	/		loviti	12 / 4,56
	/	/		prehiteti	6 / 4,44
Samostalnik + iztočnica v rodilniku	**sestanek**	**5 / 8,05**	Samostalnik + iztočnica v rodilniku	bitka	18 / 7,15
	/	/		zatočišče	9 / 7,05
	/	/		preganjanje	18 / 6,69
	/	/		jahta	6 / 6,68
	/	/		prekletstvo	8 / 6,53

Kot je razvidno iz kolokatorjev v Tabeli 1, lahko iz korpusa Gigafida razberemo tri pomene, kolokatorji nekaterih se lahko prekrivajo:

1) oseba, ki ropa ladje (npr. »somalijski«, »jahta«, »zatočišče«),
2) oseba, ki nezakonito razmnožuje zaščitene vsebine (npr. »novodoben«, »preganjati«, »kaznovati«) in
3) metaforično / knjiga, film, naslov TV-oddaje (npr. »zdelan«, »prekletstvo«, »snemati«).

V korpusu Janes na podlagi kolokatorjev v besedni skici zaznamo podobne pomene:

1) »somalski«
2) »spleten« in
3) »vesoljski«.

3 Pokritost leme je mera, ki ponazarja delež vseh pojavitev obravnavane leme v uporabljenem korpusu, ki so bile upoštevane pri gradnji besedne skice.

Poleg njih pa analiza konkordanc za kolokatorje, kot so »islandski«, »voliti«, »podreti « in »sestanek«, ki se v besedni skici iz korpusa Gigafida ne pojavijo, kažejo na nov pomen besede v tvitih, objavljenih v letih 2014 in 2015, in sicer:

4) osebe, ki so člani novih političnih strank iz Slovenije in drugih držav Evropske unije.

Ta pomen v korpusu Gigafida ni izkazan, saj korpus vključuje samo besedila, ki so bila objavljena do leta 2011, politično gibanje pa je dobilo zagon po uspehu na volitvah; v Nemčiji leta 2011, na Islandiji leta 2013 in v EU leta 2014.

Seveda pa vse razlike med korpusoma še zdaleč ne pomenijo nujno novih pomenov, temveč izkazujejo tudi subtilnejše razlike v rabi, kot sta oženje pomena in redistribucija pogostosti pomenov zaradi razlik v temah, žanrih in registrih, ki so v korpusih zastopani. Zato v ročni analizi razlikujemo med večjimi in manjšimi pomenskimi premiki, ki jih nadalje delimo tudi v podkategorije:

1) Večji pomenski premiki
   a. Vezani na dnevne dogodke
   b. Vezani na razlike v registru[4]
   c. Vezani na razlike v mediju
2) Manjši pomenski premiki
   a. Vezani na distribucijo pomenov
   b. Vezani na omejenost rabe na določene ustaljene vzorce
   c. Vezani na oženje pomena
3) Napake
   a. Zaradi šuma pri predprocesiranju korpusa
   b. Lažni kandidati

Rezultate analize večjih in manjših pomenskih premikov predstavljamo v razdelkih 4.1 in 4.2. Kot pri vsakem samodejnem postopku je tudi pri detekciji pomenskih premikov pričakovati določeno stopnjo šuma, ki se lahko pojavi v kateri koli fazi predprocesiranja korpusa ali pa zaradi pomanjkljivosti predlagane metode. Te primere obravnavamo v razdelku 4.3.

4.1 Večji pomenski premiki

Za boljši uvid v pomenski odtis besedišča, ki glede na referenčni korpus prikazuje največje razlike v rabi na Twitterju, ločujemo med pomenskimi premiki, ki so

4 Z izrazom register opredeljujemo rabo konvencionalizirane rabe jezika, skladne s specifičnimi sporočanjskimi funkcijami in družbenimi okoliščinami (Lee 2001).

vezani na dnevne dogodke, takšnimi, do katerih prihaja zaradi razlik v registru, in tistimi, ki so značilni za medij.

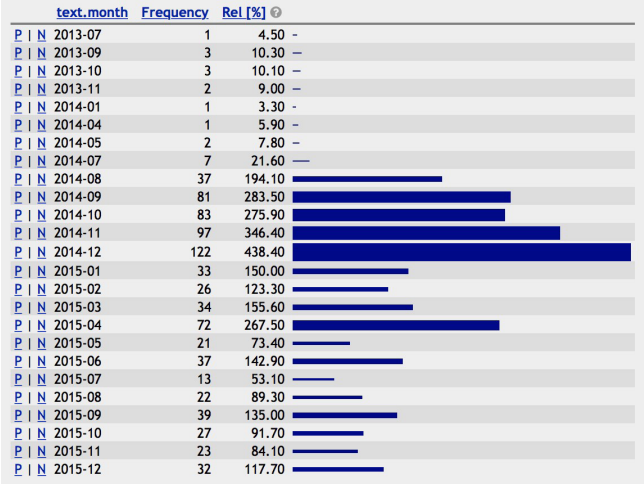
4.1.1 Pomenski premiki, vezani na dnevne dogodke

V to kategorijo uvrščamo novo rabo besed, do katere je prišlo zaradi dnevnih dogodkov, političnih razmer, naravnih katastrof in družbenih okoliščin. Eden takšnih primerov je omenjeni primer »pirat«, ki je bil včasih povezan izključno z morjem, zadnje čase pa je povezan tudi z internetom in celo s politiko, saj označuje člane nove politične stranke, in sicer v izrazito pozitivnem kontekstu (glej Tabelo 1). Drug tak primer je samostalnik »vztrajnik«, ki se v Gigafidi pojavlja precej redko in vedno s pomenom 'del stroja', medtem ko se na Twitterju uporablja precej pogosteje in se skoraj izključno nanaša na 'protestnike'. Najbolj produktivne slovnične relacije in najmočnejši kolokatorji za to iztočnico iz obeh korpusov so prikazani v Tabeli 2.

Tabela 2: Najpogostejših pet kolokatorjev za tri najproduktivnejše slovnične relacije za besedo »vztrajnik« v korpusih Janes in Gigafida. Besede v krepkem tisku označujejo nov pomen v korpusu Janes, ki v korpusu Gigafida ni izkazan.

Janes			**Gigafida**		
iztočnica: vztrajnik pogostost: 816 (7,62 na milijon) pokritost leme: 72,79 %			iztočnica: vztrajnik pogostost: 423 (0,30 na millijon) pokritost leme: 87,47 %		
Relacija	Kolokator	Frekv. / logDice	Pridevnik + iztočnica	Kolokator	Frekv. / logDice
Pridevnik + iztočnica	**drag**	**7 / 4,53**	Pridevnik + samostalnik Iztočnica v imenovalniku + glagol	dvomasen	12 / 11,41
	pravi	**5 / 2,24**		vrteč	5 / 6,24
	/	/		magneten	16 / 5,37
	/	/		lahek	15 / 2,45
	/	/		težek	7 / 0,50
Iztočnica v imenovalniku + glagol	**zapeti**	**7 / 8,81**	Osebek + glagol	skrbeti	6 / 1,97
	vztrajati	**5 / 8,05**	Samostalnik + iztočnica v rodilniku	/	/
	/	/		/	/
	/	/		/	/
	/	/		/	/
Samostalnik + iztočnica v rodilniku	**Viktor**	**8 / 10,78**	Samostalnik + samostalnik v rodilniku	motor	13 / 3,43
	Odbor	**10 / 7,33**		sistem	5 / 0,30
	/	/		/	/
	/	/		/	/
	/	/		/	/

Zanimivo je, da redki najzgodnejši primeri rabe besede »vztrajnik« iz korpusa Janes-Tweet, ki so nastali še pred obdobjem političnih in družbenih nemirov v letih 2013–2014, pripadajo tehničnemu pomenu besede, kot v referenčnem korpusu Gigafida. Na Sliki 2 lahko opazimo izrazito povečanje rabe te besede v korpusu tvitov leta 2014, ko je od začetnih 3 pojavitev od oktobra 2013 do oktobra 2014 narasla na 83, višek pa dosegla decembra 2014 (122) in nato decembra 2015 ponovno upadla na 32. Porast rabe besede sovpada z obdobjem protestov proti obsodbi Janeza Janše, ko opazimo tudi porast protestniškega pomena, ki se najprej pojavi skupaj s tehničnim in nato popolnoma prevlada, tako da med zadetki iz decembra 2015, ki je zadnji mesec, zajet v korpusu, ni zaslediti nobene rabe prvotnega pomena več.



	text.month	Frequency	Rel [%]
P | N	2013-07	1	4.50
P | N	2013-09	3	10.30
P | N	2013-10	3	10.10
P | N	2013-11	2	9.00
P | N	2014-01	1	3.30
P | N	2014-04	1	5.90
P | N	2014-05	2	7.80
P | N	2014-07	7	21.60
P | N	2014-08	37	194.10
P | N	2014-09	81	283.50
P | N	2014-10	83	275.90
P | N	2014-11	97	346.40
P | N	2014-12	122	438.40
P | N	2015-01	33	150.00
P | N	2015-02	26	123.30
P | N	2015-03	34	155.60
P | N	2015-04	72	267.50
P | N	2015-05	21	73.40
P | N	2015-06	37	142.90
P | N	2015-07	13	53.10
P | N	2015-08	22	89.30
P | N	2015-09	39	135.00
P | N	2015-10	27	91.70
P | N	2015-11	23	84.10
P | N	2015-12	32	117.70

Slika 2: Mesečna frekvenčna distribucija besede »vztrajnik« v korpusu Janes-Tweet.

Analiza računov kaže na izrazito lokalizirano uporabo te iztočnice, ki je vezana na majhen krog uporabnikov. Med korporativnimi računi (glej Erjavec et al. 2018) se skoraj vse pojavitve (157 ali 67 %) pojavijo v objavah političnega gibanja Odbor 2014, ki se je zavzemal za izpustitev Janeza Janše, ostale pa v objavah osrednjih in lokalnih pisarn stranke SDS ter časopisov, revij in portalov, ki podpirajo SDS (Demokracija, Politikis.si, Reporter). Med zasebnimi računi je najpogostejših 10 uporabnikov besede, podpornikov Janeza Janše oz. stranke SDS, poskrbelo za 24 % vseh pojavitev v korpusu.

Podrobnejši pregled čustvene zaznamovanosti tvitov, ki vsebujejo besedo *vztrajnik*, nam pokaže zelo zanimivo sliko. Izkaže se, da ima izrazito pozitivno konotacijo. Primer:

»Mnogoštevilni vztrajnice in **vztrajniki** so ponosno zapeli slovensko in evropsko himno. #svobodaJJ«

Po drugi strani pa je konotacija na videz podobnega izraza »vstajnik«, ki ga večinoma isti uporabniki uporabljajo mnogo pogosteje (frekvenca 1767; 16,51 na milijon) v pogovorih o protivladnih protestih, ki so se začeli v času, ko je bil Janša še vedno premier, izrazito negativna.

Vstajniki ste zombiji, vstajniki ste placanci, vstajniki ste levi fasisti, vstajniki ste ovce, ampak nikoli pa ni bilo receno, da ste drhal.

4.1.2 Pomenski premiki, vezani na register

V korpusu tvitov najdemo veliko pomenov, ki v referenčnem korpusu niso izkazani, saj se preko Twitterja odvija veliko neformalne komunikacije, kar prav tako vpliva na semantični odtis besed. Takšen primer je samostalnik »penzion«, ki v standardni slovenščini pomeni *gostinski obrat*, vendar se v nestandardnem jeziku uporablja tudi v pomenu *upokojitev*, kar prikazuje izvleček iz besednih skic v Tabeli 3. Zaradi minimalnega frekvenčnega kriterija petih pojavitev besedne skice iz korpusa Janes niso zelo informativne, vendar vzorec predlog + »penzion« jasno kaže, da v njem poleg pomena *gostišče* zasledimo tudi pomen *upokojitev*.

Tabela 3: Primeri konkordanc za edino produktivno slovnično razmerje za besedo »penzion« v korpusih Janes in Gigafida. Besede v krepkem tisku označujejo nov pomen v korpusu Janes, ki v besednih skicah iz Gigafide ni izkazan.[5]

Janes			**Gigafida**		
iztočnica: penzion pogostost: 1.073 (10,02 na milijon) pokritost leme: 97,48 %			iztočnica: penzion pogostost: 4.898 (3,47 na milijon) pokritost leme: 92,81 %		
Slovnična relacija	Kolokator	Freq / logDice	Slovnična	Kolokator	Freq / logDice
Predlog + iztočnica	**iz**	**127 / 5,04**	Predlog + iztočnica	izpred	107 / 8,22
	Primer: moja mam bi šla *iz Penzion* paketa na Enostavni 300			Primer: odhod bo ob 17. uri izpred *penziona* Špik	
	v	**589 / 4,25**		pred	53 / 0,72
	Primer: ker sem se odloču da se mi ne da več, grem z naslednjim letom *v penzion*			Primer: koncert narodnozabavne skupine Gašperji na plaži pred *penzionom* Tiha dolina	
	do	**9 / 1,73**		v	757 / 0,44
	Primer: vsako leto mi manjka več *do penziona*			Primer : Prenočiti je možno le v *penzionih* ali najeti počitniško hišico.	

5 Se pa pojavi v posameznih konkordancah, npr. zvezah »v penzion«.

Ko kriterij najmanj petih sopojavitev sprostimo, podoben primer rabe prikazujeta tudi naslednja vzorca:

- pridevnik + »penzion«: »(ne)zaslužen«, »priviligiran«, »invalidski«, »zaslužen«, »prisilen«, »predčasen« in
- glagoli, pri katerih iztočnica »penzion« nastopa kot predmet: »izplačevati«, »prislužiti«, »uživati«, »dočakati«, »zaslužiti«.

Tovrstna neformalna jezikovna raba je še posebej dragocena, ker ni zadostno pokrita s tradicionalnimi leksikalnimi viri, zastopana ni niti v večini obstoječih korpusov za slovenščino. Z naraščanjem obsega in pomena komunikacije na družbenih omrežjih postaja vse pomembnejše tudi proučevanje tega segmenta jezika, prav tako bi ga bilo potrebno ustrezno vključiti v sodobne leksikalne vire. Zagotavljanje pokritosti nestandardnega jezika je nujno tudi za robustno avtomatsko procesiranje šumnih spletnih besedil.

4.1.3 Pomenski premiki, vezani na medij

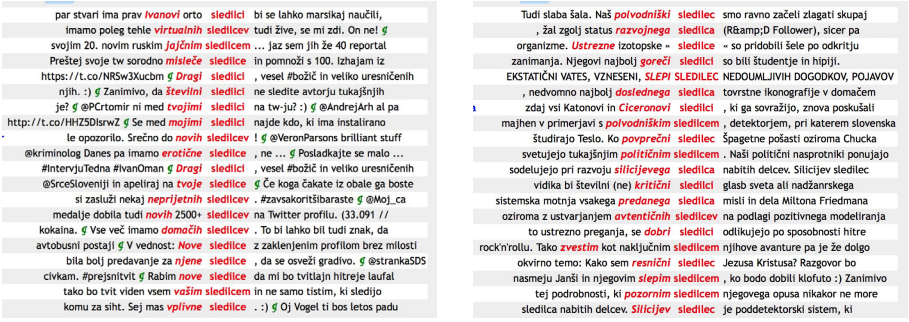
Zadnja skupina večjih pomenskih premikov so nove sporazumevalne konvencije, ki so se pojavile na družbenih omrežjih in so si nekaj obstoječega besedišča prisvojila za nove, specializirane namene. Primer tega pojava je samostalnik »sledilec« (angl. *follower*), pri katerem lahko jasno vidimo spremembo v rabi. Najprej opazimo, da se je njegova raba izrazito povečala (601 zadetkov ali 0,43 na milijon v referenčnem korpusu, ki zajema 1,2 milijarde pojavnic, v primerjavi z 2854 zadetki ali 26,65 v 10-krat manjšem korpusu tvitov). Podrobnejši pregled besednih skic iztočnic v obeh korpusih pokaže specializacijo pomena besede na mikroblogerski platformi Twitter iz enega izmed naslednjih pomenov:

1) sledilec prepričanja in dela vplivnih politikov, verskih voditeljev ali umetnikov (npr. »predan «, »zvest«, »nauk«, »ideja«, »gibanje«);
2) oseba ali organizacija, ki posnema vedenje in govorjenje drugih in ki sam ni vodja (npr. »slep«, »podrejen«, »trend«, »četica«, »prepisovalec«); in
3) sledilna naprava ali medij (npr. »izotopski«, »satelitski«, »vgrajen«, »radiokativen«, »silicijski«);

v

1a) uporabnik, ki spremlja objave drugih uporabnikov na Twitterju in drugih družbenih omrežjih (npr. »nov «, »število «, »nabirati «, »meja «, »milijon«).

Ta pomen v korpusu Janes-Tweet močno prevladuje, kar je razvidno iz 20 naključnih konkordančnih vrstic za razmerje pridevnik + »sledilec« v obeh korpusih na Sliki 3. V korpusu tvitov samo primera 1 in 17 pripadata pomenu 1) – sledilec prepričanja in dela vplivnih politikov, verskih voditeljev ali umetnikov, vsi drugi pa so primeri novega specializiranega pomena – uporabnik, ki spremlja objave drugih uporabnikov na Twitterju in drugih družbenih omrežjih. Na drugi strani pa v Gigafidi 11 (55 %) od vseh prikazanih primerov pripada pomenu 1) – sledilec prepričanja in dela vplivnih politikov, verskih voditeljev ali umetnikov, 5 (25 %) pomenu 3) – sledilna naprava ali medij in 4 (20 %) pomenu 2) oseba ali organizacija, ki posnema vedenje in govorjenje drugih in ki ni vodja. Primeri s pomenom z družbenih omrežij se v Gigafidi ne pojavijo.



par stvari ima prav *ivanovi* orto	sledilci	bi se lahko marsikaj naučili,
imamo poleg tehle *virtualnih*	sledilcev	tudi žive, se mi zdi. On ne! ↯
svojim 20. novim ruskim *jajčnim*	sledilcem	... jaz sem jih že 40 reportal
Preštej svoje tw sorodno *misleče*	sledilce	in pomnoži s 100. Izhajam iz
https://t.co/NRSw3Xucbm ↯ *Dragi*	sledilci	, vesel #božič in veliko uresničenih
njih. :) ↯ Zanimivo, da *številni*	sledilci	ne sledite avtorju tukajšnjih
je? ↯ @PCrtomir ni med *tvojimi*	sledilci	na tw-juž :) ↯ @AndrejArh al pa
http://t.co/HHZSDlsrwZ ↯ Se med *mojimi*	sledilci	najde kdo, ki ima instalirano
le opozorilo. Srečno do *novih*	sledilcev	! ↯ @VeronParsons brilliant stuff
@krimilolog Danes pa imamo *erotične*	sledilce	, ne ... ↯ Posladkajte se malo ...
#IntervjuTedna #IvanOman ↯ *Dragi*	sledilci	, vesel #božič in veliko uresničenih
@SrceSloveniji in apeliraj na *tvoje*	sledilce	↯ Če koga čakate iz obale ga boste
si zaslužil nekaj *neprijetnih*	sledilcev	. #zavsakoritšibaraste ↯ @Moj_ca
medalje dobila tudi *novih* 2500+	sledilcev	na Twitter profilu. (33.091 //
kokaina. ↯ Vse več imamo *domačih*	sledilcev	. To bi lahko bil tudi znak, da
avtobusni postaji ↯ V vednost: *Nove*	sledilce	z zaklenjenim profilom brez milosti
bila bolj predavanje za *njene*	sledilce	, da se osveži gradivo. ↯ @strankaSDS
civkam. #prejsnitvit ↯ Rabim *nove*	sledilce	da mi bo tvitlajn hitreje laufal
tako bo tvit viden vsem *vašim*	sledilcem	in ne samo tistim, ki sledijo
komu za siht. Sej mas *vpiivne*	sledilce	. :) ↯ Oj Vogel ti bos letos padu

Tudi slaba šala. Naš *polvodniški*	sledilec	smo ravno začeli zlagati skupaj
, žal zgolj status *razvojnega*	sledilca	(R&D Follower), sicer pa
organizme. *Ustrezne* izotopske »	sledilce	« so pridobili šele po odkritju
zanimanja. Njegovi najbolj *goreči*	sledilci	so bili študentje in hipiji.
EKSTATIČNI VATES, VZNESENI, *SLEPI*	SLEDILEC	NEDOUMLJIVIH DOGODKOV, POJAVOV
, nedvomno najbolj *doslednega*	sledilca	tovrstne ikonografije v domačem
zdaj vsi Katonovi in *Ciceronovi*	sledilci	, ki ga sovražijo, znova poskušali
majhen v primerjavi s *polvodniškim*	sledilcem	, detektorjem, pri katerem slovenska
študirajo Teslo. Ko *povprečni*	sledilec	Špagetne pošasti oziroma Chucka
svetujejo tukajšnjim *političnim*	sledilcem	. Naši politični nasprotniki ponujajo
sodelujejo pri razvoju *silicijevega*	sledilca	nabitih delcev. Silicijev sledilec
vidika bi številni (ne) *kritični*	sledilci	glasb sveta ali nadžanrskega
sistemska motnja vsakega *predanega*	sledilca	misli in dela Miltona Friedmana
oziroma z ustvarjanjem *avtentičnih*	sledilcev	na podlagi pozitivnega modeliranja
to ustrezno preganja, se *dobri*	sledilci	odlikujejo po sposobnosti hitre
rock'n'rollu. Tako *zvestim* kot naključnim	sledilcem	njihove avanture pa je že dolgo
okvirno temo: Kako sem *resnični*	sledilec	Jezusa Kristusa? Razgovor bo
nasmeju Janši in njegovim *slepim*	sledilcem	, ko bodo dobili klofuto :) Zanimivo
tej podrobnosti, ki *pozornim*	sledilcem	njegovega opusa nikakor ne more
sledilca nabitih delcev. *Silicijev*	sledilec	je poddetektorski sistem, ki

Slika 3: Naključne konkordančne vrstice za razmerje pridevnik + »sledilec« iz korpusa Janes-Tweet (levo) in korpusa Gigafida (desno).

4.2 Manjši pomenski premiki

Med besedami, ki v korpusu tvitov glede na referenčni korpus v svojem semantičnem odtisu izkazujejo manjše razlike, razlikujemo med naslednjimi tremi kategorijami: spremembe v frekvenčni distribuciji rabljenih pomenov, omejenost rabe na določene ustaljene vzorce in semantično oženje.

4.2.1 Spremembe v distribuciji pomenov

V prvo vrsto manjših premikov spadajo tisti primeri, pri katerih smo v obeh korpusih prepoznali enake pomene, vendar z drugačno razporeditvijo po njihovi pogostosti. Dober primer je samostalnik »sesalec«, ki v obeh korpusih pomeni tako žival kot tudi gospodinjski pripomoček, vendar v referenčnem korpusu

prevladuje pomen 'žival', v korpusu tvitov pa pomen 'gospodinjski pripomoček'. Kot prikazuje Tabela 4, samostalnik bolj izstopa v tvitih in samo dve kolokaciji (»edin« in »vrsta«) od najpogostejših petih v vseh treh najbolj produktivnih slovničnih relacijah se v korpusu tvitov nanašata na pomen *žival*, medtem ko je razmerje v referenčnem korpusu ravno obratno (samo »globinski« in »prodajati« se nanašata na napravo, ostale kolokacije se vse nanašajo na žival).

Tabela 4: Najmočnejših pet kolokatorjev za tri najproduktivnejše slovnične relacije za besedo »sesalec« v korpusih Janes in Gigafida. Besede v krepkem tisku so znak prerazporeditve pomenov v korist nestandardnega pomena besede sesalec (tj. v pomenu gospodinjskega pripomočka), ki prevladuje v korpusu Janes.

Janes			**Gigafida**		
iztočnica: sesalec pogostost: 701 (6,54 na milijon) pokritost leme: 78,17 %			iztočnica: sesalec pogostost: 7.047 (4,99 na milijon) pokritost leme: 90,02 %		
Razmerje	Kolokacija	Frekv / logDice	Razmerje	Kolokacija	Frekv / logDice
Pridevnik + samostalnik	**robotski**	**10 / 9,91**	Pridevnik + samostalnik	kopenski	81 / 8,01
	globinski	**7 / 9,53**		rastlinojed	30 / 8,00
	voden	**8 / 6,58**		morski	502 / 7,71
	edin	6 / 3,70		**globinski**	**47 / 7,68**
	nov	**11 / 1,26**		kloniran	26 / 7,49
Glagol + predmet	**imeti**	**7 / 8,81**	Glagol + predmet	klonirati	9 / 8,57
	kupiti	**5 / 8,05**		pleniti	9 / 8,52
	/	/		loviti	16 / 4,98
	/	/		napadati	6 / 4,88
	/	/		**prodajati**	**10 / 3,27**
Samostalnik + samostalnik v rodilniku	**vrečka**	**7 / 9,01**	Samostalnik + samostalnik v rodilniku	kloniranje	53 / 8,89
	zvok	**10 / 8,24**		samica	18 / 7,61
	vrsta	5 / 5,52		tkivo	19 / 7,18
	/	/		genom	10 / 7,09
	/	/		mladič	12 / 7,07

4.2.2 Omejenost na ustaljene vzorce

V to skupino uvrščamo primere, pri katerih zaznamo opazna neskladja v ustaljenih vzorcih, v katerih je iztočnica redno uporabljena in ki opredeljujejo njen pomen. Tipičen primer te kategorije je samostalnik »eter«. Glede na njegovo

besedno skico, ki je povzeta v Tabeli 5, je njegova uporaba v Gigafidi pogosta in raznolika (6.264 ali 4,44 na milijon) ter uporabljena v:

1) dobesednem (kemijskem) pomenu (npr. »molekula«, »element«, »alkohol«, »ester«, »dietil«);
2) metaforičnem pomenu, vključno z imeni podjetij, filmov, itd. (npr. »moralni«, »brazilski«, »svoboden«, »življenje«, »svetloba«); in
3) pomenu oddajanja (npr. »radio«, »televizijski«, »postaja«, »oddaja«, »v«), ki je od naštetih najmanj pogost.

Prav zadnji pomen, torej pomen oddajanja, v tvitih močno prevladuje (npr. »v«, »proti«, »iz/from«, »radijski«), zgolj peščica kolokacij pa pripada pomenu, rabljenemu v kemiji (npr. »borov«, »teorija«), prav tako identificiramo nekaj imen podjetij, ki so pravzaprav lematizacijske napake (npr. »Etra«, »eTRI«). Tudi v tem primeru je v korpusu tvitov iztočnica prominentnejša kot v referenčnem korpusu (8,12 proti 4,44 na milijon), in sicer izrazito pogosto v zvezi »v etru«. V Gigafidi je ta raba sicer zabeležena, vendar velika večina primerov sodi v kemijski ali metaforični pomen.

Tabela 5: Najpogostejših pet kolokatorjev za dve najbolj produktivni slovnični relaciji za besedo »eter« v korpusih Janes in Gigafida. Besede v krepkem tisku kažejo na prerazporeditev pomenov v korist nestandardnemu vzorcu »v etru«, ki prevladuje v korpusu Janes.

Janes			**Gigafida**		
iztočnica:	eter		iztočnica:	eter	
frekvenca:	870 (8,12 na milijon)		frekvenca:	6.264 (4,44 na milijon)	
pokritost leme:	95,86 %		pokritost leme:	85,25 %	
Relacija	Kolokacija	Frekv. / logDice	Relacija	Kolokacija	Frekv. / logDice
Pridevnik + samostalnik	85,25%	85,25%	Pridevnik + samostalnik	škroben	7 / 7,18
	Petričev	5 / 11,17		**radijski**	**279 / 6,71**
	radijski	**5 / 6,21**		toploten	11 / 4,17
	/	/		svetloben	9 / 3,71
	/	/		kemičen	9 / 3,47
Predlog + samostalnik	**izven**	**8 / 6,77**	Predlog + samostalnik	**proti**	**627 / 5,27**
	v	**707 / 4,52**		**izven**	**9 / 3,80**
	proti	**17 / 3,85**		**preko**	**18 / 3,33**
	iz	**6 / 0,63**		**zunaj**	**8 / 3,16**
	/	/		**skozi**	**29 / 2,67**

4.2.3 Semantično oženje

Tretji tip manjših pomenskih premikov, ki smo jih opazili, je oženje semantičnega repozitorija besed, ki najverjetneje ni znak zamiranja določenih pomenov, temveč je posledica omejenega števila tem, ki se pojavljajo v diskusijah na Twitterju v primerjavi s številom tem v referenčnem korpusu. Takšen primer je glagol »posodobiti«, ki se glede na besedne skice v korpusu Gigafida uporablja z zelo različnimi predmeti, kot so »infrastruktura«, »proizvodnja«, »park«, »oprema« in »flota«. Na drugi strani pa je v korpusu tvitov glagol omejen na predmete, ki se nanašajo na informacijske tehnologije: »aplikacija«, »stran«, »seznam«, »sistem«.

Še en primer oženja semantičnega odtisa v korpusu tvitov je večpomenski samostalnik »faks«, ki se v korpusu Gigafida pojavlja bodisi v pomenu naprave bodisi v pomenu ustanove, pri čemer je dominanten prvi, kar je razvidno že iz najmočnejših kolokacij v relaciji glagol + predmet: »pošiljati«, »oddajati«, »sprejemati«, »dokončati«, »vpisati«. V tvitih je raba samostalnika precej prominentnejša (frekvenca 45,78 proti 16,76 na milijon), prav tako močno prevladuje pomen ustanove: »končati«, »pustiti«, »pogrešati«, »narediti«, »imeti«, kar ni presenetljivo, saj je telefaks praktično že zastarela tehnologija.

4.3 Analiza napak

Poleg podrobne analize zaznanih pomenskih premikov analiziramo tudi napačno prepoznane kandidate. Pomenskega premika nismo zaznali pri 103 (51 %) od 200 najpogostejših kandidatov, pri čemer ločujemo dve vrsti napačno prepoznanih kandidatov. V prvo kategorijo sodijo kandidati, ki so se na vrh seznama uvrstili zaradi napak pri predprocesiranju korpusa. Če je bila npr. iztočnici v enem od korpusov pripisana napačna lema, je njen vektor za ta korpus razumljivo povsem drugačen od vektorja za isto iztočnico v drugem korpusu, saj si pravzaprav ne delita semantičnih lastnosti. Pričakujemo lahko, da se bo ta kategorija z razvojem orodij za procesiranje postopoma zmanjševala. V drugo kategorijo pa uvrščamo napačne kandidate, ki so se visoko na lestvici pojavili zaradi predlagane metode, vendar ročni pregled besednih skic zanje ni razkril pomenskih premikov. To so dejanske napake, ki razkrivajo omejitve predlaganega pristopa in jih je nujno potrebno reševati v prihodnjih izboljšavah metode.

4.3.1 Napake zaradi predprocesiranja

Z analizo lažnih kandidatov za pomenske premike, do katerih je prišlo zaradi šuma, ki ga ustvarjajo orodja za jezikoslovno procesiranje slovenščine, smo identificirali 90

(45 %) takšnih primerov. Raven šuma ne preseneča, saj se v korpusu Janes soočamo z izrazito nestandardnimi podatki, ki jih je težko procesirati z visoko stopnjo natančnosti. Drug pogost vir napak se je pojavil zaradi dejstva, da smo v raziskavi uporabili korpusa, ki sta bila označena in lematizirana z dvema različnima orodjema. Obenem pa naša analiza kaže, da je tovrsten tip šuma najvišji na vrhu seznama in nato vztrajno pada. Kar se tiče vzroka za napačno predprocesiranje, so daleč najpogostejši vir napak tuje besede iz tujejezičnih tvitov, ki so bili napačno prepoznani kot slovenski, in slovenski tviti, ki so delno napisani v tujem jeziku (34 %). Drugi najpogostejši tip napak so nestandardne ali nestandardno zapisane besede, ki jih orodje za oblikoskladenjsko označevanje in lematizacijo ni pravilno analiziralo (33 %). Na tretjem mestu so težave, povezane z osebnimi imeni (17 %), ki so še posebej izrazite, če se osebno ime prekriva z občnim. To kaže na težavnost avtomatskega procesiranja uporabniško generiranih vsebin in na to, kako pomembno je zagotoviti, da bodo orodja NLP postala robustnejša tudi za fenomene jezika družbenih medijev.

Tabela 6: Distribucija napak zaradi predprocesiranja.

Vrsta napake	Frekv.	%	Primeri
tuja	27	34 %	duda (namesto ang. »dude/frajer«) danka (namesto nem. »Danke/hvala«) kad (namesto hrv. »kad/ko«)
nestandardna	26	33 %	ajda (namesto »ajde« v pomenu gremo) bol (namesto »bol/bolj«) hod (namesto »hodu/hodil«)
ime	15	18 %	Tanko (priimek, prekriven s s pridevnikom in prislovom »tanek, tanko«) Nedelo (časopis, prekriven s samostalnikom »nedelo«) Rim (mesto, prekriven s samostalnikom »rima«)
lema	13	15 %	meni (namesto »jaz«) moa (izumrla ptica, namesto »moj«) novic (redovniški pripravnik, namesto »novica«)
besedna vrsta	6	6 %	dobro (prekrivna pridevnik in prislov) fin (prekrivna pridevnik in prislov)
diakritike	1	2 %	sel (kdor prinaša sporočilo, namesto »šel«)
orodje	2	2,5 %	nazadnje (normalizirano kot »ne nazadnje«)
Skupno	90	100 %	

4.3.2 Lažni kandidati

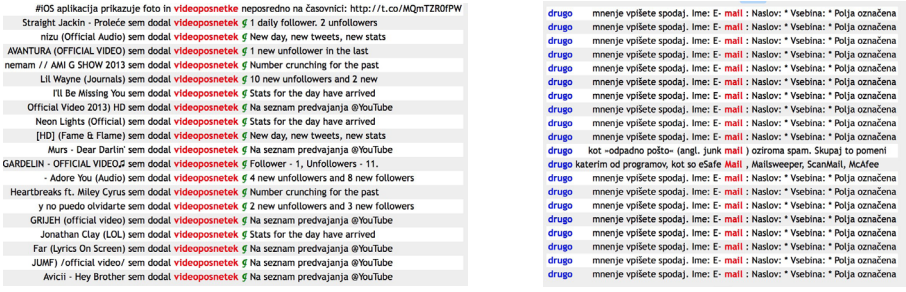
Še posebej nas zanima 13 (6,5 %) kandidatov, pri katerih glede na besedne skice pomenskega premika nismo opazili in ki niso napake, ki so se pojavile med

predprocesiranjem. S pomočjo analize konkordanc jih razvrstimo v dve skupini: besede, ki so uporabljene v samodejno ustvarjenih tvitih ali v ponavljajočih oglasnih besedilih v Gigafidi (9 oz. 69 %) in besede, ki so uporabljene v telegrafskem tviterskem ali tipičnem novinarskem slogu (4 oz. 31 %). Ocenjujemo, da je glede na kompleksnost naloge dobljeni delež lažnih kandidatov dober rezultat. Neposredno izboljšanje rezultata za prvi tip napak bi lahko dosegli z ustavljenimi postopki prepoznavanja in filtriranjem popolnih in delnih duplikatov v korpusih. Za celovitejše vrednotenje razvitega pristopa pa bi bilo treba analizirati večje število kandidatov, predvsem tistih, ki so glede na stopnjo kontekstualnih razlik med modeloma iz obeh korpusov razvrščene nižje na seznamu.

Tabela 7: Distribucija lažnih primerov.

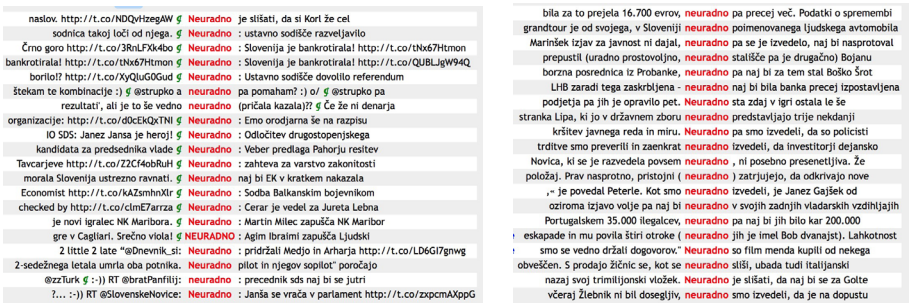
Vrsta napake	Frekv.	%	Iztočnica
Samodejno generirano besedilo	9	69 %	aktualno barometer frizerka kontakten kviz magazin mail veterinar videoposnetek
Slog	4	31 %	dopoldan ekskluzivno izjemoma neuradno
Skupaj	13	100 %	

Primere besed, ki se pojavijo v samodejno ustvarjenih tvitih ali ponavljajočih se oglasnih besedilih iz korpusa Gigafida, prikazuje Slika 4.



Slika 4: Primeri samodejno ustvarjenih besedil v korpusih Janes in Gigafida.

Na Sliki 5 podajamo primer slogovno različne rabe besede »neuradno« v primerjanih korpusih. Čeprav se beseda v Gigafidi večinoma uporablja v časopisnih člankih in revijah, je še vedno skladenjsko integrirana v stavek. Po drugi strani pa jo uporabniki – tudi zasebni uporabniki, ki niso novinarji ali predstavniki medijev – v tvitih uporabljajo v izrazito telegrafskem slogu.



Slika 5: Primeri slogovno različne rabe besede »neuradno« v korpusih Janes in Gigafida.

4.4 Rezultati in diskusija

Rezultati analize so povzeti v Tabeli 8. Iz nje je razvidno, da je bila določena vrsta pomenskega premika zaznana v nekaj manj kot polovici vseh primerov analiziranega vzorca, na podlagi česar sklepamo, da bi predlagani pristop – čeprav ni dovolj natančen za uporabo v popolnoma avtomatiziranem scenariju – lahko bil uporaben kot polavtomatski postopek, ki bi leksikografom postregel z avtomatsko generiranimi predlogi, ti pa bi potem rezultate ročno pregledali. Če upoštevamo razmeroma velik delež napak, do katerih je prišlo zaradi napak pri predprocesiranju korpusov (45 %), bi lahko z uporabo robustnejših orodij za nestandardno slovenščino rezultate še znatno izboljšali: boljša identifikacija jezika, v katerem je besedilo napisano, robustnejše procesiranje nestandardnih različic zapisa besed in nestandardnega besedišča, natančnejše prepoznavanje lastnih imen in boljše oblikoskladenjsko označevanje.

Večji in manjši pomenski premiki so bili skoraj enako pogosti (približno četrtina vzorca pri vsakem). Nepresenetljivo lahko večino pomenskih premikov pripišemo manj formalnemu registru in značilnostim tem, ki se pojavljajo v pogovorih na Twitterju (ki skupaj predstavljajo za četrtino analiziranega vzorca), kar sistematično izpostavi razlike v osredotočenosti in razponu tem v obeh korpusih. Dejstvo, da je bilo zaznanih mnogo več novih primerov rabe (novi pomeni zaradi registra, družbenega konteksta in medija) kot pa oženj (pomeni in vzorci, omejeni na

tematiko) (25 % proti 13 %) nakazuje na to, da bi lahko referenčni korpus še izboljšali s sodobnejšimi besedili in besedili z družbenih medijev ter drugimi manj formalnimi in manj standardnimi komunikacijskimi praksami, saj vsebujejo bogato in dragoceno jezikovno gradivo, ki je iz referenčnega korpusa zaenkrat skoraj popolnoma izključeno.

Še posebej zanimivi so zaznani novi pomeni kot rezultat širšega družbenega konteksta, v katerem se uporabniki izražajo na Twitterju, s čimer se prilagodijo tako nastajajočim novicam (9 %) kot tudi dinamičnim konvencijam komuniciranja na družbenih medijih (4 %). To leksikografsko gradivo je izjemno dragoceno, saj bi lahko služilo kot osnova za posodobitev obstoječih leksikosemantičnih virov slovenščine, s čemer se potrjuje potreba po spremljevalnem korpusu, ki za slovenščino zaenkrat še ni na voljo.

Tabela 8: Porazdelitev tipov pomenskih premikov v slovenski tviterščini. Vrednosti v oklepajih so izračunani za podkategorijo, odstotki pa za kategorijo.

Kategorija	Podkategorija	Frekv.	%
Večji premiki		51	25 %
	Register	(26)	(51 %)
	Dogodki	(18)	(35 %)
	Medij	(7)	(14 %)
Manjši premiki		46	23 %
	Distribucija pomenov	(24)	(52 %)
	Ožanje pomenov	(20)	(43 %)
	Ustaljeni vzorci	(2)	(4 %)
Brez premika		103	51 %
	Šum zaradi predprocesiranja	(90)	(87 %)
	Lažni kandidati	(13)	(13 %)
Skupaj		200	100 %

Z ročno analizo smo identificirali tudi nekaj kreativnega pripisovanja novih pomenov sicer vsakdanjim besedam (npr. »kahla«, ki se nanaša na politika Karla Erjavca, ki zelo prepoznavno izgovarja črko r; »pingvin«, ki se nanaša na »zamrznjenega« vodjo politične stranke Zares Zorana Jankoviča; in šaljivo rabo besede »modrec« za »modrc«).

Rezultati izvedene jezikovne analize kažejo, da lahko pristop, ki ga smo ga predstavili v pričujočem poglavju, močno pripomore k rednim polavtomatskim posodobitvam tako splošnih kot tudi specializiranih na korpusu temelječih leksikalnih virov.

5 SKLEP

V poglavju smo predstavili potencial distribucijskega simuliranja za avtomatizacijo leksikografskega dela, ki smo ga preizkusili na problemu zaznavanja pomenskih premikov v slovenščini na družbenih omrežjih. Pomenski premik besede smo izmerili kot razdaljo med reprezentacijo njenega semantičnega odtisa, naučeno iz referenčnega korpusa, in njeno ustreznico iz korpusa tvitov. Za 200 besed z izkazanimi največjimi razlikami smo opravili ročno analizo, ki je pokazala, da je – z izjemo zlahka prepoznanega šuma zaradi napak, do katerih je prišlo pri predprocesiranju (45 % analiziranih primerov) – pristop prinesel veliko dragocenih kandidatov z izkazanimi pomenskimi premiki, med katerimi so še posebej zanimivi novi pomeni, ki so posledica jezikovne rabe ob odzivanju na dnevno dogajanje ter neformalne komunikacije. Zanimivo bi bilo spremljati, ali so zaznani novi pomeni in semantični premiki kratkotrajni in kateri izmed njih bodo postali trajni del leksikosemantičnega repozitorija.

Prispevek v slovenski prostor vnaša model detektiranja pomenskih premikov, ki je neposredno uporaben pri razvoju slovarja slovenske tviterščine (Gantar et al. 2016), prav tako pa tudi pri nadgradnji in posodabljanju obstoječih splošnih slovarskih priročnikov in baz za slovenščino (Gorjanc et al. 2015). Vendar prispevek predstavljene metode sega še mnogo dlje, saj demonstrira uporabnost distribucijskih pristopov za številne leksikografske naloge za slovenščino ter tudi druge jezike, pri čemer potrebujemo le dva dovolj obsežna jezikoslovno označena korpusa; referenčni korpus in ciljni korpus, ki pokriva tisti segment jezikovne rabe, ki nas za konkretno raziskavo zanima, npr. korpus govorjenega jezika, historični korpus, različni področnospecifični korpusi ipd.

V nadaljevanju raziskav se nameravamo osredotočiti na (1) razširitev ročne analize na kandidate s spodnjega dela seznama, (2) razširitev pristopa na redkejše besedišče, (3) primerjavo predlaganega pristopa z alternativnimi metodami, kot je npr. učenje reprezentacij besed s pomočjo besednih skic oz. skladenjskih vzorcev in (4) uporabo nadzorovanega učenja za prepoznavanje pomenskih premikov, razločevanje med različnimi vrstami pomenskih premikov in filtriranje napak, do katerih pride pri predprocesiranju.

Zahvala

Zaradi mednarodne avtorske zasedbe je bil rokopis tega poglavja delno napisan v angleškem jeziku. Za pomoč pri prevodu se zahvaljujemo Lei Anžur

Literatura

Agres Kat, Stephen McGregor, Matthew Purver in Geraint Wiggins, 2015: Conceptualizing creativity: From distributional semantics to conceptual spaces. *Proceedings of the Sixth International Conference on Computational Creativity*. 118–125.

Bengio Yoshua, Aaron Courville in Pascal Vincent, 2013. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35/8: 1798–1828.

Blei, David M., Andrew Y. Ng in Michael I. Jordan, 2003: Latent Dirichlet Allocation. *The Journal of Machine Learning Research* 3: 993–1022.

Campbell, Lyle, 2013. *Historical linguistics.* Edinburg: Edinburgh University Press.

Church, Kenneth. W. in Hanks, Patrick, 1990: Word association norms, mutual information, and lexicography. *Computational Linguistics* 16/1. 22–29.

Cook, Paul in Suzanne Stevenson, 2010: Automatically Identifying Changes in the Semantic Orientation of Words. *Proceedings of the 7th LREC Conference*. 28–34.

Cook, Paul, Jey Han Lau, Michael Rundell,Diana McCarthy in Timothy Baldwin: 2013: A lexicographic appraisal of an automatic approach for detecting new word senses. *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex conference*. Tallinn, Estonia. 49–65. http://eki.ee/elex2013/proceedings/eLex2013_04_Cook+etal.pdf

Čibej, Jaka in Nikola Ljubešić, 2015: »S kje pa si?«. *Proceedings of the conference Slovene on the web and in the new media. Ljubljana: Znanstvena založba Filozofske fakultete.* 10–14.

Erjavec, Tomaž, Nikola Ljubešić in Darja Fišer, 2018: Korpus slovenskih spletnih uporabniških vsebin Janes. Fišer, Darja (ur.): *Viri, orodja in metode za analizo spletne slovenščine*. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani. 16–43.

Firth, John R, 1957: A Synopsis of Linguistic Theory. *Studies in Linguistic Analysis*: 1-32.

Fišer, Darja, Tomaž Erjavec in Nikola Ljubešić: 2016: JANES v0.4: Korpus slovenskih spletnih uporabniških vsebin. *Slovenščina 2.0* 4/2: 67–99.

Fišer, Darja in Nikola Ljubešić, 2016: Detecting Semantic Shifts in Slovene Twitterese. Proceedings of the Tenth Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2016. Brno, Czech Republic.

Gantar, Polona, Iza Škrjanec, Darja Fišer in Tomaž Erjavec, 2016: Slovar tviterščine. *Proceedings of the Conference on Language Technologies and Digital Humanities*, Ljubljana, Slovenia. 71–76.

Fodor, Imola, 2002: *A Survey of Dimension Reduction Techniques*. Center for Applied Scientific Computing, Lawrence Livermore National Laboratory. https://e-reports-ext.llnl.gov/pdf/240921.pdf

Gantar, Polona, Iztok Kosem in Simon Krek, 2015: Leksikografski proces pri izdelavi spletnega slovarja sodobnega slovenskega jezika. Gorjanc, Vojko, Polona

Gantar, Iztok Kosem in Simon Krek (ur.): *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete. 280–297.

Golub, Gene H. in Christian Reinsch, 1970: Singular value decomposition and least squares solutions. *Numerische mathematik* 14/5: 403–420.

Grčar, Miha, Simon Krek in Kaja Dobrovoljc, 2012: Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. *Proceedings of the 8th Language Technologies Conference*. Ljubljana: Institut »Jožef Stefan«.

Hamilton William L., Jure Leskovec in Dan Jurafsky, 2016: Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin, Germany. 1489–1501.

Tomaž Erjavec, Darja Fišer, Jaka Čibej in Špela Arhar Holdt, 2016: CMC training corpus Janes-Norm 1.2. *Slovenian language resource repository CLARIN.SI*.

Gulordava, Kristina in Marco Baroni, M. 2011. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. *Proceedings of the Workshop on GEometrical Models of Natural Language Semantics*. Edinburgh, Scotland. 67–71.

Ide, Nancy in Jean Véronis, 1998: Introduction to the special issue on word sense disambiguation: the state of the art. *Computational linguistics* 24/1: 2–40.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, Vít Suchomel, 2014. The Sketch Engine: ten years on. *Lexicography* 1/1. 7–36.

Krek, Simon in Adam Kilgarriff, 2006: Slovene Word Sketches. *Proceedings of the 5th Slovenian/First International Languages Technology Conference*. https://www.kilgarriff.co.uk/Publications/2006-KrekKilg-Ljub-SloveneWS.pdf

Lenci, Allessandro, 2008: Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics* 20/1: 1–31.

Lee, David, 2001: Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology* 5/3. 37–72.

Levy, Omer in Yoav Goldberg, 2014a: Neural word embedding as implicit matrix factorization. *NIPS'14 Proceedings of the 27th International Conference on Neural Information Processing Systems*. 2177–2185.

Levy, Omer in Yoav Goldberg, 2014b: Dependency-Based Word Embeddings. *Proceedings of ACL*. 302–308.

Levy, Omer in Yoav Goldberg, 2014: Linguistic Regularities in Sparse and Explicit Word Representations. *Proceedings CoNLL:* 171–180.

Ljubešić, Nikola, Darja Fišer, Tomaž Erjavec, Jaka Čibej, Dafne Marko, Senja Pollak in Iza Škrjanec, 2015: Predicting the Level of Text Standardness in User-generated Content. *Proceedings of Recent Advances in Natural Language Processing*. 371–378.

Ljubešić, Nikola in Tomaž Erjavec, 2016: Corpus vs. Lexicon Supervision in Morphosyntactic Tagging: The Case of Slovene. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. 1527–1531.

Nikola Ljubešić, Zupan, K., Darja Fišer and Tomaž Erjavec Normalising Slovene data: historical texts vs. user-generated content. *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*. 146–155.

Logar, Nataša, Miha Grčar, Tomaž Erjavec, Špela Arhar Holdt in Simon Krek, 2012: *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba.* Ljubljana: Trojina, zavod za uporabno slovenistiko.

McCulloch, Warren. S. in Pitts, Walter, 1943: A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics* 5/4. 115–133.

Mikolov, Tomas, Kai Chen, Greg Corrado in Jeffrey Dean, 2013a: Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, Tomas, Kai Chen, Greg Corrado in Jeffrey Dean, 2013b: Linguistic Regularities in Continuous Space Word Representations. *Proceedings of HLT-NAACL 2013*. 746–751.

Mitchell, T.M. 1997. *Machine Learning*. McGraw-Hill, Inc. New York, NY, USA.

Mitra, Sunny, Ritwik Mitra, Suman Kalyan Maity, Martin Riedl, Chris Biemann, Pawan Goyal In Animesh Mukherjee, 2015: An automatic approach to identify word sense changes in text media across timescales. *Natural Language Engineering* 21/5. 773–798.

Navigli, Roberto, 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)* 41/2.

Schütze, Hinrich, 1998: Automatic word sense discrimination. *Computational linguistics* 24/1. 97–123.

Sagi, Eyal, Stefan Kaufmann in Brady Clark, 2009: Semantic Density Analysis: Comparing Word Meaning across Time and Phonetic Space. *Proceedings of the EACL 2009 Workshop on GEometical Models of Natural Language Semantics*. 104–111.

Sparck Jones, Karen, 1986: *Synonymy and Semantic Classification.* Edinburg: Edinburgh University Press.

Kanerva, Pentti, Jan Kristoferson in Anders Holst, 2000: Random Indexing of Text Samples for Latent Semantic Analysis. *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*. 1036.

Salton, Gerard, Wong, Andrew, Yang, Chungshu, 1975: A vector space model for automatic indexing. *Communications of the ACM* 18/11. 613–620.

Tahmasebi, Nina, Thomas Risse in Stefan Dietze, 2011: Towards automatic language evolution tracking, a study on word sense tracking. *Proceedings of the Joint Workshop on Knowledge Evolution and Ontology Dynamics*. http://ceur-ws.org/Vol-784/evodyn2.pdf

Sparck Jones, K., 1972: A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation* 28/1. 11–21.

Zipf, George K., 1936: *The psycho-biology of language*. New York, London: Routledge.

Priloge:
Seznami besed z označenim tipom pomenskega premika

Priloga 1: Seznam besed, pri katerih smo zaznali večje pomenske premike.

Lema z besedno vrsto	Vrsta pomenskega premika	Mera razdalje
fiskalen#Ag	dogodek	0,580353749510
zombi#Nc	dogodek	0,570473162536
pirat#Nc	dogodek	0,563599872708
arhivski#Ag	dogodek	0,543735200555
bojevnik#Nc	dogodek	0,538704741919
nevtralnost#Nc	dogodek	0,528946112923
komisarka#Nc	dogodek	0,526273297109
pingvin#Nc	dogodek	0,517542639943
risa#Nc	dogodek	0,500310188561
malomaren#Ag	dogodek	0,490241581560
vstajnik#Nc	dogodek	0,476038560358
astronomski#Ag	dogodek	0,471662874089
prepih#Nc	dogodek	0,470011599897
kahla#Nc	dogodek	0,451749782663
tisa#Nc	dogodek	0,383248742949
dl#Nc	dogodek	0,367170101908
supervizor#Nc	dogodek	0,285819575689
vztrajnik#Nc	dogodek	0,143540587272
opomnik#Nc	medij	0,580180651285
sledenje#Nc	medij	0,578872731644
sledilec#Nc	medij	0,556588211346
q#Nc	medij	0,474845868899
ms#Nc	medij	0,467078140408
nm#Nc	medij	0,402784956150
rt#Nc	medij	0,211168972262
bus#Nc	register	0,579057574473
bio#Ag	register	0,578177532133
profi#Nc	register	0,577361896651
modrec#Nc	register	0,574217906729
dostavljati#Vm	register	0,571193498689
ultra#Ag	register	0,567350218824
noro#Rg	register	0,566595894804
skozi#Rg	register	0,564839986817
penzion#Nc	register	0,558843681744
jaz#Nc	register	0,557462156560
carski#Ag	register	0,553271554077

Lema z besedno vrsto	Vrsta pomenskega premika	Mera razdalje
optika#Nc	register	0,542834517374
sesalec#Nc	register	0,537911484878
depresiven#Ag	register	0,530165847517
misterij#Nc	register	0,506956166729
takisto#Rg	register	0,500157740660
karma#Nc	register	0,498301378493
spin#Nc	register	0,497109940644
info#Ag	register	0,496657400098
ajd#Nc	register	0,493257015790
glavno#Rg	register	0,490177276439
gotov#Ag	register	0,465296448441
smrkec#Nc	register	0,451190591675
fakin#Nc	register	0,391605134818
meh#Nc	register	0,361707010382
top#Ag	register	0,343414605450

Priloga 2: Seznam besed, pri katerih smo zaznali manjše pomenske premike.

Lema z besedno vrsto	Vrsta pomenskega premika	Mera razdalje
obsojenec#Nc	oženje	0,606742377344
drsati#Vm	oženje	0,589984006409
mavrica#Nc	oženje	0,585132147897
asociacija#Nc	oženje	0,569884071932
kozolec#Nc	oženje	0,563829321520
produktiven#Ag	oženje	0,563099861782
posodobiti#Vm	oženje	0,562595017512
kongresen#Ag	oženje	0,558579997685
enka#Nc	oženje	0,555646649303
ponovitev#Nc	oženje	0,544808051012
kvadrat#Nc	oženje	0,526933310826
podnapis#Nc	oženje	0,517714981669
posodobitev#Nc	oženje	0,517572595909
agregat#Nc	oženje	0,493994603878
faks#Nc	oženje	0,466049923982
malezijski#Ag	oženje	0,449709928416
album#Nc	oženje	0,439074282021
beta#Nc	oženje	0,429867341906
mol#Nc	oženje	0,420191219414
predvajanje#Nc	oženje	0,349996691863
panda#Nc	vrstni red	0,581181767675
odklop#Nc	vrstni red	0,581121382368

Lema z besedno vrsto	Vrsta pomenskega premika	Mera razdalje
burka#Nc	vrstni red	0,578919014186
domena#Nc	vrstni red	0,574300502153
tesla#Nc	vrstni red	0,566861095077
testen#Ag	vrstni red	0,565518883875
kompas#Nc	vrstni red	0,559512599605
izpiten#Ag	vrstni red	0,546894441609
odmev#Nc	vrstni red	0,530065326675
agenda#Nc	vrstni red	0,527949021220
vezava#Nc	vrstni red	0,525247135181
cifra#Nc	vrstni red	0,523292962123
recenzija#Nc	vrstni red	0,519103247498
kopitar#Nc	vrstni red	0,516488872572
x#Nc	vrstni red	0,514193345981
ciganski#Ag	vrstni red	0,512854059519
pogovoren#Ag	vrstni red	0,509486755021
krogec#Nc	vrstni red	0,494365095953
konzorcij#Nc	vrstni red	0,484658137741
profilen#Ag	vrstni red	0,480269943843
greznica#Nc	vrstni red	0,460715813930
bolha#Nc	vrstni red	0,451526694121
opeka#Nc	vrstni red	0,381554496304
tajfun#Nc	vrstni red	0,353925432855
eter#Nc	stalna besedna zveza	0,521370091208
biti#Vm	stalna besedna zveza	0,518225790812

Priloga 3: Seznam besed, pri katerih smo zaznali napake v predprocesiranju.

Lema z besedno vrsto	Vrsta napake
sel#Nc	manjkajoči diakritiki
a#Nc	tujejezična beseda
ata#Nc	tujejezična beseda
biga#Nc	tujejezična beseda
danka#Nc	tujejezična beseda
duda#Nc	tujejezična beseda
era#Nc	tujejezična beseda
gama#Nc	tujejezična beseda
jak#Nc	tujejezična beseda
kad#Nc	tujejezična beseda
let#Nc	tujejezična beseda
lik#Nc	tujejezična beseda
lina#Nc	tujejezična beseda

Lema z besedno vrsto	Vrsta napake
lot#Nc	tujejezična beseda
market#Nc	tujejezična beseda
mina#Nc	tujejezična beseda
nada#Nc	tujejezična beseda
navoj#Nc	tujejezična beseda
nem#Ag	tujejezična beseda
oda#Nc	tujejezična beseda
runa#Nc	tujejezična beseda
sita#Nc	tujejezična beseda
som#Nc	tujejezična beseda
sura#Nc	tujejezična beseda
talka#Nc	tujejezična beseda
tim#Nc	tujejezična beseda
uriti#Vm	tujejezična beseda
y#Nc	tujejezična beseda
deti#Vm	pripisana napačna lema
h#Nc	pripisana napačna lema
kola#Nc	pripisana napačna lema
logo#Nc	pripisana napačna lema
meni#Nc	pripisana napačna lema
meniti#Vm	pripisana napačna lema
mesti#Vm	pripisana napačna lema
mik#Nc	pripisana napačna lema
moa#Nc	pripisana napačna lema
novic#Nc	pripisana napačna lema
pestiti#Vm	pripisana napačna lema
stan#Nc	pripisana napačna lema
tuje#Rg	pripisana napačna lema
gin#Nc	ime
hoc#Nc	ime
hrt#Nc	ime
intelekt#Nc	ime
lek#Nc	ime
lump#Nc	ime
marka#Nc	ime
nedelo#Nc	ime
osa#Nc	ime
pikati#Vm	ime
rima#Nc	ime
tanko#Rg	ime
uk#Nc	ime

Lema z besedno vrsto	Vrsta napake
veber#Nc	ime
ajda#Nc	nestandardna raba
bat#Nc	nestandardna raba
beka#Nc	nestandardna raba
boja#Nc	nestandardna raba
bol#Nc	nestandardna raba
butati#Vm	nestandardna raba
dila#Nc	nestandardna raba
dob#Nc	nestandardna raba
hod#Nc	nestandardna raba
ke#Rg	nestandardna raba
kea#Nc	nestandardna raba
klas#Nc	nestandardna raba
koka#Nc	nestandardna raba
luk#Nc	nestandardna raba
maja#Nc	nestandardna raba
mona#Nc	nestandardna raba
plata#Nc	nestandardna raba
pona#Nc	nestandardna raba
sejati#Vm	nestandardna raba
skupiti#Vm	nestandardna raba
tele#Nc	nestandardna raba
tkati#Vm	nestandardna raba
treti#Vm	nestandardna raba
vesti#Vm	nestandardna raba
veti#Vm	nestandardna raba
vod#Nc	nestandardna raba
dobro#Nc	pripisana napačna besedna vrsta
drug#Nc	pripisana napačna besedna vrsta
fin#Ag	pripisana napačna besedna vrsta
garant#Nc	pripisana napačna besedna vrsta
halo#Nc	pripisana napačna besedna vrsta
pod#Nc	pripisana napačna besedna vrsta
nazadnje#Rg	pripisana napačna besedna vrsta
obresti#Vm	pripisana napačna besedna vrsta

Priloga 4: Seznam besed, ki so bili lažni kandidati za pomenske premike.

Lema z besedno vrsto	Razlog za napako
aktualno#Rg	avtomatsko generirane vsebine
barometer#Nc	avtomatsko generirane vsebine
frizerka#Nc	avtomatsko generirane vsebine
kontakten#Ag	avtomatsko generirane vsebine
kviz#Nc	avtomatsko generirane vsebine
magazin#Nc	avtomatsko generirane vsebine
mail#Nc	avtomatsko generirane vsebine
veterinar#Nc	avtomatsko generirane vsebine
videoposnetek#Nc	avtomatsko generirane vsebine
dopoldan#Nc	specifičen slog
ekskluzivno#Rg	specifičen slog
izjemoma#Rg	specifičen slog
neuradno#Rg	specifičen slog