»S kje pa si?« – Metapodatki o regionalni pripadnosti uporabnikov družbenega omrežja Twitter

Jaka Čibej,* Nikola Ljubešić†‡

* Oddelek za prevajalstvo, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
jaka.cibej@ff.uni-lj.si
† Odsek za tehnologije znanja, Institut »Jožef Stefan«
Jamova cesta 39, 1000 Ljubljana
‡ Odsek za informacijske in komunikacijske znanosti, Filozofska fakulteta, Univerza v Zagrebu
Ivana Lučića 3, 10000 Zagreb
nljubesi@ffzg.hr

Povzetek

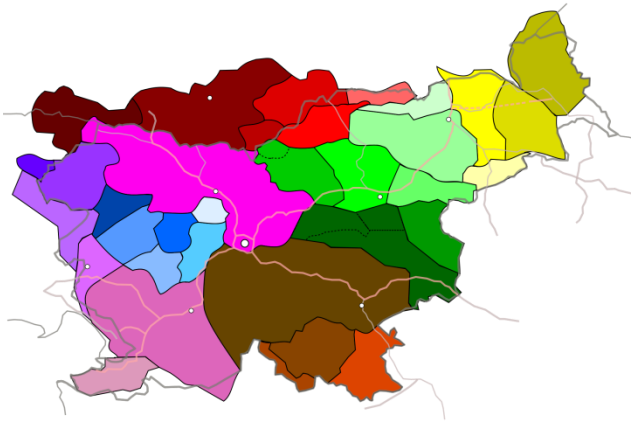
Slovenščina je kot narečno razčlenjen jezik z dialektološkega vidika že zelo dobro raziskana, a le v govoru. V dobi vsesplošne uporabe družbenih medijev se je slovenska regionalna jezikovna produkcija razširila tudi v računalniško posredovano komunikacijo prek številnih (pisnih) platform za sporazumevanje, kot sta npr. družbeni omrežji Facebook in Twitter. To nakazuje, da bo tudi spletno sporazumevanje odigralo vlogo v nadaljnjem razvoju slovenskih regionalnih jezikovnih različic, zato je nujno, da sodobne dialektološke raziskave upoštevajo tudi ta vidik jezikovne rabe. V prispevku zato predstavljamo prvo stopnjo v raziskavi slovenskih regionalnih jezikovnih različic na spletu, in sicer postopek dodajanja metapodatkov o regionalni pripadnosti uporabnikov Twitterja v korpus spletne slovenščine Janes. To bo omogočilo medregionalno kontrastivno primerjavo jezikovne produkcije in ugotavljanje specifik regionalnih jezikovnih različic na spletu.

"Where you from?" – Metadata on regional origin of Twitter users

From a dialectological perspective, Slovene as a dialectally diverse language has been researched to a considerable extent, but only in speech. Due to the pervasive presence of social media today, Slovene regional language production has expanded into computer-mediated communication through numerous (written) communication platforms, such as the social networks Facebook and Twitter. This indicates that internet communication will play a part in the future development of Slovene regional language variants, which is why modern dialectological research should also take this aspect of language use into account. In this paper, we present the first step in the research of Slovene regional language variants on the web: the addition of metadata on regional origin of tweeters in the Janes corpus of internet Slovene. The metadata will enable an inter-regional contrastive analysis of language production and a definition of specific characteristics of Slovene regional language variants on the web.

1 Uvod

Slovenščina je v jezikoslovnem smislu svojevrsten fenomen, saj je kljub relativno majhnemu številu govorcev in geografskemu ozemlju narečno zelo razčlenjena. Že ob začetku obširnejših dialektoloških raziskav v prvi polovici 20. stoletja je Fran Ramovš (1931) govorce slovenščine razdelil v 7 narečnih skupin s skupno več kot 30 narečji (Slika 1).



Slika 1: Slovenske narečne skupine.

Temu primerno je tudi dialektološki vidik slovenščine precej obširno raziskan, a kljub temu ni brez pomanjkljivosti. Tradicionalna dialektologija se je namreč pogosto opirala na tezo, da narečja izumirajo in se počasi zlivajo v standardno različico jezika (Kolarič, 1954) oziroma v najboljšem primeru v nestandardno različico, ki prevladuje v mestnih središčih (Ramovš, 1951). Raziskave so se zato omejevale na proučevanje t. i. »čistih« narečij, tj. na govorice tistih govorcev, ki so bili v svojem življenju čim bolj izolirani od drugih jezikovnih zvrsti in so kot taki predstavljali idealne govorce narečja (Kenda Jež, 2002: 26). Kriteriji, ki so predstavljali idealnega govorca, so bili v nekaterih primerih precej natančni, saj so poleg kraja bivanja določali tudi spol in stopnjo izobrazbe govorca, izpostavljenost drugim govoricam, govorico staršev ipd.[1]

Ob takšnem pristopu se porajajo številne kritike o reprezentativnosti vzorcev, obenem pa je iz nekaterih sodobnejših dialektoloških raziskav razvidno, da teza o izumiranju narečij ni povsem aksiomatična in da jezikovni razvoj narečja pelje v drugo smer, o kateri pa si dialektologi niso enotni: npr. da se bo raba narečij

[1] Logar (1958: 129) in Unuk (1997: 310) npr. zagovarjata, da so ženske boljši informatorji od moških, ker so večinoma bolj doma in tudi bolj konservativne v svoji govorici. Chambers in Trudgill (1994: 33) navajata, da je večina informatorjev v dialektoloških raziskavah starejših kmečkih moških, ki stalno živijo doma (ang. *NORM → Non-mobile Older Rural Male*).

omejevala sprva na ruralno okolje in nazadnje le še na kontekst folklornega kulturnega udejstvovanja ali da bodo zemljepisne jezikovne različice zamenjale družbene (Sgall et al., 1992), da se bodo lokalna narečja strnjevala v večja regionalna narečja (Niebaum in Macha, 1999; Kenda Jež, 2004) ali pa da bo v jezikovnem razvoju prišlo do t. i. nove dialektizacije (Labov, 1994), pri kateri bodo ključno vlogo odigrale jezikovne inovacije v govoricah jezikovnih skupnosti v urbanih središčih, ki se bodo postopoma razširile tudi v manjše jezikovne skupnosti na podeželju in tako ustvarile nov nabor narečij.[2] Tudi položaj narečnih jezikovnih različic ni povsem samoumeven. Jezikovne razmere na Norveškem so npr. zelo naklonjene rabi narečij v vseh funkcijskih zvrsteh (Jahr, 1997), govorci nasploh pa lahko jezikovne različice tudi zavestno kultivirajo ter jih uporabljajo v vedno večji meri (Reichan, 1999).

Pri proučevanju sprememb regionalnih jezikovnih različic v prihodnosti pa je treba v današnjih razmerah upoštevati še en vidik, ki prej ni bil prisoten. Z vzponom spleta in informacijske tehnologije 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 sta Facebook in Twitter. Jezik v računalniško posredovani komunikaciji (še posebej v klepetu in v drugih neformalnih kontekstih) pa se od standarda precej razlikuje (Crystal, 2001; Baron, 2010; Myslin in Gries, 2010; Erjavec in Fišer, 2013) in vsebuje tudi narečne prvine (Ueberwasser, 2013; Fišer et al., 2015).

Spletna komunikacija v jezikovni produkciji že dolgo več ne zajema zanemarljivega deleža, o čemer pričajo tudi podatki Statističnega urada Republike Slovenije:[3] v prvem četrtletju 2014 je v družbenih omrežjih sodelovalo skoraj 60 odstotkov oseb, 41 odstotkov pa je splet uporabljalo tudi za telefoniranje ali videotelefoniranje. Deleži iz leta v leto rastejo, kar nakazuje, da bo tudi spletna komunikacija odigrala vlogo pri nadaljnjem razvoju slovenskih regionalnih različic. Ključno je torej, da sodobne dialektološke raziskave upoštevajo tudi ta vidik jezikovne produkcije, proučevanje rabe regionalnih prvin v računalniško posredovani komunikaciji pa bo pripomoglo tudi k razvoju novih (ali izboljšanju že obstoječih) jezikovnih tehnologij za slovenščino, kot so označevalniki, lematizatorji, strojni prevajalniki ipd.

V prispevku zato predstavljamo prvi korak v raziskavi regionalne členjenosti v spletnem kontekstu. Najprej opravimo kratek pregled sorodnih raziskav, nato pa opišemo postopek, po katerem smo kategorizirali uporabnike Twitterja glede na regionalno pripadnost. Pridobljene metapodatke smo dodali v korpus spletne slovenščine Janes (Fišer et al., 2014) in tako ustvarili podkorpuse za proučevanje specifik rabe regionalnih jezikovnih različic[4] v pisnem sporazumevanju na spletu. Nazadnje še na kratko predstavimo preliminarne regionalne podkorpuse in navedemo predloge za prihodnje delo.

2 Pregled sorodnih raziskav

Za razliko od slovenščine so bile za številne tuje jezike že opravljene obsežne korpusne dialektološke raziskave, a najpogosteje na podlagi transkripcij govora v govornih korpusih, ki pa pogosto prvotno niso bili namenjeni dialektološkim raziskavam – British National Corpus npr. kljub razdelani taksonomiji vsebovanih narečij nudi samo standardizirano transkripcijo govora brez posnetkov. Za nizozemska narečja je bil zgrajen korpus DynaSAND (Kunst in Wesseling, 2010). Podoben projekt za nordijske jezike je Nordic Dialect Corpus (Johanessen et al., 2009), za angleščino pa Freiburg Corpus of English Dialects (Hernández, 2006).

Raziskovanje dialektalnih prvin v uporabniških spletnih vsebinah pa je tudi v tujini še precej sveže, deloma najbrž tudi zato, ker so bila jezikovnotehnološka orodja naučena na standardnih jezikovnih različicah in so se šele pred nedavnim začela prilagajati tudi šumnim besedilom, med katera lahko štejemo tudi narečno jezikovno produkcijo na spletu. Obenem je bilo to področje prej domena jezikovnih tehnologov kot jezikoslovcev. Predvsem z namenom gradnje novih jezikovnih tehnologij (za oblikoskladenjsko označevanje, strojno prevajanje, avtomatsko detekcijo regionalnih različic ipd.) je bilo opravljenih že veliko raziskav spletnih regionalnih različic arabščine (Harrat et al., 2013; Harrat et al., 2014; Cotterell in Callison-Burch, 2014) in ameriške angleščine (Eisenstein et al., 2010; Eisenstein et al., 2015), pa tudi manjših, jezikovnotehnološko manj podprtih jezikov, kot so npr. tatarsko narečje mišar (Khakimov et al., 2015), švicarska nemščina (Ruef in Ueberwasser, 2013) in alzaščina (Bernhard in Ligozat, 2013).

To kaže na svetovni trend, ki potrjuje, da bi bilo orodja za obdelavo regionalnih jezikovnih različic na spletu koristno razviti tudi za slovenščino, ki v tem smislu še ni jezikovnotehnološko podprta.

3 Metapodatki o regionalni pripadnosti uporabnikov Twitterja

V tem razdelku opisujemo postopek, po katerem smo klasificirali uporabnike Twitterja glede na njihovo regionalno pripadnost ter metodologijo in vire, ki so bili pri tem uporabljeni.

3.1 Korpus spletne slovenščine Janes

Trenutna različica[5] korpusa spletne slovenščine Janes, ki ga sestavljajo tviti, forumska sporočila, blogovski zapisi in komentarji na spletne novice, vsebuje približno 160 milijonov pojavnic. Od tega je skoraj 61 milijonov oz. 40 odstotkov tvitov, ki jih je spisalo približno 7.500 različnih avtorjev. Tviti v korpusu so že opremljeni z nekaterimi metapodatki (npr. sentiment, spol, ali gre za zasebnega uporabnika ali organizacijo, stopnja standardnosti besedila (Ljubešić et al., 2015)), kar omogoča natančnejše določanje ustreznega gradiva za številne jezikoslovne

[2] Labov (1994) predpostavlja, da so današnja narečja pravzaprav ostanki jezikovnega razvoja, ki se je začel v mestih in se nato postopoma razširil na podeželje.

[3] http://www.stat.si/StatWeb/glavnanavigacija/podatki/prikazistaronovico?IdNovice=6560

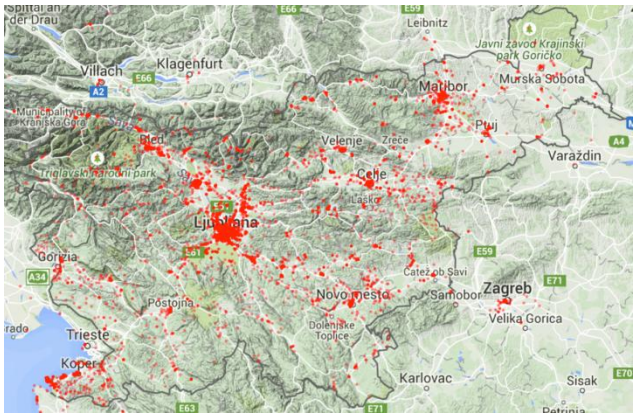
[4] Ker gre v primeru našega gradiva za pisni diskurz, ker se ne osredotočamo na konkretna narečja in ker pričakujemo, da se bo raba jezika na spletu precej razlikovala v primerjavi z rabo v govoru, se bomo izognili poimenovanju *narečje* in namesto tega za opazovane jezikovne zvrsti uporabljali manj specifični termin *regionalne jezikovne različice*, saj želimo jezik opisovati neodvisno od uveljavljene tipologije zvrsti brez vnaprejšnjega kategoriziranja.

[5] Različica 0.3 je bila zgrajena 5. marca 2015.

raziskave. Da bo korpus Janes uporaben tudi z geolingvističnega in dialektološkega vidika, smo uporabnikom Twitterja določili še regijo, iz katere najpogosteje pošiljajo tvite. To smo dosegli s pomočjo zbirke tvitov s podatki o geolokaciji, ki jo podrobneje predstavljamo v nadaljevanju.

3.2 Tviti s podatki o geolokaciji

Januarja 2015 smo začeli s pomočjo namenskega orodja TweetCat (Ljubešić et al., 2014) zajemati slovenske tvite s podatki o geolokaciji, tj. s koordinatami kraja, s katerega je bil tvit poslan. Do avgusta 2015 je bilo zajetih približno 130.000 tvitov, ki jih je napisalo 1.661 tviterašev po vsej Sloveniji (glej Sliko 2).



Slika 2: Razporeditev zajetih tvitov po Sloveniji. Vsaka rdeča pika predstavlja en tvit.

Izločili smo uporabnike, ki niso vključeni v korpus Janes, in tiste, ki niso zasebni uporabniki, saj organizacije na Twitterju v prevladujoči meri objavljajo v standardni slovenščini in bi v naših regionalnih podkorpusih predstavljale šum. Ostalo nam je 119.236 tvitov, ki jih je napisalo 1.461 uporabnikov. V korpusu Janes je uporabnikov, ki so označeni kot zasebni, 5.806. Z zbiranjem tvitov z geolokacijo smo do avgusta 2015 torej zajeli približno četrtino v korpus vključenih uporabnikov.

3.3 Razdelitev Slovenije na regije

V naslednjem koraku smo Slovenijo s pomočjo orodja Google Maps API v3 Tool[6] razdelili na 9 koordinatnih poligonov, ki predstavljajo 7 narečnih skupin[7] (gorenjsko, dolenjsko, štajersko, panonsko, koroško, rovtarsko in primorsko; glej Sliko 3) ter Ljubljano in Maribor.

Ljubljano in Maribor smo se 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).

[6] http://www.birdtheme.org/useful/v3tool.html

[7] Kategorizacija po Toporišiču (2000: 23–24) sicer v dolenjski skupini loči tudi posebno osmo, kočevsko skupino (na območju nekdanje nemške poselitve), a smo se v tem prispevku zaradi omejenega števila podatkov osredotočili le na 7 narečnih skupin po Ramovšu (1931), v katere smo vključili tudi slovenske manjšine v Italiji, v Avstriji in na Madžarskem.

Slika 3: Razdelitev Slovenije na koordinatne poligone.

3.4 Določitev regionalne pripadnosti uporabnikov

Za vsak tvit od preostalih 1.461 tviterašev iz baze tvitov z geolokacijo smo nato v programskem jeziku Python z metodo metanja žarka (ang. *ray-casting method*) preverili, iz katere regije je bil poslan. Metoda iz podane točke (v našem primeru so to koordinate tvita) pošlje žarek in preveri število presečišč med žarkom in robovi podanega poligona (regije) – če je število liho, točka leži v notranjosti poligona. Razporeditev po regijah je prikazana v Tabeli 1.

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

Tabela 1: Razporeditev tvitov po regijah.

Največ tvitov (36 %) je bilo poslanih iz Ljubljane, najmanj pa iz rovtarske (slaba 2 %) in panonske regije (dobra 2 %), ki sta tudi po površini med najmanjšimi.

Uporabnikom, ki so več kot 90 % tvitov z geolokacijo poslali iz ene same regije in so obenem poslali vsaj 3 tvite, smo pripisali metapodatek o regionalni pripadnosti. Takšnih uporabnikov je bilo 364, končni rezultati njihove kategorizacije pa so prikazani v Tabeli 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.

Regija	Število tviterašev	Delež (%)
Gorenjska	48	13,19
Dolenjska	22	6,04
Štajerska	42	11,54
Panonska	14	3,85
Koroška	5	1,37
Primorska	31	8,52
Rovtarska	7	1,92
Ljubljana	116	31,87
Maribor	14	3,85
Tujina	65	17,86
Skupno	364	100,00

Tabela 2: Število uporabnikov po regijah.

V povprečju je vsak uporabnik poslal približno 74 tvitov, mediana pa je 18 tvitov. Uporabnikov, ki so poslali zgolj 3 tvite, je bilo skupno 39 (2 dolenjska, 4 gorenjski, 8 ljubljanskih, 1 mariborski, 3 panonski, 4 primorski, 1 rovtarski, 6 štajerskih in 10 iz tujine). Več kot 74 tvitov je poslalo 89 uporabnikov, najproduktivnejši pa je poslal kar 1188 tvitov, in sicer iz Ljubljane.

Zanimivo je, da je koroških uporabnikov le 5, spisali pa so skupno 167 tvitov. Število tvitov iz te regije ni bilo majhno (približno 4.200), zato lahko sklepamo, da večina tamkajšnjih uporabnikov tvita tudi iz drugih regij (oziroma da so velik delež tamkajšnjih tvitov prispevali uporabniki iz drugih regij), zato zaradi strogih kriterijev (najmanj 90-odstotna pripadnost eni sami regiji) niso bili vključeni v končni nabor. Podobno je z rovtarsko skupino, pri kateri je uporabnikov le 7, poslali pa so skupno 956 tvitov. Število rovtarskih tvitov je bilo primerljivo s panonsko skupino, pri kateri pa je uporabnikov 14.

4 Regionalni podkorpusi

Uporabnikom Twitterja v korpusu Janes v smo pripisali metapodatke o regionalni pripadnosti in tako izdelali 9 regionalnih podkorpusov ter zabeležili število pojavnic v njih. Preverili smo tudi število pojavnic, če iščemo samo po tvitih, ki so bili v korpusu označeni kot nestandardni (L2 in L3), ter izračunali deleže nestandardnih tvitov. Rezultati so predstavljeni v Tabeli 3.

Regionalni podkorpus	Število pojavnic	Število pojavnic (L2, L3)	Delež L2 in L3 (%)
Gorenjska	37.683	16.679	44,26
Dolenjska	17.364	5.503	31,69
Štajerska	41.712	14.091	33,78
Panonska	5.020	1.345	26,79
Koroška	6.207	2.644	42,60
Primorska	13.917	3.579	25,72
Rovtarska	4.823	1.778	36,87
Ljubljana	92.104	27.036	29,35
Maribor	4.789	1.205	25,16

Tabela 3: Velikost regionalnih podkorpusov.

Kot je bilo pričakovano, je po številu pojavnic največji ljubljanski podkorpus, najmanjši pa so panonski, koroški, rovtarski in mariborski (kar je morda nekoliko presenetljivo, saj smo na začetku pričakovali, da bo kot drugo največje slovensko mesto v zbirko zajetih tvitov doprinesel mnogo več).

Po deležu nestandardnosti izstopata gorenjski in koroški podkorpus, oba z dobrimi 40 % nestandardnih tvitov. Najbolj standardni so mariborski, primorski in panonski podkorpus, pri katerih je kot nestandardnih označenih le dobrih 25 % tvitov. Ti podatki nam podajo grobo oceno, kateri podkorpusi vsebujejo največ nestandardnih (in potencialno tipično regionalnih) jezikovnih prvin, za podrobnejši vpogled v njihovo vsebino pa bo potrebna še temeljita jezikoslovna analiza.

5 Zaključek

V prispevku smo opisali postopek, po katerem smo uporabnikom s pomočjo zbirke tvitov s podatki o geolokaciji pripisali metapodatke o regionalni pripadnosti. V prihodnjem delu bomo poskušali obseg nastalih podkorpusov povečati z zajemanjem novih tvitov z geolokacijo (in novih uporabnikov), tiste podkorpuse, ki se bodo zaradi premajhnega števila podatkov izkazali za neuporabne, pa bomo po potrebi izključili iz nadaljnjih raziskav.

Poleg tega bomo temeljito preučili sestavo in vsebino regionalnih podkorpusov ter poskušali ugotoviti značilne razlike med regionalnimi jezikovnimi različicami spletne slovenščine, npr. z izdelavo seznamov ključnih besed za vsak regionalni podkorpus glede na celotni podkorpus tvitov korpusa Janes ter s kvalitativnim pregledom materiala z vidika regionalnih jezikovnih značilnosti (npr. regionalne razširjenosti variantnih različic zapisa besed, npr. *kaj*, *ka*, *kva*, *kwa*, *kuga*). Novi metapodatki bodo med drugim omogočili tudi primerjavo z govornim korpusom GOS, odkrite značilke pa bomo nato uporabili pri razvoju in učenju modela za avtomatsko prepoznavanje regionalnih jezikovnih različic slovenščine na spletu. Za primerjavo pa bomo avtorje poskušali razvrstiti tudi z gručenjem, ki ne bo odvisno od vnaprej določenih regij.

6 Zahvala

Raziskava, opisana v prispevku, je bila opravljena v okviru nacionalnega temeljnega projekta “Viri, orodja in metode za raziskovanje nestandardne spletne slovenščine” (J6-6842, 2014-2017), ki ga financira ARRS.

Avtorja se za pomoč in nasvete pri pripravi prispevka iskreno zahvaljujeta Tomažu Erjavcu in Darji Fišer ter anonimnim recenzentom za konstruktivne opombe.

7 Literatura

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

Janne Bondi Johannessen, Joel Priestley, Kristin Hagen, Tor Anders Åfarli in Øystein Alexander Vangsnes. 2009. The Nordic Dialect Corpus – an Advanced Research Tool. V: K. Jokinen in E. Bick (ur.): Proceedings of the 17th Nordic Conference of

Computational Linguistics NODALIDA 2009. NEALT Proceedings Series Volume 4.

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

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

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

Jacob Eisenstein. 2015. Written dialect variation in online social media. V: C. Boberg, J. Nerbonne in D. Watt (ur.): Handbook of Dialectology. Wiley.

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

Darja Fišer, Tomaž Erjavec, Ana Zwitter Vitez in Nikola Ljubešić. 2014. JANES se predstavi: metode, orodja in viri za nestandardno pisno spletno slovenščino. V: Zbornik Devete konference Jezikovne tehnologije. Ljubljana: Institut Jožef Stefan.

Darja Fišer, Tomaž Erjavec, Jaka Čibej in Nikola Ljubešić. 2015. Gradnja in analiza korpusa spletne slovenščine JANES. Obdobja 2015.

Ernst Håkon Jahr. 1997. On the Use of Dialects in Norway. V: Heinrich Ramisch in Kenneth Wyne (ur.): Language in Time and Space: Studies in Honour of Wolfgang Viereck on the Occasion of his 60th Birthday, str. 363–369. Stuttgart: Franz Steiner Verlag.

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

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

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

Jack K. Chambers in Peter Trudgill. 1994. Dialectology. Cambridge: University Press.

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

Karmen Kenda Jež. 2004. Narečje kot jezikovnozvrstna kategorija v sodobnem jezikoslovju. V: E. Kržišnik (ur.): Obdobja 22. Ljubljana: Filozofska fakulteta Univerze v Ljubljani, Center za slovenščino kot drugi/tuji jezik, Oddelek za slovenistiko, str. 263–276.

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

Rudolf Kolarič. 1954. Die slowenische Mundartsforschung. V: Orbis: Bulletin International de Documentation Linguistique 3/1, str. 182–188. Louvain.

William Labov. 1994. Principles of Linguistic Change 1: Internal Factors. Oxford/Cambridge: Blackwell.

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

Nikola Ljubešić, Darja Fišer, Tomaž Erjavec, Jaka Čibej, Dafne Marko, Senja Pollak in Iza Škrjanec. 2015. Predicting the level of text standardness in user-generated content. RANLP 2015.

Tine Logar. 1959. Iz priprav za lingvistični atlas. V: Jezik in slovstvo 4, str. 129–135.

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

Herman Niebaum in Jürgen Macha. 1999. Einführung in die Dialektologie des Deutschen. Tübingen: Max Niemeyer Verlag.

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

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

Fran Ramovš. 1951. Osnovna črta v oblikovanju slovenskega vokalizma. V: Slavistična revija 4, str. 1–9.

Jerzy Reichan. 1999. Gwary polskie w końcu XX w. V: Polszczyzna 2000, str. 262–278.

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

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

Petr Sgall, Jiři Hronek, Alexandr Stich in Ján Horecký. 1992. Variation in Language: Code Switching in Czech as a Challenge for Sociolinguistics. Amsterdam/Philadelphia: John Benjamins Publishing Company.

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

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

Drago Unuk. 1997. Dialektologija kot jezikoslovna disciplina. V: Jezik in slovstvo 43, str. 307–313.

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