Zapisovalne prakse v spletni slovenščini

Darja Fišer, Maja Miličević Petrović, Nikola Ljubešić

Izvleček

V poglavju obravnavamo značilnosti nestandardnega zapisa besed v slovenskih objavah na družbenem omrežju Twitter. Analiza temelji na ročno normaliziranih, lematiziranih in oblikoskladenjsko označenih vzorcih tvitov v slovenščini. Podrobneje proučimo distribucijo pretvorb iz standardnega v nestandardni zapis besed glede na besedne vrste in leme ter distribucijo treh različnih vrst transformacij: izpust, dodajanje in zamenjavo črk. Rezultati kažejo, da je največ transformacij med polnopomenskimi besedami, vendar so transformacije slovničnih besed najpogostejše. Najpogostejši tip transformacij so izpusti, predvsem samoglasnikov, do česar največkrat prihaja na koncu besed, s čimer se neformalna komunikacija v tvitih približuje govoru.

Ključne besede: nestandardni zapis besed, računalniško posredovana komunikacija, Twitter, slovenščina

1 UVOD

Nekonvencionalno, nekanonično oz. nestandardno zapisovanje besed je ena od najopaznejših značilnosti računalniško posredovane komunikacije (RPK), prav tako pa povzroča tudi največ težav pri avtomatski obdelavi takšnih besedil, zato ni presenetljivo, da temu fenomenu RPK tako jezikoslovci kot jezikovni tehnologi posvečajo veliko pozornosti. Zgodnje raziskave so bile večinoma opravljene na SMS-sporočilih, predvsem v angleščini (Shortis 2007, Tagg et al. 2012) in francoščini (Anis 2007), danes pa so za številne jezike v ospredju družbena omrežja Facebook (Chariatte 2014), Twitter (Van Halteren in Oostdijk 2012, Sidarenka et al. 2013) in WhatsApp (Ueberwasser 2013).

Metodološko in jezikovnogradivno heterogenim raziskavam so skupne ugotovitve, da nestandardno zapisovanje besed v RPK, če izvzamemo tipkarske napake, večinoma ni kaotično in kataklizmično, temveč v veliki meri namerno, predvsem pa »funkcionalno, načelno in osmišljeno« (Tagg et al. 2012: 367). Funkcionalno, ker se pojavi v interakciji znotraj določene družbene skupine kot odziv na konkretne komunikacijske potrebe in tipično nima negativnega vpliva na razumevanje sporočila med pripadniki te skupine, načelno, ker navadno odseva splošne vzorce ortografske variacije v tem jeziku, in osmišljeno, ker prispeva k udejanjanju družbenih identitet uporabnikov. Thurlow in Brown (2003) navajata tri osnovne značilnosti nestandardno zapisanih besed: kratkost in hitrost (krajšanje, nadomeščanje črk s številčnimi homofoni, opuščanje kapitalizacije, ločil, presledkov, nadomeščanje črk s števil), nadomeščanje paralingvističnih sredstev (raba velikih črk in ponavljanje ločil za nadomeščanje prozodičnih in čustvenih komponent) in fonološka aproksimacija (približevanje zapisa (neformalnemu) govoru).

Zapisovalnim praksam v slovenskih tvitih je bilo posvečenih že kar nekaj študij. Analiza strategij krajšanja v tvitih (Goli et al. 2016) je pokazala močno tendenco po krajšanju, ki je izrazito pogosta v nestandardnih sporočilih in se kaže predvsem kot redukcija na nivoju zapisa besed (*mam* za *imam, anglesk* za *angleško, lah* za *lahko*). Analiza novih skovank v različnih tipih slovenskih uporabniških vsebin iz korpusa Janes, ki vsebujejo homofone s črkami in številkami (Marko 2016), je pokazala, da so ti izrazito značilni za Twitter, da se pojavljajo enako pogosto v tujih in slovenskih besedah in da se isti simbol lahko uporablja grafično (*g33k* za *geek*) ali fonetično (*u3nek* za *utrinek*). Analiza platform, ki omogočajo interaktivno in takojšnjo komunikacijo, kot je na primer Twitter, pa je pokazala, da se v njih brišejo meje med govornim in pisnim diskurzom (Zwitter Vitez in Fišer 2015), kar je med drugim razvidno iz pogoste rabe fonetičnega zapisa besed, pogovornih izrazov, deiktike in nestandardne leksike.

Vsi ti rezultati jasno kažejo, da so nekanonične zapisovalne prakse pri neformalnem komuniciranju slovenskih uporabnikov na družbenih omrežjih zelo razširjene, manjkajo pa raziskave, ki bi jih empirično preverile ter sistematično opisale, kako se razlikujejo od standarda. To je cilj pričujočega prispevka, v katerem se posvečamo značilnostim nestandardnega zapisa besed v slovenskih objavah na družbenem omrežju Twitter. Predstavljena analiza sodi v okvir večjezične raziskave (Miličević et al. 2017), kjer smo primerjali zapisovalne prakse v slovenskih, hrvaških in srbskih tvitih, pri čemer je v pričujočem prispevku poudarek na interpretaciji rezultatov za slovenščino.

Analiza temelji na ročno normaliziranih,[1] lematiziranih in oblikoskladenjsko označenih vzorcih tvitov v slovenščini, ki so bili izdelani za razvoj orodij za avtomatično normalizacijo in označevanje besedil računalniško posredovane komunikacije (Čibej et al. 2018). V nadaljevanju najprej predstavimo vzorec, ki je bil uporabljen za analize, nato pa opišemo postopek in rezultate analize. V prvem delu se osredotočimo na analizo odmikov od standardnega zapisa glede na besedno vrsto in lemo, v drugem pa se posvetimo pregledu vrst odmikov.

2 OPIS VZORCA

V prispevku uporabljamo vzorec, ki smo ga izluščili iz korpusa Janes-Norm in vsebuje slovenske uporabniško generirane vsebine, ki so bile ročno normalizirane, lematizirane in označene (Čibej et al. 2018). Glede na to, da poglavje obravnava nestandardni zapis besed v tvitih, smo v vzorec za analizo zajeli zgolj jezikovno nestandardne tvite (Erjavec et al. 2018), ki predstavljajo 1983 tvitov oz. 54.688 pojavnic.

Primer tvita z nestandardnimi prvinami je prikazan na Sliki 1. Te prvine vključujejo fenomene, značilne za računalniško posredovano komunikacijo na splošno, kot so fonetični zapisi tujih besed (npr. *lajk* za *like*), izpust strešic (*razrednicarka* za *razredničarka*) ali okrajšave (npr. *yt* za *YouTube*), fenomene, značilne za Twitter, kot so ključniki, omembe imen z znakom @ ali emotikoni/emojiji, in fenomene, ki se pogosto uporabljajo v neformalnih komunikacijskih situacijah, kot je uporaba pogovornih in narečnih nestandardnih oblik (npr. *tko* za *tako* ipd.).

Smernice za označevanje so bile oblikovane v okviru projekta JANES, tviti pa so bili označeni na petih ravneh: pojavnica (popravljanje mej med besedami), stavek (popravljanje stavčne segmentacije), normalizacija (standardizacija nestandardnih jezikovnih prvin), lematizacija (pripisovanje osnovne oblike vsaki

1 Normalizacija je postopek pripisovanja standardne ustreznice nestandardni pojavnici v korpusu (npr. *js* – *jaz*, glej Čibej et al. 2018).

besedi v tekočem besedilu, npr. *objavili – objaviti*) in oblikoskladenjski opis (pripisovanje oblikoskladenjskih oznak vsaki besedi v tekočem besedilu glede na standard MULTEXT-East v5.0,[2] npr. *demona* > *Sometd* za *samostalnik*, *občno ime*, *moški spol*, *ednina*, *tožilnik*, *živost*) (Čibej et al. 2018).

@user99 vrjamm [Verjamem] ja :) nm [Nam] pa rece [reče] razrednicarka [razredničarka], da je naj do 6ihne [6-ih ne] budimo, in tko [tako] npr [npr.] smo bli [bili] ze [že] enkrt [enkrat] ob 4 zjutri [zjutraj] pred Louvrom :D

Slika 1: Primer nestandardnega tvita s pripisanimi normaliziranimi oblikami (Tvit [standardna oblika besede]).

Od vseh ravni označevanja, s katerimi je bil korpus Janes-Norm označen, je za pričujočo raziskavo najpomembnejša raven z jezikovno normalizacijo. Normalizacija je bila omejena na nivo besede, kar pomeni, da besedni red, skladnja, uporaba ločil, elipse, uporabniška imena, ključniki, emotikoni/emojiji in leksikalne izbire (npr. pogovorni izraz *mobi* za *mobitel*) niso bile normalizirane. Je pa normalizacija vključevala standardiziranje tako nestandardnih različic črkovanja (npr. *jst* > *jaz*) kot tudi napak v črkovanju in tipkanju (npr. *popoldme* > *popoldne*) ter rediakritizacijo (npr. *vceraj* > *včeraj*). Pri normalizaciji so se označevalci držali načela minimalne intervencije. Z drugimi besedami – osredotočili smo se na nestandardne oblike, ki jih lahko razumemo kot odstopanje od standardnega črkovanja, pri tem pa nismo vplivali na slog, slovnico in fenomene, značilne za Twitter. Po drugi strani pa smo za razliko od nekaterih sorodnih raziskav variantnosti zapisovanja besed (npr. van Halteren and Oostdijk 2012) normalizirali nestandardno morfologijo (npr. *hodu* > *hodil*), saj sta cilja naših raziskav dvojna: zagotavljanje gradiva za razvoj orodij za avtomatsko procesiranje nestandardnega jezika in za raziskovanje pojavov v nestandardni spletni slovenščini.

Pri razreševanju nejasnih in dvoumnih primerov (npr. *k* > *ki* v *stvar k je postala »slavna«*, ampak *k* > *kot* v *ameriške jopice izgledajo, k da so jih babice spletle*) so označevalci upoštevali kontekst, če pa primera niso mogli razrešiti z uporabo danega konteksta, beseda ni bila normalizirana. Nestandardna pojavnica je bila v večini primerov normalizirana v eno standardno pojavnico, v redkih primerih pa je ena nestandardna pojavnica morala biti razdeljena v več standardnih pojavnic (1:n, *nevem – ne vem*) in obratno (n:1, *vse eno – vseeno*). Delež pojavnic z 1:n v vzorcu znaša 0,47 %, z n:1 pa 0,06 %.

Pri določanju smernic za normalizacijo se je bilo potrebno tudi natančno opredeliti do oblik, ki jih obravnavamo kot nestandardne, kar nujno vključuje tudi vprašanje jezikovne norme in referenčnih jezikovnih virov, ki jih pri

2 http://nl.ijs.si/ME/V5/msd/html/

označevanju upoštevamo. Vloga jezikovne norme in njen odnos do jezikovne rabe (preskriptiven : deskriptiven) sta v slovenistiki neusahljiv vir diskusij in polemik (glej Verovnik 2004 in Smolej 2015). V okviru predstavljene raziskave smo zavzeli deskriptivno stališče in nestandardnih oblik (z izjemo tipkarskih napak) ne obravnavamo kot napake, temveč kot variante, ki so v okoliščinah, v katerih so uporabljene, pretežno funkcionalne, načelne in osmišljene. Zato bi bilo zmotno našo normalizacijo interpretirati kot »popravljanje«, temveč nam služi kot pripomoček za lažjo avtomatsko obdelavo in analizo slovenščine, kot se uporablja v računalniško posredovani komunikaciji, kjer je standard razumljen kot skupni imenovalec, ne pa kot nabor neizpodbitnih pravil.

V smernicah se opiramo na splošno sprejete referenčne vire za standardno slovenščino, hkrati pa si prizadevamo upoštevati tudi realno jezikovno rabo. Zato smo anotatorje prosili, da se pri obravnavi povsem jasnih in neproblematičnih primerov, kot so manjkajoče strešice in očitne tipkarske napake, zanašajo na lastno intuicijo, v vseh ostalih primerih pa uporabijo referenčne priročnike v naslednjem vrstnem redu: (1) spletni portal Fran,[3] na katerem sta dostopna Slovar slovenskega knjižnega jezika in Slovenski pravopis, (2) oblikoslovni leksikon Sloleks,[4] (3) konkordančnik Gigafida[5] in (4) korpus Janes v0.4.[6] Uporaba korpusov je bila potrebna za pojavnice, ki niso zajete v referenčnih virih standardne slovenščine, še posebej pa za tiste, ki se pojavljajo v več različicah (npr. *fouš – fauš – favš*). V teh primerih smo anotatorje prosili, da jih normalizirajo v najpogostejšo obliko (npr. *fouš* v zgornjem primeru).

3 ANALIZA PODATKOV

V tem razdelku predstavimo rezultate analiz, ki so bile opravljene na normaliziranih slovenskih tvitih. Glede na to, da so smernice za normalizacijo temeljile na opisnih kategorijah, ki jih je avtomatsko težko identificirati (npr. fonetična transkripcija ali napačen zapis), smo se v tem prispevku omejili na analizo po avtomatsko določljivih kriterijih. S tem namenom smo se osredotočili na transformacije, tj. modifikacije, ki so se kazale z uporabo nestandardnega jezika v nasprotju s standardnim. Gre torej za nasprotni proces od ročne normalizacije, predstavljene v tretjem razdelku, kjer nestandardnim oblikam pripisujemo standardne različice (npr. *reko* > *rekel*). V naši analizi ta fenomen obravnavamo kot transformacijo standardne oblike *rekel* v nestandardno obliko *reko* z zamenjavo znakov.

3 http://www.fran.si

4 http://www.slovenscina.eu/sloleks

5 http://www.gigafida.net

6 Glej Erjavec et al. (2018).

Analizo smo izvedli za štiri ravni: (1) zapis izvornih pojavnic v primerjavi z (2) normaliziranimi,[7] (3) oblikoskladenjske oznake, pripisane normaliziranim pojavnicam in (4) leme, pripisane normaliziranim pojavnicam. Porazdelitev transformacij opazujemo glede na besedne vrste, prav tako pa izluščimo najpogosteje transformirane leme in pregibne oblike. Ko opazujemo pregibne oblike normaliziranih in izvornih pojavnic, klasificiramo razlike glede na Levenshteinove vrste transformacij (izpust, dodajanje, zamenjava; Levenshtein 1966), prav tako pa smo pozorni na položaj specifične transformacije znotraj besed.

3.1 Skupna pogostost transformacij

Skupni delež transformiranih pojavnic znaša 17,39 % (9.555 pojavnic). Pri nekaterih transformacijah gre zgolj za izpust strešic (č, ć, š, ž, đ > c, c, s, z, dj), ki so posledica tehničnih in ne jezikovnih razlogov (tipkanje na mednarodnih tipkovnicah je hitrejše brez uporabe strešic). Če te izločimo, ostane 15,56 % (8.552) transformiranih pojavnic. Rezultati so v skladu s predhodnimi raziskavami, ki kažejo, da je v slovenščini močnejša tendenca uporabe nestandardnih oblik kot izpuščanja diakritikov (Fišer et al. 2015, Miličević in Ljubešić 2016).

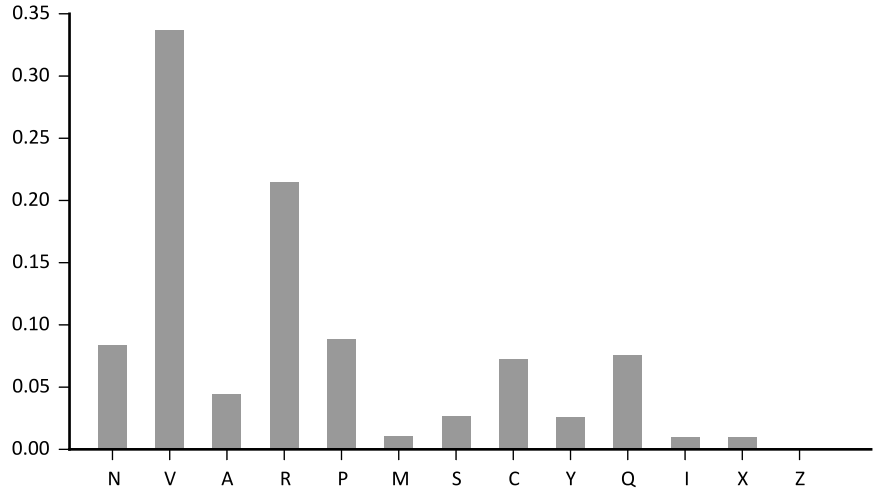
3.2 Analiza glede na besedne vrste

V tej analizi opazujemo porazdelitev transformacij glede na besedno vrsto (tj. koliko transformacij pripada določeni besedni vrsti). Prav tako izračunamo delež oblik, ki so bile transformirane za vsako besedno vrsto (tj. koliko besed izmed vseh, ki pripadajo določeni besedni vrsti, je bilo transformiranih). Obe analizi sta omejeni na pojavnice, kjer transformacija ni zajemala izpusta strešic.

3.2.1 Pogostost transformacij glede na besedno vrsto

Relativne frekvence transformacij glede na besedno vrsto so prikazane na Sliki 2. S slike je razvidno, da so najpogosteje transformirani glagoli, sledijo jim prislovi, zaimki in samostalniki.

7 Zaradi tehničnih omejitev platforme, na kateri je označevanje potekalo, je ena izvorna pojavnica lahko normalizirana v največ štiri pojavnice (npr. 1 > 2: *nevem* > *ne vem*, 1 > 3: *anede* > *a ne da*, 1 > 4: *norostinividetikonca* > *norosti ni videti konca*), prav tako je več izvirnih pojavnic lahko normaliziranih v eno samo pojavnico.



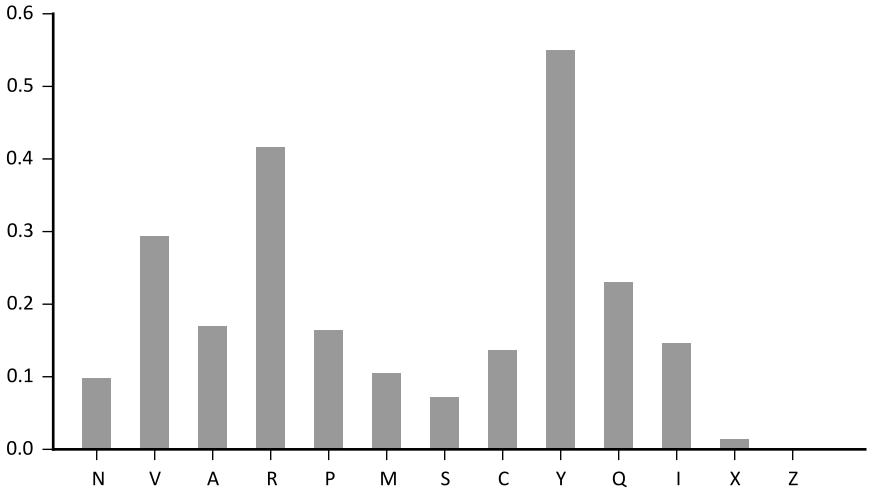
Slika 2: Distribucija transformiranih oblik glede na besedno vrsto[8] in relativno frekvenco.

Transformacije glagolov se v večini primerov nanašajo na pomožni glagol *biti*, predvsem pri obliki za prvo osebo ednine *sem* (pogosto zapisano kot *sm*) in pri obliki za tretjo obliko ednine v pretekliku *bilo* (skrajšano v *blo*). Pogoste so tudi transformacije drugih glagolov s krajšanjem nedoločnika, npr. *gledat* za *gledati*. Prislovi so večinoma okrajšani (npr. *tako* je pogosto skrajšan v *tko*), pojavljajo pa se tudi druge vrste transformacij. Zanimiv primer je *zdaj*, ki se v vzorcu kaže v treh različnih transformacijah, in sicer *zdej*, *zdj* in *zj*. Transformacije medmetov se nanašajo predvsem na ponovitev samoglasnikov ali zlogov (npr. *hahahaha*).

3.2.2 Deleži transformiranih oblik znotraj besedne vrste

Deleži oblik, ki so bile transformirane znotraj določene besedne vrste, kažejo, da so slovnične besedne vrste pogosteje transformirane kot polnopomenske. Najvišji delež transformiranih pojavnic najdemo med okrajšavami (pogosto gre za izpust končne pike, npr. *slo* namesto *slo.* za *slovenski*). Členki in vezniki so večinoma skrajšani z izpustom zadnjega samoglasnika, npr. *al* za *ali* in *kak* za *kako*. Najpogostejša transformirana oblika je osebni zaimek za prvo osebo ednine *jaz*, pogosto zapisana kot *jst*, *js*, *jest* ali *jz*.

8 Oznake besednih vrst so: N – samostalnik, V – glagol, A – pridevnik, R – prislov, P – zaimek, M – številnik, S – predlog, C – veznik, Y – okrajšava, Q – členek, I – medmet, X – neuvrščeno, Z – ločilo.



Slika 3: Deleži transformiranih oblik znotraj posameznih besednih vrst.

Med polnopomenskimi besednimi vrstami smo največ transformacij zasledili med prislovi, glagoli in pridevniki, kar sovpada s tendencami transformacij glede na besedne vrste, opisane v razdelku 3.2.1.

Pri prvi primerjavi zajemajo polnopomenske besede večino vseh transformacij, pri drugi pa vodijo funkcijske besede. Z drugimi besedami – čeprav so leksikalne besede pogostejše, jih transformiramo v manjši meri. To je tudi razlog, da leksikalne besede dominirajo na Sliki 2, na Sliki 3 pa ne.

3.3 Analiza glede na leme in pregibne oblike

V tem razdelku predstavimo analizi glede na najpogosteje transformirane leme (3.3.1) in pregibne oblike (3.3.2).

3.3.1 Analiza lem

V Tabeli 1 so predstavljene najpogosteje transformirane leme z deležem transformiranih oblik, ki jih pokriva določena lema (% skupaj, npr. *biti*) in deležem vseh oblik te leme, ki so bile transformirane (% lema, npr. *sm, blo, bla, nism, bit, bli, nebi, biu, sn, sm, neb, ble*). Transformacije z izpustom strešic ponovno niso upoštevane.

Tabela 1: 20 najpogosteje transformiranih lem v slovenskih tvitih.

Lema	% skupaj	% lema
biti#V	8,33 %	17,02 %
jaz#P	3,24 %	33,9 %
tudi#Q	3,13 %	82,21 %
imeti#V	3,09 %	66,5 %
saj#C	1,61 %	79,77 %
potem#R	1,49 %	73,41 %
tako#R	1,39 %	74,38 %
zdaj#R	1,34 %	76,16 %
malo#R	1,3 %	82,22 %
samo#Q	1,29 %	61,45 %
lahko#R	1,2 %	52,82 %
toliko#R	1,09 %	91,18 %
ne#Q	1,06 %	11,15 %
kaj#P	1,05 %	36,29 %
kar#R	1,04 %	70,08 %
ali#C	1,03 %	63,77 %
videti#V	0,83 %	76,34 %
misliti#V	0,81 %	62,73 %
kot#C	0,72 %	32,46 %
danes#R	0,70 %	61,86 %

Najpogosteje transformirana lema je pomožni glagol *biti*, sledijo ji funkcijske besede in medmeti. Med leksikalnimi besedami prednjačijo prislovi, med glagoli pa je največ nedoločnikov, kjer gre za izpust končnega *i*-ja. Samostalniki in pridevniki se na seznam ne uvrščajo.

3.3.2 Analiza pregibnih oblik

V Tabeli 2 podajamo 20 najpogostejših parov standardnih oblik in njihovih transformacij, pri tem pa izpuščamo tiste, pri katerih smo zabeležili le izpust strešic. Specifične transformacije so zapisane v oklepajih, prav tako so podani deleži teh oblik glede na celotno število transformacij.

Tabela 2: 20 najpogosteje transformiranih oblik v slovenskih tvitih.

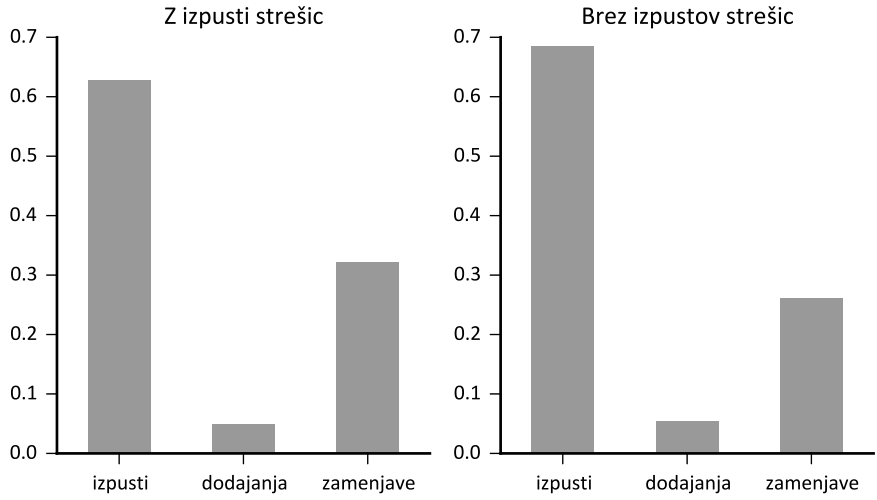
Oblika	% skupaj
sem (sm)	3,37 %
tudi (tud)	2,29 %
samo (sam)	1,93 %
bilo (blo)	1,68 %
potem (pol)	1,39 %
saj (sej)	1,30 %
tako (tko)	1,28 %
jaz (jst)	1,21 %
malo (mal)	1,21 %
kar (kr)	1,10 %
ali (al)	1,07 %
jaz (js)	1,03 %
zdaj (zdej)	0,97 %
tudi (tut)	0,89 %
imam (mam)	0,76 %
pri (pr)	0,70 %
ko (k)	0,70 %
kaj (kej)	0,70 %
nekaj (neki)	0,66 %
toliko (tolk)	0,66 %

Zelo pogosti sta obliki *js* in *jst* za *jaz*, druge transformacije pa zajemajo zamenjavo samoglasnika (tipično *a* > *e*) ali izpust samoglasnika na različnih položajih v besedi. Glede na besedno vrsto je med najpogostejšimi 20 pari največ prislovov.

3.4 Analiza glede na vrsto transformacije

V tem razdelku predstavimo verjetnostno porazdelitev treh vrst Levenshteinovih transformacij (Levenshtein 1996): izpust (npr. *tudi* > *tud*), dodajanje (npr. *super* > *suuuper*) in zamenjava (*zdaj* > *zdej*). Pri tem transformacije opazujemo v smeri od normaliziranih oblik k izvornim oblikam, ki jih najdemo v tvitih. Rezultati so povzeti na Sliki 4. Na levi strani slike so zajete vse transformacije. Najpogostejši so izpusti, sledijo zamenjave, najmanj pa je dodajanj. Na desni strani slike so zajete transformacije brez izpustov strešic, kjer tendence ostajajo podobne.

Najpogostejša vrsta transformacij je opuščanje znakov, sledijo zamenjave, dodajanja pa so v nestandardnem jeziku na Twitterju najredkejši fenomen.



Slika 4: Primerjava distribucij transformacij z upoštevanimi transformacijami zaradi golega izpuščanja strešic (levo) in brez njih (desno).

V naslednjem koraku analiziramo najpogostejše specifične transformacije, kjer ponovno izpuščamo besede, pri katerih gre zgolj za izpuščanje diakritičnih znamenj. V Tabeli 3 je prikazanih najpogostejših 10 transformacij za vsako izmed treh Levenshteinovih vrst, transformacije pa so podkrepljene tudi s pogosto uporabljenimi primeri.

Tabela 3: 10 najpogostejših transformacij za vsako vrsto (s primeri).

Izpust			Dodajanje			Zamenjava		
i	35,04 %	tudi > tud	a	25,8 %	pa > paa	l > u	14,65 %	mogel > mogu
e	17,83 %	sem > sm	h	14,97 %	haha > hahah	a > e	13,32 %	zdaj > zdej
o	13,30 %	lahko > lahk	e	14,17 %	ne > neee	j > i	5,21 %	zjutraj > zjutri
a	11,23 %	tako > tko	j	9,24 %	ne > nej	o > u	4,37 %	ono > uno
j	3,88 %	skoraj > skor	_	4,62 %	odkar > od kar	a > s	4,19 %	jaz > jst
_	3,10 %	ne bi > neb	o	4,14 %	zelo > zelooo	m > l	4,09 %	potem > pol
.	2,79 %	npr. > npr	s	3,98 %	imate > maste	a > o	3,98 %	danes > dons
t	2,73 %	potem > pol	i	3,82 %	vsak > saki	z > s	3,95 %	jaz > js
d	1,77 %	tudi > tut	u	3,82 %	super > suuuper	z > t	3,88 %	jaz > jst
u	1,26 %	tule > tle	m	2,71 %	bi > bim	i > t	3,57 %	tudi > tut

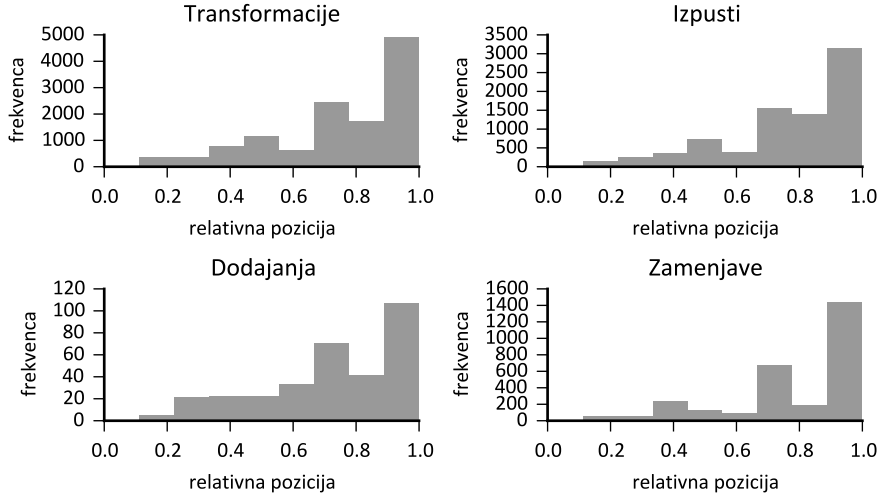
Uporabniki najpogosteje izpuščajo samoglasnike in presledke. Najpogostejši izpust, ki je hkrati tudi najpogostejša transformacija v celotnem korpusu, je samoglasnik *i (tudi > tud)*, ki zavzema več kot tretjino vseh izpustov. Sledijo mu izpusti samoglasnikov *e (sem > sm)*, *o* (*lahko > lahk*) in *a (tako > tko)*, pri čemer je izpuščanje samoglasnika *u* za cel velikostni razred redkejše (*tule > tle*). Samoglasniki so izpuščeni tako na koncu (*tudi > tud*) kot znotraj besede (*tako > tko*). Občutno redkejši so izpusti soglasnikov, med katerimi je največkrat izpuščen *j (skoraj > skor).* Med najpogostejših 10 izpustov se uvršča tudi izpuščanje presledka (*ne bi > neb*) in pike (*npr. > npr*).

Dodajanja so večinoma posledica ekspresivnega ponavljanja zlogov (npr. *hahahahaha*) ali samoglasnikov (npr. *zelooo*) v medmetih in polnopomenskih besedah. Druga najpogostejša kategorija je niz dveh besed, napačno zapisanih kot ena sama ali obratno (npr. *nebo* za *ne bo* ali *od kar* za *odkar*). Sledijo idiosinkratičen zapis domačih besed (npr. *sobitza* za *sobica*), neuveljavljen zapis krajšav (npr. *esemes* za *sms*) ali narečne oblike (npr. *imaste* za *imate*).

Pri zamenjavah je najpogostejša transformacija *l > u* pri deležnikih (*napisal > napisu*, *mogel > mogu*, *mislil > mislu* itd.), sledi pa ji zamenjava samoglasnikov *a > e*, s katero uporabniki zapis besed približujejo izgovoru (*kaj > kej*, *zdaj > zdej* itd.).

3.5. Analiza glede na položaj transformacije

V tem razdelku obravnavamo položaj transformacij (izpusta, dodajanja ali zamenjave) v besedi. Na Sliki 5 je prikazana skupna porazdelitev položaja transformacij, slike 6, 7 in 8 pa prikazujejo relativne položaje izpustov, dodajanj in zamenjav.



Slike 5–8: Distribucije transformacij glede na relativno pozicijo.

S Slike 5 je razvidno, da se transformacije najpogosteje pojavljajo na koncu besede, zelo redko pa na začetku. Podoben trend se pojavlja tudi pri specifičnih vrstah transformacij. Kot je razvidno s Slike 6, so izpusti večinoma vezani na konec besede, predvsem kot posledica izpusta zadnjega samoglasnika (npr. pri funkcijskih besedah in nedoločnikih, kot je prikazano v razdelkih 3.2 in 3.3). Dodajanja (Slika 7) in zamenjave (Slika 8) nakazujejo še močnejšo tendenco pojavljanja na koncu besede (npr. *ne* > *neee*). Podrobnejši pregled dodajanj razkriva, da gre v večini primerov za ponovitev zadnjega samoglasnika. Zamenjave na koncu besede so v veliki meri posledica zamenjave znakov *l* > *u* pri glagolih.

5 SKLEP

V poglavju smo obravnavali zapisovalne prakse v spletni slovenščini. V ta namen smo analizirali vzorec ročno normaliziranih, lematiziranih in oblikoskladenjsko označenih slovenskih tvitov, pri čemer smo se osredotočili na analizo transformacij nestandardnih besednih oblik glede na njihove standardne ustreznice. Analiza transformacij glede na besedne vrste je pokazala, da je teh največ pri polnopomenskih besedah, med katerimi prvo mesto zasedajo prislovi. Analiza znotraj posamičnih besednih vrst pa je pokazala obratno sliko, saj so najpogostejše transformacije slovničnih besed, kar potrjuje tudi ročna analiza najpogosteje transformiranih lem, ki razkriva, da med najpogosteje transformiranimi lemami najdemo največ pomožnih glagolov, medmetov in veznikov.

Z računanjem Levenshteinovih transformacij smo ugotovili, da so daleč najpogostejši tip transformacij izpusti. Glede na to, da smo za analizo uporabili tvite zasebnih uporabnikov, ki vsebujejo nestandardne prvine, je bil tak rezultat pričakovan, ne samo zaradi splošnega načela jezikovne ekonomičnosti, ampak tudi zaradi neformalnega, interaktivnega okolja komunikacije, ki se za povrh pogosto odvija na majhnih prenosnih napravah, ki so opremljene z neergonomičnimi tipkovnicami. Med izpusti prevladuje izpuščanje samoglasnikov, s čimer uporabniki posnemajo govor (glej Zwitter Vitez in Fišer 2018), dodajanja pa so v veliki meri posledica ekspresivnega ponavljanja črk in zlogov, predvsem pri medmetih. Zamenjave so raznorodnejše, vključujejo pa predvsem transformacije v pogovorne oblike in regionalne/narečne variante. Ugotovili smo, da se na začetku besede transformacije pojavljajo le redko, najpogostejše pa so na koncu besede, kar je sicer značilno za nestandardno govorjeno slovenščino (Može 2013), ki se ji neformalno, interaktivno komuniciranje na družbenih omrežjih približuje s fonetiziranim zapisom.

Identificirani fenomeni so primerljivi z raziskavami, opravljenimi na drugih jezikih (glej Miličević et al. 2017 za hrvaščino in srbščino, van Halteren in Oostdijk 2012 za nizozemščino, Sidarenka et al. 2013 za nemščino in Eisenstein 2013 za angleščino),

kjer prav tako prevladujeta težnja po krajšanju besed in prisotnost oblik, ki so značilne za (geografsko in demografsko) različne družbene skupine. Rezultati na nivoju zapisa besed še posebej izrazito kažejo brisanje meja med govorjenim in pisnim jezikom (glej Eisenstein 2013 za angleščino, Zwitter Vitez in Fišer 2015 za slovenščino), kar pomeni, da pojavi, na katere smo naleteli, niso novi, temveč zgolj nekoliko bolj opazni zaradi množičnejše komunikacije in trajnejšega medija v primerjavi z večino neformalnih govornih situacij. Rezultati prav tako nakazujejo povezavo med variantnostjo zapisa z udejanjanjem identitete uporabnikov (glej Tagg 2012), kjer odstopanje od norme z rabo regionalno in demografsko obarvanih različic igra pomembno vlogo. Vprašanje, kako se tovrstne jezikovne prakse vpenjajo v širšo razpravo o demokratizaciji jezika, jezikovne izbire in standardizacije, je kompleksno in bo zagotovo predmet zanimivih prihodnjih raziskav.

Glede na pomanjkanje empiričnih podatkov za računalniško posredovano komunikacijo v slovenščini pričujoča analiza predstavlja dragocen vpogled v naravo odstopanj od jezikovnih norm, prav tako pa služi kot osnova za prihodnje bolj poglobljene raziskave tega jezikoslovnega fenomena, ki bodo osredotočene na preverjanje specifičnih hipotez. Nadaljnje raziskave bi lahko vključevale analizo vpliva sociodemografskih faktorjev na opazovane transformacije, kot so starost, geografsko poreklo, izobrazba uporabnikov ipd. V prihodnosti bi prav tako lahko opravili leksikalno analizo nestandardnih prvin v računalniško posredovani komunikaciji. Tovrstni primeri v uporabljenem označenem korpusu niso zajeti, so pa predhodne raziskave (Fišer et al. 2015) že pokazale, da so ti primeri zelo pomembni za primerjave med jeziki.

Zahvala

Ker je avtorska zasedba tega poglavja mednarodna, smo rokopis pripravili v angleščini. V slovenščino ga je prevedla Dafne Marko, ki se ji za natančen in tekoč prevod ter skrbno upravljanje s terminologijo iskreno zahvaljujemo.

Literatura

Anis, Jacques, 2007: Neography: Unconventional spelling in French SMS. Danet, Brenda in Susan C. Herring (ur.): *The Multilingual Internet: Language, culture, and communication online*. Oxford: Oxford University Press. 87–115.

Chariatte, Nadine, 2014: "Facebook Style": The use of non-standard features in virtual speech conditioned by the medium Facebook. Brumme, Jenny in Sandra Falbe (ur.): *The Spoken Language in a Multimodal Context: Description, Teaching, Translation*. 93–114. Berlin: Frank & Timme.

Čibej, Jaka, Darja Fišer in Tomaž Erjavec, 2016: Normalisation, tokenisation and sentence segmentation of Slovene tweets. *Proceedings of Normalisation and Analysis of Social Media Texts (NormSoMe) 2016*, *LREC 2016*. 5–10. http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-NormSoMe_Proceedings.pdf.

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.

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

Eisenstein, Jacob, 2013: What to do about bad language on the Internet. Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics. 359–369. http://www.cc.gatech.edu/~jeisenst/papers/naacl2013-badlanguage.pdf.

Fišer, Darja, Tomaž Erjavec, Nikola Ljubešić, and Maja Miličević (2015): Comparing the nonstandard language of Slovene, Croatian and Serbian tweets. Smolej, Mojca (ur.): *Obdobja 34. Slovnica in slovar - aktualni jezikovni opis (1. del)*. Ljubljana: Znanstvena založba filozofske fakultete. 225–231.

Goli, Teja, Eneja Osrajnik in Darja Fišer, 2016: Analiza krajšanja slovenskih sporočil na družbenem omrežju Twitter. *Proceedings of the Language Technologies and Digital Humanities Conference*. Ljubljana, Slovenia. 77–82.

Levenshtein, Vladimir I, 1966: Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10/8: 707–710.

Marko, Dafne, 2016: The Use of Alphanumeric Symbols in Slovene Tweets. *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities*. Ljubljana, Slovenia. 48–53.

Miličević, Maja, Nikola Ljubešić in Darja Fišer, 2017: Birds of a feather don't quite tweet together: An analysis of spelling variation in Slovene, Croatian and Serbian Twitterese. Fišer, Darja in Michael Beißwenger (ur.): *Investigating Computer-Mediated Communication: Corpus-Based Approaches to Language in the Digital World*. Ljubljana: Znanstvena založba Filozofske fakultete. 14–43.

Miličević, Maja in Nikola Ljubešić, 2016: Tviterasi, tviteraši or twitteraši? Producing and analysing a normalised dataset of Croatian and Serbian tweets. *Slovenščina 2.0* 4/2. 156–188. http://dx.doi.org/10.4312/slo2.0.2016.2.156-188

Može, Sara, 2013: Raba kratkega nedoločnika: korpusni pristop. *Slovenščina 2.0* 1/1: 155–175. http://www.trojina.org/slovenscina2.0/arhiv/2013/1/Slo2.0_2013_1_08.pdf

Shortis, Tim, 2007: Revoicing Txt: Spelling, vernacular orthography and 'unregimented writing'. Posteguillo, Santiago, María José Esteve in M. Lluïsa Gea-Valor (ur.): *The Texture of Internet: Netlinguistics in Progress*. Newcastle: Cambridge Scholars Publishing. 2–23.

Sidarenka, Uladzimir, Tatjana Scheffler in Manfred Stede, 2013: Rule-based normalization of German Twitter messages. *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology*. https://gscl2013.ukp.informatik.tu-darmstadt.de/fileadmin/user_upload/Group_UKP/conferences/ gscl2013/workshops/sidarenka_scheffler_stede.pdf

Smolej, Mojca (ur.), 2015: *Slovnica in slovar – aktualni jezikovni opis*. *Obdobja 34*. Ljubljana: Znanstvena založba Filozofske fakultete.

Tagg, Caroline, Alistair Baron in Paul Rayson, 2012: "i didn't spel that wrong did i. Oops": Analysis and normalisation of SMS spelling variation. *Lingvisticæ Investigationes* 35/2. 367–388.

Thurlow, Crispin in Alex Brown, 2003: Generation Txt? The sociolinguistics of young people's text-messaging. *Discourse Analysis Online* 1/1.

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.

Halteren, Hans van in Nelleke Oostdijk, 2012: Towards identifying normal forms for various word form spellings on Twitter. *CLIN Journal* 2. 2–22.

Verovnik Tina, 2004: Norma knjižne slovenščine med kodifikacijo in jezikovno rabo v obdobju 1950–2001. *Družboslovne razprave XX*, 46/47: 241–258.

Zwitter Vitez, Ana in Darja Fišer, 2015: From mouth to keyboard: the place of non-canonical written and spoken structures in lexicography. *Electronic lexicography in the 21st century: linking lexical data in the digital age: proceedings of eLex 2015 Conference*, 11-13 August 2015, Herstmonceux Castle, United Kingdom. Ljubljana: Trojina, zavod za uporabno slovenistiko; Brighton: Lexical Computing. 250–267.

Zwitter Vitez, Ana in Darja Fišer, 2018: Govorne prvine v nestandardni spletni slovenščini. Fišer, Darja (ur.): *Viri, orodja in metode za analizo spletne slovenščine*. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani. 254–273.