Maja Miličević Petrović[*]
Filološki fakultet, Beograd[**]
Nikola Ljubešić
Institut „Jožef Stefan”, Ljubljana
Darja Fišer
Filozofski fakultet, Ljubljana

УДК 004.773:316.472.4
ДОИ https://doi.org/10.18485/analiff.2017.29.2.8

NESTANDARDNO ZAPISIVANJE SRPSKOG JEZIKA NA TVITERU: MNOGO BUKE OKO MALO ODSTUPANJA?[***]

U radu se analiziraju varijante nestandardnog zapisa reči srpskog jezika koje se javljaju na društvenoj mreži Tviter. Građu za analizu čini uzorak automatski prikupljenih tvitova za koje je utvrđeno da sadrže nestandardne odlike. Uzorak je ručno normalizovan, označen morfosintaksičkim informacijama i lematizovan. U analizi se nestandardni oblici porede sa standardnim oblicima na koje su normalizovani i utvrđuje se kakvim transformacijama su nastali. Analiza pokazuje da su transformacije posebno česte kod zatvorenih klasa reči, poput pomoćnih glagola, uzvika i skraćenica. Brisanje karaktera je češće od njihovog dodavanja i zamene, a sve transformacije najčešće se javljaju na kraju reči. Uprkos nesumnjivom prisustvu nestandardnih odlika u jeziku Tvitera, zaključuje se da su one ukupno gledano nedovoljno učestale i suviše funkcionalno specijalizovane da bi se iz njih izveo zaključak o negativnom uticaju komunikacije posredovane računarom na standardnojezičku normu.

Ključne reči: komunikacija posredovana računarom, jezik društvenih mreža, Tviter, korpusna lingvistika, nestandardni jezik

---

[*] Katedra za opštu lingvistiku, Filološki fakultet, Studentski trg 3, 11000 Beograd, m.milicevic@fil.bg.ac.rs.

[**] Rad je nastao u okviru projekata *Standardni srpski jezik: sintaksička, semantička i pragmatička proučavanja*, koji finansira Ministarstvo prosvete, nauke i tehnološkog razvoja Republike Srbije (projekat br. 178004), *Regional Linguistic Data Initiative*, koji je finansirao Švajcarski nacionalni fond za nauku (projekat br. 160501, 2015-2017) i *Jezikoslovna analiza nestandardne slovenščine*, finansiranog od strane Agencije za istraživanja Republike Slovenije (projekat br. J6-6842, 2014-2017).

[***] Deo rada prethodno je pod naslovom „(Ne)standardni jezik Tvitera” izložen na skupu „Jezici i kulture u vremenu i prostoru 7”, u tematskoj sekciji „Srpski jezik na prelazu vekova: novine u sistemu i upotrebi”, 19.11.2017. godine u Novom Sadu.

1. Uvod

Komunikacija posredovana računarom (KPR; eng. *computer-mediated communication*, CMC) obeležila je brojne sfere ljudskog života od kraja XX veka do danas, do te mere da se često govori o „komunikacijskoj revoluciji" (Ćosić, 2005).[1] Novi vidovi komunikacije nužno su se odrazili i na jezičku upotrebu. U domenu privatnog komuniciranja veliki značaj dobijaju elektronska pošta i SMS poruke, kao i ćaskaonice i aplikacije za dopisivanje kakve su *Viber* i *WhatsApp.* Sa druge strane, blogovi, forumi, mogućnost pisanja komentara na tuđe tekstove, a posebno društvene mreže poput Fejsbuka (*Facebook*) i Tvitera (*Twitter*), omogućili su svakome ko to želi da ostavi pisani trag i u javnom prostoru. Otvorenost velikom broju govornika snizila je stepen formalnosti javne pisane komunikacije, koja je ranije bila podložna praćenju norme i gotovo nužno je prolazila filter intervencije lektora, čime je doprinela jednom vidu demokratizacije jezika. Očekivana posledica je pojava odstupanja od jezičkog standarda i u javnoj sferi.

Paralelno sa pojavom komunikacije posredovane računarom pojavljuju se i rasprave o uticaju koji ona ima, ili se očekuje da će dugoročno imati na standardni jezik. Uočava se bliskost velikog dela KPR sa govornim jezikom, usled toga što ovakva komunikacija, iako pisana, u velikoj meri zapravo preuzima funkcije govora. Na primer, elektronska pošta ne zamenjuje samo tradicionalna pisma, već i telefonske razgovore (Ćosić, 2005). Sa druge strane, naglašavaju se generacijske razlike i jezik „novih medija" uglavnom se povezuje sa mlađom populacijom, za čiju pismenost postoji zabrinutost i u većim i u manjim jezicima (v. Verheijen, 2013 za engleski; Verheijen i Spooren, 2017 za holandski; Vlajković, 2010 i Stamenković i Vlajković, 2012 za srpski).

Međutim, istraživanja pokazuju da odstupanja od standarda proističu iz posebnih tehničkih i komunikacijskih okolnosti u kojima se KPR odvija, pa kao takva ne predstavljaju haotičnu i nekontrolisanu pojavu koja je pretnja standardnom jeziku. Na celokupnu KPR mogu se primeniti zaključci do kojih dolaze Thurlow i Brown (2003), koji smatraju da upotrebu engleskog jezika u SMS porukama oblikuju so-

1 Iako naziv to ne odražava direktno, u komunikaciju posredovanu računarom ubraja se i komunikacija putem mobilnih telefona, tableta i drugih sličnih prenosnih uređaja.

ciolingvistička načela kratkoće i brzine, iskazivanja paralingvističkih informacija i fonološke aproksimacije (približavanja zapisa neformalnom govoru). Rezultat načela kratkoće poruke i brzine njenog pisanja mogu biti oglušenja o pravopisnu normu, poput pisanja isključivo malim slovima, izostavljanja interpunkcije ili upotrebe slova bez dijakritičkih znakova. Drugi čest proizvod ovog načela jeste skraćivanje reči, koje pored relativno ustaljenih skraćenica i akronima vrlo često uključuje kontrakcije (poput *vrv* za *verovatno* u srpskom), ili upotrebu brojeva umesto niza slova (up. eng. *2nite* < *tonight*). Zatim, potreba za iskazivanjem paralingvističkih informacija dovodi do pojave novih sredstava za ostvarivanje funkcija koje pisani jezik izražava manje direktno nego govor: izraz lica i gestikulacija nadoknađuju se upotrebom emotograma (eng. emoticon), a velika slova i ponavljanje karaktera iskazuju naglašavanje. Najzad, približavanje zapisa govoru ogleda se u načinu pisanju koje više odgovara govornim formama (eng. *dunno* < *don't know*, u srpskom *desi* < *gde si*).

Iako se neke od ovih pojava mogu objasniti tehničkim okolnostima u kojima se KPR odvija, pre svega odlikama tastatura koje se koriste i ograničenjem broja karaktera u poruci,[2] dodatni faktor koji treba uzeti u obzir jeste svestan izbor načina izražavanja koji odslikava identitet govornika, bilo geografski (što vodi ka upotrebi dijalekatskih oblika) ili demografski (što može voditi upotrebi kolokvijalnih izraza). Radić-Bojanić (2007) i Popović (2009) navode i da odstupanje od jezičke norme može biti sredstvo privlačenja i zadržavanja pažnje sagovornika, dok Radić-Bojanić (2007) i Filipan-Žignić (2012) ukazuju na povezanost odstupanja od norme sa većom kreativnošću u upotrebi jezika. U svakom slučaju, upotreba nestandardnih oblika može se opisati kao „funkcionalna, principijelna i smislena" (Tagg, Baron i Rayson, 2012: 367). Funkcionalna je zbog toga što se ovi oblici javljaju u komunikaciji određene društvene grupe, iz konkretnih komunikacijskih potreba i bez negativnih posledica po međusobnu razumljivost. Principijelna je zbog toga što ne narušava sistemska pravila koja bi ugrozila razumevanje i slična je drugim situacijama u kojima se javljaju nestandardni

2 Čini se da ograničenje broja karaktera ipak nema presudan uticaj na pojavu nestandardnih elemenata (v. pregled u Eisenstein, 2013). U srpskom jeziku na to ukazuje povremena upotreba digrama ili trigrama umesto pojedinačnih karaktera sa djakriticima (na primer, *ss, sh* ili *sch* za *š*).

oblici u istom jeziku.[3] Najzad, smislena je zbog toga što doprinosi izgradnji i prikazu identiteta govornika. Karakteristike koje se tiču uspostavljanja identiteta posebno su uočljive u domenu leksike, koja je u KPR često žargonska ili geografski obeležena, i morfofonologije, koja može biti dijalekatska, poput upotrebe ikavskih oblika u Hrvatskoj (v. Miličević i Ljubešić, 2016).

U pogledu mogućeg uticaja KPR na pismenost mladih, istraživanja daju donekle protivrečne rezultate (v. pregled za engleski jezik u Verheijen, 2013). Međutim, sve su brojnije korpusne i eksperimentalne studije koje ukazuju ne samo na izostanak negativnih posledica KPR na opštu pismenost (v. npr. Verheijen i Spooren, 2017 za holandski), već i na pozitivnu povezanost upotrebe KPR sa pismenošću i uspesnošću u gramatičkim zadacima (v. Van Dijk, van Witteloostuijn, Vasić, Avrutin i Blom, 2016).

Polazeći od viđenja prema kome KPR sadrži nestandardne oblike koji su upotrebljeni sa određenim ciljem i stoga ne predstavljaju greške (uz izuzetak grešaka u kucanju), u radu se koncentrišemo na odlike nestandardnog srpskog jezika na Tviteru. Tviter je jedna od najčešće korištenih i najviše proučavanih društvenih mreža, na kojoj se u proseku dnevno objavi preko 500 miliona tvitova (kratkih poruka, dužine do 140 karaktera) vrlo raznolikog sadržaja, od vesti i zvaničnih informacija koje objavljuju kompanije i institucije, do ličnih razmišljanja i privatne komunikacije.[4] U radu predstavljamo analizu zasnovanu na uzorku tvitova koji je ručno normalizovan, a zatim lematizovan i označen morfosintaksičkim informacijama. U nastavku rada prvo se osvrćemo na relevantna prethodna istraživanja, zatim opisujemo sastav korpusa i način izdvajanja uzorka, nakon čega predstavljamo tok procesa normalizacije, metod i rezultate analize nestandardnih oblika. Analiza je delom kvantitativna, a delom kvalitativna. U kvantitativnom delu navode se podaci o učestalosti i raspodeli nestandardnih oblika prema vrstama reči, daje se pregled reči koje se često javljaju u nestandardnom obliku i ispituju se procesi koji dovode do formiranja nestandardnih oblika. Kvalitativna analiza daje lingvističko tumačenje dobijenih kvantitativnih rezultata.

3 Tagg i saradnici navode kao primer to da se u engleskom jeziku umesto *ch* ili *q* ne sreće neko proizvoljno slovo, već isključivo *k* (*skool* < *school*, *kwik* < *quick*), koje se koristi i u fonetskoj transkripciji. Uz to se ovakvi primeri beleže i daleko pre pojave RPK.

4 http://www.internetlivestats.com/twitter-statistics/

2. Prethodna istraživanja KPR na srpskom i srodnim jezicima

Na području nekadašnjeg srpskohrvatskog jezika istraženi su različiti vidovi komunikacije posredovane računarom: SMS poruke (Filipan-Žignić, Sobo i Velički, 2012; Vrsaljko i Ljubomir, 2013; Jelić i Polovina, 2015), ćaskaonice (Radić-Bojanić, 2007), društvene mreže (Fejsbuk: Vrsaljko i Ljubomir, 2013; Vlajković 2010; Stamenković i Vlajković 2012; Tviter: Miličević i Ljubešić, 2016; Miličević, Ljubešić i Fišer, 2017). Vršena su i poređenja različitih žanrova (Popović, 2009; Filipan-Žignić, 2012). Najviše pažnje posvećeno je leksici (anglicizmima, vulgarizmima, žargonu mladih) i odstupanjima od pravopisnih pravila (v. sledeći pasus). U manjoj meri su proučeni planovi morfologije (tj. sporadična odstupanja od standarda u fleksiji), sintakse (koja je pojednostavljena, često eliptična i uglavnom zasnovana na prostim rečenicama), diskursa i pragmatike (gde pojedinačni žanrovi imaju sopstvene komunikacijske norme).

Za naš rad najrelevantnija su istraživanja odstupanja od pravopisnih pravila, pri čemu pod pravopisom ovde podrazumevamo sve pojave vezane za različite načine zapisivanja teksta, odnosno obuhvatamo i neke od pojava koje prvenstveno pripadaju domenu leksike (npr. različiti vidovi skraćivanja reči) ili morfologije (nestandardna fleksija). Citirani radovi s obzirom na ovako shvaćen pravopis uglavnom uočavaju slične tendencije. Ključne pravopisne odlike KPR na srpskom jeziku rezimirane su i ilustrovane primerima u tabeli 1 (izvori iz kojih su preuzeti primeri navedeni su u legendi ispod tabele).

Pojava	**Primer**
Izostavljanje interpunkcije	*<JANJA_beba> Vlado jos se ti lozis na Sandru* *<iLLeGaL> ma jok* (1)
Izostanak upotrebe velikih slova	*to deki tooo* (1)
Mešanje malih i velikih slova unutar reči	*lJaKsE Te nApRaViO* (1)
Pogrešno sastavljeno ili rastavljeno pisanje reči	*neznam [ne znam], skim [s kim], dali [da li]* (1)
Izostavljanje dijakritičkih znakova	*pazljivo [pažljivo] pise [piše] u log* (1)
Alternativni zapis slova sa dijakriticima[5]	*U utorak tje [će] dotji [doći] do rashchishtjavanja [raščišćavanja]* (2)

5 Ova i sledeća stavka spadaju u način pisanja koji se ponekad označava kao „internetica" (v. Popović, 2009).

Zamena domaćih slova ili kombinacija slova stranim	*exkurzija [ekskurzija]* (2), *hwala [hvala], gloop [glup]* (3)
Zamena slova simbolima i brojevima koji liče na njih	*tr38@ [treba] mi kint@ [kinta]* (3)
Upotreba ćiriličnog pisma samo za pojedine reči ili delove reči	*tjixu tjixu tjixu xuuuu [ćihu ćihu ćihu hu]* (3)
Naglašavanje pisanjem velikim slovima	*Da li je to bila ONA exkurzija?* (2)
Naglašavanje ponavljanjem slova i slogova	*Grrrrr, Hahahahahaha, pa kreeeeeeeeeeten* (2)
Naglašavanje ponavljanjem upitnika i uzvičnika	*Cao svima!!!!!!!!!!! Neko za chat??????????* (1)
Naglašavanje pisanjem reči sa razmakom između slova	*s o r i* (1)
Skraćivanje reči izgovornim oblicima brojeva	*o5 [opet], 3k [trik]* (2)
Skraćivanje reči i većih celina izostavljanjem samoglasnika	*bzv [bezveze]* (2), *Fnmnln! [fenomenalno], nshvts [ne shvataš]* (3)
Upotreba postojećih standardnih skraćenica	*bg [Beograd], ns [Novi Sad]* (1)
Skraćivanje reči inicijalnom elizijom[6]	*dja [sviđa]* (3), *si [jesi] tu* (2)
Skraćivanje reči finalnom elizijom[7]	*pozz [pozdrav]* (3), *kanc [kancelarija], dž [džabe]* (2)
Skraćivanje akronimizacijom	*nh [ne, hvala]* (1), *vtnns [volim te najviše na svetu]* (3)
Upotreba stranih reči, izraza i rečenica	*kakav offensive nick* (1)
Fonetizovana transkripcija stranih reči	*d fors iz vid ju jang skajvoker [the force is with you young skywalker]* (3)
Upotreba šatrovačkog	*Aj lipa [pali] tebra [brate]!* (2)
Nestandardna fleksija	*ja cu ih zabraniti u australiju [australiji]* (1)

Tabela 1: Najčešće nestandardne pravopisne odlike KPR na srpskom jeziku [sa standardnim oblikom]. Izvori: (1) Radić-Bojanić (2007), (2) Popović (2009), (3) Vlajković (2010).

Grupisanjem nabrojanih pojava može se uočiti nekoliko osnovnih tendencija u odstupanjima od standardnojezičkog zapisa: (1) narušavanje pravopisnih pravila vezanih za interpunkciju, upotrebu velikih i malih slova i sastavljeno i rastavljeno pisanje reči, (2) alternativni način zapisivanja pojedinih slova, (3)

6 Poredeći srpski jezik sa engleskim, Radić-Bojanić (2007: 61) navodi da je elizija reči relativno retka i manje raznovrsna – uglavnom uključuje govorne oblike koji su preneti u pisani jezik (poput *odma, ajd, al*).

7 U konsultovanoj literaturi nisu pronađeni primeri medijalne ili kombinovane medijalne i finalne elizije za srpski jezik (up. eng. *b-day [birthday], hub [husband]*, iz Radić-Bojanić, 2007: 60).

upotreba nekih vidova nestandardnog zapisa, posebno ponavljanja, u svrhu naglašavanja, (4) skraćivanje. Ove tendencije potpuno su u skladu sa tehničkim zahtevima KPR (posebno 1 i 2), komunikativnim potrebama korisnika (posebno 3) i/ili principom jezičke ekonomije (posebno 4).[8]Ostvarenje komunikativnih potreba korisnika i iskazivanje njihovog identiteta ostvaruju se i kroz poslednje četiri pojave iz tabele 1, koje ne pripadaju izdvojenim tendencijama. Dodatno se u nekim pojavama može primetiti uticaj engleskog jezika.

Istraživanja su se bavila i dodatnim temama, između ostalog poređenjem sa drugim jezicima. Filipan-Žignić i dr. (2012) su za SMS poruke utvrdili da su nestandardne odlike hrvatskog jezika kakav se koristi u njima slične odlikama engleskog i nemačkog i uglavnom proističu iz potrebe za skraćivanjem, koje u slučaju SMS poruka dovodi do uštede ne samo prostora i vremena, već i novca. Radić-Bojanić (2007) poredi srpski i engleski jezik ćaskaonica i takođe pronalazi brojne sličnosti, ali i određene razlike. Neke od razlika proizilaze iz sistemskih odlika ovih jezika (na primer, veća varijacija u zapisu u engleskom jeziku usled njegove manje transparentne ortografije), dok neke nemaju jasno objašnjenje (na primer učestalija upotreba elizije u engleskom jeziku). Ukupno gledano, utvrđuje se procentualno viši broj ogrešenja o pravopisna pravila u srpskom nego u engleskom jeziku, što autorka objašnjava tradicijom pisanog testiranja na anglosaksonskom, a usmenog ispitivanja na srpskom govornom području, što bi moglo da utiče na svest o pravopisnim pravilima.

Još jedna vrlo zastupljena tema jeste uticaj KPR na jezičku normu i pismenost govornika, posebno mlađe populacije i posebno u pogledu uticaja engleskog jezika. Baveći se hrvatskim jezikom, Vrsaljko i Ljubomir (2013) analiziraju uzorak SMS poruka i komentara na Fejsbuku učenika IV razreda osnovne škole i utvrđuju niz odstupanja (na primer u pogledu razlikovanja glasova *č* i *ć* i pravila alternacije *je/ije*) koja vide kao nepoželjna. Zbog toga predlažu da se u školama radi na tome da se standard poštuje i u neformalnoj komunikaciji, posebno na ranom uzrastu, dok se pravopisna pravila još usvajaju. Vlajković (2010) i Stamenković i Vlajković (2012) imaju sličan pristup srpskom jeziku. Ovi autori proučavaju korpus od 500 iskaza sa društvene mreže Fejsbuk i iznose tvrdnju da se u njima koristi „hibrid" engleskog i srpskog jezika, odnosno da su odstupanja od norme

8 Zanimljivo je i da neke od nabrojanih pojava korisnici ponekad sami ispravljaju (v. Jelić i Polovina, 2015).

koja se sreću u manjoj meri nastala pod uticajem konvencija koje sam tip komunikacije nosi sa sobom, a u većoj meri pod uticajem engleskog jezika. Neke od karakterističnih pojava koje uočavaju jesu pisanje naziva (filmova, muzičkih grupa, itd.) i etničkih prideva velikim slovom, u skladu sa engleskim pravopisom. Ovi autori smatraju da su pojave koji se sreću u njihovim podacima zabrinjavajuće, da prete da pređu i u druge vidove komunikacije i time dovedu do „erozije" i „siromašenja" srpskog jezika, zbog čega predlažu reforme u obrazovnom sistemu.

Sa druge strane, Filipan-Žignić, Legac, Pahić i Sobo (2015) i Filipan-Žignić i Turk Sakač (2016) smatraju da nema razloga za ovakve bojazni. Njihova istraživanja bila su anketnog tipa i tražila su od učenika završnih razreda hrvatskih gimnazija i osnovnih škola da odgovore na pitanja vezana za to da li i u kojoj meri u radovima pisanim za školu i u slobodno vreme koriste anglicizme, skraćenice, emotograme, vulgarizme i druga sredstva karakteristična za KPR. Pokazalo se da uprkos čestoj upotrebi nestandardnih oblika u neformalnom pisanju, učenici izuzetno retko prenose ovakav jezik u školske zadatke. Ovi autori stoga zaključuju da su maternji govornici sposobni da razluče različite komunikativne situacije i konvencije koje one nameću, odnosno da nema razloga da se govori o „pogoršanju" ma čijeg jezika.

Nešto drugačiji pristup problemu pručavanja nestandardnih odlika u komunikaciji posredovanoj računarom nalazimo u nizu radova posvećenih odlikama srpskog, hrvatskog i slovenačkog jezika na društvenoj mreži Tviter. Za razliku od prethodno prikazanih radova, koji su u velikoj meri kvalitativni i zasnovani na isključivo ručno sprovedenim analizama, u pitanju su kvantitativna istraživanja u kojima je deo analize sproveden automatski. Prvi od radova, Fišer, Erjavec, Ljubešić i Miličević (2015), poredi nivo lingvističke i tehničke standardnosti tvitova napisanih na slovenačkom, hrvatskom i srpskom jeziku. Utvrđeno je da je 67% tvitova u srpskom, 70% u slovenačkom i 73% u hrvatskom vrlo standardno, kao i to da je u ostatku tvitova u slovenačkom jeziku najviše nestandardnih odlika bilo vezano za ortografiju, dok je u hrvatskom, a posebno srpskom, bila dominantnija nestandardna leksika.

Istraživanje predstavljeno u ovom radu posebno tesno je povezano sa radovima Miličević i Ljubešić (2016) i Miličević i dr. (2017), koji su zasnovani na poređenju nestandardnih oblika sa standardnim oblicima na koje su (ručno) normalizovani (npr. *veceras* > *večeras*). Ovi radovi takođe

su utvrdili da postoje određene razlike između nestandardnog srpskog, hrvatskog i slovenačkog, delom lingvistički uslovljene (na primer, elizija završnih vokala češća je u slovenačkom i hrvatskom nego u srpskom, velikom delom zbog čestog ispuštanja finalnog *i* u infinitivima), a delom verovatno proistekle iz demografskih varijabli poput uzrasta korisnika (za srpski jezik je karakterističan viši broj vulgarizama). Najzad, sva tri rada su pokazala da se dijakritički znakovi najviše izostavljaju u srpskom jeziku.

3. Nestandardno zapisivanje srpskog jezika na Tviteru

3.1 Ciljevi istraživanja i opis uzorka

Studija predstavljena u ovom radu zasnovana je na automatski sastavljenom korpusu tvitova na srpskom jeziku i deo je šireg istraživanja u okviru koga su korpusi poruka sa Tvitera izrađeni za slovenački, hrvatski i srpski jezik. Krajnji cilj formiranja ovih korpusa jeste izrada alata za automatsko označavanje nestandardnog jezika (v. npr. Ljubešić, Erjavec i Fišer, 2017), koja zahteva da deo korpusa prvobitno bude ručno označen, između ostalog normalizovanjem nestandardnih oblika, kako bi se omogućilo treniranje automatskih alata. Naravno, dodatni značajan cilj je i olakšavanje lingvističke analize nestandardnog jezika. U radu se konkretnije bavimo analizom manjeg ručno normalizovanog uzorka sa ciljem utvrđivanja najčešćih nestandardnih odlika srpskog jezika na Tviteru.

Korpus čiji uzorak analiziramo prikupljen je upotrebom programa TweetCat (Ljubešić, Fišer i Erjavec, 2014) i obuhvata tvitove na srpskom jeziku objavljene između 2013. i 2015. godine. Ukupna veličina korpusa iznosi približno 205 miliona tokena, iz kojih je za potrebe ovog istraživanja izdvojen manji uzorak od 1.856 tvitova, odnosno 45.134 tokena (po završetku procesa normalizacije 45.322 tokena).[9] Odabrani su isključivo tvitovi koji sadrže nestandardne odlike i imaju dužinu od najmanje 100 karaktera.

Primeri tvitova sa nestandardnim elementima dati su u (1) i (2). Neke od ilustrovanih nestandardnih pojava tipične su za KPR uopšte, poput izostavljanja dijakritika (*veceras* < *večeras*) i upotrebe skraćenice (*yt* za *YouTube, mob* za *mobilni*), ali prisutni su i fenomeni specifični za

9 Kao zasebni tokeni tretiraju se ne samo reči, već i znakovi interpunkcije i emotogrami.

Tviter, poput upotrebe tarabe (*hashtag*) za označavanje teme i znaka @ za navođenje imena.

(1) Bad Copy i Sasa [Saša] Kovacevic [Kovačević] su skoro istovremeno objavili spotove veceras [večeras], a Bad Copy imaju vise [više] lajkova do sad na yt #geto #kvalitet

(2) Meni @ts_ts_ts kada nešto hoće da poruči, on ne uzme svoj mob [mobilni] da mi pošalje esemes [SMS], već moj i napiše mi u noutsima.

3.2 Proces normalizacije

Uzorak je ručno normalizovan na platformi Webanno (Eckart de Castilho, Biemann Gurevych i Yimam, 2014). Svaki tvit normalizovala su dva anotatora, nakon čega je treći anotator poredio rezultate i razrešavao neslaganja. Pored lingvističke normalizacije urađene su i ispravke automatske podele tvitova na tokene i rečenice, lematizacija (dodeljivanje svakom tokenu leme, odnosno osnovnog oblika, npr. *objavili* > *objaviti*) i morfosintaksički opis (dodeljivanje oznake svakoj reči u tekstu, prema smernicama MULTEXT-East v5.0,[10] npr. *koleginica* > *Ncfsn* za *noun, common, feminine, singular, nominative*, odnosno zajedničku imenicu ženskog roda u nominativu jednine).

Normalizacija je proces koji značajno olakšava kasniju lingvističku analizu, budući da korisnicima korpusa pruža mogućnost pretrage prema standardnim oblicima.[11] Međutim, glavni cilj normalizacije u ovom slučaju bilo je kasnije treniranje alata za automatsku normalizaciju, pa da su odluke u pogledu toga koje nestandardne pojave će biti normalizovane bile podređene tome. Drugim rečima, normalizovane su pojave za koje je procenjeno da mogu biti naučene od strane programa za automatsku normalizaciju. Smernice za normalizaciju, zajedničke za srpski i hrvatski jezik, nastale su prilagođavanjem uputstva za slovenački (v. Čibej, Fišer i Erjavec, 2016), pri čemu su uzete u obzir razlike između ortografskih i gramatičkih sistema ovih jezika. Smernice su dostupne na stranici korpusa u repozitorijumu CLARIN.SI.[12]

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

11 Na primer, upitom „fenomenalno" mogu se dobiti sve varijante zapisa reči – *fenomenalno*, *fnmln*, *fenom*, itd.

12 https://www.clarin.si/repository/xmlui/handle/11356/1171

Normalizacija je rađena na nivou reči, što znači da red reči i druge pojave vezane za sintaksu (npr. elipsa) nisu obuhvaćeni. Od normalizacije su izuzeti i emotogrami, korisnička imena, tarabe, kao i leksičke pojave poput upotrebe nestandardne reči umesto standardne (npr. kolokvijalno *fotka* nije normalizovano).[13] Sa druge strane, proces normalizacije obuhvatio je sve nestandardne zapise, bilo da se radilo o očiglednim greškama u kucanju, izostanku dijakritika ili nekom drugom vidu izmenjenog pisanja. U tabeli 2 navedene su pojave koje su normalizovane (sa primerima).[14]

Pojava	**Primer (nestandardni > standardni oblik)**
Ispušteni dijakritički znakovi	*dzabe > džabe, predjasnjeg > pređašnjeg, noc > noć, oci > oči*
Zapisi domaćih reči stranim slovima ili kombinacijama slova	*fax > faks, teshko > teško, chak > čak, commarac > komarac, devojq > devojku, Ezweene > Izvini*
Greške u kucanju, elidirani vokali, fonetska prilagođavanja, regionalne varijante, šatrovački, drugi vidovi nestandarndnog zapisa	*ughapsili > uhapsili, al > ali, jel > je li, Udrogiro > Udrogirao, pogin'o > poginuo, jope > opet, PreCednika > PreDSednika, nailaksi > najlakši, povatao > pohvatao, nemo'> nemoj, Kaće > Kad će, ljankase > seljanka, ladan > hladan, dvajs > dvadeset, ojra > evra, vidla > videla*
Nestandardni flektivni nastavci	*obrisat > obrisan, bi > bih/bismo/biste, reko > rekoh, dnojeva > dna, hejtao > hejtovao, mislEla > mislIla*
Izostale glasovne promene	*krugiće > kružiće, odpušač > otpušač, Predhodnih > Prethodnih, konbinacija > kombinacija*
Ponavljanja slova u punoznačnim rečima	*riboo > ribo, kiseooooniiiiik > kiseonik, zelimm > želim*
Ponavljanja slova i slogova u uzvicima	*hahaha > haha, Jaooo > Jao, UUUUUHHHH > UH*

13 Razlog za izostanak normalizacije leksike leži pre svega u složenosti zadatka. Granice između standardne i nestandardne leksike mnogo je teže povući nego granice između standardnog i nestandardnog zapisa reči, a vrlo se često nameće i problem izbora standardne lekseme na koju će nestandardna reč biti normalizovana – na primer, kolokvijalno *fotka* moglo bi biti normalizovano na *fotografija* ili *slika*. Detaljnije o temi nestandardne leksike piše Durbaba (2015).

14 Navedene su pojave obuhvaćene smernicama za anotatore. Tokom anotacije beležene su i rešavane i dodatne situacije, poput brisanja crtica upotrebljenih za naglašavanje (*o-ga-vno > ogavno*).

Zapisi skraćeni upotrebom brojeva	*o5 > opet*
Reči pogrešno napisane sastavljeno/rastavljeno	*ustvari > u stvari, nebi > ne bismo, ubicuse > ubiću se, ni jedan > nijedan*
Nestandardni zapis flektivnih oblika skraćenica	*smsa > sms-a, RTSu > RTS-u, BBCiju > BBC-ju*
Standardne skraćenice sa izostavljenom ili suvišnom tačkom	*god > god., dr. Stefanovica > dr Stefanovića, ps > p.s.*
Nestandardne skraćenice i akronimi (sa izuzetkom vlastitih imena)	*mng > mnogo, nzm > ne znam, osn. > osnovnim, NG > Novu godinu, teveu > TV-u, nnc > nema na čemu*

Tabela 2: Pojave koje su normalizovane prema smernicama za normalizaciju.

Pravila su se odnosila i na izvorno napisane strane reči (*twit > tweet*, *hipstaaa > hipster, lolololol > lol*), dok fonetski transkribovane strane reči nisu normalizovane na izvorni zapis (npr. zadržani su oblici *fejv*, *menšn*, *on d edž*). U slučajevima postojanja varijantnih oblika, oba oblika tretirana su kao standardna (npr. *ujutro* i *ujutru*).

Anotatorima je naglašeno da treba da uzmu u obzir i kontekst, posebno kod dvosmislenih slučajeva (npr. *sto* bi moglo zaista biti imenica ili broj *sto*, ali i *što*, a značenje i flektivni oblik skraćene reči nije uvek moguće odrediti samostalno – *več. škole > večernje škole*; *u osn. školama > u osnovnim školama*), kao i da ne normalizuju ukoliko dvosmislenost nije moguće razrešiti. U pogledu odnosa prema normi, glavni zahtev iskazan u smernicama bio je „U celom tvitu treba proveriti da li su pojedinačne reči u skladu sa standardnim jezikom, a u slučaju da odstupaju od standarda, treba im pripisati normalizovane verzije", ali anotatori (po obrazovanju lingvisti) nisu upućeni na unapred određenu normativnu literaturu, već im je ostavljena sloboda da se oslanjaju na intuiciju i rešavaju problematične slučajeve u međusobnoj i saradnji sa autorima smernica, konsultujući literaturu po potrebi.

Primeri pokazuju da nije bilo intervencija vezanih za upotrebu velikih i malih slova. Ova odluka, iako u suprotnosti sa pravopisnim pravilima, doneta je zbog toga što bi ovakve intervencije značajno usložnile normalizaciju i otežale treniranje automatskih alata, a bez dobiti za njihovu uspešnost. Iz primera se takođe uočava da, iako je vršena na nivou reči, normalizacija nije nužno bila 1:1. Tačnije, dešavalo se da se jedan nestandardni token deli na više standardnih (*ustvari > u stvari*), ili obratno (*ni jedno > nijedno*). Najzad, u tabeli 2 treba uočiti razliku između skraćenica koje imaju standardni, ili barem ustaljeni oblik (npr. *god.* za *godina*), i različitih idiosinkratičnih

skraćivanja (poput *mng* za *mnogo*). U procesu normalizacije, prvi tip skraćenica nije proširivan u punu reč, dok nestandardni skraćeni oblici jesu. Pored toga, skraćenice su označene kao posebna vrsta reči, dok su nestandardnim skraćenim oblicima dodeljivane oznake reči na koje su normalizovani (npr. *msm,* normalizovano n*a mislim,* označeno je kao glagol).

3.3 Metod analize

U tabeli 3 dat je primer jednog anotiranog tvita. Podaci su u vertikalnom formatu u kome svaka kolona predstavlja jedan nivo anotacije. Za tokene koji odgovaraju standardnom jeziku vrednosti u kolonama „Izvorni token" i „Normalizovani token" se poklapaju, dok se u slučaju normalizovanja u drugoj koloni nalazi standardizovani oblik.

Izvorni token	**Normalizovani token**	**Lema (osnovni oblik)**	**Morfosintaksička oznaka**
Gde	gde	gde	Rgp
tacno	tačno	tačno	Rgp
ta	ta	taj	Pd-fsn
frustracija	frustracija	frustracija	Ncfsn
lezi	leži	ležati	Vmr3s
,	,	,	Z
who	who	who	Xf
cares	cares	cares	Xf
.	.	.	Z

Tabela 3: Primer tvita u vertikalnom formatu sa pridruženom anotacijom.

Budući da se smernice za normalizaciju velikim delom zasnivaju na pojavama i kategorijama koje je teško ili čak nemoguće automatski prepoznati (npr. glasovne promene ili skraćenice), u analizi je bilo neophodno oslanjanje na drugačije kriterijume. Centralni pojam stoga je **transformacija**, odnosno izmena na nivou pojedinačnog karaktera koju je moguće identifikovati automatski. Polazišna tačka u analizi transformacija jeste standardni jezik. Drugim rečima, podrazumeva se da do transformacija dolazi prilikom prelaska iz standardnog u nestandardni oblik, čime transformisanje postaje proces suprotan normalizaciji opisanoj u prethodnom odeljku. Na primer, u tabeli 2 navodi se da je oblik *dzabe* normalizovan u

džabe, dok se u analizi isti primer tretira kao transformaciju standardnog *džabe* u nestandardno *dzabe* zamenom karaktera (*ž* > *z*).

U prvom koraku analize sve izvorne tokene poredimo sa odgovarajućim normalizovanim tokenima. Za tokene kod kojih je utvrđeno nepoklapanje prve dve kolone koristimo i morfosintaksičke opise i leme dodeljene normalizovanim tokenima (treća i četvrta kolona u tabeli 3), odnosno računamo raspodelu učestalosti transformacija prema vrstama reči, izdvajamo najčešće transformisane leme i najčešće transformisane oblike. Pored toga, transformisane oblike klasifikujemo prema Levenštajnovim transformacijama (brisanje, dodavanje, zamene; Levenshtein 1966), izdvajamo najčešća brisanja, dodavanja i zamene, i proučavamo položaj različitih tipova transformacija unutar reči.

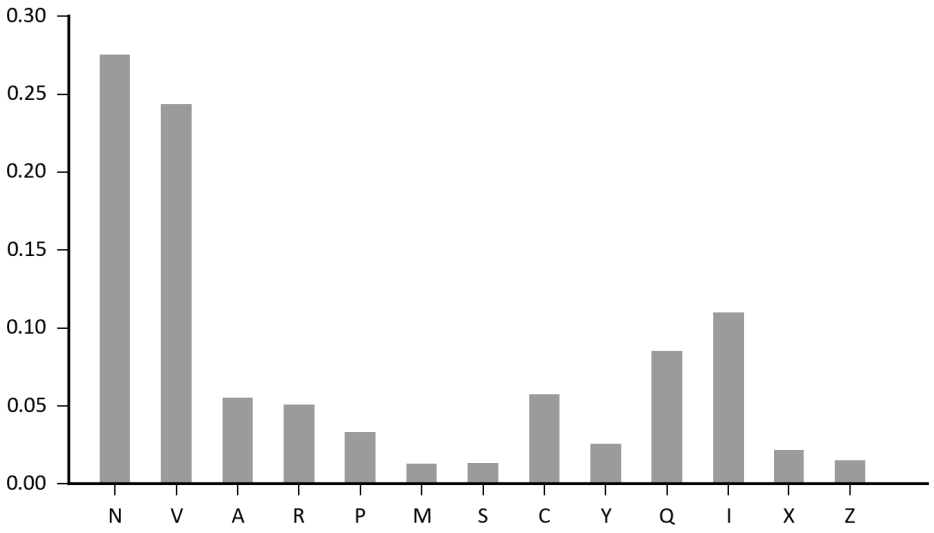
4. Rezultati

Do nekog vida transformacije došlo je u 10,32%, odnosno 4.679 tokena. Procenat transformacija 1:n iznosio je 0.39%, dok je za n:1 procenat bio 0.07%. Međutim, kod velikog broja tokena transformacije se sastoje isključivo u izostavljanju dijakritika (*č, ć, š, ž, đ* > *c, c, s, z, dj*), do koga dolazi u većoj meri iz tehničkih nego iz lingvističkih razloga (jedan mogući razlog je česta upotreba engleske tastature, posebno na manjim uređajima). Ukoliko se izuzmu ovakvi tokeni, preostaje 3,96%, odnosno 1.793 transformisanih tokena. Sa izuzetkom pregleda učestalosti različitih tipova transformacija, u daljoj analizi uzimamo u obzir samo tokene kod kojih postoji i neka dodatna vrsta transformacije (pri čemu se kod takvih tokena analiziraju i izostavljanja dijakritika).

4.1 Transformacije prema vrstama reči

Relativna učestalost transformacija prema vrstama reči (procenat ukupnog broja transformacija koji otpada na imenice, glagole, prideve, itd.) prikazana je na slici 1. Lako se može uočiti da se više od jedne polovine transformacija tiče imenica i glagola. Za njima slede uzvici i rečce, sa znatno manjim udelom od približno 10% transformacija, dok je udeo drugih vrsta reči još niži i uglavnom ne prelazi 5%. Drugim rečima, veći deo transformacija dešava se na otvorenim nego na zatvorenima klasama reči.[15]

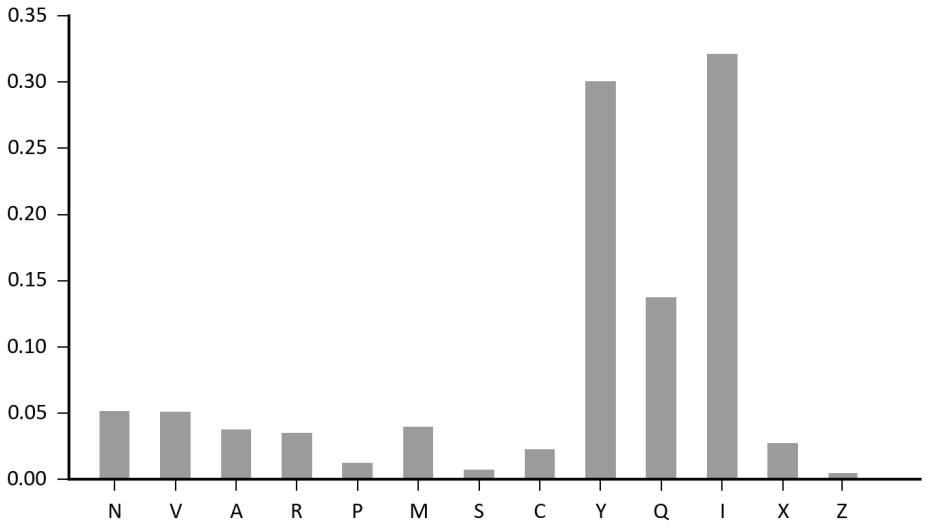
15 Ne uzimamo u obzir broj transformacija unutar pojedinačnih tokena.



Slika 1: Raspodela transformisanih tokena prema vrstama reči (N – imenica, V – glagol, A – pridev, R – prilog, P – zamenica, M – broj, S – predlog, C – veznik, Y – skraćenica, Q – rečca, I – uzvik, X – ostalo, Z – interpunkcija).

Nešto drugačiji rezultati vezani za vrste reči prikazani su na slici 2, gde se može videti procenat tokena koji su transformisani unutar svake od vrsta reči (na primer, od svih imenica transformisano je približno 5%). Najveći udeo transformisanih oblika javlja se kod uzvika i skraćenica (gde se transformacije dešavaju kod približno jedne trećine tokena), za njima dolaze rečce, dok je kod svih ostalih vrsta reči udeo transformacija vrlo nizak. Tendencija koja se ovde uočava suprotna je onoj sa slike 1 – više transformacija nalazi se kod zatvorenih nego kod otvorenih klasa reči.[16]

16 Podsećamo da su kao skraćenice označena samo standardna skraćivanja poput *godina* > *god.*, za koja se može smatrati da čine zatvorenu klasu.



Slika 2: Udeo transformisanih oblika u ukupnom broju tokena koji pripadaju različitim vrstama reči.

Ukupno gledano, otvorene klase reči sadrže većinu transformacija u prvoj raspodeli, a zatvorene klase u drugoj. Do ove razlike dolazi prvenstveno zbog toga što je za prvu analizu značajna i raspodela učestalosti različitih vrsta reči u uzorku (bez obzira na transformacije) – imenice i glagoli predstavljaju najčešće vrste reči, koje zajedno čine približno 40% uzorka (imenice 21,03%, glagoli 18,92%), dok uzvici, skraćenice i rečce u zbiru ne dostižu ni 5% (uzvici 1,35%, skraćenice 0,34%, rečce 2,45%). Drugim rečima, iako su reči iz zatvorenih klasa podložnije transformacijama, njihova ukupna učestalost je suviše niska da bi se to videlo u udelu u zbiru svih transformacija.

Ovde, međutim, treba istaći i određene ograde, budući da se neke od zatvorenih klasa koje su uključene u našu analizu u lingvistici zapravo ne tretiraju kao posebne vrste reče – takve su skraćenice, interpunkcija i kategorija „ostalo" (u kojoj se nalaze, na primer, strane reči). Uzvici se takođe mogu smatrati posebnim slučajem, budući da predstavljaju vrstu reči koja je tradicionalno uključena u podele, ali su u našim podacima transformisani na drugačiji način i često sa drugačijim ciljem nego druge vrste reči – uglavnom su podložni ponavljanju pojedinačnih ili grupa karaktera, što je pragmatička i paralingvistička pojava vezana za naglašavanje, a ne za fonetiku ili neki drugi nivo unutar jezičkog sistema u užem smislu.

4.2 Najčešće transformisane leme i površinski oblici

Leme kod kojih je zabeležena najviša učestalost transformacija prikazane su u tabeli 4, dok je dvadeset najčešće transformisanih površinskih oblika navedeno u tabeli 5. Za svaku lemu navedeno je koji procenat unutar svih transformisanih oblika pokrivaju njene transformacije (% transformacija), na čemu je zasnovano rangiranje u tabeli, kao i procenat tokena koji pripadaju toj lemi koji su transformisani (% leme); ova komplementarna „dvostruka" analiza slična je onoj iz odeljka 4.1. Uz leme je naveden i podatak o vrsti reči (npr. #V za glagol). U tabeli 5 je nestandardni oblik naveden u zagradi, a prikazan je i procenat transformisanih oblika koji odlazi na svaku od navedenih transformacija.

	Lema	**% transformacija**	**% leme**		**Lema**	**% transformacija**	**% leme**
1.	biti#V	7.53%	6.12%	11.	da#C	0.84%	1.07%
2.	li#Q	6.53%	61.26%	12.	jebiga#I	0.84%	83.33%
3.	haha#I	2.90%	81.25%	13.	moći#V	0.78%	8.19%
4.	hajde#I	2.84%	92.73%	14.	min.#Y	0.78%	77.78%
5.	hteti#V	2.01%	9.78%	15.	ja#P	0.73%	1.35%
6.	ali#C	1.73%	19.38%	16.	u#S	0.67%	1.36%
7.	kao#C	1.51%	14.21%	17.	.#Z	0.61%	0.62%
8.	jebati#V	1.45%	27.08%	18.	?#Z	0.61%	3.3%
9.	ne#Q	1.34%	4.86%	19.	ili#C	0.56%	8.85%
10.	jebote#I	1.23%	68.75%	20.	odmah#R	0.56%	50.0%

Tabela 4: 20 najčešće transformisanih lema.

	Oblik	**% transformacija**		**Oblik**	**% transformacija**
1.	je li (jel)	3.99%	11.	bismo (bi)	0.62%
2.	li (l')	1.81%	12.	hajde (ae)	0.62%
3.	ali (al)	1.56%	13.	haha (hahaha)	0.56%
4.	hajde (aj)	1.50%	14.	odmah (odma)	0.50%
5.	jebote (jbt)	1.31%	15.	haha (hahah)	0.44%
6.	jebiga (jbg)	0.87%	16.	bih (bi)	0.44%
7.	min. (min)	0.87%	17.	ili (il)	0.44%
8.	kao (k'o)	0.81%	18.	jebao (jebo)	0.44%
9.	kao (ko)	0.78%	19.	u stvari (ustvari)	0.44%
10.	hajde (ajde)	0.75%	20.	li (l)	0.37%

Tabela 5: 20 najčešće transformisanih površinskih oblika.

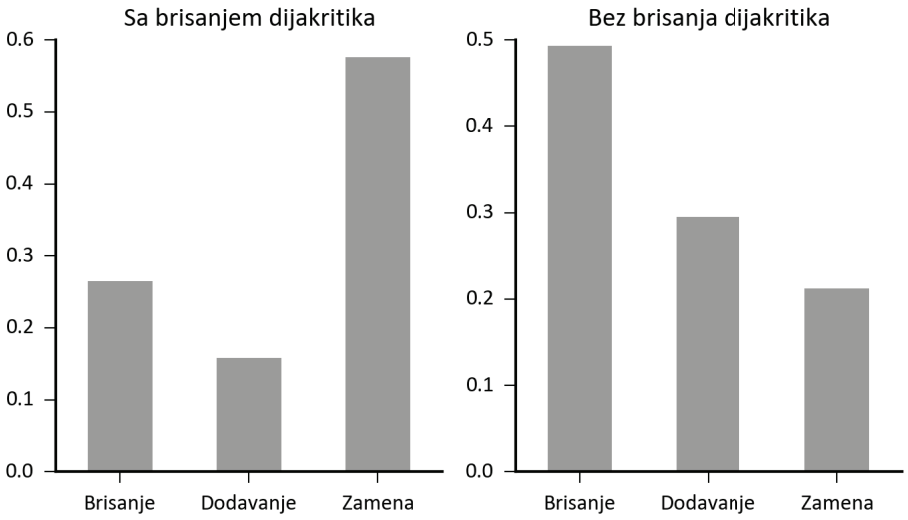
Prvo mesto na listi lema u tabeli 4 zauzima pomoćni glagol *biti* (zajedno sa *jesam*), kod koga se najveći broj transformacija tiče oblika aorista *bih* i *bismo*, koji se u našem uzorku vrlo često skraćuju na *bi* (v. tabelu 5). Za glagolom *biti* slede rečca *li* (često skraćena na *l* ili *l'*) i uzvici *haha* (sa čestim ponavljanjem slova ili slogova, posebno u obliku *hahaha* i *hahah*) i *hajde* (skraćeno na *ajde* i *ae*). Za glagol *hteti*, visoko rangiran u tabeli 4, interesantno je da nijedan njegov pojedinačni oblik nije zastupljen u tabeli 5; uvid u podatke potvrđuje intuiciju da se oblici ovog glagola transformišu izostavljanjem ili zamenom početnog *h* – *oću*, *'oću*, *oćeš*, i sl. *Kao* se često koristi u oblicima *k'o* i *ko*, *ali* i *ili* se skraćuju na *al* i *il*, a *ne* je uglavnom pogrešno napisano sastavljeno ili rastavljeno (i izostaje iz tabele 5). Za znake interpunkcije uočavaju se različiti razlozi čestih transformacija – tačka se uglavnom transformiše izostavljanjem u skraćenicama (up. lemu/oblik *min.*), dok se uzvičnik često ponavlja više puta.

Iako je naše polazište i cilj opis nestandardnog zapisivanja, gornje tabele daju uvid i u leksiku. Naime, vrlo naglašeno prisustvo u obe tabele imaju vulgarizmi, pre svega glagol *jebati*, koji se u uzorku posebno često javlja u obliku *jebo*. Na listama su se uz njega našle još dve povezane leme označene kao uzvici, *jebote* i *jebiga*, usled čestog skraćivanja na *jbt* i *jbg*.

Ukupno se u ovim analizama ponovo uočava dominantno prisustvo zatvorenih klasa reči. Posebno je zanimljivo primetiti da se na listama ne nalazi nijedna imenica, dok su od glagola prisutni pomoćni i modalni, plus svega jedan leksički (i to vulgarizam).

4.3 Analiza transformacija prema tipu

Dalju pažnju posvećujemo raspodeli tri tipa transformacija koje je definisao Levenštajn (Levenshtein, 1966), brisanja, dodavanja i zamene. Prvi deo rezultata prikazan je na slici 3, gde se poredi raspodela ova tri tipa transformacija u slučaju kada se uzimaju u obzir i izostavljanja dijakritika i u slučaju kada se ona zanemaruju. Grafikoni jasno pokazuju da su zamene daleko najčešća vrsta transformacija ukoliko se uzmu u oblik izostavljanja dijakritika, odnosno zamena slova sa dijakriticima slovima bez njih. Međutim, kada se izuzmu tokeni koji sadrže samo ove transformacije, najčešći fenomen postaje brisanje, a zamena postaje najređa.



Slika 3: Raspodela različitih tipova transformacija, sa brisanjem dijakritika (slika levo) i bez brisanja dijakritika (desno).

Zatim, u tabeli 6 je za svaki tip prikazano 10 najčešćih transformacija, uz primer koji ga ilustruje. Prilikom tumačenja rezultata iz ove tabele treba imati na umu da se u analizi transformacija uzimaju u obzir pojedinačni karakteri (tako da se digrami poput *lj* tretiraju kao dva odvojena slova, kao i strane slovne kombinacije upotrebljene umesto domaćih slova). Neizbežna posledica ovakvog pristupa jeste ta da se kod nekih tokena dobija niz transformacija čija priroda je više tehnička nego lingvistička. Na primer, prelazak iz *šiša* u *shisha* opisan je transformacijama „insert_s|replace_š-h|insert_s|replace_š-h", koje u manjoj meri odražavaju jezičke pravilnosti, a u većoj tehničke odluke.[17] Za lingvistička objašnjenja zato je u ovom slučaju često potrebna dodatna kvalitativna analiza.

17 Posebna pravila formulisana su za dve situacije – transformacije *đ* > *dj* i *ks* > *x* tretiraju se kao zamene umesto kao zamena i dodavanje ili brisanje.

Brisanje			Dodavanje			Zamena		
i	13.62%	ali > al	a	22.51%	jao > jaao	i-’	7.49%	ali > al’
e	10.95%	se > s	h	12.63%	hehe > heheheh	a-’	5.05%	ostao > ost’o
a	10.67%	kao > ko	e	11.59%	je > jee	ks-x	3.06%	faks > fax
	10.33%	je li > jel	.	9.97%	… > ……	i-e	2.45%	zaspi > zaspe
h	5.96%	odmah > odma	o	6.36%	Alo > Aloo	š-h	2.29%	šiša > shisha
o	5.90%	ovako > vako	i	5.03%	ima > iiima	h-’	2.14%	hoće > ’oće
d	4.03%	kad će > kaće		3.89%	trebaće > treba će	e-i	2.14%	mirišem > mirišim
j	3.97%	mi je > mie	!	3.61%	!!! > !!!!	a-e	1.99%	šerovao > šeroveo
u	3.58%	umrem > mrem	u	3.04%	juhu > juhuuuu	h-x	1.83%	hehe > xexe
-	3.46%	sms-a > smsa	?	2.85%	?! > ??!!	r-v	1.53%	smrde > smvde

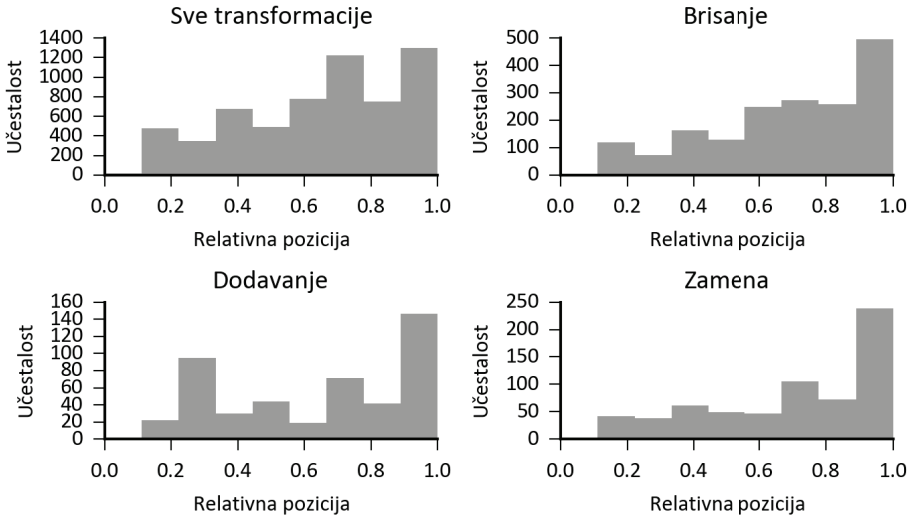
Tabela 6: 10 najčešćih transformacija prema tipu.

Među brisanjima dominira brisanje vokala i razmaka. Najčešće se ispuštaju *i* (*je li* > *jel*), *e* (u skraćenicama poput *aj* za *ajde*, ili *jbg* za *jebiga*) i *a* (u skraćenim oblicima kakvi su *ko* za *kao* ili *reko* za *rekao*). Tendencija ka brisanju vokala, i to posebno navedena tri, vrlo je očekivana, budući da se oni najčešće izostavljaju i u govoru. Međutim, primećuje se da nijedan pojedinačni vokal nije izrazito dominantan (up. u Miličević i dr., 2017 rezultate za slovenački i hrvatski, gde se *i* briše znatno češće od drugih vokala). Razmak izostaje u rečima pogrešno napisanim zajedno (*jel* < *je li*, *ustvari* < *u stvari*).

Dodavanje se uglavnom sreće u uzvicima i interpunkciji, u vidu ekspresivnog ponavljanja slogova (npr. *hahahahaha*) ili karaktera (*jaao*, *juhuuuu*), dok se druga najčešća situacija tiče reči pogrešno napisanih odvojeno (*treba će* umesto *trebaće*). Dodavanje je, iz tehničkih razloga, takođe prisutno kod idosinkratičnih načina pisanja domaćih reči (*shisha* < *šiša*, *vishe* < *više*), i proširenja akronima (*esemes* < *sms*). Zamene su u većini slučajeva proizvod upotrebe apostrofa umesto ispuštenog karaktera (*je l’* < *je li*, ili *ost’o* < *ostao*), što suštinski ponovo ukazuje na lingvističko (iako ne i tehničko) skraćivanje i bliskost sa govornim jezikom.

4.4 Analiza prema poziciji transformacija

U poslednjem koraku analize fokus je na relativnom položaju transformacija u reči. Rezultati relevantni za ovu analizu prikazani su na slici 4, gde prvi deo grafikona prikazuje raspodelu učestalosti svih transformacija zajedno, dok su preostala tri posvećena pojedinačnim tipovima. Relativna pozicija u reči iskazana je procentualno (npr. vrednost od 0,5 predstavlja sredinu reči).



Slika 4: Raspodela transformacija prema poziciji.

Sa slike se vidi da se transformacije uglavnom javljaju na kraju reči, iako ova tendencija nije preterano naglašena. Brisanje odlikuje najujednačenija raspodela, sa blagom tendencijom ka drugoj polovini, a posebno samom kraju reči. Slabost tendencije po svojoj prilici je posledica toga što je u srpskom jeziku na Tviteru često skraćivanje koje uključuje brisanje karaktera na različitim pozicijama (up. *jbg* < *jebiga*, *nzm* < *ne znam*), kao i toga što se često koriste oblici iz govornog jezika u kojima izostaje početno *h* (*ladan*, *oću*, *ajde*).[18]

Dodavanje i zamena karaktera naglašenije su prisutni na kraju reči. Kod dodavanja situacija je takva pre svega usled ekspresivnih dodavanja u uzvicima i interpunkciji, donekle i leksičkim rečima. Orijentisanost zamena ka kraju reči uglavnom je vezana za zamenu izostavljenih finalnih karaktera apostrofom.

18 Miličević i dr. (2017) pokazuju da je koncentracija obrisanih finalnih karaktera viša u hrvatskom i slovenačkom jeziku, između ostalog zbog čestog izostavljanja finalnog *-i* u infinitivima glagola.

5. Diskusija i zaključak

U radu je predstavljena analiza normalizovanog uzorka tvitova na srpskom jeziku, sa naglaskom na transformacijama koje su dovele do formiranja pronađenih nestandardnih elemenata. Prva pojava koja je uočena jeste vrlo naglašena sklonost ka izostavljanju dijakritičkih znakova, koje je zbog svoje pretežno tehničke prirode (upotrebe engleske tastature i/ili manje preglednih tastatura na prenosnim uređajima) isključeno iz dalje analize. Pregled učestalosti transformacija prema vrstama reči pokazao je da se vrlo visok procenat transformacija događa kod otvorenih klasa reči, pre svega imenica i glagola. Međutim, komplementarna analiza zasnovana na utvrđivanju učestalosti transformacija unutar vrsta reči pokazala je upravo suprotno – viši stepen transformisanosti zatvorenih klasa, posebno uzvika, skraćenica i rečci. Ovu drugu tendenciju potvrđuje i analiza prema lemama i površinskim oblicima, gde se vidi da su najčešće transformacije one na pomoćnim glagolima, rečcama, uzvicima i veznicima. Drugim rečima, visoka učestalost imenica i glagola u uzorku dovodi do njihovog velikog udela u ukupnom broju transformacija, ali zapravo se u unutar ovih vrsta reči transformacije dešavaju ređe i na većem broju različitih tokena (usled čega ne dospevaju među najčešće pojedinačne transformacije) nego kod zatvorenih klasa. Ovakav rezultat sasvim je očekivan kada se uzme u obzir relativno mali broj pripadnika zatvorenih klasa, njihova relativna kratkoća i velika predvidivost – poznato je da je reči iz zatvorenih klasa lakše predvideti iz rečeničnog konteksta i da su one stoga podložnije skraćivanju i u govoru (v. npr. rad Bell, Brenier, Gregory, Girand i Jurafsky, 2009 o engleskom jeziku).

Analizom Levenštajnovih transformacija utvrđeno je da je (ne računajući izostavljanja dijakritičkih znakova) najčešći tip transformacije brisanje. Ovakav rezultat je takođe očekivan, ne samo prema opštem principu jezičke ekonomije, već i usled interaktivnog i neformalnog komunikacijskog okruženja (iako se Tviter koristi i za formalnu komunikaciju, naš uzorak sadržao je isključivo tvitove sa nestandardnim odlikama, kakve se sreću pre svega u neformalnom komuniciranju), kao i upotrebe prenosnih uređaja na kojima je kucanje otežano. Brisanje uglavnom pogađa vokale i slovo/glas *h*, slično govornom jeziku. Dodavanje je najčešće kod uzvika i interpunkcije, gde ima ulogu naglašavanja. Nešto drugačiji slučajevi od prethodnih jesu oni bliže vezani za pravopisnu normu – pogrešno brisanje i dodavanje razmaka između reči ili u njih. Najzad, situacija je najraznovrsni-

ja kod zamene, do koje dolazi usled umetanja apostrofa na mesto ispuštenih slova, ali i iz različitih drugih razloga i u vrlo različitim rečima. Sve transformacije se nešto češće javljaju na kraju reči nego na početku.

Kada se osvrnemo i uporedimo naše rezultate sa istraživanjima prikazanim u odeljku 2, možemo uočiti da naša analiza potvrđuje prisustvo većine pojava iz tabele 1. Posle izostavljanja dijakritičkih znakova, daleko najčešće se javljaju različiti vidovi skraćivanja (sa izuzetkom upotrebe brojeva, koja gotovo da se ne sreće). Naglašavanje ponavljanjem sledeće je po učestalosti, a za njim sledi pogrešno sastavljeno i rastavljeno pisanje reči. Pojave vezane za upotrebu malih i velikih slova i izostavljanje interpunkcije nisu bile obuhvaćene našom analizom, tako da za njih ne možemo izvesti nikakve zaključke.

Ono što se dodatno može uočiti u našim rezultatima, čak i bez detaljnije leksičke analize, jeste vrlo naglašeno prisustvo vulgarizama. Jedan od mogućih pravaca daljih istraživanja svakako je detaljnije bavljenje ovim fenomenom, kao i uključivanje u istraživanje sociodemografskih varijabli poput uzrasta korisnika. Sa druge strane, naše istraživanje jedno je od retkih koja u centar pažnje stavljaju učestalost različitih nestandardnih oblika u KPR, i premda je ono zbog nedostatka prethodnih sličnih podataka pretežno deskriptivne i uvodne prirode, očekujemo da će iz prikazanog opisa proisteći brojne uže lingvističke hipoteze koje će biti testirane na istim ili novim podacima.

Na samom kraju, važno je osvrnuti se na pitanje norme i njenog mogućeg trajnijeg narušavanja kroz komunikaciju posredovanu računarom. Iz naših rezultata može se zaključiti da odstupanja od norme u KPR svakako postoje. Međutim, podsećamo da naš uzorak čine isključivo nestandardni tvitovi i da su čak i u njima nestandardne odlike prisutne u približno jednoj desetini (10,32%) tokena računajući izostavljanja dijakritičkih znakova, odnosno u svega 3,96% tokena ukoliko se ova izostavljanja izuzmu. Kada se tome doda informacija da je procenat tvitova sa nestandardnim odlikama za srpski jezik procenjen na 33% (v. Fišer i dr., 2105), jasno je da su odstupanja zapravo retka. Ako uzmemo u obzir i funkcionalnu specijalizovanost nestandardnih oblika, odnosno njihovu specifičnu ulogu u skraćivanju poruke, iskazivanju paralingvističkih informacija i gradnji identiteta korisnika KPR, čini se da nema previše osnova za pretpostavku da će se ovi oblici proširiti na kontekste u kojima ove funkcije nije potrebno ili nije poželjno ostvariti, nestandardnim sredstvima ili uopšte. Drugim rečima, sma-

tramo da nestandardne elemente u jeziku ne treba po svaku cenu osuđivati i proterivati, već ih pre svega proučavati. U skladu sa tim, rad možemo zaključiti primenom na srpski jezik citata sa bloga koji govori o demokratizaciji engleskog jezika: „Breaking news! That thud you hear in the background isn't the sound of standards falling. It's the sound of language remaining fit for purpose" (Najnovija vest! Buka koju čujete ne potiče od rušenja standarda. To je zvuk jezika koji se prilagođava funkcijama koje obavlja).[19]

Literatura

Bell, A., Brenier, J. M. , Gregory, M., Girand, C. i Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language* 60, 92–111.

Čibej, J., Fišer, D. i Erjavec, T. (2016). Normalisation, tokenisation and sentence segmentation of Slovene tweets. U U. Andrius, J. Vaičenonienė i R. Butkienė (ur.), *Proceedings of Normalisation and Analysis of Social Media Texts (NormSoMe)* (str. 5–10). http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-NormSoMe_Proceedings.pdf

Durbaba, O. (2015). O nestandardnoj leksici u leksikografskim izvorima. *Anali Filološkog fakulteta* 27/2, 211–223.

Eckart de Castilho, R., Biemann, C., Gurevych, I. i Yimam, S. M. (2014). WebAnno: a flexible, web-based annotation tool for CLARIN. *Proceedings of the CLARIN Annual Conference (CAC) 2014*, Soesterberg, Netherlands. https://www.clarin.eu/sites/default/files/cac2014_submission_6_0.pdf

Eisenstein, J. (2013). What to do about bad language on the Internet. *Proceedings of HLT-NAACL 2013* (str. 359–369). http://www.cc.gatech.edu/~jeisenst/papers/naacl2013-badlanguage.pdf

Filipan-Žignić, B. (2012). *O jeziku novih medija. Kvare li novi mediji suvremeni jezik?* Split: Matica hrvatska.

Filipan-Žignić, B. i Turk Sakač, M. (2016). Utjecaj novih medija na jezik mladih u pisanim radovima. *Slavistična revija* 4, 463–474.

Filipan-Žignić, B., Legac, V., Pahić, T. i Sobo, K. (2015). New literacy of young people caused by the use of new media. *Procedia – Social and Behavioral Journal* 192, 172–179.

Filipan-Žignić, B., Sobo, K. i Velički, D. (2012). SMS communication – Croatian SMS language features as compared with those in German and English speaking countries. *Revija za elementarno izobraževanje* 5, 5–22.

19 http://blog.sfep.org.uk/internet-democratisation-english-part-1-power-people/

Fišer, D., Erjavec, T., Ljubešić, N. i Miličević, M. (2015). Comparing the non-standard language of Slovene, Croatian and Serbian tweets. U M. Smolej (ur.), *Simpozij Obdobja 34. Slovnica in slovar - aktualni jezikovni opis (1. del)* (str. 225–231). Ljubljana: Filozofska fakulteta.

Jelić, G. i Polovina, V. (2015). Samoispravka i ispravka u jeziku kratkih poruka. *Anali Filološkog fakulteta* 27/2, 419–436.

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

Ljubešić, N., Erjavec, T. i Fišer, D. (2017). Adapting a state-of-the-art tagger for South Slavic languages to non-standard text. U *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing* (str. 60-68). Valencia: Association for Computational Linguistics.

Ljubešić, N., Fišer, D. i Erjavec, T. (2014). TweetCaT: a tool for building Twitter corpora of smaller languages. U N. Calzolari i dr. (ur.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (str. 2279–2283). http://www.lrec-conf.org/proceedings/lrec2014/pdf/834_Paper.pdf

Miličević, M. i Ljubešić, N. (2016). Tviterasi, tviteraši or twitteraši? Producing and analysing a normalised dataset of Croatian and Serbian tweets. *Slovenščina 2.0* 4/2, 156–188.

Miličević, M., Ljubešić, N. i Fišer, D. (2017). Birds of a feather don't quite tweet together: An analysis of spelling variation in Slovene, Croatian and Serbian Twitterese. U D. Fišer i M. Beißwenger (ur.), *Investigating Computer -Mediated Communication: Corpus-Based Approaches to Language in the Digital World* (str. 14–43). Ljubljana: University Press, Faculty of Arts.

Radić-Bojanić, B. (2007). *neko za chat?! Diskurs elektronskih ćaskaonica na engleskom i srpskom jeziku*. Novi Sad: Filozofski fakultet.

Stamenković, D. i Vlajković, I. (2012). Jezički identitet u komunikaciji na društvenim mrežama u Srbiji. U B. Mišić-Ilić i V. Lopičić (ur.), *Jezik, književnost, komunikacija: zbornik radova. Jezička istraživanja* (str. 212–224). Niš: Filozofski fakultet.

Tagg C., Baron A. i Rayson P. (2012). "i didn't spel that wrong did i. Oops": Analysis and normalisation of SMS spelling variation. *Lingvisticæ Investigationes* 35, 367–388.

Thurlow, C. i Brown, A. (2003). Generation Txt? The sociolinguistics of young people's text-messaging. *Discourse Analysis Online* 1(1). [https://extra.shu.ac.uk/daol/articles/v1/n1/a3/thurlow2002003-paper.html]

Van Dijk, C., van Witteloostuijn, M., Vasić, N., Avrutin, S. i Blom, E. (2016). The influence of texting language on grammar and executive functions in primary school children. *PLoS ONE* 11, e0152409. doi: 10.1371/journal.pone.0152409

Verheijen, L. (2013). The effects of text messaging and instant messaging on literacy. *English Studies* 94, 582–602.

Verheijen, L. i Spooren, W. (2017). The impact of WhatsApp on Dutch youths' school writing. U E. W. Stemle i C. R. Wigham (ur.), *Proceedings of the 5th Conference on CMC and Social Media Corpora for the Humanities (cmccorpora17)* (str. 65–69). Bolzano.

Vlajković, I. (2010). Uticaji engleskog jezika na srpski na planu pravopisa, leksike i gramatke u komunikaciji na Fejsbuku. *Komunikacija i kultura online* 1, 183–196.

Vrsaljko, S. i Ljubomir, T. (2013). Narušavanje pravopisne norme u ranojezičnoj neformalnoj komunikaciji (na primjeru SMS poruka i internetske društvene mreže Facebook). *Magistra Iadertina* 8/1, 155–163.

Maja Miličević Petrović
Nikola Ljubešić
Darja Fišer

Summary

NON-STANDARD SPELLING ON SERBIAN TWITTER: MUCH ADO ABOUT LITTLE DEVIATION?

This paper deals with non-standard spelling variants of Serbian words on Twitter. The analysis is based on a sample of automatically collected tweets identified as containing non-standard features. The sample was manually normalised, tagged with morhosyntactic descriptions and lemmatised. In the analysis, we compare non-standard forms to the standard ones they are related to and we establish the types of transformations that led to their formation. The results show that transformations are particularly frequent in closed-class words, such as auxiliary verbs, interjections and abbreviations. Character deletions are more common than character insertions and replacements. More transformations tend to occur at word end than in other positions. Despite the indisputable presence of non-standard forms on Serbian Twitter, we conclude that these forms are overall too infrequent and too functionally specialised to justify any claims about a negative influence of computer-mediated communication on the standard language norm.

Key words: computer-mediated communication, social networks, Twitter, corpus linguistics, non-standard language