Regionalne jezikovne različice v slovenski računalniško posredovani komunikaciji: korpusni pristop z ročno označenim korpusom Janes-Geo

Jaka Čibej

Izvleček

V poglavju predstavljamo gradnjo in analizo ročno označenega korpusa Janes-Geo, ki predstavlja prvi korak h korpusnemu proučevanju slovenskih regionalnih jezikovnih različic v spletni slovenščini. Korpus Janes-Geo vsebuje približno 64.000 pojavnic, ki jih je prispevalo približno 270 uporabnikov Twitterja, ki glede na avtomatsko pripisane metapodatke o regionalni pripadnosti spadajo v eno od devetih regij (primorska, gorenjska, rovtarska, ljubljanska, dolenjska, štajerska, koroška, mariborska in panonska). V korpusu so bile ročno označene nestandardne jezikovne prvine v skladu z izdelano tipologijo. Namen korpusa Janes-Geo je dvojni: ugotoviti, v kakšnih oblikah se (najpogosteje) izraža jezikovna nestandardnost v spletni slovenščini, in primerjati razlike v rabi nestandardnih jezikovnih prvin med uporabniki iz različnih regij. Poleg postopka avtomatskega pripisovanja metapodatkov o regionalni pripadnosti uporabnikov opišemo tudi označevanje korpusa, njegovo sestavo in nekatere poglavitne razlike med njegovimi regionalnimi podkorpusi, npr. pogostost izpustov soglasnikov in samoglasnikov, različne nestandardne oblikoslovne prvine, najpogostejše nestandardno besedje in najpogostejše transformacije grafemov.

Ključne besede: regionalne jezikovne različice, slovenščina, tviti, geolokacija, računalniško posredovana komunikacija

1 UVOD

Z vzponom spleta in računalniško posredovane komunikacije (RPK) v zadnjih 20 letih, še zlasti pa v zadnjem desetletju, so govorci pridobili številne nove platforme za pisno sporazumevanje, npr. spletne forume, novičarske portale in družbena omrežja, kot so Facebook, Twitter, WhatsApp in Snapchat. Jezik v RPK (še posebej v klepetih in v drugih podobno neformalnih kontekstih) se od standarda precej razlikuje (Crystal 2011, Baron 2010, Myslin in Gries 2010), ena od njegovih ključnih značilnosti pa so tudi regionalno specifične jezikovne prvine (Ueberwasser 2013, Huang et al. 2016), kar velja tudi za slovensko RPK. Slovenščina je kljub relativno majhnemu številu govorcev in geografskemu ozemlju zelo razčlenjena: Ramovš (1931) je npr. govorce slovenščine razdelil v 7 narečnih skupin s skupno več kot 40 narečji in podnarečji (glej tudi Škofic et al. 2011: 11). Temu primerno so tudi regionalne jezikovne različice[1] slovenščine precej obširno raziskane, a le v govoru. Sistematičnih raziskav o tem, kako se slovenska regionalna jezikovna členjenost odseva v spletnem sporazumevanju, ki je za razliko od tradicionalne govorjene narečne rabe najpogosteje pisno, pa še ni na voljo, čeprav pisna spletna komunikacija že dolgo več ne zajema zanemarljivega deleža: kot poroča Valicon (2016), ima v Sloveniji profil na Facebooku že več kot 830.000 oseb v starosti od 15 do 75 let, skoraj 600.000 (oz. 70 %) oseb pa ga uporablja vsak dan. Podobno je tudi na Twitterju, kjer je profilov več kot 200.000, dnevnih uporabnikov 33.000, tedenskih pa 100.000.

V pričujočem poglavju povzemamo in nadaljujemo prve korake (Čibej in Ljubešić 2015, Čibej 2016) h korpusnemu proučevanju slovenskih regionalnih jezikovnih različic v spletni slovenščini in še dodatno razširimo nabor raziskav značilnosti slovenske RPK, ki so bile v okviru projekta JANES opravljene npr. o nestandardni skladnji (Arhar Holdt 2018), rabi vejic (Popič in Fišer 2018), pojavih krajšanja (Goli et al. 2016) ter preklapljanju med jeziki (Reher in Fišer 2018). Namen naše raziskave je predstaviti eno od metod za proučevanje regionalnih jezikovnih različic na Twitterju, ponuditi sistematičen uvid v načine, na katere se nestandardnost kaže v slovenski spletni komunikaciji, in ugotoviti, ali se izražanje nestandardnosti razlikuje med različnimi regijami.

Poglavje začnemo s kratkim pregledom sorodnih raziskav o regionalni jezikovni variantnosti v računalniško posredovani komunikaciji (razdelek 2). Nato v razdelku 3 opišemo postopek avtomatskega pripisovanja metapodatkov o regionalni pripadnosti uporabnikom, način vzorčenja korpusa in izdelavo tipologije nestandardnih jezikovnih prvin v slovenskih tvitih ter smernic za označevanje. V razdelku 4 predstavimo poglavitne razlike med regionalnimi

1 V članku uporabljamo termin regionalne jezikovne različice v smislu jezikovnega sistema, ki je odvisen od geografskega oz. regionalnega izvora jezikovnega uporabnika.

podkorpusi glede na šest glavnih kategorij nestandardnih jezikovnih prvin iz tipologije: izpusti, transformacije, nestandardno besedje, različice pogostih besed, nestandardno oblikoslovje in drugo. V zaključku strnemo ugotovitve, orišemo uporabnost korpusa za jezikoslovne raziskave in navedemo načrte za prihodnje delo.

2 REGIONALNA JEZIKOVNA VARIANTNOST V GOVORU IN RAČUNALNIŠKO POSREDOVANI KOMUNIKACIJI

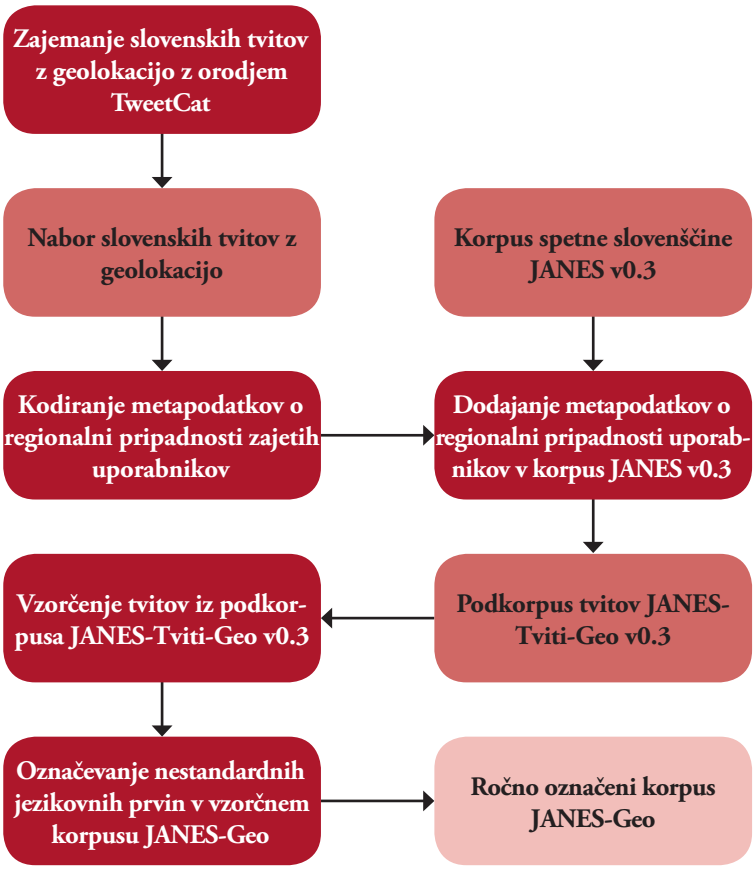
Raziskava, opisana v tem prispevku, se umešča na področje sociolingvistike, natančneje v disciplini korpusne dialektologije in korpusne dialektometrije (Szmrecsanyi 2011), ki temelji na analizi korpusnih besedil za sistematično merjenje razdalj in razlik med jezikovno rabo uporabnikov iz različnih geografskih regij. Za številne tuje jezike so bile že opravljene korpusne dialektološke raziskave, najpogosteje na podlagi transkripcij govora v govornih korpusih, kot je npr. korpus DynaSAND za nizozemska narečja (Kunst in Wesseling 2010), Nordic Dialect Corpus za nordijske jezike (Johanessen et al. 2009) in Freiburg Corpus of English Dialects za regionalne jezikovne različice angleščine (Hernández 2006). Slovenska dialektologija se je doslej zanašala predvsem na terenske raziskave z informanti (glej npr. Logar 1981, Kenda Jež 2002), ustrezen dialektološki korpus, ki bi omogočal korpusni pristop k problemu, pa še ni bil zgrajen. Korpus govorjene slovenščine GOS (Verdonik in Zwitter Vitez 2011) sicer vsebuje posnetke govorcev iz vseh regij, a primarno ni bil zasnovan za dialektološke namene. Na tej točki je treba omeniti tudi to, da so se slovenske dialektološke raziskave do zdaj opirale na t. i. idealnega narečnega govorca (Bitenc 2016: 180), pri katerem ni prišlo do prilagajanja drugim jezikovnim različicam. Ta pristop pa zanemarja širok nabor govorcev in jezikovnih različic, ki v različnih vidikih odstopajo tako od standarda kot od t. i. čistega narečja. V okviru raziskave v tem prispevku se ne osredotočamo na določen tip govorca, temveč jezikovno variantnost opisujemo empirično in brez predpostavk, tudi zato, ker gre pri računalniško posredovani komunikaciji za drug medij (pisni).

Raziskovanje regionalnih jezikovnih prvin v uporabniških spletnih vsebinah je tudi v tujini še precej sveže. S tem področjem so se do zdaj ukvarjali pretežno jezikovni tehnologi, jezikoslovci pa v mnogo manjši meri. Raziskave so se osredotočale predvsem na gradnjo novih orodij npr. za avtomatsko prepoznavanje regionalnih jezikovnih različic (Harrat et al. 2013, Cotterell in Callison-Burch 2014 za arabščino; Eisenstein et al. 2010, Eisenstein et al. 2015 za ameriško angleščino; Ljubešić in Kranjčić 2014 za hrvaščino, srbščino, bosanščino in

črnogorščino), za strojno prevajanje (Harrat et al. 2014 med regionalnimi jezikovnimi različicami arabščine in sodobno standardno arabščino; Haddow et al. 2013 med dunajsko regionalno jezikovno različico nemščine in standardno avstrijsko nemščino) in oblikoskladenjsko označevanje (Khakimov et al. 2015 za tatarsko narečje mišar; Ruef in Ueberwasser 2013 za švicarsko nemščino; Bernhard in Ligozat 2013 za alzaščino). Gre torej za raziskovalno področje, ki pokriva zelo raznovrsten nabor jezikov, tudi neinstitucionalnih in takšnih z majhnim številom govorcev. To kaže na svetovni trend, ki potrjuje, da bi bilo tudi slovenska jezikovnotehnološka orodja smiselno prilagoditi, da bodo dovolj robustna za obdelavo regionalnih jezikovnih različic na spletu, slovenščini pa za nadaljnji korak v to smer manjka neobremenjen jezikovni opis spletne slovenščine in rabe regionalnih jezikovnih različic v njej. Raziskovanje značilnosti regionalnih jezikovnih različic na spletu se je začelo v jezikoslovni skupnosti razraščati prav zdaj: tudi za angleščino so bila namreč šele pred kratkim objavljena obširnejša dela s tega področja. Grieve (2016) npr. predstavi sodobni korpusni in statistično podprti pristop k dialektologiji in dialektometriji na primeru regionalnih jezikovnih različic pisne ameriške angleščine, o proučevanju regionalne jezikovne členjenosti na družbenih medijih na splošno pa pišejo Eisenstein (2015) in Jørgensen et al. (2015). Za raziskavo v tem prispevku je še posebej relevantnen prispevek Huang et al. (2016), ki obravnava regionalne jezikovne različice ameriške angleščine na Twitterju s pomočjo tvitov z geolokacijo.

3 IZDELAVA ROČNO OZNAČENEGA KORPUSA JANES-GEO

V naslednjih podrazdelkih opisujemo izdelavo ročno označenega korpusa Janes-Geo, ki je bil vzorčen iz korpusa slovenske računalniško posredovane komunikacije Janes v0.3 (Fišer et al. 2015). Izdelava je potekala v več stopnjah, ki jih prikazuje delotok na Sliki 1. Zeleni okvirčki predstavljajo postopke, modri uporabljene podatkovne zbirke, oranžni okvirček pa končni rezultat. Najprej smo zajeli tvite s podatki o geolokaciji in na njihovi podlagi kodirali metapodatke o regionalni pripadnosti zajetih uporabnikov. Metapodatke smo nato dodali v že obstoječi korpus Janes v0.3, na njihovi podlagi pa smo iz njega nato vzorčili besedila za ročno označeni korpus Janes-Geo. Temu sta sledila še izdelava tipologije nestandardnih jezikovnih prvin in ročno označevanje korpusa.

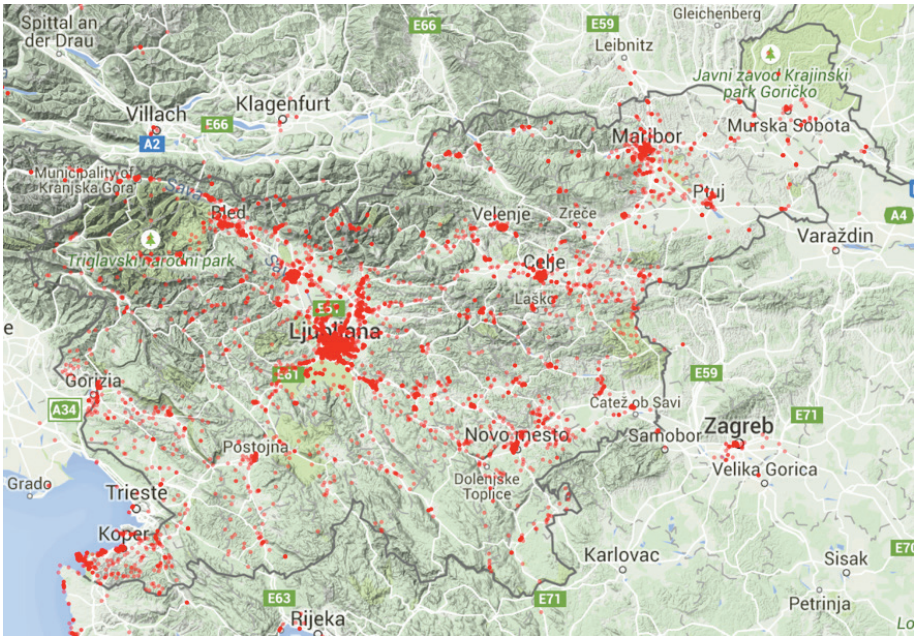


Slika 1: Postopek izdelave ročno označenega korpusa Janes-Geo.

3.1 Zbiranje tvitov s podatki o geolokaciji

Na prvi stopnji smo zbirali slovenske tvite s podatki o geolokaciji, tj. s podatki o zemljepisni širini in dolžini, s katerih je bil tvit poslan. Tvite smo začeli zajemati januarja 2015 z orodjem TweetCat (Ljubešić et al. 2014) in v približno pol leta (do avgusta 2015) zajeli 130.143 tvitov, ki jih je objavilo 1.661 uporabnikov. Razporeditev zajetih tvitov glede na njihove koordinate prikazuje Slika 2.

Iz razporeditve je razvidno, da zajeti tviti pokrivajo dobršen del Slovenije. Največ tvitov je bilo sicer poslanih iz mestnih središč in njihove neposredne okolice, npr. iz Ljubljane, Maribora, Celja in Kranja, a so zastopana tudi manj urbana območja.



Slika 2: Razporeditev zajetih tvitov z geolokacijo.

Na tej stopnji smo upoštevali samo zasebne uporabnike, ki so bili že vključeni v korpus Janes v0.3, ostale pa smo izločili. Za ta poseg smo se odločili ob predpostavki, da lahko od uporabniških računov organizacij, kot so agencije, medijske hiše in podjetja, s precejšnjo verjetnostjo pričakujemo, da na Twitterju v prevladujoči meri objavljajo tvite v standardni slovenščini, obenem pa objavljajo mnogo več avtomatsko generiranih tvitov, kar bi v naš korpus vneslo šum.

Po izločitvi nezasebnih uporabniških računov je ostalo 119.236 tvitov (približno 92 % vseh zajetih), ki jih je napisalo 1.524 uporabnikov. V korpusu Janes v0.3 je zasebnih uporabnikov 5.806, torej smo z zbiranjem tvitov z geolokacijo do avgusta 2015[2] zajeli približno četrtino (26 %) v korpus vključenih uporabnikov.

3.2 Kodiranje metapodatkov o regionalni pripadnosti

V naslednjem koraku smo z orodjem Google Maps API v3 Tool[3] območje Slovenije (vključno z zamejskimi regijami v Furlaniji, na avstrijskem Koroškem in v

2 Z raziskavo smo začeli avgusta 2015 in takrat prvič izvozili zajete tvite z geolokacijo, tvite pa smo zajemali še naprej in novopridobljene podatke (do aprila 2016) uporabili za preverjanje zanesljivosti metode avtomatskega pripisovanja metapodatkov o geolokaciji (več o tem v razdelku 3.2.1).

3 Google Maps API v3 Tool: http://www.birdtheme.org/useful/v3tool.html.

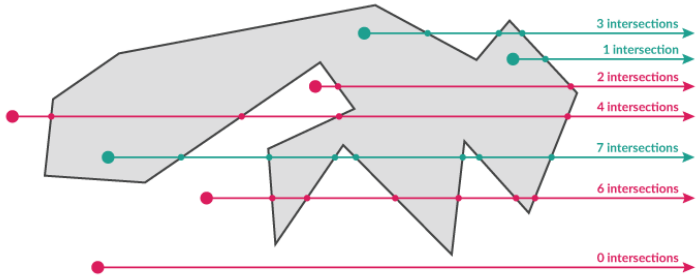
Porabju) razdelili na koordinatne poligone, za kar smo uporabili orodje Google Maps API v3 Tool. Glede na dosedanje raziskave smo imeli na voljo več načinov delitve, ki odsevajo bodisi narečne skupine (Ramovš 1931; Logar in Rigler 1986; Toporišič 2000: 23) ali statistične regije (Zemljarič Miklavčič 2008). Za namene te raziskave smo izbrali delitev na regije[4] v skladu s sedmimi glavnimi narečnimi skupinami po Ramovšu (1931), saj je bila ta kategorizacija med vsemi najbolj robustna, podrobnejša delitev pa bi zaradi neenakomerne razporeditve zbranih podatkov povzročala dodatne težave pri vzorčenju. Slovensko govoreče območje smo tako razdelili na skupno devet koordinatnih poligonov. Prvih sedem predstavlja narečne skupine (gorenjsko, dolenjsko, štajersko, panonsko, koroško, rovtarsko in primorsko), dodatna poligona pa predstavljata Ljubljano in Maribor, ki smo se ju odločili obravnavati posebej kot urbani središči, h katerima gravitira prebivalstvo iz številnih drugih krajev (tako okoliških kot bolj oddaljenih) in ki bi kot taki vnesli precejšnjo mero šuma v druge regije. Tak pristop zagovarja tudi Zemljarič Miklavčič (2008: 79) pri zasnovi govornih korpusov. Tako nastale koordinatne poligone predstavlja Slika 3 (ljubljanski in mariborski poligon zaradi majhnosti nista prikazana).

Slika 3: Razdelitev slovensko govorečega področja na koordinatne poligone.

S pomočjo metode metanja žarka (ang. *ray-casting method*; Sparks in Krishnan 2012) smo za vsak zajeti tvit z geolokacijo preverili, iz katerega koordinatnega

4 V članku se pri gradnji in analizi korpusa Janes-Geo s terminom »regije« ne nanašamo na narečne skupine, temveč na koordinatne poligone, na podlagi katerih smo določili metapodatke o regionalni pripadnosti uporabnikov. Za ločevanje terminov narečna skupina in regija smo se odločili, ker metapodatki o regionalni pripadnosti uporabnikov ne odsevajo nujno njihovega jezikovnega ozadja v skladu s kategorizacijo po narečnih skupinah, obenem pa se naša členitev ne prekriva povsem s členitvijo na narečne skupine, saj Ljubljano in Maribor obravnavamo posebej.

poligona je bil poslan. Metoda metanja žarka iz podane točke (v našem primeru so to koordinate tvita) zariše premico oz. žarek ter preveri število presečišč med žarkom in robovi podanega poligona. Če je število liho, točka leži v notranjosti poligona (Slika 4).



Slika 4: Prikaz metode metanja žarka.

Tabela 1 prikazuje razporeditev zajetih tvitov po regijah.[5] V razporeditvi tvitov se že v tem pogledu kaže precejšnja razlika med regijami, saj je bilo daleč največ tvitov (36 %) poslanih iz Ljubljane, najmanj pa iz rovtarske (slaba 2 %) in panonske regije (dobra 2 %), ki sta tudi po površini med najmanjšimi (če ne štejemo Ljubljane in Maribora). Zanimivo je, da poligona Maribora in Ljubljane po količini tvitov nista primerljiva, saj je bilo kar desetkrat več tvitov poslanih iz Ljubljane kot iz Maribora. Dobro zastopana je tudi Gorenjska (slabih 19 %), precejšen delež tvitov pa je bil poslan tudi iz tujine (slabih 16 %).

Tabela 1: Razporeditev tvitov z geolokacijo po regijah.

Regija	Število tvitov	Delež (%)
Gorenjska	22.070	18,51
Dolenjska	6.922	5,81
Štajerska	9.284	7,79
Panonska	2.512	2,11
Koroška	4.203	3,52
Primorska	5.748	4,82
Rovtarska	2.348	1,97
Ljubljana	43.018	36,08
Maribor	4.340	3,64
Tujina	18.791	15,76
Skupno	119.236	100,00

5 Tabela 1 prikazuje tudi tvite, ki so bili poslani iz tujine. Ti ne vključujejo tvitov, poslanih iz Furlanije, avstrijske Koroške in Porabja – te smo dodali primorski, koroški in panonski regiji. Tvitov, ki so bili poslani iz tujine, pri izdelavi in vzorčenju korpusa za ročno označevanje nismo upoštevali.

V naslednjem koraku smo za vsakega od uporabnikov izračunali, kolikšen delež svojih tvitov je poslal iz vsakega od devetih poligonov, glede na deleže pa smo vsakemu uporabniku pripisali metapodatek o regionalni pripadnosti. Ob predpostavki, da so uporabniki pogosto tudi mobilni in objavljajo iz vsaj dveh regij (zlasti če upoštevamo, da so bili tviti zajeti v polletnem obdobju, ki vključuje tudi poletne počitniške mesece), smo določili hevristiko, s katero smo zmanjšali medregionalni šum in se osredotočili samo na podatke, značilnejše za regijo. Metapodatke smo tako pripisali samo uporabnikom, ki so več kot 90 % tvitov poslali iz ene same regije in so obenem poslali vsaj 3 tvite.

Uporabnikov, ki so izpolnjevali oba kriterija, je bilo skupno 269, razrez čistih uporabnikov[6] po regijah pa prikazuje Tabela 2. Uporabniki, ki so tvite pošiljali večinoma iz tujine, za našo raziskavo niso relevantni, a jih kljub temu navajamo v tabeli, saj predstavljajo nezanemarljiv delež.

Tabela 2: Razporeditev uporabnikov po regijah.

Regija	Število vseh zasebnih uporabnikov	Delež (%)	Število čistih uporabnikov	Delež (%)	Razmerje med čistimi in vsemi zasebnimi uporabniki v regiji
Gorenjska	208	13,65	50	12,95	0,24
Dolenjska	92	6,04	23	5,96	0,25
Štajerska	170	11,15	46	11,92	0,27
Panonska	43	2,82	17	4,40	0,40
Koroška	33	2,17	6	1,55	0,18
Primorska	99	6,50	33	8,55	0,33
Rovtarska	37	2,43	7	1,81	0,19
Ljubljana	506	33,20	125	32,38	0,25
Maribor	59	3,87	14	3,63	0,24
Tujina	277	18,18	65	16,84	0,23
Skupno	1524	100,00	386	100,00	0,25

Skoraj tretjina vseh čistih uporabnikov je spadala v ljubljansko regijo, sledijo pa ji gorenjska (13 %), štajerska (12 %) in primorska regija (8,5 %). Najmanj čistih uporabnikov imata koroška in rovtarska regija (manj kot 2 % odstotka).

V zadnjem stolpcu je podan količnik razmerja med čistimi uporabniki in vsemi zasebnimi uporabniki znotraj regije. Večji količnik pomeni večji delež čistih uporabnikov (in posledično manjšo mobilnost uporabnikov znotraj regije). Pri večini regij je čistih uporabnikov približno četrtina. Izstopajo panonska regija z nekoliko

6 V prispevku uporabnike, ki izpolnjujejo kriterije za pripis metapodatka o regionalni pripadnosti (tj. 90-odstotni prag za delež tvitov iz dominantne regije in najmanj 3 poslani tviti), imenujemo *čisti uporabniki*.

manjšo mobilnostjo ter panonska in koroška regija, kjer je čistih uporabnikov nekoliko manj.

Metapodatke o regionalni pripadnosti smo nato vnesli v korpus Janes v0.3 in izdelali podkorpus Janes-Tweet-Geo v0.3.4, iz katerega smo v naslednjem koraku vzorčili besedila za vzorčni korpus Janes-Geo (glej razdelek 3.3).

3.2.1 Ponovitev kodiranja metapodatkov

Tvite z geolokacijo smo ponovno izvozili aprila 2016. Na tej točki je bilo vseh zajetih tvitov 160.888, kar je 20 % več kot v prvi fazi zbiranja.[7] Na novozajetih tvitih smo ponovili avtomatsko pripisovanje metapodatkov, razporeditev tvitov in uporabnikov po regijah pa se je med fazama večinoma ohranila, kot je razvidno iz Tabele 3 in Tabele 4.

Tabela 3: Razporeditev in porast tvitov po regijah v prvi in drugi fazi zajemanja.

Regija	Število tvitov (1. faza)	Delež (%)	Število tvitov (2. faza)	Delež (%)	Porast tvitov (%)
Gorenjska	22.070	18,51	27.399	17,03	+24 %
Dolenjska	6.922	5,81	9.864	6,13	+43 %
Štajerska	9.284	7,79	15.989	9,94	+72 %
Panonska	2.512	2,11	3.873	2,41	+54 %
Koroška	4.203	3,52	5.170	3,21	+23 %
Primorska	5.748	4,82	8.383	5,21	+46 %
Rovtarska	2.348	1,97	2.873	1,79	+22 %
Ljubljana	43.018	36,08	57.008	35,43	+33 %
Maribor	4.340	3,64	6.116	3,80	+41 %
Tujina	18.791	15,76	24.213	15,05	+29 %
Skupno	119.236	100,00	160.888	100,00	+35 %

V večini regij se je število tvitov povečalo za približno 25–30 %, največji porast pa so zabeležile štajerska (72 %), panonska (54 %) in primorska regija (46 %).

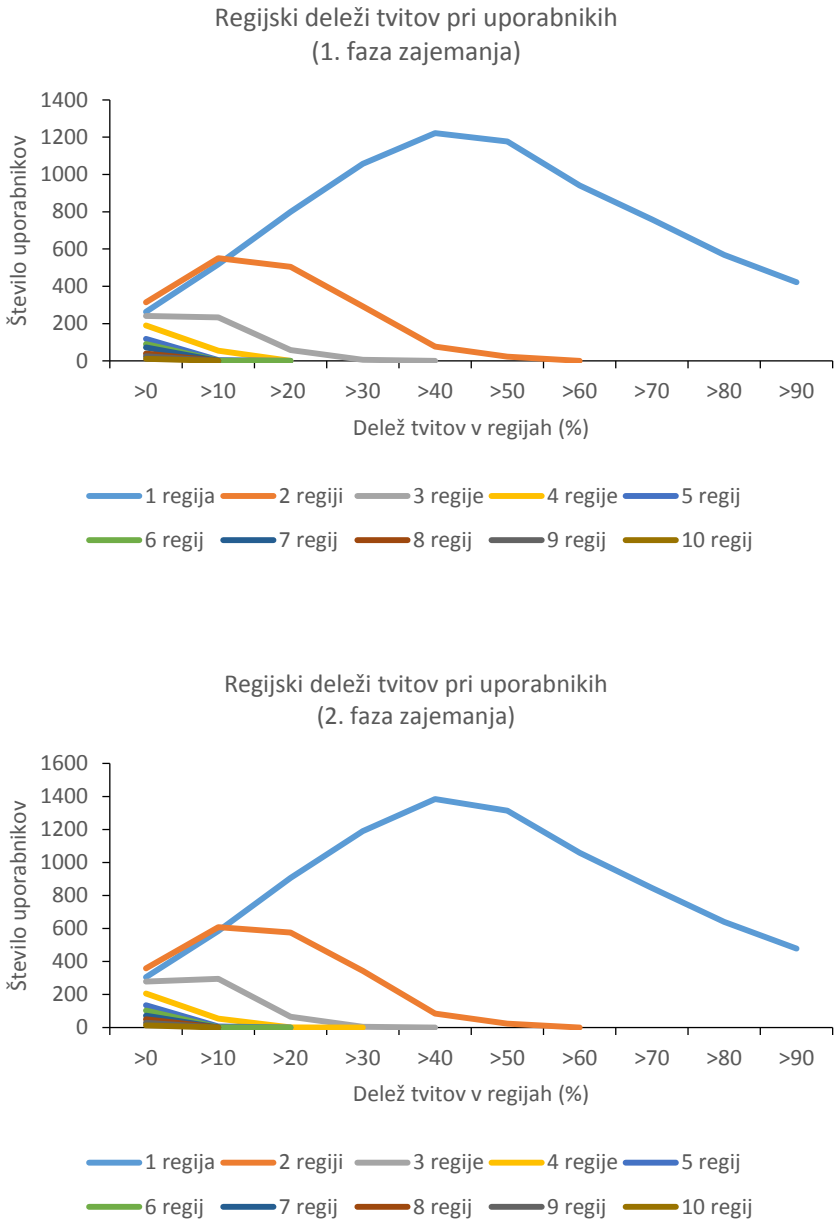
7 Na tej točki je treba omeniti, da so nekateri uporabniki med prvo in drugo fazo zajemanja svoje tvite (ali celo uporabniške račune) izbrisali.

Tabela 4: Razporeditev in porast uporabnikov po regijah v prvi in drugi fazi zajemanja.

Regija	Število čistih uporabnikov (1. faza)	Delež (%)	Število čistih uporabnikov (2. faza)	Delež (%)	Porast uporabnikov (%)
Gorenjska	50	12,95	50	13,89	0 %
Dolenjska	23	5,96	21	5,83	-9 %
Štajerska	46	11,92	43	11,94	-7 %
Panonska	17	4,40	15	4,17	-12 %
Koroška	6	1,55	4	1,11	-33 %
Primorska	33	8,55	30	8,33	-9 %
Rovtarska	7	1,81	7	1,94	0 %
Ljubljana	125	32,38	109	30,28	-13 %
Maribor	14	3,63	15	4,17	+7 %
Tujina	65	16,84	66	18,33	+2 %
Skupno	386	100,00	360	100,00	-7 %

Da bi preverili, do kolikšne mere so avtomatsko pripisani metapodatki stabilni in zanesljivi, smo primerjali spremembe v razporeditvi deležev tvitov pri uporabnikih med prvo in drugo fazo. Rezultati primerjave kažejo, da je populacija zelo stabilna, saj so uporabniki med fazama v veliki večini ohranili svojo dominantno regijo: od čistih uporabnikov sta namreč samo dva (0,12 % vseh čistih uporabnikov v prvi fazi) ob prehodu v drugo fazo spremenila dominantno regijo, omeniti pa je treba tudi, da sta se iz tujine premaknila v ljubljansko regijo, kar pomeni, da ju v raziskavi kot uporabnikov iz tujine nismo upoštevali. 32 čistim uporabnikom iz prve faze se je delež v dominantni regiji nekoliko zmanjšal in padel pod 90-odstotni prag, zato v drugi fazi niso bili več obravnavani kot čisti uporabniki. Iz tega razloga je število čistih uporabnikov med fazama nekoliko upadlo, kar smo pričakovali, saj so nekateri uporabniki v prvi fazi poslali le majhno količino tvitov (med 3 in 10), ki po vsej verjetnosti niso ustrezno predstavljali njihove regionalne razporeditve.

Slika 5 prikazuje število uporabnikov glede na število regij, iz katerih so poslali določen delež svojih tvitov. Vsaka črta predstavlja število regij, v katerih so uporabniki prisotni. Kot vidimo, je zelo malo uporabnikov (okrog 150), ki so tvite pošiljali iz več kot treh regij, obenem pa je tudi malo uporabnikov (okrog 50), ki imajo v več kot dveh regijah delež tvitov nad 20 %. Večina jih ima eno dominantno regijo, iz katere so poslali 50–90 % tvitov, preostali delež pa je razdeljen med druge regije. Med grafoma prve in druge faze ni bistvenih razlik, iz česar lahko sklepamo, da podatki precej zanesljivo predstavljajo regionalno dinamiko uporabnikov, vključenih v pričujočo raziskavo.



Slika 5: Primerjava regijske razporeditve uporabnikov v prvi in drugi fazi.

3.3 Vzorčenje korpusa Janes-Geo

Janes-Tweet-Geo v0.3.4 je podkorpus korpusa Janes v0.3, ki zajema približno 268.000 tvitov in 4,5 milijona pojavnic, vanj pa je vključenih vseh 386 uporabnikov, ki so bili med kodiranjem metapodatkov o regionalni pripadnosti opredeljeni kot čisti (glej opombo 5). Tabela 5 prikazuje prerez podkorpusa Janes-Tweet-Geo v0.3.4 po regionalnih metapodatkih uporabnikov.

Tabela 5: Sestava podkorpusa Janes-Tweet-Geo v0.3.4 po regionalnih podkorpusih.

Regija	Tviti	Delež (%)	Pojavnice	Delež (%)	Pojavnice L1	Pojavnice L2	Pojavnice L3	Delež L1 (%)	Delež L2 (%)	Delež L3 (%)
Gorenjska	39.961	15	620.386	14	337.157	180.769	102.460	54	29	17
Dolenjska	18.204	7	312.783	7	220.718	69.227	22.838	71	22	7
Štajerska	43.787	16	751.658	17	497.089	202.650	51.919	66	27	7
Panonska	5.198	2	76.048	2	55.814	18.253	1.981	73	24	3
Koroška	6.518	2	129.104	3	74.382	47.256	7.466	58	37	6
Primorska	14.211	5	241.023	5	176.326	53.370	11.327	73	22	5
Rovtarska	5.215	2	81.844	2	52.660	22.767	6.417	64	28	8
Ljubljana	96.217	36	1.569.946	36	1.113.182	369.921	86.843	71	24	6
Maribor	5.001	2	81.276	2	60.417	18.013	2.846	74	22	4
Tujina	33.632	13	547.655	12	358.477	151.928	37.250	65	28	7
Skupno	267.944	100	4.411.723	100	2.946.222	1.134.154	331.347	67	26	8

Tudi v korpusu Janes-Tweet-Geo v0.3.4 se je ohranila podobna razporeditev tvitov po regijah kot pri zajetih tvitih z geolokacijo, na podlagi katerih smo pripisali metapodatke. Korpus Janes-Tweet-Geo v0.3.4 je bil avtomatsko označen tudi s stopnjo jezikovne nestandardnosti (Ljubešić et al. 2015). V povprečju je 67 % vsakega regionalnega podkorpusa označenega kot standardnega (L1), okrog 26 % delno nestandardnega (L2) in le okrog 7 % nestandardnega (L3). Pojavnic, ki so se pojavljale v nestandardnih tvitih (L3), je torej le okrog 330.000.

V večini tvitov, ki so vključeni v korpus Janes-Tweet-Geo v0.3.4, torej ne pričakujemo velike mere nestandardnih jezikovnih prvin, zato smo kot primarni vir gradiva za vzorčni korpus Janes-Geo vzeli tvite z najvišjo stopnjo jezikovne nestandardnosti (L3). V vzorčni korpus smo hoteli vključiti skupno 4.500 tvitov, tj. po 500 tvitov iz vsake od devetih regij (tvite iz tujine smo izključili). Iz korpusa Janes-Tweet-Geo v0.3.4 smo najprej izvozili vse tvite z oznako L3, nato pa glede na število uporabnikov v regiji določili maksimalno količino tvitov (med 25 in 50), ki jih je uporabnik lahko prispeval v vzorec. Na ta način smo poskrbeli, da je vzorec karseda uravnotežen, saj so razlike v produktivnosti med uporabniki lahko

zelo velike (nekateri so poslali le po 3 tvite, drugi pa tudi po 2000). Pri regijah z zelo majhno količino podatkov (Koroška, Rovtarska, Maribor in Panonska), v katerih po tovrstnem vzorčenju ni bilo mogoče zbrati 500 nestandardnih tvitov, smo vključili tudi tvite s stopnjo jezikovne standardnosti L2. Končno sestavo vzorčnega korpusa Janes-Geo prikazuje Tabela 6.

Tabela 6: Sestava vzorčnega korpusa Janes-Geo.

Regija	Število tvitov	Delež tvitov (%)	Število pojavnic	Delež pojavnic (%)
Gorenjska	500	12,58	8.502	13,28
Dolenjska	500	12,58	7.957	12,43
Štajerska	500	12,58	8.209	12,82
Panonska	500	12,58	7.222	11,28
Koroška	258	6,49	4.630	7,23
Primorska	500	12,58	8.560	13,37
Rovtarska	383	9,64	5.948	9,29
Ljubljana	500	12,58	7.702	12,03
Maribor	333	8,38	5.281	8,25
Skupno	3.974	100,00	64.011	100,00

Vzorčni korpus Janes-Geo vsebuje približno 4.000 tvitov oz. 64.000 pojavnic, kar predstavlja okrog 1,5 % celotnega korpusa Janes-Tweet-Geo v0.3.4 oziroma slabo petino (19 %) vseh nestandardnih tvitov (L3) v njem. Večina regij zajema po približno 12–13 % korpusa, izjeme pa so Koroška (7 %), Rovtarska (9 %) in Maribor (8 %). V vseh treh regijah je bilo uporabnikov premalo, da bi po kriterijih vzorčenja (tudi z vključitvijo tvitov L2) prispevali 500 tvitov.

3.4 Izdelava tipologije in smernic za označevanje nestandardnih jezikovnih prvin v tvitih

Na podlagi ročnega pregleda 200 tvitov iz vsake regije (skupno torej 1800 tvitov) smo zabeležili vse za raziskavo relevantne nestandardne jezikovne prvine. Te so večinoma zajemale nivoja zapisa in besedišča, v manjši meri pa tudi oblikoslovje in nekatere druge, bolj priložnostne in manj sistematične spremembe.[8] Vse zabeležene nestandardne prvine smo nato hierarhično kategorizirali ter izdelali tipologijo in smernice za označevanje nestandardnih jezikovnih prvin v tvitih (Čibej

8 V raziskavi nismo upoštevali rabe ločil, šumnikov in velike začetnice, saj se v računalniško posredovani komunikaciji pogosto opuščajo, kar je lahko tudi posledica naprave, s katere uporabnik pošilja tvit (npr. telefonski zapisi brez šumnikov). V prvotni različici tipologije smo predvidevali tudi skladenjski nivo, a smo v vzorčnih tvitih odkrili zanemarljivo malo nestandardnih skladenjskih pojavov, zato smo kategorijo odstranili.

2017).[9] Tipologija v trenutni različici (v1.0) vsebuje 6 glavnih kategorij (izpuste, transformacije, nestandardno oblikoslovje, nestandardno besedje, nestandardne različice pogostih besed z variantnimi oblikami, drugo) in skupno 292 različnih oznak, podrobnejši prerez s številom oznak na kategorijo pa prikazuje Tabela 7. Več kot polovico (58 %) različnih oznak zavzemajo izpusti, skoraj petino pa transformacije.

Tabela 7: Število oznak na kategorijo v tipologiji nestandardnih jezikovnih prvin v slovenskih tvitih.

Kategorija	Število različnih oznak	Delež (%)
Izpusti	170	58,22
Transformacije	58	19,86
Nestandardno oblikoslovje	9	3,08
Nestandardno besedje	13	4,45
Nestandardne različice pogostih besed z variantnimi oblikami	27	9,25
Drugo	15	5,14
Skupno	292	100,00

3.4.1 Pregled glavnih kategorij tipologije

V tem razdelku na kratko predstavimo glavne kategorije tipologije nestandardnih jezikovnih prvin in njihove oznake.

Kategorija, v katero spada največ nestandardnih jezikovnih prvin v tvitih, so **izpusti**, ki jih v kontekstu te raziskave definiramo kot izpuščanje grafemov pri zapisu v primerjavi z neposredno standardno različico besede, delimo pa jih na dve podkategoriji, in sicer na izpuste soglasnikov (*glej* → *lej*) ter izpuste samoglasnikov (*sovražim* → *sovražm*). Oznake pri izpustih lahko vsebujejo naslednje podatke:

a) ali gre za izpust soglasnika (*Ik*) ali samoglasnika (*Iv*),

b) ali gre za izpust končnega (*Ikk*, *Ivk*) ali nekončnega grafema (*Ikn*, *Ivn*),

c) katera je besedna vrsta besede, v kateri je prišlo do izpusta (*G* – glagol, *S* – samostalnik, *P* – pridevnik, *R* – prislov in *D* – drugo),

d) oblikoskladenjske značilnosti besede, v kateri je prišlo do izpusta (npr. število, spol, sklon),

e) kateri grafem je bil izpuščen.

9 Označevalne smernice in tipologija so prosto dostopne na uradni spletni strani projekta JANES: http://nl.ijs.si/janes/viri/.

Oznaka *IvnSmei.e* npr. označuje izpust (*I-*) nekončnega samoglasnika *e* (*-vn-* in *-.e*) v samostalniku (*-S-*) moškega spola (*-m-*) v ednini (*-e-*) in imenovalniški obliki (*-i-*), npr. *teden* → *tedn*, *konec* → *konc*.

Pri kategoriji **transformacij** smo zabeležili vse grafeme, ki jih je uporabnik v besedi zapisal drugače, kot bi to zahtevala standardnoslovenska različica besede (*všeč* → *ušeč*, *mislila* → *mislala*). Oznake transformacij vsebujejo podatke o izvirnem grafemu (ali sklopu grafemov) in o ciljnem grafemu (ali sklopu grafemov). Oznaka *Ta.o* npr. označuje transformacijo grafema *-a* v grafem *-o*, npr. *prav* → *prov*.

Jezikovne prvine, ki smo jih uvrstili v kategorijo **nestandardnega oblikoslovja**, so se v vzorčnem korpusu pojavljale zelo sporadično in mnogo redkeje v primerjavi z ostalimi prvinami, a smo za razliko od skladnje kategorijo v tipologiji ohranili, saj so bili pojavi v njej bolj sistematični. V kategorijo smo uvrstili nestandardna obrazila pri glagolih (*greva* → *grema*, *bova* → *boma*, *morava* → *morve*) in samostalnikih (*Bučku* → *Bučkotu*, *s penziči* → *s penzičmi*). Pri ostalih pregibnih besednih vrstah oblikoslovnih posebnosti nismo zaznali.

V kategorijo **nestandardnega besedja** smo dodali vse besede, ki smo jih v kontekstu dojemali kot nestandardne,[10] pri čemer smo se opirali predvsem na kvalifikatorje (npr. *pogovorno*, *narečno*) v referenčnih virih, kot so Slovar slovenskega knjižnega jezika, Slovar novejšega besedja, Sprotni slovar slovenskega jezika, Slovenski pravopis in Slovenski etimološki slovar).[11] Omeniti je treba, da je bilo ocenjevanje nestandardnosti pri pojavnicah, ki niso bile opisane v referenčnih virih, do določene mere nujno subjektivno – to še zlasti velja za besede, privzete iz tujih jezikov. Kategorijo nestandardnega besedja smo razdelili na devet podkategorij: pogovorne/slengovske/narečne/žargonske besede (*NSB.Pog*, npr. *gujdek*, *kafič*, *spizditi*), germanizmi (*NSB.Ger*, npr. *cajt*, *zihr*), anglizmi[12] (*NSB.Ang*, npr. *kjut*, *appov*), kroatizmi/srbizmi (*NSB.Srb*, npr. *svašta*, *rukohvatskim*), italianizmi (*NSB.Ita*, npr. *mona*, *birca*), hispanizmi (*NSB.Špa*, npr. *el clasico*), francizmi (*NSB.Fra*, npr. *passé*), besede iz spletnega jezika (*NSB.Net*, npr. *jbg – jebiga*, *s5 – spet*) ter priložnostne/ustvarjalne tvorjenke (*NSB.Kre*, npr. *butljezik*, *paradajzkomunajzar*).

V kategorijo **različic pogostih besed** smo uvrstili omejen nabor besed, ki so se med označevanjem v besedilih pogosto pojavljale v več različicah zapisa, npr. osebni zaimek *jaz*, ki se je v vzorčnem korpusu pojavljal v oblikah *jst*, *js*, *jz*, *jest*, *ist* in *ject*, ali prislov *zdaj*, ki se je pojavljal kot *zdj*, *zj*, *zdej*, *zej* in *zaj*. Označili smo samo

10 Pri tej kategoriji nismo upoštevali besed, ki imajo neposredne standardne ustreznice, njihova nestandardnost pa se izkazuje na način, ki spada pod druge kategorije te tipologije – to so npr. besede z izpuščenimi samoglasniki (*tudi* → *tud*) ali besede, v katerih je prišlo do transformacij samoglasnikov (*utrgalo* → *ftrgalo*) ali soglasnikov (*bog* → *boh*).

11 http://fran.si/

12 Pri anglizmih smo označevali tudi stopnjo podomačitve, za kar smo uporabili oznake *NSB.Ang.N* (povsem nepodomačeno, npr. *shopping*), *NSB.Ang.P.F* (podomačitev, razvidna iz fonetiziranega zapisa, npr. *imidž*), *NSB.Ang.P.K* (podomačitev, razvidna iz končnice, npr. dumplingi) in *NSB.Ang.P.O* (podomačitev, razvidna tako iz fonetiziranega zapisa kot iz končnice).

nestandardne oblike. V končni različici označenega korpusa nam te oznake omogočajo, da opazujemo, ali so uporabniki pri rabi različic dosledni oz. ali izbirajo med več različicami, obenem pa lahko preverimo, ali so nekatere različice značilnejše za uporabnike iz določene regije. Oznake iz te kategorije vsebujejo samo normalizirano obliko označene besede, npr. *V.zakaj* za besede *zakej*, *zaka*, *zakva* ipd.

V kategorijo **drugo** smo vključili vse ostale nestandardne jezikovne prvine, ki jih nismo mogli uvrstiti v nobeno od prej naštetih kategorij, npr. pisanje skupaj (*Dskupaj*, npr. *ne bi* → *nebi*), sklapljanje besed (*Dsklop*, npr. *to je* → *toj*), vrivanje dodatnega grafema (*Dvrivanje*, npr. *zajtrk* → *zajterk*), raba nestandardnih besed, ki delujejo kot vezniki (*Dk*, *Dki*, *Dka* za *k*, *ki*, *ka*) in nestandardna raba kategorije živosti *(Dživost,* npr. *[matram] iphone* → *[matram] iphona*).

Poudariti je treba, da tipologija sicer vsebuje oznake za vse proučevane nestandardne prvine, ki smo jih zaznali v korpusu, a je kljub temu ne moremo obravnavati kot izčrpno, saj je vzorčni korpus, na podlagi katerega je bila izdelana, relativno majhen. S precejšnjo zanesljivostjo pa lahko trdimo, da vsebuje primere vseh najpogostejših kategorij nestandardnih jezikovnih pojavov, ki se pojavljajo v slovenskih tvitih, obenem pa je zastavljena tako, da jo je mogoče nadgraditi z novimi oznakami, zaradi česar je primerna tudi za morebitno prihodnje označevanje večjih vzorcev.

3.5 Označevanje korpusa Janes-Geo

V tvitih smo označevali samo pojavnice, ki smo jih po vsaj enem od kriterijev v smernicah dojemali kot nestandardne. Vsaki nestandardni pojavnici smo pripisali oznake za vse prisotne nestandardne jezikovne prvine, razen če ni bilo v smernicah določeno drugače. Pojavnici, v kateri se je npr. poleg dveh izpustov pojavila še ena transformacija, smo pripisali tri ustrezne oznake iz tipologije.

```
@user     jah, men so ble ušeč, zato sm jih tut kupu. pojamram
pa zato k niso ble lih zastonj, pa sm pač probu mal informacijo
razširt :/

@user       [jah]{NSB.Pog},   [men]{IvkD.i}   so   [ble]{IvnGd.i}
[ušeč]{Tv.u},   zato   [sm]{IvnGs1e.e}   jih   [tut]{IvkD.i}{Td.t}
[kupu]{Tl.u}{IvnGd.i}. [pojamram]{NSB.Ger} pa zato [k]{Dk} niso
[ble]{IvnGd.i} [lih]{NSB.Ger} zastonj, pa [sm]{IvnGs1e.e} pač
[probu]{NSB.Ger}{Tl.u}{IvnGd.a}     [mal]{IvkR.o}     informacijo
[razširt]{IvnGn.i}{IvkGn.i} :/
```

Slika 6: Primer tvita brez oznak in z oznakami za nestandardne jezikovne prvine.

Slika 6 prikazuje primer tvita, v katerem so bile označene vse nestandardne jezikovne prvine v skladu s smernicami. Vsaki nestandardni pojavnici je pripisana ena ali več ustreznih oznak. Pojavnica *lih* je npr. označena kot germanizem (*NSB.Ger*), pojavnica *probu* pa ima tri oznake (*NSB.Ger* za germanizem, *Tl.u* za transformacijo *-l* v *-u* (*probal* → *probau*) in *IvnGd.a* za izpust nekončnega samoglasnika *-a* v preteklem deležniku glagola (*probau* → *probu*).

4 ANALIZA OZNAČENEGA KORPUSA JANES-GEO

Iz končne različice ročno označenega korpusa Janes-Geo smo s pomočjo regularnih izrazov izvozili oznake in besede, ki so bile z njimi označene, in sicer na več nivojih: po regionalnih podkorpusih, po posameznih uporabnikih in za vsak posamezen tvit. V tem prispevku se osredotočamo samo na analizo razlik med posameznimi regionalnimi podkorpusi in v manjši meri na razlike med posameznimi uporabniki.

Tabela 8 prikazuje število besed,[13] ki so bile označene kot nestandardne, ter število oznak v korpusu Janes-Geo. Tretji stolpec podaja delež nestandardnih besed v primerjavi z vsemi besedami, zadnji stolpec pa delež oznak glede na število vseh oznak v korpusu.

Tabela 8: Števila in deleži nestandardnih besed ter oznak v korpusu Janes-Geo.

Regija	Število besed	Število nestandardnih besed	Delež (%)	Število oznak	Delež (%)
Gorenjska	7.834	1.836	23,44	2.552	20,56
Dolenjska	7.447	1.610	21,62	2.195	17,69
Štajerska	7.516	1.273	16,94	1.659	13,37
Panonska	6.539	436	6,67	499	4,02
Koroška	4.232	454	10,73	569	4,59
Primorska	7.823	1.325	16,94	1.758	14,17
Rovtarska	5.547	868	15,65	1.169	9,42
Ljubljana	6.930	1.148	16,57	1.504	12,12
Maribor	4.820	428	8,88	505	4,07
Skupno	58.688	9.378	15,98	12.410	100,00

Od približno 59.000 besed v celotnem korpusu jih je bilo okrog 9.300 označenih kot nestandardnih, kar predstavlja približno 16 % vseh besed. Že na tem nivoju lahko identificiramo razlike med regijami: z nekoliko višjo nestandardnostjo

13 Kot besede smo obravnavali vse pojavnice, ki niso ločila, številke, emotikoni, URL-naslovi, sklici na uporabniška imena (@avtor) ali ključniki (#kljucnik).

(21–23 %) izstopata gorenjska in dolenjska regija, manj nestandardne (7–10 %) pa so manjše regije, tj. panonska, mariborska in koroška. To je pričakovano, saj smo zaradi pomanjkanja podatkov vanje vključili tudi tvite z nižjo stopnjo jezikovne nestandardnosti (L2).

Tudi po številu in deležu oznak najbolj izstopata gorenjska in dolenjska regija, zanimivo pa je, da gorenjska v primerjavi z dolenjsko vsebuje nekoliko večji delež oznak (20,5 % in 17,5 %), čeprav je delež nestandardnih besed pri obeh regijah primerljiv. Ker se na eni besedi najpogosteje verižijo oznake za izpuste in transformacije, razlika v deležih nakazuje, da lahko v gorenjski regiji pričakujemo več izpustov in transformacij.

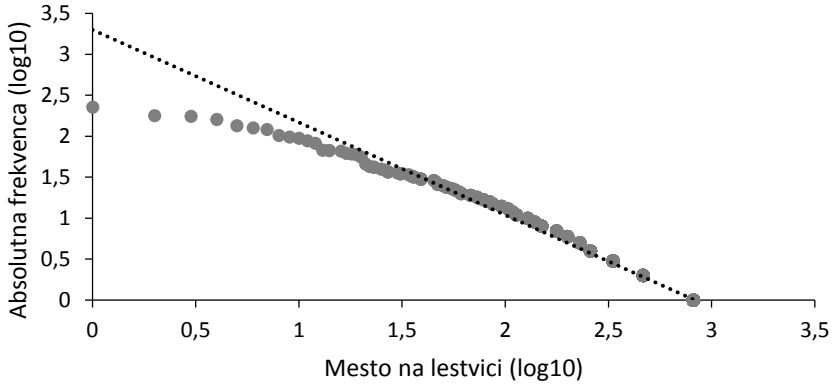
Tabela 9 prikazuje prvih 60 najpogostejših označenih besed iz korpusa Janes-Geo, ki imajo absolutno frekvenco večjo od 20. Da bi zmanjšali razpršenost podatkov, smo besede pri izvozu pretvorili v zapis z malimi začetnicami in brez šumnikov. Absolutna frekvenca za besedo *drgac* npr. zajema tudi različice *Drgac*, *Drgač*, *drgač* ipd. Relativna frekvenca je izračunana glede na število vseh besed v korpusu in normalizirana na 1.000 besed. Omeniti je treba tudi, da izvoz najpogostejših besed ni upošteval oznak, zato so nekatere oblike besed prekrivne (npr. *ka* kot različica vprašalnega zaimka *kaj* ali kot veznik; *ma* kot členek ali kot sedanjiška oblika glagola *imeti* v tretji osebi; *dobr* kot pridevnik moškega spola *dober*, kot pridevnik srednjega spola *dobro* ali kot prislov *dobro*).

Tabela 9: Prvih 60 najpogostejših nestandardnih besed iz korpusa Janes-Geo.

Mesto	Nestandardna beseda	Absolutna frekvenca	Relativna frekvenca	Mesto	Nestandardna beseda	Absolutna frekvenca	Relativna frekvenca
1	sm	227	3,87	31	cist	35	0,60
2	tud	179	3,05	32	fajn	35	0,60
3	a	176	3,00	33	itak	35	0,60
4	blo	161	2,74	34	tolk	34	0,58
5	sam	135	2,30	35	u	33	0,56
6	kr	127	2,16	36	lol	32	0,55
7	sej	121	2,06	37	lahk	32	0,55
8	jst	103	1,76	38	morm	32	0,55
9	k	98	1,67	39	dost	30	0,51
10	ma	95	1,62	40	tist	30	0,51
11	mal	88	1,50	41	skor	30	0,51
12	pol	82	1,40	42	jz	30	0,51
13	tko	68	1,16	43	evo	30	0,51
14	al	67	1,14	44	nism	30	0,51
15	dobr	67	1,14	45	zihr	29	0,49
16	mam	66	1,12	46	nevem	28	0,48
17	nc	62	1,06	47	bli	26	0,44

Mesto	Nestandardna beseda	Absolutna frekvenca	Relativna frekvenca	Mesto	Nestandardna beseda	Absolutna frekvenca	Relativna frekvenca
18	zdej	61	1,04	48	omg	26	0,44
19	js	60	1,02	49	bit	26	0,44
20	kej	56	0,95	50	tak	25	0,43
21	zdj	46	0,78	51	drgac	25	0,43
22	pr	43	0,73	52	zarad	24	0,41
23	ful	42	0,72	53	dej	24	0,41
24	bla	42	0,72	54	kul	24	0,41
25	neki	40	0,68	55	kva	23	0,39
26	tut	39	0,66	56	tok	23	0,39
27	ka	37	0,63	57	vidla	22	0,37
28	mas	37	0,63	58	okol	22	0,37
29	dons	37	0,63	59	wtf	22	0,37
30	men	36	0,61	60	vseen	21	0,36

Kot je razvidno iz Tabele 9, so tudi najpogostejše nestandardne besede v korpusu relativno redke, saj se pojavljajo v povprečju manj kot enkrat na 1.000 besed (povprečna relativna frekvenca prvih 60 najpogostejših besed je 0,95). Med najpogostejšimi nestandardnimi besedami najdemo nekatere pogoste glagole večinoma v sedanjiški obliki (*sm*, *nism*, *mam*, *mas*, *morm*) oz. kot pretekle deležnike ali nedoločnike (*bla*, *bli*, *vidla*, *bit*), osebne (*jst*, *js*, *jz*, *men*) in vprašalne zaimke (*ka*, *kej*, *kva*). Pogosti so tudi členki (*tud*, vprašalni členek *a*, *ma*, *evo*), vezniki (*sam*, *sej*, *k*, *ka*) ter prislovi (*zdej*, *zdj*, *zj*, *pol*, *dons*). Nestandardnost slovenskih uporabnikov Twitterja se torej pogosto in najbolj sistematično izkazuje v omejenem naboru zaprtih besednih vrst, pri odprtih besednih vrstah, kot so samostalniki, pridevniki in glagoli, pa je raba



Slika 7: Frekvenčna razporeditev nestandardnih besed v korpusu Janes-Geo.

nestandardnih oblik bolj sporadična in razpršena. Samostalniški besedi z najvišjo relativno frekvenco sta npr. *dnarja* (0,15; 149. mesto) in *cajt* (0,14; 151. mesto).

Slika 7 prikazuje razporeditev nestandardnih besed v korpusu Janes-Geo glede na njihovo absolutno frekvenco in mesto na lestvici. Razporeditev je razmeroma linearna in se v večjem delu dobro sklada z idealno Zipfovo distribucijo (koeficient trendne premice je -1,14), iz česar lahko sklepamo, da korpus Janes-Geo kljub majhnemu številu pojavnic relativno dobro predstavlja raznolikost nestandardnega besedišča uporabnikov Twitterja. Razporeditev od idealne Zipfove distribucije odstopa le pri približno 10 najpogostejših besedah, pri katerih je frekvenca nekoliko manjša od pričakovane, in pri besedah na zadnjem mestu lestvice, ki se v korpusu pojavijo zgolj enkrat. Takšnih besed je 2.686, kar znaša slabih 29 % vseh nestandardnih besed.

4.1 Pregled glavnih kategorij nestandardnih jezikovnih prvin

V tem razdelku predstavljamo analizo označenega korpusa z vidika vseh šestih glavnih kategorij oznak iz tipologije. Tabela 10 prikazuje število in delež vseh oznak znotraj različnih kategorij glede na regijo.

Tabela 10: Oznake v korpusu Janes-Geo po glavnih kategorijah in regijah.

Regija	Izpusti	Transformacije	Različice pogostih besed	Nestandardno besedje	Nestandardno oblikoslovje	Drugo
Gorenjska	1.321	348	302	435	5	141
	51,76 %	13,64 %	11, 83%	17,05 %	0,20 %	5,53 %
Dolenjska	1.152	300	214	415	8	106
	52,48 %	13,67 %	9,75 %	18,91 %	0,36 %	4,83 %
Štajerska	786	212	174	427	3	57
	47,38 %	12,78 %	10,49 %	25,74 %	0,18 %	3,44 %
Panonska	151	51	32	224	3	38
	30,26 %	10,22 %	6,41 %	44,89 %	0,60 %	7,62 %
Koroška	230	65	69	167	4	34
	40,42 %	11,42 %	12,13 %	29,35 %	0,70 %	5,98 %
Primorska	741	288	221	419	9	80
	42,15 %	16,38 %	12,57 %	23,83 %	0,51 %	4,55 %
Rovtarska	566	191	120	228	0	64
	48,42 %	16,34 %	10,27 %	19,50 %	0,00%	5,47 %
Ljubljana	695	198	128	394	8	81
	46,21 %	13,16 %	8,51 %	26,20 %	0,53 %	5,39 %
Maribor	217	66	29	178	7	8
	42,97 %	13,07 %	5,74 %	35,25 %	1,39 %	1,58 %
Skupno	5.859	1.719	1.289	2.887	47	609
	47,21 %	13,85 %	10,39 %	23,26 %	0,38 %	4,91 %

Najpogostejša kategorija nestandardnih jezikovnih prvin v korpusu so izpusti, ki zajemajo skoraj polovico oz. 47 % vseh oznak. Izpustom sledijo kategorije nestandardnega besedja (23 %), transformacij (14 %), različic pogostih besed (10 %) in drugo (5 %). Najmanj oznak je v kategoriji nestandardnega oblikoslovja (le 0,38 %).

4.1.1 Izpusti

Kot prikazuje Tabela 11, v korpusu močno prevladujejo izpusti samoglasnikov, ki v posamezni regiji zajemajo do 95 % vseh izpustov (oz. v celotnem korpusu približno 92 % vseh izpustov). Izpusti soglasnikov se pojavljajo mnogo redkeje (od 6 do 8 %). Z nekoliko večjim deležem izpustov soglasnikov izstopajo Panonska, Koroška in Maribor, a je to po vsej verjetnosti predvsem posledica vzorčenja.

Tabela 11: Izpusti samoglasnikov in soglasnikov v korpusu Janes-Geo.

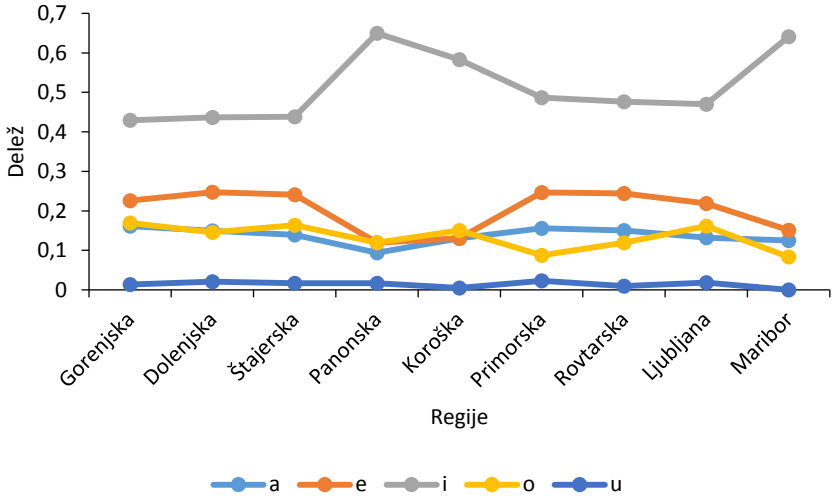
Regija	Izpust soglasnika	Delež (%)	Izpust samoglasnika	Delež (%)
Gorenjska	105	7,95	1216	92,05
Dolenjska	76	6,60	1076	93,40
Štajerska	61	7,76	725	92,24
Panonska	34	22,52	117	77,48
Koroška	31	13,48	199	86,52
Primorska	43	5,80	698	94,20
Rovtarska	54	9,54	512	90,46
Ljubljana	50	7,19	645	92,81
Maribor	25	11,52	192	88,48
Skupno	479	8,18	5380	91,82

Iz Tabele 12 lahko razberemo, da so izpusti nekončnih samoglasnikov v korpusu nekoliko pogostejši (okrog 60 %) od izpustov končnih samoglasnikov (okrog 40 %). Pri porazdelitvi najbolj izstopa Koroška (56 % izpustov končnih samoglasnikov in le 43 % nekončnih), a glede na zelo majhno število uporabnikov številke po vsej verjetnosti niso reprezentativne.

Kot prikazuje Slika 8, se v korpusu Janes-Geo najpogosteje izpušča samoglasnik *-i*, najredkeje pa samoglasnik *-u*. Pri nekaterih regijah so opazne razlike v deležih – Panonska, Koroška in Maribor v primerjavi z drugimi samoglasniki mnogo pogosteje izpuščajo *-i* (okrog 60 % vseh izpustov samoglasnikov), precej redkeje pa samoglasnik *-e* (okrog 10 % v primerjavi s približno 25 % na Gorenjskem, Dolenjskem in Štajerskem). Na Primorskem, Rovtarskem in v Mariboru je opazno tudi redkejše izpuščanje samoglasnika *-o* (manj kot 10 %).

Tabela 12: Izpusti končnih in nekončnih samoglasnikov v korpusu Janes-Geo.

Regija	Izpusti končnih samoglasnikov	Delež (%)	Izpusti nekončnih samoglasnikov	Delež (%)
Gorenjska	485	39,88	731	60,12
Dolenjska	411	38,20	665	61,80
Štajerska	312	43,03	413	56,97
Panonska	45	38,46	72	61,54
Koroška	113	56,78	86	43,22
Primorska	213	30,52	485	69,48
Rovtarska	192	37,50	320	62,50
Ljubljana	269	41,71	376	58,29
Maribor	94	48,96	98	51,04
Skupno	2134	39,67	3246	60,33



Slika 8: Izpusti samoglasnikov v korpusu Janes-Geo.

4.1.2 Transformacije

Tabela 13 prikazuje razporeditev pojavitev 10 najpogostejših transformacij v korpusu Janes-Geo po regijah.

Tabela 13: Najpogostejše transformacije v korpusu Janes-Geo po regijah.

Regija	Tl.u	Taj.ej	Tv.u	Ta.o	Tj.i	Tpri.pr	Tz.s	To.u	Td.t	Tl.o
Gorenjska	79	62	26	58	24	16	11	6	12	0
Dolenjska	89	42	28	22	24	26	7	18	9	0
Štajerska	46	35	9	4	14	14	15	5	4	12
Panonska	3	2	3	3	0	1	1	3	0	7
Koroška	0	30	2	1	3	2	6	2	3	2
Primorska	80	64	22	5	18	7	11	17	6	1
Rovtarska	24	36	26	12	13	6	17	9	1	0
Ljubljana	51	29	10	14	17	11	10	13	8	1
Maribor	7	6	0	2	5	2	3	2	3	7
Skupno	379	306	126	121	118	85	81	75	46	30

Najpogostejša je transformacija *-l* v *-u*, do katere največkrat pride pri moški obliki preteklega deležnika glagola (*naredil* → *naredu*, *spomnil* → *spomnu*). Najredkejša je v panonski, mariborski in koroški regiji, zanimivo pa je, da se sorodna transformacija *-l* v *-o* (*naredil* → *naredo*, *spomnil* → *spomno*) najpogosteje pojavlja prav v teh regijah (in na Štajerskem).

Sledita transformaciji *-aj* v *-ej* (*daj* → *dej*, *kaj* → *kej*, *zdaj* → *zdej*), ki jo prav tako redko zasledimo v panonski in mariborski regiji (a je pogostejša v koroški), in *-v* v *-u* (*všeč* → *ušeč*, *v* → *u*), ki je najpogostejša v gorenjski, dolenjski, primorski in rovtarski regiji. Precejšnje razlike se kažejo tudi pri transformaciji *-a* v *-o* (*prav* → *prov*, *gledal* → *gledov*), ki je zelo pogosta na Gorenjskem in (v nekoliko manjši meri) na Dolenjskem, v drugih regijah pa mnogo redkejša. Nasprotno je na Dolenjskem in Primorskem nekoliko pogostejša transformacija *-o* v *-u* (*blo* → *blu*, *okno* → *oknu*).

4.1.3 Nestandardno besedje

Besed, ki so bile uvrščene v kategorijo nestandardnega besedja (NSB), je bilo v celotnem korpusu 2.887, kar predstavlja slabo tretjino (31 %) vseh označenih besed. Tabela 14 prikazuje števila in deleže NSB po regijah v primerjavi z vsemi označenimi besedami in vsemi besedami.

Zanimiv podatek je, da je pri vseh regijah delež nestandardnega besedja v primerjavi z vsemi besedami primerljiv (giblje se med približno 3 in 6 %), nasprotno pa so razlike v deležih nestandardnega besedja glede na vse označene besede precej višje: nestandardno besedje ima najmanjši delež v gorenjski regiji (24 %), najvišjega pa v panonski (51 %). Na tem mestu je treba znova upoštevati, da so bili pri regijah z manj podatki (Panonska, Koroška, Maribor in Rovtarska) uporabljeni tudi tviti s stopnjo jezikovne nestandardnosti L2, zaradi česar je lahko višji delež nestandardnega besedja tudi posledica manjše količine izpustov, ki jih avtomatska

klasifikacija nestandardnosti lažje zazna in jih je zato v tvitih z nižjo stopnjo jezikovne nestandardnosti manj. Kljub temu vseh razlik ne moremo pripisati le vzorčenju: podobnost med Panonsko in Mariborom, ki sta tudi po geografski legi blizu, po vsej verjetnosti ni naključna. To še dodatno podpirajo manj izrazite razlike med Koroško in Rovtarsko (ki vsebujeta tudi tvite L2) na eni ter regijami z več podatki (ki vsebujejo samo tvite L3) na drugi strani. V mariborski in panonski regiji se torej na prvi pogled kaže tendenca, da se nestandardnost v jeziku pogosteje izraža z besediščem kot z izpusti. Razlika med Gorenjsko (24 % NSB) in Panonsko (51 % NSB) je tudi statistično veljavna (χ^2 (2, N = 2.272) = 131,12; p < 0,01). Podobno razmerje najdemo tudi med npr. Dolenjsko (25 % NSB) in Mariborom (41 % NSB) – χ^2 (2, N = 2.038) = 40,98; p < 0,01.

Tabela 14: Števila in deleži nestandardnega besedja po regijah v korpusu Janes-Geo.

Regija	NSB	Vse označene besede	NSB/Vse označene besede (%)	Vse besede	NSB/Vse besede (%)
Gorenjska	435	1.836	23,69	7.834	5,55
Dolenjska	415	1.610	25,78	7.447	5,57
Štajerska	427	1.273	33,54	7.516	5,68
Panonska	224	436	51,38	6.539	3,43
Koroška	167	454	36,78	4.232	3,95
Primorska	419	1.325	31,62	7.823	5,36
Rovtarska	228	868	26,27	5.547	4,11
Ljubljana	394	1.148	34,32	6.930	5,69
Maribor	178	428	41,59	4.820	3,69
Skupno	2.887	9.378	30,78	58.688	4,92

Tabela 15: Najpogostejše nestandardno besedje v korpusu Janes-Geo.

Pojavnica	f_A	Pojavnica	f_A
ma	60	btw	17
ful	42	cool	17
fajn	35	jao	17
itak	35	skos	17
lol	32	glih	16
evo	30	brezveze	15
zihr	29	jap	15
omg	26	un	12
kul	24	app	11
wtf	22	lih	11
kao	20	matr	11
folk	19	wow	11
jp	19	ziher	11
sorry	18		

Tabela 15 prikazuje vse besede NSB v korpusu, ki so se pojavljale z absolutno frekvenco, večjo od 10. Takšnih besed je skupno 28, kar znaša le 1,7 % od 1.745 različnih besed NSB v korpusu. Podatki torej izkazujejo visoko stopnjo razpršenosti, kar je posledica majhnega števila pojavnic v korpusu nasploh, v manjši meri pa tudi dejstva, da med analizo še nismo imeli na voljo normalizirane in lematizirane različice korpusa. Na podlagi preliminarnega pregleda seznama nestandardnega besedja smo sicer sklepali, da lematizacija razpršenosti ne bi znatno izboljšala, saj se le manjši del besed pojavlja v več oblikah. Približno 1.440 (83 %) vseh različnic NSB se v korpusu pojavi le enkrat, zato je korpus premajhen za zanesljivo medregionalno statistično primerjavo posameznih besed, omogoča pa primerjavo različnih podkategorij nestandardnega besedja, kot prikazuje Tabela 16.[14]

Tabela 16: Najpogostejše podkategorije nestandardnega besedja v korpusu Janes-Geo.

Regija	NSB.Pog	NSB.Ang	NSB.Ger	NSB.Srb	NSB.Ita	NSB.Net
Gorenjska	126	153	103	14	3	26
Dolenjska	137	153	80	18	3	23
Štajerska	133	151	96	24	8	11
Panonska	85	63	37	20	2	10
Koroška	69	37	33	9	1	16
Primorska	148	119	64	32	26	25
Rovtarska	73	94	37	7	0	13
Ljubljana	112	160	59	26	0	26
Maribor	57	68	27	8	0	14
Skupno	940	998	536	158	43	164

Najpogostejši kategoriji v korpusu sta *NSB.Pog*, ki vsebuje mdr. pogovorne, narečne in slengovske besede, in *NSB.Ang*, ki vsebuje anglizme z različnimi stopnjami podomačitve na ravni zapisa in oblikoslovja. Sledijo germanizmi (*NSB.Ger*), kroatizmi in srbizmi (*NSB.Srb*) ter italianizmi (*NSB.Ita*), ki pa jih je opazno manj, a je zanimivo, da so večino prispevali uporabniki iz primorske regije. Besede iz spletnega jezika (*NSB.Net*) so razporejene nekoliko bolj enakomerno.

4.1.4 Variantne različice pogostih besed

V korpusu so kot variantne označene različice naslednjih besed: *da*, *danes*, *domov*, *jaz*, *kaj*, *kako*, *koliko*, *kolikokrat*, *kolikor*, *kot*, *lahko*, *nekaj*, *potem*, *prav*, *saj*, *tako*,

14 Tabela 16 prikazuje samo podkategorije z največjimi frekvencami. Dodatne kategorije so še francizmi, hispanizmi in priložnostne tvorjenke, a jih je v korpus vključenih le peščica, zato jih tu ne navajamo.

takoj, *takole*, *toliko*, *tukaj*, *tule*, *včeraj*, *zakaj*, *zdaj*, *zdajle* in *zjutraj*. Seznam je bil med označevanjem sproti dopolnjevan z novimi različicami. Za primer si oglejmo regionalno razporeditev različic besede *koliko*, ki je prikazana v Tabeli 17.

Tabela 17: Različice besede koliko v korpusu Janes-Geo.

Regija	kok	kolk	kolko	kuk
Gorenjska	6	4	0	0
Dolenjska	7	3	0	0
Štajerska	0	0	0	0
Panonska	0	0	4	0
Koroška	0	0	0	0
Primorska	0	2	7	0
Rovtarska	2	1	1	2
Ljubljana	2	4	0	1
Maribor	0	0	3	0
Skupno	17	14	15	3

V korpusu smo zabeležili štiri različice: *kok*, *kolk*, *kolko* in *kuk*. Kljub majhnemu številu pojavitev že lahko opazimo nekatere vzorce. Različici *kok* in *kolk* se pojavljata na Gorenjskem in Dolenjskem ter v osrednjem delu Slovenije (Rovtarska in Ljubljana), različica *kolko* pa je skupna uporabnikom iz primorske, panonske in mariborske regije.

4.1.5 Nestandardno oblikoslovje

V korpusu je bilo v kategorijo nestandardnega oblikoslovja uvrščenih le nekaj nestandardnih prvin, skupaj le 47 (od 0 do največ 9 v vsaki regiji oz. pri skupno 36 uporabnikih, kar znaša dobrih 7 % vseh uporabnikov v korpusu). Tovrstni jezikovni pojavi so torej v našem vzorcu razmeroma redki – kar 125-krat redkejši od izpustov. Kar 37 primerov oz. 79 % vsega nestandardnega oblikoslovja smo zaznali pri glagolih, preostalih 10 primerov oz. 21 % pa pri samostalnikih.

Tabeli 18 in 19 prikazujeta vse zaznane nestandardne jezikoslovne posebnosti pri glagolih in samostalnikih ter podajata frekvenco, regije, v katerih so bile prvine prisotne, in ponazoritvene primere.

Tabela 18: Nestandardno oblikoslovje pri glagolih v korpusu Janes-Geo.

Nestandardno glagolsko obrazilo	f_A	Regije	Primer
Podaljšava kratkega nedoločnika na *-č* s *-t*	21	Gorenjska, Dolenjska, Ljubljana, Štajerska, Primorska	*reči* → *rečt, teči* → *tečt, obleči* → *obležt*
Obrazilo *-ma* v 1. osebi dvojine	8	Maribor, Koroška, Panonska	*greva* → *grema, zmeniva* → *zmenma, bova* → *boma*
Podaljšava s *-s* v 2. osebi množine	2	Primorska	*morate* → *moreste, imate* → *maste*
Obrazilo *-te* v 2. osebi množine	2	Primorska, Dolenjska	*boste* → *bote*
Obrazilo *-ta* v 2. osebi dvojine	2	Maribor	*bosta* → *bota*
Obrazilo *-ve* v 1. osebi dvojine	1	Dolenjska	*morava* → *morve*
Obrazilo *-ava* namesto *-uje* v 3. osebi ednine	1	Koroška	*opravičuje* → *opravičava*

Tabela 19: Nestandardno oblikoslovje pri samostalnikih v korpusu Janes-Geo.

Nestandardno samostalniško obrazilo	f_A	Regije	Primer
Podaljšava samostalnika moškega spola s *-t*	8	Primorska, Dolenjska, Panonska, Maribor	*deme* → *demote, psiha* → *psihota, nona* → *nonota*
Podaljšava samostalnika moškega spola z *-m*	2	Dolenjska, Primorska	*pri bogu milemu* → *pr bogmi milmi, s penziči* → *s penzičmi*

Zaradi nizkih frekvenc korpus Janes-Geo natančnejše medregionalne statistične primerjave na nivoju nestandardnega oblikoslovja ne dopušča, a ponuja zanimivo izhodišče za nadaljnje morfološke raziskave na večji količini podatkov.

4.1.6 Drugo

Pri kategoriji Drugo se v tem prispevku osredotočamo le na dve najpogostejši kategoriji, in sicer na pisanje skupaj (*Dskupaj*) ter sklapljanje besed (*Dsklop*).

Pisanje skupaj smo zabeležili pri 82 različnicah (oz. 175 pojavnicah), le šest pa se jih v korpusu pojavi več kot petkrat (oz. le 17 več kot enkrat). Najpogostejši zapisi skupaj in njihove absolutne frekvence so prikazani v Tabeli 19.

Tabela 20: Najpogostejši zapisi skupaj v korpusu Janes-Geo.

Zapis skupaj	f_A
nevem	28
pomoje	15
ane	11
ubistvu	11
nebi	10
vredu	6

Sklapljanje besed se je pojavljalo pri omejenem naboru besed (18 različnic in 33 pojavnic). Sklope, ki se v korpusu pojavijo več kot enkrat, prikazuje Tabela 21. Najpogosteje gre za kombinacijo naslonk. Zanimivo je, da so večino (20 pojavnic oz. 61 %) prispevali uporabniki z Gorenjske, preostanek pa uporabniki iz Dolenjske (4), Štajerske (5), Rovtarske (1) in Ljubljane (3).

Tabela 21: Sklapljanje besed v korpusu Janes-Geo.

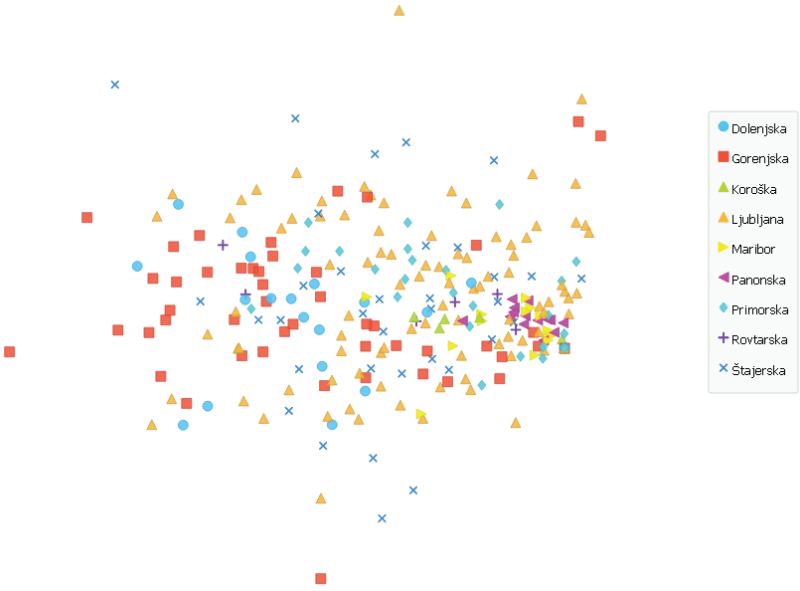
Sklop	Normalizirana oblika	f_A
nau	ne bo	6
sej	saj je	6
toj	to je	4
daj	da je	2
as	a si	2

4.2 Vizualizacija uporabnikov glede na uporabljene nestandardne jezikovne prvine

Da bi preverili, ali je uporabnike iz korpusa Janes-Geo mogoče razvrstiti v skupine tudi na podlagi njihove (ne)rabe nestandardnih jezikovnih prvin (in ne zgolj na podlagi geolokacije njihovih tvitov oz. pripisanih metapodatkov o regionalni pripadnosti), smo rabo nestandardnih jezikovnih prvin vsakega uporabnika vizualizirali. Za vsakega uporabnika smo izvozili relativne frekvence $(f)_r$[15] vseh nestandardnih jezikovnih prvin, ki jih je uporabil v tvitih, in jih nanizali v *n*-dimenzionalne vektorske reprezentacije $\vec{v}_u = (f_{r1}, f_{r2}, f_{r3},...,f_{rn})$, pri čemer je *n* število vseh kategorij in podkategorij nestandardnih jezikovnih prvin iz tipologije.

15 Relativne frekvence smo izračunali tako, da smo absolutno frekvenco določenega nestandardnega jezikovnega pojava iz tipologije (npr. IvkGn.i, izpust končnega samoglasnika *-i* v nedoločniku, *delati* → *delat*) delili s številom pojavnic, ki jih je uporabnik prispeval v korpus Janes-Geo.

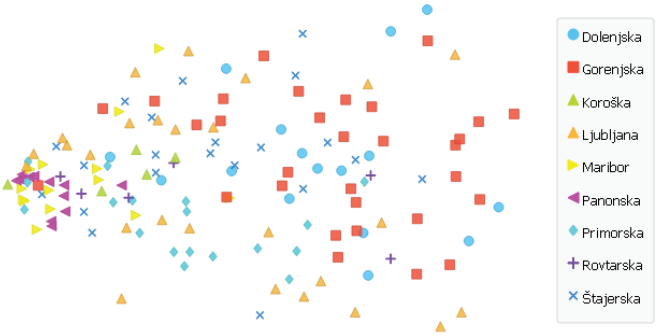
Vektorjem smo nato s pomočjo metode večdimenzionalnega skaliranja (Borg in Groenen 2005) zmanjšali dimenzionalnost ter jih s pomočjo paketa za podatkovno analizo Orange (Demšar et al. 2013) vizualizirali v dvodimenzionalni razsevni grafikon, ki v čim večji možni meri ohrani razdalje med izvirnimi vektorji. Slika 9 prikazuje vizualizacijo vseh uporabnikov v korpusu Janes-Geo glede na nestandardne jezikovne prvine v njihovih tvitih.



Slika 9: Vizualizacija uporabnikov v korpusu Janes-Geo glede na rabo nestandardnih jezikovnih prvin.

Čeprav je korpus relativno majhen in vsebuje zelo kratka besedila, so pri vizualizaciji uporabnikov glede na njihovo rabo nestandardnih jezikovnih prvin že opazne gruče. Največji vpliv na položaj uporabnikov v grafikonu ima količina izpustov, ki so najpogostejša in najbolj razširjena kategorija nestandardnih jezikovnih pojavov v korpusu. Najbolj razpršeni so ljubljanski uporabniki, ki jih je tudi največ, njihovo razpršenost pa gre pripisati več razlogom: nekateri uporabniki so v korpus prispevali manj gradiva, zaradi česar tudi relativne frekvence njihovih nestandardnih prvin ne predstavljajo ustrezno njihove dejanske jezikovne rabe. Obenem je mogoče sklepati, da se v ljubljansko regijo glede na njen centralni kulturni in gospodarski položaj uvrščajo uporabniki z zelo raznorodnimi jezikovnimi ozadji. Vizualizacija torej potrjuje ustreznost naše odločitve, da ljubljansko regijo obravnavamo posebej. Najbolj očitne so razlike med Gorenjsko in Dolenjsko na eni strani ter Panonsko in Mariborom na drugi, vmes pa prihaja tudi do nekoliko manj očitnega gručenja štajerskih, koroških in primorskih uporabnikov. Da bi

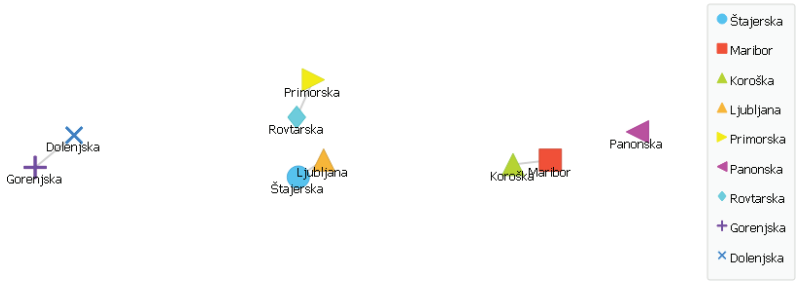
nekoliko zmanjšali šum, smo vizualizacijo ponovili samo z uporabniki, ki so v korpus Janes-Geo prispevali vsaj 100 pojavnic. To mejo smo določili na podlagi mediane števila pojavnic vseh uporabnikov (108 pojavnic). Rezultat vizualizacije prikazuje Slika 10.



Slika 10: Vizualizacija uporabnikov z več kot 100 pojavnicami v korpusu Janes-Geo.

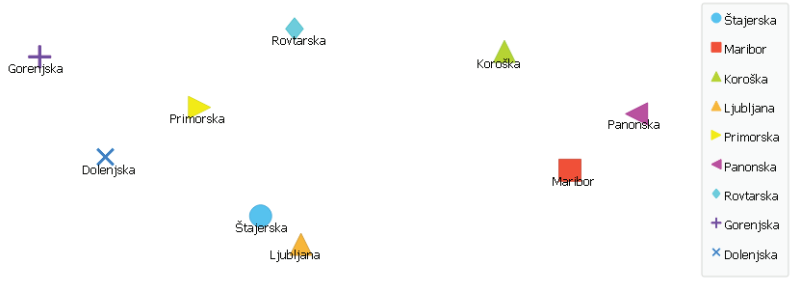
Z določitvijo minimalnega praga v pojavnicah smo nekoliko zmanjšali razpršenost ljubljanskih uporabnikov, ki so zdaj nekoliko bolj zgoščeni, a je zanimivo, da se ohranjata dve jasni veji (v spodnjem in zgornjem delu grafikona). Večjo zgoščenost lahko opazimo tudi pri uporabnikih iz primorske regije. Kot je pričakovano, bi še jasnejšo in natančnejšo sliko dobili z več podatki o jezikovni rabi posameznega uporabnika.

Preverili smo tudi, ali so razlike v jezikovni rabi podobne tudi med celotnimi regionalnimi podkorpusi, ne samo med posameznimi uporabniki. Vizualizacijo prikazuje Slika 11.



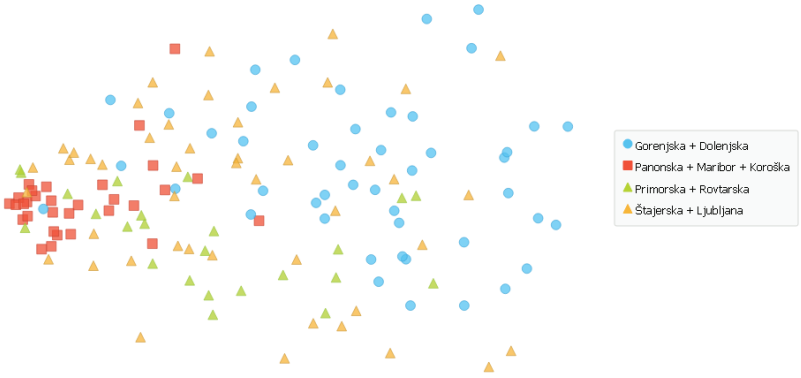
Slika 11: Vizualizacija regionalnih podkorpusov glede na vsebovane nestandardne jezikovne prvine.

Glede na podobnost se kaže jasna delitev na tri oz. štiri skupine, ki so skladne z ugotovitvami iz prejšnjih vizualizacij in primerjav. Ker je ta delitev v veliki meri odvisna od količine izpustov kot najpogostejše kategorije nestandardnih prvin (v gorenjski in dolenjski regiji je bilo izpustov največ, v mariborski in panonski pa najmanj), smo preverili podobnost regionalnih podkorpusov samo ob upoštevanju vseh drugih kategorij iz tipologije. Rezultat prikazuje Slika 12.



Slika 12: Vizualizacija regionalnih podkorpusov glede na vsebovane nestandardne jezikovne prvine (brez izpustov).

Kot vidimo, se razdalje in položaji tudi v primeru neupoštevanja izpustov v precejšnji meri ohranjajo. Nekoliko večje razlike je mogoče opaziti le med rovtarsko in primorsko ter med koroško in mariborsko regijo. Zanimivo je, da podobnosti med regijami v grobem sledijo tudi geografski razporeditvi. Na podlagi podobnosti med podkorpusi lahko regije torej združimo v štiri makroregije: Panonska + Maribor + Koroška, Štajerska + Ljubljana, Gorenjska + Dolenjska ter Primorska + Rovtarska. Slika 13 prikazuje vizualizacijo vseh uporabnikov, ki so v korpus prispevali več kot 100 pojavnic, ob upoštevanju novih, makroregionalnih metapodatkov.



Slika 13: Vizualizacija jezikovne rabe uporabnikov glede na makroregionalno delitev.

Najbolj razpršeni so še vedno uporabniki iz ljubljanske in štajerske regije, pri ostalih makroregijah pa se že nekoliko jasneje kažejo razmejitve. Glede na podobnosti med podkorpusi bi torej lahko posplošili, da so glede na kategorije, ki smo jih upoštevali pri označevanju, v korpusu Janes-Geo v grobem prisotne štiri jezikovne različice nestandardne spletne slovenščine, ki predstavljajo štiri makroregije: severovzhodno Slovenijo (Panonska + Maribor + Koroška), vzhodni del z Ljubljano (Štajerska + Ljubljana), osrednji del v smeri severozahod-jugovzhod (Gorenjska + Dolenjska) ter zahodni in jugozahodni del (Primorska + Rovtarska).

5 SKLEP

V poglavju smo predstavili poglavitne vidike ročno označenega korpusa Janes-Geo, ki omogoča korpusni pristop k proučevanju regionalnih jezikovnih različic v spletni slovenščini in je prosto dostopen pod licenco Creative Commons – Priznanje avtorstva (CC BY-SA 4.0) na repozitoriju CLARIN.SI (Čibej et al. 2018). Poleg postopka avtomatskega pripisovanja regionalnih metapodatkov na podlagi tvitov z geolokacijo smo predstavili gradnjo in označevanje korpusa Janes-Geo ter prvi vpogled v regionalne jezikovne različice v slovenski računalniško posredovani komunikaciji. Poleg samega korpusa kot pomemben rezultat raziskave velja omeniti tudi označevalne smernice in tipologijo nestandardnih jezikovnih prvin v spletni slovenščini, ki je bila sicer izdelana na podlagi tvitov, a lahko predpostavljamo, da je uporabna tudi za označevanje drugih besedilnih tipov v spletni komunikaciji (npr. forumska sporočila, blogovski zapisi in komentarji na novice).

Čeprav je korpus po številu pojavnic zelo majhen in ga zato ne moremo obravnavati kot povsem reprezentativnega, dobro prikazuje regionalno jezikovno variantnost v spletni slovenščini, kar potrjuje, da je uporabljena metoda uspešna. V prihodnje bi bilo raziskavo smiselno ponoviti na večjem vzorcu, ki vsebuje več zajetih uporabnikov in obenem ponuja čim večje število pojavnic na uporabnika, saj je le na ta način slika rabe nestandardnih jezikovnih prvin realna. Poleg tega zdajšnji vzorec za številne kategorije nestandardnih prvin zaradi pomanjkanja podatkov oz. prevelike razpršenosti ne omogoča učinkovite primerjave, vendar ponuja dobro izhodišče za nadaljnje raziskave tudi na drugih pisnih besedilnih tipih, v katerih lahko pričakujemo nestandardno slovenščino.

Kot zamisel za prihodnje delo velja izpostaviti tudi evalvacijo metode avtomatskega kodiranja metapodatkov o regionalni pripadnosti s pomočjo anketiranja uporabnikov, rabo nestandardnih jezikovnih prvin v spletni slovenščini pa bi bilo dobro raziskati tudi z bolj sociolingvističnega vidika, npr. v katerih situacijah uporabniki uporabljajo bolj regionalno obarvano jezikovno različico, v kolikšni

meri se jezikovno prilagajajo sogovorcem in v kolikšni meri na to vpliva tip komunikacije (javno, zasebno).

Zaznane medregionalne razlike bodo uporabljene tudi kot značilke za razvoj klasifikatorja, ki bo uporabnika uvrstil v ustrezno regijo. Klasifikator bo nato, če bo uspešen, omogočil gradnjo večjega korpusa, v katerem bo mogoča tudi statistična primerjava pojavov, ki so bili v korpusu Janes-Geo preredki.

Zahvala

Za tehnično podporo pri zasnovi raziskave se zahvaljujem Nikoli Ljubešiću in Tomažu Erjavcu.

Literatura

Arhar Holdt, Špela, 2018: Korpusni pristop k skladnji računalniško posredovane slovenščine. Fišer, Darja (ur.): *Viri, orodja in metode za analizo spletne slovenščine*. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani. 228–253.

Baron, Naomi S., 2010: *Always On: Language in an Online and Mobile World*. Oxford: Oxford University Press.

Bernhard, Delphine in Anne-Laure Ligozat, 2013: Hassle-free POS-Tagging for the Alsatian Dialects. Zampieri, Marcos in Sascha Diwersy (ur.). *Non-standard Data Sources in Corpus-based Research*. Aachen: Shaker Verlag. 85–92.

Bitenc, Maja, 2016: *Z jezikom na poti med Idrijskim in Ljubljano*. Ljubljana: Znanstvena založba Filozofske fakultete.

Borg, Ingwer in Patrick Groenen, 2005: *Modern Multidimensional Scaling: theory and applications (2nd ed.)*. New York: Springer-Verlag. 207–212.

Cotterell, Ryan in Chris Callison-Burch, 2014: A Multi-Dialect, Multi-Genre Corpus of Informal Written Arabic. *Zbornik konference Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik: ELRA.

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

Čibej, Jaka in Nikola Ljubešić, 2015: "S kje pa si?" – Metapodatki o regionalni pripadnosti uporabnikov družbenega omrežja Twitter. Fišer, Darja (ur.): *Zbornik konference Slovenščina na spletu in v novih medijih*. Ljubljana: Znanstvena založba Filozofske fakultete. 10–14.

Čibej, Jaka, 2016: Framework for an Analysis of Slovene Regional Language Variants on Twitter. *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities*. Ljubljana: Znanstvena založba Filozofske fakultete. 17–21.

Čibej, Jaka, 2017: *Tipologija in smernice za označevanje nestandardnih jezikovnih prvin v slovenskih tvitih*. https://nl.ijs.si/janes/viri

Čibej, Jaka, Tomaž Erjavec in Darja Fišer, 2018: *Tweet corpus of Slovene regional language variants Janes-Geo v1.0*. http://hdl.handle.net/11356/1174

Čibej, Jaka, Špela Arhar Holdt, Tomaž Erjavec in Darja Fišer, 2018: Ročno označeni korpusi Janes za učenje jezikovnotehnoloških orodij in jezikoslovne raziskave. Fišer, Darja (ur.): *Viri, orodja in metode za analizo spletne slovenščine*. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani. 44–73.

Demšar, Janez, Tomaž Curk, Aleš Erjavec, Črt Gorup, Tomaž Hočevar, Mitar Milutinovič, Martin Možina, Matija Polajnar, Marko Toplak, Anže Starič, Miha Štajdohar, Lan Umek, Lan Žagar, Jure Žbontar, Marinka Žitnik in Blaž Zupan, 2013: Orange: Data Mining Toolbox in Python. *Journal of Machine Learning Research* 14/2013. 2349–2353. http://eprints.fri.uni-lj.si/2267/1/2013-Demsar-Orange-JMLR.pdf

Eisenstein, Jacob, 2015: Written dialect variation in online social media. Boberg, Charles, John Nerbonne in Dominic Watt (ur.): *Handbook of Dialectology*. New York: Wiley.

Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith in Eric P. Xing, 2010: A latent variable model for geographic lexical variation. *Zbornik konference Empirical Methods for Natural Language Processing (EMNLP)*. Stroudsburg, Pennsylvania: Association for Computational Linguistics. 1277–1287.

Fišer, Darja, Ljubešić, Nikola, Erjavec, Tomaž. 2015. The JANES corpus of Slovene user generated content: construction and annotation. *International Research Days: Social Media and CMC Corpora for the e-humanities: Book of Abstracts. 23.-24. October 2015*. Rennes, France. 11.

Goli, Teja, Eneja Osrajnik in Darja Fišer, 2016b: Analiza krajšanja slovenskih sporočil na družbenem omrežju Twitter. Erjavec, Tomaž in Darja Fišer (ur.): *Proceedings of the Conference on Language Technologies and Digital Humanities*. Ljubljana: Znanstvena založba Filozofske fakultete. 77–82.

Grieve, Jack, 2016: *Regional Variation in Written American English*. Cambridge: Cambridge University Press.

Haddow B., A. Hernandez-Huerta, F. Neubarth, H. Trost, 2013: Corpus Development for Machine Translation between Standard and Dialectal Varieties. Zbornik konference *Workshop 'Adaptation of Language Resources and Tools for Closely Related Languages and Language Variants' of the 9th Int. Conf. on Recent Advances in Natural Language Processing (RANLP 2013)*. Hissar, Bulgaria. 7–14.

Harrat, Salima, Karima Meftouh, Mourad Abbas in Kamel Smaili, 2014: Building Resources for Algerian Arabic Dialects. Zbornik konference *15th Annual Conference of the International Speech Communication Association (INTERSPEECH 2014)*. Singapur.

Harrat, Salima, Mourad Abbas, Karima Meftouh in Kamel Smaili, 2013: Diacritics restoration for Arabic dialect texts. *Zbornik konference 14th Annual Conference of the International Speech Communication Association (INTERSPEECH 2013)*. Francija.

Hernández, Nuria, 2006: *User's Guide to FRED*. Freiburg: University of Freiburg. http://www.freidok.uni-freiburg.de/volltexte/2489/

Huang, Yuan, Diansheng Guo, Alice Kasakoff in Jack Grieve, 2016: Understanding US regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems 59*. 244–255.

Johanessen, Janne Bondi, Joel Priestley, Kristin Hagen, Tor Anders Åfarli in Øystein Alexander Vangsnes, 2009: The Nordic Dialect Corpus – an Advanced Research Tool. Jokinen, K. in E. Bick (ur.): *Zbornik konference 17th Nordic Conference of Computational Linguistics NODALIDA 2009. NEALT Proceedings Series Volume 4*.

Jørgensen, Anna Katrine, Dirk Hovy in Anders Søgaard, 2015: Challenges of studying and processing dialects in social media. *Proceedings of the ACL 2015 Workshop on Noisy User-generated Text*. Peking, Kitajska. 9–18.

Kenda Jež, Karmen, 2002: *Cerkljansko narečje: teoretični model dialektološkega raziskovanja na zgledu besedišča in glasoslovja*. Doktorska disertacija. Ljubljana: Filozofska fakulteta.

Khakimov, Bulat, Farid Salimov in Dariya Ramazanova, 2015: Building dialectological corpora for Turkic languages: Mishar dialect of Tatar. *Procedia – Social and Behavioral Sciences* 198. 218–225.

Kunst, Jan Pieter in Franca Wesseling, 2010: Dialect Corpora Taken Further: The DynaSAND corpus and its application in newer tools. *Zbornik konference 24th Pacific Asia Conference on Language, Information and Computation*. 759–767.

Ljubešić, Nikola in Denis Kranjčić, 2014: Discriminating between VERY similar languages among Twitter users. *Zbornik konference Language technologies: 17th International Multiconference Information Society IS2014*. Ljubljana: Institut »Jožef Stefan«.

Ljubešić, Nikola, Darja Fišer in Tomaž Erjavec, 2014: TweetCaT: a tool for building Twitter corpora of smaller languages. *Zbornik konference Ninth International Conference on Language Resources and Evaluation (LREC 2014)*. Reykjavik, Iceland. 2279–2283. http://www.lrec-conf.org/proceedings/lrec2014/pdf/834_Paper.pdf

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. *10th International Conference on Recent Advances in Natural Language Processing: Proceedings of RANLP 2015 Conference*. Hissar, Bolgarija. 371–378.

Logar, Tine, 1981: Govor kraja Podlešče na Banjšicah (Glasoslovna študija). *Goriški letnik* 8/1981. 275–283.

Myslín, Mark in Stefan T. Gries, 2010: k dixez? A corpus study of Spanish Internet orthography. *Literacy and Linguistic Computing* 25/1. 85–104.

Popič, Damjan in Darja Fišer, 2018: (Ne)normativnost računalniško posredovane komunikacije v slovenščini: merilo vejice. Fišer, Darja (ur.): *Viri, orodja in metode za analizo spletne slovenščine*. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani. 140–159.

Ramovš, Fran, 1931: *Dialektološka karta slovenskega jezika*. Ljubljana: Rektorat univerze kralja Aleksandra I. in J. Blaznika nasl., Univerzitetna tiskarna.

Reher, Špela in Darja Fišer, 2018: Kodno preklapljanje v objavah slovenskih uporabnikov Twitterja. Fišer, Darja (ur.): *Viri, orodja in metode za analizo spletne slovenščine*. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani. 294–323.

Ruef, Beni in Simone Ueberwasser, 2013: The Taming of a Dialect: Interlinear Glossing of Swiss German Text Messages. Zampieri, Marcos in Sascha Diwersy (ur.): *Non-standard Data Sources in Corpus-based Research*. Aachen: Shaker Verlag. 61–68.

Sparks, Evan in Sanjay Krishnan, 2012: *TweetLocalize: Inferring Author Location in Social Media*. University of Berkeley. http://bid.berkeley.edu/cs294-1-spring13/images/c/c2/TweetLocalizeReport.pdf

Szmrecsanyi, Benedikt, 2011: Corpus-based dialectometry: a methodological sketch. *Corpora* 6/1. 45–76.

Škofic, Jožica in drugi, 2011: *Slovenski lingvistični atlas 1: Človek (telo, bolezni, družina)*. Ljubljana: Založba ZRC.

Toporišič, Jože, 2000: *Slovenska slovnica: četrta, prenovljena in razširjena izdaja*. Maribor: Založba Obzorja.

Ueberwasser, Simone, 2013: Non-standard data in Swiss text messages with a special focus on dialectal forms. Zampieri, Marcos in Sascha Diwersy (ur.): *Non-standard Data Sources in Corpus-based Research*. Aachen: Shaker Verlag. 7–24.

Valicon, 2016: *Raziskava MEDIA+. Maj/junij 2016*. http://www.valicon.net/files/Sporocilo%20za%20javnost%202016-06-23%20(1).pdf

Verdonik, Darinka in Ana Zwitter Vitez, 2011: *Slovenski govorni korpus Gos*. Ljubljana: Trojina, zavod za uporabno slovenistiko.

Zemljarič Miklavčič, Jana, 2008: *Govorni korpusi*. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani.

Logar, Tine in Jakob Rigler, 1986: Karta slovenskih narečij. Avtotehna Zveza organizacij za tehnično kulturo Slovenije. https://www.dlib.si/?URN=URN:NBN:SI:IMG-VSVHWWS9