Korpusni pristop k skladnji računalniško posredovane slovenščine

Špela Arhar Holdt

Izvleček

Poglavje predstavlja označevanje in analizo skladenjskih značilnosti računalniško posredovane slovenščine. Na ravni besednega reda najprej predstavimo pripravo korpusa Janes-Syn in prilagoditve označevalnega sistema specifikam računalniško posredovane slovenščine, nato pa preverimo, kako trije neodvisni označevalci razumejo besednoredno zaznamovanost v danem gradivu, kolikšno je ujemanje med njihovimi odločitvami, katere vrste besednorednih problemov se glede na pripisane oznake v podatkih pojavljajo in kako so ti problemi razporejeni glede na avtomatsko pripisano kategorijo jezikovne (ne)standardnosti. Raziskava pokaže, da besednoredno zaznamovanost označevalci opredeljujejo zelo različno, pogled na podatke z vidika pripisane (ne)standardnosti pa postavlja v razmerje koncepta zaznamovanosti in nestandardnosti, kar bi bilo v nadaljevanju smiselno raziskati z vključitvijo podatkov govorjenega jezika. Kategorizacija označenih besednorednih značilnosti osvetljuje segment skladnje računalniško posredovane slovenščine, ki je bil do sedaj neraziskan, in nakazuje naslednje korake za korpusnojezikoslovne raziskave besednega reda v slovenščini.

Ključne besede: računalniško posredovana slovenščina, korpusno jezikoslovje, skladenjsko označevanje, tviti, besedni red

1 UVOD

Uvodno zadržanost do jezikovnih realnosti, ki jih prinaša ali prvič v zgodovini širše javno izpostavlja računalniško posredovana komunikacija, v zadnjih letih nadomešča rastoč raziskovalni interes, tako na področju jezikoslovja kot tudi obdelave naravnih jezikov, podatkovnega rudarjenja in drugih disciplin, ki jim je v interesu opisati oz. uporabljati podatke sodobne jezikovne rabe. Študijam na večjih jezikih (npr. Crystal 2011; Myslin in Gries 2010; Storrer 2013; Chanier 2015) so se pridružile študije na slovenščini, večina kot rezultat projekta JANES.

Pregled literature ob začetku projektnih aktivnosti je razkril, da je skladnja slovenske računalniško posredovane komunikacije raziskovalno komajda dotaknjeno področje. Omeniti je mogoče nekatere pilotske (zdaj je že mogoče reči tudi pionirske) kvalitativne študije na avtentičnem gradivu različnih vrst, h katerim se vračamo v nadaljevanju prispevka: Kranjc (2003) analizira jezik spletnih klepetov, Dobrovoljc (2008) e-pošta sporočila, Jakop (2008) forumske zapise, Kalin Golob (2008) SMS-e, Michelizza (2015) članke Wikipedije in bloge. Zahtevnejši problem za načrtovanje raziskav je bilo dejstvo, da ob začetku projekta JANES virom navkljub tudi celovitejši korpusni opis značilnosti standardne slovenske skladnje še ni bil na voljo;[1] od obsežnejših korpusnih skladenjskih študij je mogoče izpostaviti monografijo Ledinek (2014), ki se pred analizo izbranega nabora glagolov in njihove vezljivosti ukvarja tudi z vprašanjem skladenjskega označevanja slovenščine, zasnovo FrameNeta za slovenščino (Može 2013) in razprave, vezane na pripravo Leksikalne baze za slovenščino (Gantar 2011; 2015). Odsotnost sintetičnih referenčnih informacij, ki so predpogoj za obsežnejše korpusnojezikoslovne primerjave in sistematično identifikacijo nestandardnih skladenjskih prvin v odnosu do standardnih,[2] je usmerila delo v razvoj metodologije za rabo obstoječih virov pri preučevanju specifičnih skladenjskih vprašanj, na drugi strani pa v pripravo novih virov, ki bodo celovitejši vpogled lahko omogočili v prihodnosti.

Prvi cilj projektne aktivnosti je bil tako izdelati, preizkusiti in evalvirati metodologijo, ki omogoča raziskave trendov rabe nestandardnih jezikovnih prvin za potrebe slovenske normativistike. Ker korpus računalniško posredovane

1 Vrzel naslavlja nacionalni projekt Nova slovnica sodobne standardne slovenščine: viri in metode (ARRS J6-8256, vodja Simon Krek), ki se pričenja v času priprave prispevka.

2 Računalniško posredovana slovenščina je zbirni pojem za različne načine komunikacije oz. raznovrstne besedilne vrste. Za slednje je mogoče reči, da *v splošnem* prinašajo opazen delež nestandardnih jezikovnih prvin, če slednje razumemo kot prvine, ki se razlikujejo od standar(dizira)nega dela jezika (Krek 2015), pogosto v smeri nenamernih ali namernih odstopov od obstoječih jezikovne norme. Ker se projekt JANES ciljno posveča nestandardnim prvinam v slovenščini (tudi z vidika identifikacije, kako »nestandardno« sploh opredeljevati, glej Stabej et al. 2016), se na slednje osredotočamo tudi v prispevku, pri čemer pa se je treba na vseh mestih zavedati, da je na ravni posameznih računalniško posredovanih besedil stopnja in vrsta nestandardnosti zelo različna.

slovenščine Janes (Erjavec et al. 2018) prinaša besedila, ki za razliko od večine gradiva v referenčnih korpusih večinoma niso jezikovno korigirana, realneje izkazuje tendence rabe oz. (ne)intuitivnost obstoječih jezikovnih pravil v širši jezikovni skupnosti. Metodološko premišljena primerjava podatkov korpusa Janes in referenčnega korpusa lahko razkrije tisti del jezikovnih sprememb, ki so širše oz. sistemske, in jih loči od redkejših, sporadičnih ali avtorsko/žanrsko vezanih odklonov od norme. Za preizkus metode smo izbrali zveze samostalnika z neujemalnim levim prilastkom (npr. *solo petje, RTV prispevek* proti *solopetje, RTV-prispevek*). Dosedanja slovenistična polemika o tem jezikovnem problemu se na eni strani dotika vprašanja, katere od tovrstnih zvez zapisovati skupaj in katere narazen, na drugi strani pa, kako v primeru zapisa narazen besednovrstno uvrščati prvi del besedne zveze. Študija se posveti tem vprašanjem, pri čemer je pomembno odkritje, da je praksa zapisovanja v korpusu Janes prepričljivo konsistentnejša od prakse zapisovanja v korpusu Kres, kar pomeni, da jezikovna regulacija obravnavanega problema krepi oz. povišuje variantnost v jezikovni rabi. To dejstvo je v nasprotju s pričakovanim ter odpira ključna vprašanja o namenu ter načinu lektoriranja v slovenskem prostoru, kot tudi stanju in vlogi jezikovnih priročnikov za slovenščino in obstoječih standardizacijskih teles ter praks. Opis metode in rezultati so bili predstavljeni v Arhar Holdt in Dobrovoljc (2015; 2016) ter diskutirani v Stabej et al. (2016).

V tem prispevku se osredotočamo na drugi korak projektnih aktivnosti, tj. pripravo skladenjsko označenega korpusa Janes-Syn, ki služi kot izhodišče za nadaljnje odvisnostno označevanje računalniško posredovane slovenščine, s fokusom na njenih nestandardnih značilnostih. Glavne označevalne odločitve so že bile opredeljene v kratkem prispevku (Arhar Holdt et al. 2016), ki ga na tem mestu nadgradimo z natančnejšo analizo podatkovnega seta in primeri gradiva. Nato predstavimo rezultate jezikoslovne analize skladenjskih značilnosti označenih podatkov s poudarkom na besednem redu. K vprašanju pristopimo z označevanjem zaznamovanega besednega reda, ki so ga neodvisno izvedli trije označevalci, oznake kategoriziramo ter prikažemo ujemanje med označevalci po kategorijah, v diskusijo pa pritegnemo tudi pregled identificiranih kategorij glede na njihovo pojavljanje v primerih, ki so bili v izhodiščnem korpusu avtomatsko označeni kot jezikovno standardni ali nestandardni. Prispevek zaključujemo s strnitvijo projektnih spoznanj in nalog za nadaljnje delo.

2 IZDELAVA IN OZNAČEVANJE KORPUSA JANES-SYN

Če so primerjave med korpusom Janes in korpusom Kres (ali Gigafida) dobro izhodišče za obravnavo posameznih skladenjskih problemov, zlasti če gre za primere, kjer je podatke mogoče pridobiti z uporabo oblikoskladenjskih oznak, je za celovitejše primerjave potrebno zagotoviti označenost korpusov na skladenjskem nivoju. Za to nalogo smo pripravili pilotsko množico skladenjsko označenih tvitov, ki lahko služi za nadaljnje učenje razčlenjevanja slovenske računalniško posredovane komunikacije.[3]

V sklopu projekta je bilo veliko pozornosti namenjene razvoju oz. prilagoditvam postopkov jezikoslovnega označevanja slovenščine specifikam nestandardnega jezika. V te namene je bil razvit učni korpus Janes, v katerem so pojavnice ročno popravljene na ravni segmentacije in tokenizacije ter normalizirane na besedni ravni, ročno pa so pregledane tudi pripisane leme ter oblikoskladenjske oznake (Čibej et al. 2016a; 2016b). Iz korpusa smo vzorčili množico 200 tvitov (475 stavkov), in sicer na način, da vsebuje enakomerne deleže tvitov, ki so avtomatsko označeni kot jezikovno in tehnično (ne)standardni (Ljubešić et al. 2015), ter obenem vključuje primere, ki so daljši od 120 znakov in v avtorstvu zasebnih uporabnikov. Zadnja pogoja sta omogočila, da smo v korpus zajeli besedila, ki vsebujejo dovolj za označevanje relevantnih specifik. Rezultati so bili pripravljeni v tabeli (Slika 1), ki poleg besedila tvita vsebuje razpoložljive metapodatke (ID, čas nastanka, spol in ime uporabnika, reakcije uporabnikov na tvit (všečkanje, deljenje) ter avtomatsko pripisane kategorije jezikovne in tehnične standardnosti ter sentimenta sporočila.

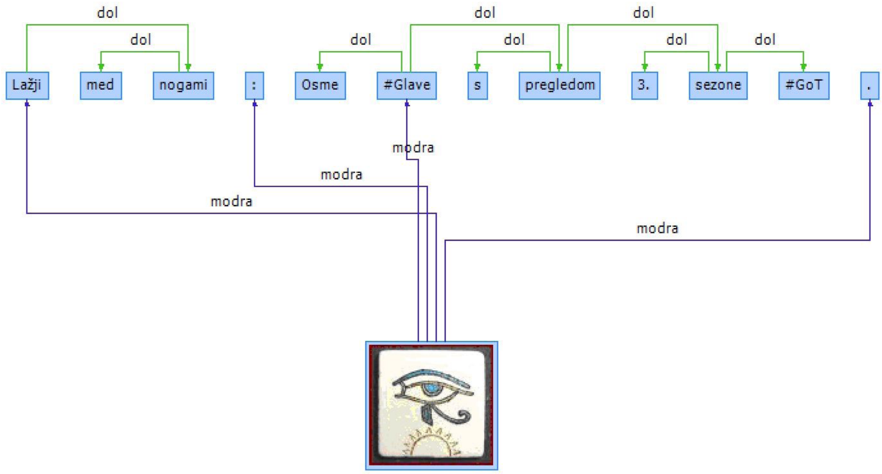
	A	B	C	D	E	F	G	H	I	J	K	L	M
1		id	name	sex	text	created	favorited	retweeted	std_tech	td_tech_n	std_ling	sentiment	source
2	*	tid.7487658297135104		male	Manj kot dan je trajalo, da smo dobili prvi jailbreak za Applovo najnovejšo različico operacijskega sistema iOS, 4.2.1. Kdo bi si mislil.	2010-11-24T17:36:11	0	0	T1	42095	L1	negative	private
3	*	tid.266122536432062464		male	Nekateri zvesti podporniki… Še vam ni jasno, da če bi želeli videti vsak tweet kandidatov, bi enostavno sledili njim? #predsednik12	2012-11-07T10:18:51	0	0	T1	42095	L1	negative	private
4	*	tid.311838678605512704		male	@petrasovdat V priemerjavi s svojim predhodnikom nedvomno. Okoliščine in pogoji dela pa so mu bili vse prej kot naklonjeni.	2013-03-13T13:58:29	0	0	T1	42095	L1	negative	private
5	*	tid.336054660693643264		female	Delam, delam, delam, odstranil bom plevel, prekopal bom vrtiček, prepeval ves vesel… #gardening inspired by Palček Primož #nowsinging	2013-05-19T09:44:09	1	0	T1	42095	L1	positive	private
6	*	tid.342718195666399232		male	Čeferin: sodišče podlega javnemu mnenju. Ni samo obsodilna sodba tista, ki kaže na to, da pravna država funkcionira. #pogledislovenije	2013-06-06T19:02:39	1	3	T1	42005	L1	negative	private
7	*	tid.352691378020548608		male	Lažji med nogami: Osme #Glave s pregledom 3. sezone #GoT. Gobcamo @anzet @BokiNachbar @WiC_HmR @matevzluzar http://t.co/LWHogSK9nj	2013-07-04T07:32:31	2	4	T1	1.0	L1	neutral	private

Slika 1: Vzorčenje korpusa Janes-Syn (anonimizirani prikaz).

Za označevanje vzorca smo izbrali sistem odvisnostne skladnje JOS (Erjavec et al. 2010), ki je bil razvit posebej za slovenski jezik in v slovenskem prostoru uspešno uporabljen za označevanje učnega korpusa ssj500k (Krek et al. 2015). Na podlagi slednjega je bil v sklopu projekta Sporazumevanje v slovenskem

3 Kot vsi drugi rezultati projekta JANES je tudi Janes-Syn prosto dostopen za uporabo, zaradi tehničnih težav s pretvorbo med formati sicer v nekoliko skrajšani različici (4.000 pojavnic oz. 170 besedil). Kot podatkovna množica je na voljo na repozitoriju CLARIN.SI (Arhar Holdt et al. 2017, http://hdl.handle.net/11356/1086), iskanje po korpusu pa je mogoče tudi v konkordančniku noSkE: http://nl.ijs.si/noske/sl.cgi/corp_info?corpname=janes.syn.

jeziku[4] razvit tudi razčlenjevalnik za slovenščino (Dobrovoljc et al. 2012).[5] Vzorec 200 tvitov je bil s tem programom avtomatsko razčlenjen in nato uvožen v program za vizualizacijo drevesnic SSJ (avtor J. Brank, glej Sliko 2). Pripisane skladenjske oznake oz. povezave so bile nato ročno popravljene skladno z označevalnimi smernicami (Holozan et al. 2008), prilagoditve sistema specifikam nestandardnega jezika pa so bile zabeležene v nadgrajeni različici smernic (Arhar Holdt 2016). Dopolnitve so na petih ravneh: označevanje žanrsko specifičnih elementov; raba tujejezičnih prvin; obravnava eliptičnosti in fragmentarnosti jezika; nestandardna raba ločil; in druge skladenjske posebnosti, h katerim se vračamo v nadaljevanju prispevka.



Slika 2: Primer označenega tvita v Označevalniku SSJ.

V procesu vzorčenja in pretvorbe smo iz obravnave izpustili 4 besedila, končni nabor, o katerem pišemo v prispevku, obsega torej 196 tvitov. Vsi v nadaljevanju navedeni primeri besedil so, kot rečeno, normalizirani na besedni ravni (Čibej et al. 2016b); poleg tega uporabljamo različice z odstranjenimi nerelevantnimi žanrsko specifičnimi elementi (glej razdelek 2.1), saj se s tem izognemo navajanju uporabniških imen, ki bi lahko razkrila identiteto pišočih.[6] Razlike med izvornim besedilom in različico v prispevku prikazuje naslednji primer:

4 Projektna stran: www.slovenscina.eu.

5 Alternativna izbira, prav tako že preizkušena za slovenščino (Dobrovoljc in Nivre 2016, Dobrovoljc et al. 2016), bi bil sistem Universal Dependencies, katerega prednost je medjezikovna primerljivost rezultatov. Odločitev za sistem JOS (v danem trenutku in za dano nalogo) utemeljujejo: dobra predhodna seznanjenost s sistemom in obstoj označevalnih smernic, na katerih je bilo mogoče osnovati dopolnitve za nestandardni jezik, ter obstoj zmogljivega programa za označevanje in pregledovanje drevesnic, v katerem je bilo mogoče med delom raziskovati odločitve, aplicirane v korpusu ssj500k.

6 V primerih, kjer imena služijo kot nujni del ponazoritve, smo ohranili imena inštitucij ali znanih oseb iz sveta politike ali športa.

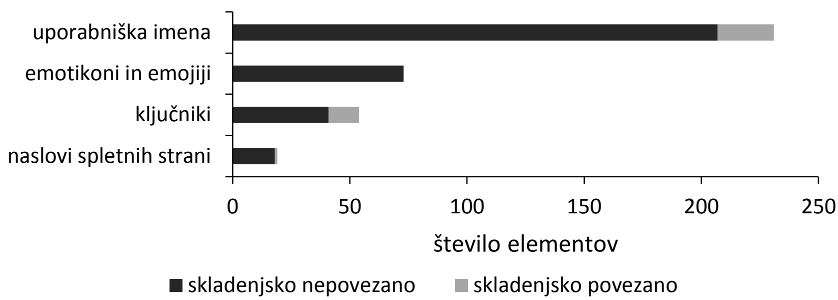
[A] izvorno besedilo: @union_pivo pizda, zakaj morm zmer uniona pit z laško kozarca? A to je taka politika al nimate kozarcev al vam je vseeno za kulturo piva??!!

[B] normalizacija na besedni ravni: @union_pivo pizda, zakaj moram zmeraj Uniona piti z Laško kozarca? A to je taka politika ali nimate kozarcev ali vam je vseeno za kulturo piva??!!

[C] izpust tviterskih elementov: pizda, zakaj moram zmeraj Uniona piti z Laško kozarca? A to je taka politika ali nimate kozarcev ali vam je vseeno za kulturo piva??!!

2.1 Žanrsko specifični elementi

Pri označevanju korpusa Janes-Syn je bila sprejeta odločitev, da bodo žanrsko specifični elementi (emotikoni, emojiji, naslovi spletnih strani, sklici na uporabniška imena in ključniki) povezani v strukturo drevesnic, kadar so del stavčne skladnje, sicer bodo izpuščeni iz obravnave, pri čemer se pri elementih na začetku in koncu tvitov nagibamo k nevključevanju. Rezultati kažejo, da je izpustljivih elementov bistveno več, v smislu povezovanja pa so relevantna predvsem uporabniška imena in ključniki, izjemoma tudi naslovi spletnih strani. Emotikoni in emojiji so v skladenjskem smislu v obravnavanem vzorcu vsi izpustljivi. Slika 3 prikazuje razmerje med povezanimi in nepovezanimi elementi v korpusu Janes-Syn.



Slika 3: Skladenjsko povezani in nepovezani tviterski elementi v Janes-Syn.

Med označevanjem je bilo v skladenjske strukture vključenih 24 uporabniških imen (10 % vseh). V Janes-Syn imena običajno nastopajo v vlogi osebka (14 primerov) ali predmeta (7 primerov), redkeje v prislovnih vlogah:

[1] Hmm, **@Delo** je zbrisalo razmislek ob Tugomerju da smo spet pod Franki, ker sta Bxl in Lux. frankovski središči, pa tudi Juncker je Frank ...?

[2] Zanimivo, da vaša **@strankaSDS** in njeni člani nabirajo točke z **@JJansaSDS** ter komunisti, ko jih sami toliko omenjate, več kot program.

[3] Ko berem komentarje pod tekstom o plebiscitu na **@rtvslo**, mi je žal, da večina njih, ne bo nikoli v rokah kakšnega polpismenega desetarja v JLA.

Kar se tiče ključnikov, je bilo skladenjsko povezanih 13 primerov (24 % vseh). Tudi ključniki v strukturah nastopajo samostojno ali kot del besednih zvez, pretežno v vlogi osebka (6 primerov) ali predmeta (5 primerov):

[4] @LakovicJaka hvala za vse kar si naredil za reprezentanco. Čeprav ti letos ni šlo brez tebe ne bi bili **#junaki.** Rečem ti lahko le SREČNO.

[5] Več kot očitno sta risa hotela ujeti **avion za #sochi.** Nista mogla zraven zaradi Massijeve prtljage.

[6] @bota112 reciva, da je **#krneki** … tako kot večina državne inf., z funkcionalno nepismenimi IT managerji - ki je **#btw** kazenska funkcija v **#du** :(

Primer skladenjsko vpetega naslova spletne strani je v obravnavanem vzorcu samo eden:

[7] jah saj za manj recimo tudi jaz ne bi peljal ;) samo dobro oni jih malo več peljejo :) probaj še **na http://t.co/YaVQdnaN5p** :)

Opisane označevalne odločitve so primerljive delu na angleščini pri Kaufmann in Kalita (2010), kjer so ločevanje skladenjsko relevantnih tviterskih elementov od nerelevantnih tudi avtomatizirali, in sicer z upoštevanjem besednorednih oznak konteksta: če uporabniškemu imenu denimo sledi veznik, predlog ali glagol, to s precejšnjo zanesljivostjo nakazuje njegovo skladenjsko relevantnost. Ključniki so v tem smislu nekoliko zahtevnejša naloga, saj so različnih besednih vrst.[7]

2.2 Tujejezični elementi

Kot ugotovljeno (Michelizza 2015: 161–65, Reher in Fišer 2018), vsebuje računalniško posredovana slovenščina (v splošnem) opazen delež tujejezičnih prvin. V označenem gradivu se slednji pojavljajo v 26 % tvitov, od tega 20 % iz angleščine in 6 % iz sorodnih južnoslovanskih jezikov. V primerih je mogoče opaziti različne stopnje prilagojenosti slovenskemu črkovanju in oblikoskladnji (Čibej et al. 2016b), potrdi se torej ugotovitev, da uporabniki tuje besedišče samoiniciativno »ne le oblikoslovno in skladenjsko, temveč tudi pisno podomačijo« (Jakop

7 Svojevrsten označevalni izziv, ki ga zaenkrat puščamo za prihodnost, predstavlja notranja struktura ključnikov, ki so lahko sestavljeni iz ene ali več besed v enem ali več različnih jezikih in z raznovrstnimi zapisovalnimi specifikami.

2008: 322). Kar se tiče dolžine tujejezičnih elementov, se pojavljajo tako posamezne besede (46 primerov) kot tudi besedne zveze (18 primerov) in daljše stavčne strukture (17 primerov).

[8] Optimiziran je tako, da čim manj porabi brez veze. Več kot pošiljaš, več bo porabil. Če **šeraš** slike in »**stickyje**« bo šlo orenk gor.

[9] Na Kongrescu smo, **meanwhile,** v popolnoma drugi dimenziji. Nek **hardcore band** se dere. sliši se boljše kot **basket, just so you know.**

[10] videla par sekund posnetka na Fb, ko ena (vstavi poljubno žaljivko) tepe parmesečnega dojenčka. **Da bog da joj majka ljubila sliku na banderi!**

Vpenjanje tujejezičnih elementov v skladenjsko označevanje je pri Janes-Syn potekalo tako, da se posamezne besede in tiste besedne zveze, kjer je odvisnost med elementi enostavno določljiva (in primerljiva s slovensko skladnjo, npr. zveza samostalnika in določujočega pridevnika), vpenja v skladenjsko drevo. Daljših struktur, zlasti stavčnih, ne vpenjamo oz. jih povezujemo kot fragmente neposredno na označevalno jedro. Označevalna izkušnja kaže, da je tovrstno ločevanje v praksi dovolj izvedljivo in smiselno, saj na tak način v drevesu ohranimo lastna imena tipa *Creative Commons, Candy Crush*, kot tudi občno besedišče različnih besednih vrst (npr. *fake, prpa, chatati*). Odločitev je v nadaljevanju treba preveriti na večji količini gradiva in v luči konkretiziranih označevalnih namenov po potrebi prilagoditi.

2.3 Nestandardna raba ločil

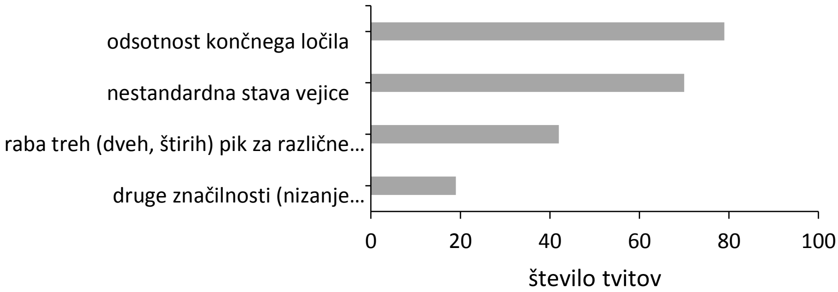
Tudi o nestandardni rabi ločil v računalniško posredovani slovenščini se je že pisalo, pogosto z normativističnega vidika (npr. Jakop 2008: 323, Dobrovoljc 2008: 309–311, Michelizza 2015: 133–140, Popič et al. 2016). V obravnavanem vzorcu je najti raznovrstne specifike rabe ločil v 69 % označenih tvitov. Pojavljajo se: izpuščanje končnih ločil oz. njihovo nadomeščanje z emotikoni, ključniki ipd.; odstopi od norme na ravni rabe vejice, predvsem njeno izpuščanje; raba treh (ali dveh, štirih) pik za nakazovanje premorov ali kot nadomestilo drugih ločil; in druge težave, npr. pri rabi ločil za poročani govor, upoštevanju razlik pri rabi vezajev in pomišljajev ter nizanju klicajev in/ali vprašajev za izražanje stopnje čustvene obarvanosti sporočila:

[11] nekje sem bral da obstajajo študije o smrtonosnosti cepljenj in hudi škodljivosti chemtrailov …

[12] težave z elektriko **…** pokličeš kolega ki te ne pusti na cedilu **…** spiješ enega ali dva **…** nice Saturday

[13] Janez.Janša se bo vrnil v velikem SLOGU, Z VSEM SLOVENSKIM NARODOM NA ČELU IN Z milijoni EVROV V ŽEPU ZA STORJENE MU KRIVICE, JA.!!

Številčni podatki za prisotnost navedenih elementov so prikazani na Sliki 4 (posamezni tvit lahko vsebuje eno ali več navedenih značilnosti).



Slika 4: Značilnosti glede rabe ločil v Janes-Syn.

Pri razčlenjevanju standardne slovenščine so ločila lahko koristna informacija za določanje stavčnih meja in razmerij med deli povedi. Kot je razvidno (tudi) iz obravnavanih podatkov, se je na ločila v računalniško posredovani komunikaciji mogoče zanašati v manjši meri. Značilnosti nestandardne rabe, ki so sistemske (npr. odsotnost končnih ločil pred emotikoni), je mogoče pri avtomatski predpripravi korpusnih besedil upoštevati, nekateri drugi načini rabe so manj predvidljivi. Glede na ugotovljeno se zdi, da je mogoče delo nadaljevati v dve smeri, bodisi s poskusi učenja razčlenjevanja brez upoštevanja ločil ali z vključitvijo koraka njihove normalizacije.[8]

2.4 Fragmentarnost jezika in izpusti

V podatkih korpusa Janes-Syn se pojavljajo tako krajšanja in izpusti kot drobljenje sporočila po vzoru govorjenega jezika (glej primer [12]). Tovrstne značilnosti, ki so jih identificirali mdr. tudi Kaufman in Kalita (2010) ter Schneider (2015) na angleških tvitih, pri nas pa Kranjc v spletnih klepetih (2003: 76), Kalin Golob v SMS-ih (2008: 292), Michelizza v blogih in Wikipediji (2015: 216–220), predstavljajo za skladenjsko označevanje poseben izziv.

Analiza označenega vzorca pokaže, da se elipsa pojavi v 20 primerih, večinoma kot izpust pomožnega glagola (11 primerov), redkeje izpust modalnega glagola

8 Druga možnost bi ponudila tudi možnost za izboljšavo orodij, kot so slovnični pregledovalniki, kot je bilo nakazano npr. v Holozan (2013) ter Kranjc in Robnik Šikonja (2015).

(2 primera), zaimka (2 primera), simbola (1 primer) ali polnopomenske besede, npr. v sklopu fraz, kot so *ni druge* ali *potem pa ti meni* (4 primeri). Rezultati so primerljivi izsledkom Goli et al. (2016), ki ugotavljajo, da je med krajšanji v 800 slovenskih tvitih na skladenjski ravni sicer bistveno manj posegov kot na drugih ravneh, prevladujejo pa predvsem izpusti pomožnika *biti*.

[14] Šele zdaj videl kakšna drama je bila v CONCACAF, ko so ZDA v zadnjih minutah priigrale Mehiki dvoboj z N. Zelandijo. Konec velikega rivalstva?

[15] tudi jaz .. zato pa prvo v glavi porihtati, da je šejk namesto obroka .. in ker je sladek, boš sčasoma zgubila sugar rush

[16] haha saj se mi je zdelo, premalo si na Štajerskem!!! potem pa ti meni o manjkajočih j-jih in i-jih v besedah. fff :D

Za označevanje skladnje in skladenjsko razčlenjevanje so izpusti večji problem, kadar vplivajo na strukturo drevesnice, tj. kadar gre za elemente, kot so npr. jedra besednih zvez ali stavčni povedek, na katere se tipično vežejo drugi elementi. Označevalni sistemi se z vprašanjem tovrstnih izpustov soočajo na dva načina, ena od možnosti je povišanje odvisnega elementa na mesto manjkajočega (Dobrovoljc in Nivre 2016), na drugi strani povezovanje fragmentov neposredno na označevalno jedro pohitri proces označevanja in odločanja (Kong et al., 2012), seveda ob določeni izgubi informacij. Ker sistem označevanja JOS to omogoča, se pri označevanju Janes-Syn odločamo za drugi način, v naslednjem koraku pa bi bilo smiselno primerjati rezultate obeh pristopov z vidika označevalnih stroškov v primerjavi s pridobitvijo oz. izgubo na ravni kakovosti rezultatov za določen raziskovalni namen.

3 BESEDNI RED RAČUNALNIŠKO POSREDOVANE SLOVENŠČINE

Do sedaj obravnavane skladenjske značilnosti imajo neposreden vpliv na proces označevanja, zato smo jih obravnavali v ločenem poglavju. Razen naštetega pa se v povezavi z računalniško posredovano slovenščino pojavljajo tudi druge ugotovitve. Med pogosteje obravnavanimi temami je vprašanje skladenjske kompleksnosti: Kranjc (2003: 76) denimo na jeziku spletnih klepetalnic opaža (tudi) strukturno zapletene večstavčne povedi, ob čemer so v rabi predvsem predmetni odvisniki. Dobrovoljc (2008: 309) zapiše, da je skladenjska zgradba e-poštnih sporočil »brez zapletenih povedi, veliko je kratkih stavkov, ki so povezani v poved brez veznikov«. Michelizza (2015: 213–234) zaključi, da skladnja blogov in wikipedijskih člankov ni okrnjena in da skladenjske specifike, kolikor jih je zaznati, ne vplivajo na razumevanje pomena. Redkejši so izsledki o drugih skladenjskih lastnostih računalniško posredovane slovenščine, vključno z vprašanjem besednega reda, ki nas zanima v nadaljevanju prispevka.

Da je besedni red za dano raziskovalno področje relevantno vprašanje, so navedene študije sicer nakazale, niso pa tematike podrobneje analizirale. Kranjc (2003) v povzetku navaja, da se v jeziku spletnega klepeta kaže govorna forma v »spremenjenem besednem redu in neupoštevanju načela členitve po aktualnosti«, vendar se k vprašanju v sami razpravi ne vrne. Kalin Golob (2008: 292) med značilnostmi analiziranih SMS-ov omenja »spontani besedni red«, ki ob preostalih identificiranih značilnostih kaže »zapis po govoru«, vendar konkretni primeri niso diskutirani. Michelizza (2015: 228–230) se členitve po aktualnosti dotakne prek tipologije leksemov, ki se pogosto pojavljajo v remi, ne posveča pa se vprašanju samega besednega reda. Vprašanje besednega reda v slovenščini, tako nezaznamovanega kot zaznamovanega, sicer velja za zahtevno in slabo raziskano področje, na kar avtorji opozarjajo že od Breznika (1908) naprej. Kasnejša dela (npr. Toporišič 1967, Jug Kranjec 1981, Vidovič Muha 2000, Toporišič 2008, Toporišič 2004 – v nadaljevanju SS 2004) se osredotočajo na členitev po aktualnosti na eni strani in stalno stavo na drugi, predvsem v nizu določujočih pridevnikov ter naslonskem nizu. Kadar je v literaturi govor o zaznamovanosti besednega reda, se slednja obravnava predvsem na jeziku leposlovnih del, zato ugotovitve na obravnavo računalniško posredovane komunikacije niso neposredno prenosljive.

Sledeč Toporišičevi smernici, da je kriterij za ločevanje običajnega in zaznamovanega besednega reda »takrat, kadar ga tako ali drugače občutimo« (Toporišič 2008: 31), smo se odločili, da potencialno zaznamovanost besednega reda v korpusu Janes-Syn preverimo s sodelovanjem treh neodvisnih označevalcev, kot je razloženo v nadaljevanju prispevka. V raziskavi nas zanima, kako označevalci razumejo besednoredno zaznamovanost v računalniško posredovani slovenščini, kolikšno je ujemanje med njihovimi odločitvami, katere vrste besednorednih problemov se glede na pripisane oznake pojavljajo v podatkih in kako so ti problemi razporejeni glede na avtomatsko pripisano kategorijo jezikovne (ne)standardnosti (o metodologiji pripisa glej Ljubešić et al. 2018).

3.1 Označevanje in kategorizacija problemov

Označevalci so za svoje delo prejeli tabelo z besedili korpusa Janes-Syn in nalogo označiti tvite, v katerih se zdi besedni red z vidika standardne pisne slovenščine zaznamovan. Ob tem so primere morali tudi popraviti, kot bi jih, denimo, v procesu lektoriranja, vendar po principu minimalne intervencije, tj. samo s spremembo zaporedja besed (in po potrebi ločil, začetnic), ne pa s spreminjanjem ostalih jezikovnih značilnosti. Smernice za označevanje problemov so nalogo namenoma opredeljevale ohlapno, saj je bil eden od ciljev raziskave preveriti intuitivnost »besednoredne zaznamovanosti« in ujemanje med označevalci glede slednje.

Za potrebe prispevka smo rezultate označevanja razvrstili v vsebinske skupine, in sicer od spodaj navzgor na osnovi gradiva. Kot je običajno za tovrstna razvrščanja, so se nekatere kategorije pokazale zelo hitro in so povsem jasno ločljive (npr. primeri z naslonkami na prvem mestu stavka, nesklonljivim levim prilastkom, vprašanja postavitve členka), medtem ko so druge težje določljive in deloma medsebojno prekrivne (npr. členitev po aktualnosti, vprašanja razporeditve stavčnih členov v stavku). Številčne rezultate v nadaljevanju je treba razumeti z upoštevanjem navedenih metodoloških značilnosti.

3.2 Rezultati

V procesu označevanja je bilo opravljenih 131 popravkov besednega reda v 94 različnih tvitih. Popravki so bili nato razvrščeni v 14 skupin. Pri kategorizaciji so bili upoštevani vsi primeri, ki so bili označeni vsaj enkrat, torej tudi tisti, ki jih je označil samo en od označevalcev, preostala dva ne. To dejstvo je pomembno, ker je ujemanje med označevalci za dano nalogo zelo nizko, k čemur se vrnemo v razdelku 3.3. Da je ponazoritev rezultatov jasnejša, v primere tvitov dodajamo podatek o tem, kateri od označevalcev je posamezni primer izpostavil kot zaznamovanega (*označ. 1, 2* ali *3*).

Rezultati označevanja prinašajo večji nabor primerov, kjer bi kot osrednji problem lahko izpostavili **členitev po aktualnosti,** tj. vprašanje razvrščanja informacij od znanega k manj znanemu ali postavljanje osrednje, najbolj bistvene informacije v remi na koncu stavka oz. povedi (SS 2004: 660). Brez poznavanja avtorjevega namena in besedilnega konteksta je v praksi (ne)zaznamovanost členitve po aktualnosti težje presojati. Tviti so v tem smislu dodatno problematični,[9] ker so izvorno lahko (ne pa nujno) del dialoške komunikacije oz. se nanašajo/sklicujejo na vsebine, ki pri označevanju gradiva niso bile več na voljo. Na drugi strani se vprašanje razporejanja delov povedi v temo, prehod in remo mestoma prekriva z drugimi skladenjskimi vprašanji, kot opisujemo v nadaljevanju. Kjer je bilo vzrok identificirane in popravljene težave mogoče pripisati v kako drugo skupino, so pri kategorizaciji le-te imele prednost. Na koncu je v skupini členitev po aktualnosti ostalo 23 primerov (17,6 % vseh), ki so primarno pomensko-poudarjalne narave:

9 V primerjavi s celovitimi in zaključenimi besedili, ki se običajno uporabljajo za zgled členitve po aktualnosti, prim. Trdinovo bajko o grofu in medvedu v SS 2004: 660.

[17] G: »Od kdaj je nama tako kul **hladno vreme?** po mojem zato, ker se lahko stiskava, brez da bi naju švic lepil en na drugega.« > (*Označ. 2*) G: »Od kdaj nama je hladno vreme **tako kul**? Po mojem zato, ker se lahko stiskava, ne da bi naju švic lepil enega na drugega.«

[18] glej, ni druge kot, da jo unfollowamo **vsi,** če bo še naprej toliko nesramna … P.s. jaz bi raje onega ta temnega > (*Označ. 1, 3*) glej, ni druge, kot da jo vsi **unfollowamo,** če bo še naprej tako nesramna … P.S. Jaz bi raje onega ta temnega.

Z omenjeno deloma prekrivna kategorija so popravki besednega reda **glagolskih besednih zvez s prislovno sestavino** (21 primerov, 16,0 % vseh); prekrivnost je pri primerih, kjer je prislovna sestavina v remi, torej bi izvorni besedni red bilo mogoče potencialno razumeti s stališča vsebinskega poudarjanja. Skupina vsebuje tako zveze glagola s prislovom (11 primerov) kot zveze glagola s predložno samostalniško zvezo (10 primerov), pri čemer v tem prispevku vprašanje (ne)obveznosti prislovnega dela (SS 2004: 592) puščamo ob strani:[10]

[19] začnite jih **spreobračati na polno** v tisto kar ne marajo. Ali pa kimate in jim date prav, naredite pa po svoje. > (*Označ. 1, 2, 3*) Začnite jih **na polno spreobračati** v tisto, česar ne marajo. Ali pa kimate in jim date prav, naredite pa po svoje.

[20] Kolega, čisti Ljubljančan, je to za joke zinil v lokalu v centru Lj.. Konec na urgenci, so se lj. idioti **med seboj sfajtali** > (*Označ. 1, 2, 3*) Kolega, čisti Ljubljančan, je to za joke zinil v lokalu v centru Lj. Konec na urgenci, lj. idioti so se **sfajtali med seboj.**

Z vsebinskim poudarjanjem je povezana tudi skupina, ki smo jo nekoliko pavšalno poimenovali **vrivanje stavčnih členov** (10 primerov, tj. 7,6 % vseh), ker gre za postavitev (pretežno) stavčnega osebka ali predmeta med dele sestavljenega povedka oz. med povedek in prislovna določila. Ta skupina prinaša primere zaznamovanega besednega reda, ki se v govoru izraža v stavčni intonaciji in poudarku (Jug Kranjec 1981):

[21] ne vem, da ni @RomanLeljak -a dobila **Udba** v kremplje? Že nekaj dni ne čivka. Kaj mislite? Bi bilo treba tiralico? > (*Označ. 1, 3*) Da ni @RomanLeljak -a dobila v kremplje **Udba?** Že nekaj dni ne čivka. Kaj mislite? Bi bilo treba tiralico?

[22] šetamo po Lj. in pride **Zoki** mimo, se ustavi pa da Emanuelu petko. ta malemu nič jasno. rečem to je Zoki kralj in se ta mali zadere: Zoki kralj! > (*Označ. 1, 3*) Šetamo po Lj. in pride mimo **Zoki,** se ustavi in da Emanuelu petko. Malemu nič jasno. Rečem: »To je Zoki kralj,« in mali se zadere: »Zoki kralj!«

10 Zdi se, da označevalci prislove pogosteje premikajo na mesto desno od glagola, predložne zveze pa levo, vendar je podatkov za posplošitve premalo. Prav tako ni dovolj podatkov za ugotavljanje potencialne drugačne stave načinovnih določil oz. prislovov (prim. Toporišič 1967: 257–258).

Podoben poudarek si je mogoče misliti tudi pri primerih, kjer se po tovrstni stavi **povedek pojavlja na koncu stavka** (8 primerov, tj. 6,1 % vseh). Te primere smo obdržali kot ločeno skupino, ker je zanje značilno, da so za označevalce posebej opazni in da se obenem pojavljajo izključno v tvitih, avtomatsko označenih za jezikovno nestandardne (več o tem v razdelkih 3.3 in 3.4):

[23] Noben ISP ne ponuja dnevnega / tedenskega zakupa ? To bi bilo kul. jaz nimam TV-ja, zdajle bi pa plačala, da bi lahko **tekme** gledala. > (*Označ. 1, 2, 3*) Noben ISP ne ponuja dnevnega ali tedenskega zakupa ? To bi bilo kul. Jaz nimam TV-ja, zdajle bi pa plačala, da bi lahko gledala **tekme.**

[24] jaz : Bu. sodelavec : *krik, ki ga je **cela bajta** slišala * jaz :* nekontroliran izbruh smeha, ki ga še vedno cela bajta posluša* > (*Označ. 1, 2*) Jaz : Bu. Sodelavec : * Krik, ki ga je slišala **cela bajta.** * Jaz : *Nekontroliran izbruh smeha, ki ga še vedno posluša cela bajta.*

Ločeno so obravnavani tudi drugi primeri, vezani na mesto **povedka** oz. njegovih delov (7 primerov, tj. 5,3 % vseh). Gre za primere, kjer se del povedka, npr. modalni glagol, pojavlja na prvem mestu stavka, ali pa je popravek označevalca vezan na postavitev povedka v glavnem stavku, ki sledi odvisniku:

[25] Pa še 3 - 4 leta nazaj sem bil tako ponosen na svojo kondicijo (resda zlasti kar se tiče hoje) **Moram malo** spremeniti način življenja! * > (*Označ. 1*) Pa še 3-4 leta nazaj sem bil tako ponosen na svojo kondicijo (resda zlasti kar se tiče hoje). **Malo moram** spremeniti način življenja!

[26] če vprašate mene (in mislim, da se @anakobal_kobe strinja), Tini do res odličnega rezultata **manjka** le malo teže na spodnji smučki. > (*Označ. 3*) Če vprašate mene (in mislim, da se @anakobal_kobe strinja), **manjka** Tini do res odličnega rezultata le malo teže na spodnji smučki.

Za razliko od do sedaj naštetih kategorij, ki so deloma prekrivne, je enoznačna za identifikacijo kategorija problemov s **postavitvijo členka** (15 primerov oz. 11,5 % vseh). Popravki so v primerih, kjer členek v izvirniku (po presoji označevalca) ne stoji pred delom stavka, ki naj bi ga pomensko modificiral (SS 2004: 675):

[27] padanje se je **tudi** pričelo že kar nekaj časa nazaj in je dokaj konstantno od cca. začetka UA krize > (*Označ. 1*) **Tudi** padanje se je pričelo že kar nekaj časa nazaj in je dokaj konstantno od cca. začetka krize UA.

[28] Pri nas v Ustavi pa **seveda** vladavino imamo. Demokracija je ena od oblik vladavine nad državno - kapitalsko državo! > (*Označ. 2*) Pri nas v Ustavi pa vladavino **seveda** imamo. Demokracija je ena od oblik vladavine nad državno - kapitalsko državo!

Prav tako enostavno ločljiva skupina so primeri, kjer se **naslonka** pojavlja na prvem mestu stavka (15 primerov oz. 11,5 % vseh), kar se v standardnem jeziku utemeljuje v primerih izpusta vezniške ali naglašene sestavine pred njo (SS 2004: 676). V obravnavanem vzorcu se na atipičnem prvem mestu najpogosteje pojavita pomožni glagol (9 primerov) in povratni zaimek (5 primerov):

[29] Kolega, čisti Ljubljančan, je to za joke zinil v lokalu v centru Lj.. Konec na urgenci, **so se lj. idioti** med seboj sfajtali. > (*Označ. 1, 2, 3*) Kolega, čisti Ljubljančan, je to za joke zinil v lokalu v centru Lj. Konec na urgenci, **lj. idioti so se** sfajtali med seboj.

[30] so Ready To Go. **Se veselim** dobre letine pristankov. > (*Označ. 1, 2, 3*) So Ready To Go. **Veselim se** dobre letine pristankov.

Na vprašanje naslonskega niza se veže skupina primerov (8 primerov oz. 6,1 %), kjer označevalci popravijo **stavo naslonk** ob veznikih. Tipičen primer je veznik *pa*, ki naj bi nezaznamovano stal pred prostimi naslonkami (SS 2004: 676), v gradivu pa se pojavlja na mestu za pomožnikom (5 primerov):

[31] ma dva dedca čvekata na polno, vsi ostali **smo pa** hoteli dremati. potem se pa začneta meniti o kinih pa o filmih... > (*Označ. 1, 2, 3*) Ma, dva dedca čvekata na polno, vsi ostali **pa smo** hoteli dremati. Potem se pa začneta meniti o kinih pa o filmih ...

[32] Pomagajte mi sem sin odvisnika in odvisnice. oče je že 2 leti na Besedovnjaku mama **je pa** na Candy Crushu in Flappy Birdu. > (*Označ. 2*) Pomagajte mi, sem sin odvisnika in odvisnice. Oče je že 2 leti na Besedovnjaku, mama **pa je** na Candy Crushu in Flappy Birdu.

Ločeno smo obravnavali primere z **neujemalnim levim prilastkom** (5 primerov, 3,8 % vseh), in sicer tudi tiste, kjer je prilastek kratični. Pri slednjih gre sicer primarno za vprašanje zapisovanja z vezajem ali brez njega (Arhar Holdt in Dobrovoljc 2016), ker pa je bila označevalna naloga usmerjena v besedni red, so označevalci v tem okviru identificirani problem tudi reševali:

[33] pizda, zakaj moram zmeraj Uniona piti z **Laško kozarca**? A to je taka politika ali nimate kozarcev ali vam je vseeno za kulturo piva??!! > (*Označ. 1, 3*) Pizda, zakaj moram zmeraj Union piti z **kozarca Laško**? A je to taka politika ali nimate kozarcev ali vam je vseeno za kulturo piva??!!

[34] Plus, premikanje na **SD kartico** sem probala že stokrat. Tudi telefon to ponudi kot možnost. Rezultat? »Ni predmetov za premikanje« (*Označ. 1, 3*) Plus, premikanje na **kartico SD** sem probala že stokrat. Tudi telefon to ponudi kot možnost. Rezultat? »Ni predmetov za premikanje.«

Ločena skupina so še primeri (4 primeri, 3,1 % vseh), kjer je bila kot zaznamovana označena stava **osebnega zaimka,** ki je v izvornem besedilu v izpostavljeni poziciji:

[35] G : »Od kdaj je **nama** tako kul hladno vreme? po mojem zato, ker se lahko stiskava, brez da bi naju švic lepil en na drugega.« > (*Označ. 1, 2*) G : »Od kdaj **nama je** tako kul hladno vreme? Po mojem zato, ker se lahko stiskava, ne da bi naju švic lepil en na drugega.«

[36] Meni sta pa oba tako kjut v tem dialogu, da vama bom zdaj kar takole na daljavo rekla: jaz imam pa **vaju** oba rada! Eto! > (*Označ. 2*) Meni sta v tem dialogu oba tako kjut, da vama bom zdaj kar takole na daljavo rekla : Jaz **vaju** imam pa oba rada!

Redkeje sta zastopani skupini primerov, kjer je težava s **postavitvijo določujočega elementa** znotraj besedne zveze (2 primera oz. 1,5 % vseh) ali **zaporedja elementov znotraj glagolske zveze** (3 primeri ali 2,3 % vseh):

[37] ja no, meni se to tudi skozi dogaja! še večkrat pa s senčkami na Instagramu. naredim smokey zgleda pa **čisto nekaj nežnega** > (*Označ. 1, 2*) Ja no, meni se to tudi skozi dogaja! Še večkrat pa s senčkami na Instagramu. Naredim smokey, a zgleda **nekaj čisto nežnega.**

[38] Pripomba »PS« mi je všeč. O tej temi morava še kdaj kaj reči PS 2 pa ne, nisem videla. **Moram poguglati** > (*Označ. 1*) Pripomba »P. S.« mi je všeč. O tej temi morava še kdaj kaj reči. P. S. 2 pa ne, nisem videla. **Poguglati moram.**

Ostanejo še primeri (9 primerov, 6,9 % vseh), ki smo jih uvrstili v skupino **Drugo**. Gre za popravke posameznih označevalcev, izvirajočih iz različnega razumevanja vsebine obravnavanih tvitov in tudi zadane naloge, kot prikazuje spodnji popravek besednega reda v citirani pesmi:

[39] Kdo pozna to pesmico? "Takole se prične: Po belih in črnih tipkah tja v svet odjadralo skupaj **je prstkov deset** ..." > (*Označ.* 3) Kdo pozna to pesmico? "Takole se prične: Po belih in črnih tipkah **je** tja v svet odjadralo skupaj **deset prstkov** ..."

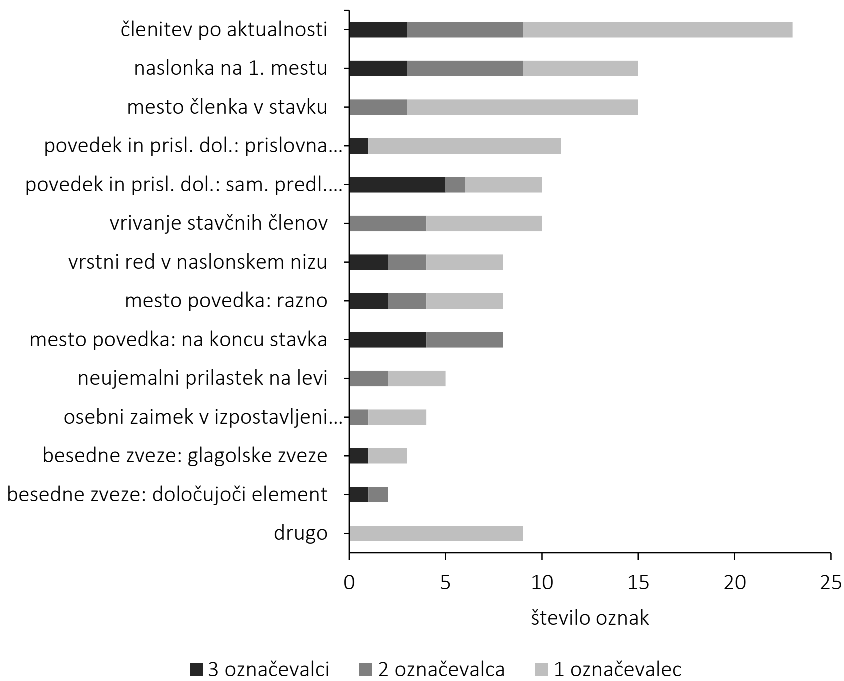
3.3 Ujemanje med označevalci

Ko govori o stilni vrednosti besednega reda, zapiše Toporišič (2008: 30) takole:

> Da si besede v govoru sledijo po določenem zaporedju, je znano. Za taka tipična besedna zaporedja imamo natančno razvit čut, ki nam pove, ali

govoreči razvršča besede primerno navadam danega jezika ali ne. Če jih kdaj ne razvršča pravilno, nam navadno ni prav nič težko govorečemu besedni red popraviti.

Raziskava, ki jo predstavljamo v tem prispevku, prinaša povsem drugačne rezultate. Od 131 popravkov besednega reda v korpus Janes-Syn je samo 23 primerov (17,6 %) takih, da so jih primerljivo označili vsi trije označevalci. 32 primerov (24,4 %) sta primerljivo označila dva od treh označevalcev, preostalih 77 (58,8 %) označb pa je individualnih. Ujemanje med označevalci v posamezni kategoriji kaže Slika 5.



Slika 5: Ujemanje med označevalci po kategorijah problemov.

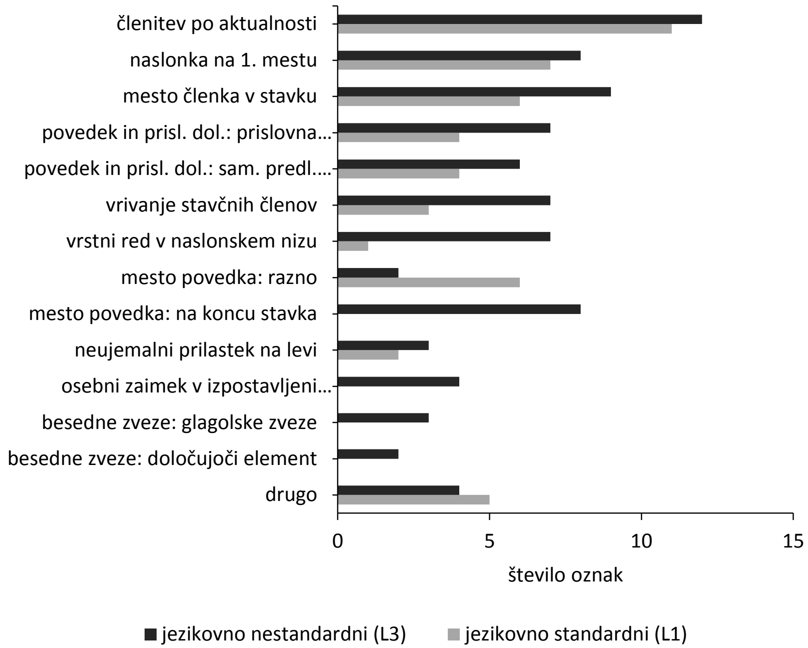
Zelo povedna ugotovitev je, da se pri nobeni od kategorij ne pojavljajo izključno primeri, kjer bi bilo označevanje enotno. Tudi če iščemo primere, ki sta jih enotno označila vsaj dva od treh, je mogoče izpostaviti samo dve kategoriji: primere, kjer se povedek pojavlja na koncu stavka (*zdajle bi pa plačala, da bi lahko **tekme gledala***) in manjšo skupino problemov pri postavitvi določujočega elementa znotraj besednih zvez (*zgleda pa **čisto nekaj nežnega***). V smislu same pogostnosti označenih primerov je nekoliko višjo skladnost opaziti še pri primerih, kjer je v ospredju mesto prislovnega določila v obliki predložne samostalniške besedne zveze (*začnite jih **spreobračati na polno** v tisto kar ne marajo*), pri členitvi po aktualnosti (*glej, ni druge kot, da jo unfollowamo* ***vsi***) in primerih, kjer se naslonka pojavlja na prvem mestu stavka (***so se*** *lj. idioti med seboj sfajtali*). Veliko individualnih oznak imajo na drugi strani – poleg kategorije členitve po aktualnosti, kjer so pričakovane – še kategorija mesta členka v stavku (*padanje se je* ***tudi*** *pričelo že kar nekaj časa nazaj*) in mesto prislovnega določila v obliki prislova (*jaz sem sicer bil tako zaspan, da sem skoraj* ***skup padel***).

Da ujemanje med označevalci za določene kategorije ne bo popolno, smo pričakovali, visoka raven in prisotnost odstopanj pri skoraj vseh kategorijah pa nas je vseeno presenetila. Neujemanje bi lahko deloma pripisali neizkušenosti označevalcev za tovrstno delo, čeprav bi glede na Toporišičev citat intuicija naravnega govorca (pri označevalcih gre dodatno za jezikoslovce) pri presojanju besednega reda morala voditi v primerljive rezultate.[11] Druga možna razlaga za dobljeni rezultat bi bila, da se pri obravnavi jezikovnega gradiva, ki prinaša opazen nabor specifičnih, tudi nestandardnih jezikovnih značilnosti, presojanje zaznamovanosti besednega reda prilagodi. Kar bi bilo vidno zaznamovano v sopostavitvi s standardnim jezikovnim gradivom, v množici tvitov, ki prinašajo drugačen tip jezika na vseh ravninah, postane manj opazno. Ali pa se označevalci, podobno kot učitelji pri popravljanju šolskih pisnih izdelkov, odločajo samo še za popravke, ki so res temeljni, manj moteče pa samodejno preskočijo. Samodejno prilagajanje kriterijev je vsekakor vznemirljivo vprašanje, ki bi se mu bilo smiselno v prihodnosti natančneje posvetiti. Na tem mestu pa je potrebno poudariti še, da je označevalna naloga v praksi zahtevnejša, kot se zdi po prebiranju že kategoriziranih rezultatov, kjer so za prikaz namenoma izbrani najbolj reprezentativni, jasni in nedvoumni primeri.

3.4 Nestandardno specifične kategorije

V okviru projekta JANES je bila razvita metodologija za avtomatski pripis tehnične in jezikovne (ne)standardnosti besedilom računalniško posredovane komunikacije. Značilke za določanje (ne)standardnosti posegajo na znakovno raven, npr. nestandardno rabo ločil in presledkov, ponovitev znakov, razmerje med abecednimi in neabecednimi znaki itd., in besedno raven, npr. rabo nestandardnega oz. atipičnega besedišča, zapis z velikimi ali malimi črkami itd. (Ljubešić et al. 2015). Skladenjska raven oz. raven besednega reda v metodologijo ni vključena, zato je zanimiva primerjava zastopanosti identificiranih besednorednih problemov glede na avtomatsko oceno (ne)standardnosti omenjenih dveh ravnin. Primerjave z jezikovno (ne)standardnostjo prikazuje Slika 6.

11 Za preverjanje tega izhodišče smo namenoma ohranili označevalne smernice ohlapne, brez povzetka besednorednih pravil oz. ugotovitev iz referenčnih jezikovnih priročnikov, na kaj naj bodo označevalci pozorni. Če bi označevalne poskuse nadaljevali, bi lahko smernice dopolnili, vendar se s tem proces spremeni v pripisovanje kategorij od zgoraj navzdol, s čimer tvegamo, da določenih pojavljajočih se jezikovnih realnosti v gradivu ne identificiramo.



Slika 6: Zastopanost kategorij glede na pripisani oznaki za jezikovno (ne)standardnost.

Avtomatski pripis jezikovne (ne)standardnosti seveda ni povsem zanesljiv, kljub temu pa rezultati kažejo zanimivo sliko, iz katere je mogoče sklepati o razmerju med zaznamovanimi in potencialno nestandardnimi značilnosti korpusa Janes--Syn. Kot nestandardni so denimo označeni vsi primeri, kjer se povedek pojavlja na koncu stavka (*zdajle bi pa plačala, da bi lahko **tekme gledala***), kjer se pojavlja izpostavljanje osebnih zaimkov (*od kdaj **je nama** tako kul hladno vreme?*) ter oznake razvrstitve elementov znotraj besednih zvez (***Moram poguglati***). K skupinam, kjer nestandardni primeri znatno prevladujejo, je mogoče prišteti še težave vrstnega reda v naslonskem nizu (*vsi ostali **smo pa** hoteli dremati*) in vrivanje stavčnih členov v sestavljeni povedek (*šetamo po Lj. in pride **Zoki** mimo*). Najbolj uravnoteženi glede na oznake (ne)standardnosti sta kategoriji členitve po aktualnosti (*glej, ni druge kot, da jo unfollowamo **vsi***) in naslonke na prvem mestu stavka (***so se** lj. idioti med seboj sfajtali*). Edina kategorija, kjer je prisotnih več označb iz jezikovno standardnih primerov, je atipično mesto povedka, predvsem na prvem mestu stavka (***Moram se danes zvečer** dol usesti*).

S primerjavo Slik 5 in 6 vidimo, koliko je določena identificirana besednoredna značilnost opazna v svoji zaznamovanosti ter kakšna je njena distribucija glede na avtomatsko identificirane nestandardne prvine na drugih jezikovnih ravneh. Za najpogosteje zastopano kategorijo, členitev po aktualnosti, je značilno, da prinaša pretežno pomensko-poudarjalne probleme. To dejstvo pojasni visok delež individualnih označb ter enakomerno pojavljanje v jezikovno nestandardnih, kot tudi standardnih tvitih. Visoko število oznak na drugi strani priča o tem, da izbira centralne vsebinske točke pri upovedovanju ni enoznačna naloga oz. da je možnih več interpretacij, ki pa s stališča označevalcev niso enakovredno nevtralne. Rezultati so torej zelo zanimivi tudi v luči konceptov proste in stalne stave: v rezultatih se pojavlja veliko označb na ravni proste stave, na drugi strani pa neskladnosti z jezikoslovno identificiranimi »stalnostmi« niso označevane dosledno.

Zanimiva ugotovitev je, da se med najpogosteje označenimi značilnostmi pojavlja raba naslonke na prvem mestu stavka. Distribucija pojavitev na drugi strani kaže, da se takšna stava pojavlja enakomerno v nestandardnih ter standardnih tvitih. To jezikovno značilnost bi bilo torej mogoče (nekoliko provokativno) izpostaviti kot tisto, ki je v komunikaciji, ki skuša biti standardna, najboljši indikator, da temu ni tako. Prav tako zanimiva je kategorija postavljanja členkov, ki je (vsaj v teoriji) precej enoznačno, v podatkih pa se kaže precejšnja označevalna permisivnost, skupaj z visokim deležem pojavitev v standardnih besedilih ob sicer prevladujočih nestandardnih. Na tej osnovi lahko oblikujemo tezo, da v jezikovni rabi stava členka manj vpliva na razumevanje pomena, kot bi si morda mislili. Naj členek stoji ob delu stavka, ki ga modificira, ali nekje drugod v neposredni bližini, v sporočilu ga osmislimo po principu največje verjetnosti, skladenjsko neustrezne pozicije pa pri tem pogosto sploh ne opazimo.

Obe izpostavljeni vprašanji bi bilo v razpravi mogoče povezati z značilnostmi govorjenega jezika, za kar pa bi bilo treba vključiti podatke iz govornega korpusa, kar presega domet poglavja. Za primerjavo in nadaljnje analize so relevantne tudi identificirane kategorije, kjer si je mogoče misliti poseben stavčni poudarek v primeru, da bi bilo besedilo govorjeno, pri čemer se osebek in predmet izpostavljata na atipičnih pozicijah. Glede na rezultate raziskave se stava povedka na zadnje mesto stavka, tj. za (predvideno) intonacijsko izpostavljeni osebek ali predmet, pokaže kot najboljši indikator besednoredne nestandardnosti: pojavlja se izključno v besedilih, označenih za nestandardne, obenem pa so ti primeri za označevalce opazno zaznamovani.

4 SKLEP

V prispevku smo opisali pripravo korpusa skladenjsko označenih tvitov Janes-Syn in odločitve, ki smo jih pri tem sprejeli glede označevanja nestandardnih jezikovnih značilnosti. Korpus je mogoče uporabiti kot učno množico za učenje razčlenjevanja računalniško posredovane slovenščine. Prvi poskus, na osnovi katerega bo mogoče pripraviti večjo količino ročno pregledanega gradiva, bo izpostavil morebitne pomanjkljivosti glede označevanja in omogočil optimizacijo smernic. Izbiro jezikovnospecifičnega označevalnega sistema JOS utemeljuje obstoj in lahka dostopnost potrebne infrastrukture za izvedbo dane naloge, v prihodnje pa bo treba več pozornosti posvetiti pretvorljivosti in izmenljivosti podatkov. Predvideno je, da se bodo v prihajajočem obdobju vprašanja sistema označevanja rešila na ravni obravnave standardnega jezika, nujno pa je zagotoviti, da se odločitve aplicirajo tudi na nestandardni ravni.

Raziskava besednega reda v korpusu Janes-Syn je potrdila, da gre za tematiko, ki brez dvoma potrebuje dodatno jezikoslovno pozornost. Izkazalo se je, da besednoredne zaznamovanosti označevalci ne razumejo enotno, da so njihovi zaznamki v več kot pol primerih individualni, da pogosto identificirajo zaznamovanost na ravni proste stave in obenem izpuščajo besednoredne atipičnosti na ravni stalne stave, vključno z jezikoslovno jasno definiranimi, npr. stava členka ob modificirani del stavka ali razvrstitev naslonk v nizu. Pogled na podatke z vidika avtomatsko pripisane oznake jezikovne (ne)standardnosti razkrije dodatne ugotovitve v razmerju zaznamovanosti in nestandardnosti, ki bi jih bilo treba v nadaljevanju raziskati z vključitvijo podatkov govorjenega jezika.

Kategorizacija označenih značilnosti osvetljuje področje skladnje računalniško posredovane slovenščine in je relevantno izhodišče za podrobnejše analize identificiranih problemov. Izsledki omogočajo premik razumevanja prototipske besednoredne zaznamovanosti in raziskovalnega fokusa od primerov, kot sta *domovina naša* in *dedec prebrisani* (Toporišič 2008: 31), in v splošnem iz konteksta literarnega jezika v realnost, ki jo izkazuje vsakdanja jezikovna produkcija širše populacije. V tem smislu je jasno, da je edini pristop, ki lahko v resnici stre oreh besednega reda v slovenščini, korpusnojezikoslovni. V nadaljevanju je treba zagotoviti skladenjsko označenost relevantnega korpusnega gradiva (referenčni korpus, govorni korpus, korpus računalniško posredovane komunikacije) ter podobnosti in razlike ugotoviti statistično, s premišljeno metodo, ki upošteva označevalne specifike in sestavo obravnavanih virov. Delo na projektu JANES, ki smo ga predstavili v tem prispevku, je korak v opisano želeno smer.

Zahvala

Priprava korpusa Janes-Syn je ekipno delo, pri katerem gre zahvala za izvedbo predvsem Tomažu Erjavcu. Pri označevanju besednega reda sta sodelovali Aleksandra Rajković in Polona Logar.

Literatura

Arhar Holdt, Špela in Kaja Dobrovoljc, 2015: Zveze samostalnika z nesklonljivim levim prilastkom v korpusih Janes in Kres. Fišer, Darja (ur.): *Zbornik konference Slovenščina na spletu in v novih medijih*. Ljubljana: Znanstvena založba Filozofske fakultete. 4–9.

Arhar Holdt, Špela in Kaja Dobrovoljc, 2016: Vrednost korpusa Janes za slovensko normativistiko. *Slovenščina 2.0* 4/2. 1–37.

Arhar Holdt, Špela, 2016: *Smernice za označevanje z odvisnostnim sistemom JOS: nestandardna slovenščina, v1.0*. Ljubljana: Specifikacije projekta Jezikoslovna analiza nestandardne slovenščine. Dostop: http://nl.ijs.si/janes/wp-content/uploads/2014/09/Janes-skladnja-v1.0.pdf

Arhar Holdt, Špela, Tomaž Erjavec in Darja Fišer, 2017*: CMC training corpus Janes-Syn 1.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1086.

Breznik Anton, 1908. Besedni red v govoru. *Dom in svet* 21. 258–267.

Chanier, Thierry, Céline Poudat, Benoit Sagot, Georges Antoniadis, Ciara R. Wigham, Linda Hriba, Julien Longhi in Djamé Seddah, 2014: The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. *Journal for Language Technology and Computational Linguistics* 29/2. 1–30.

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

Čibej, Jaka, Darja Fišer in Tomaž Erjavec, 2016a: Normalisation, Tokenisation and Sentence Segmentation of Slovene Tweets. *Proceedings of the Workshop on Normalisation and Analysis of Social Media Texts (NormSoMe)*. Portorož: ELRA. 5–10.

Čibej, Jaka, Špela Arhar Holdt, Tomaž Erjavec in Darja Fišer, 2016b: Razvoj učne množice za izboljšano označevanje spletnih besedil. Erjavec, Tomaž in Darja Fišer (ur.): *Proceedings of the Conference on Language Technologies and Digital Humanities.* Ljubljana: Znanstvena založba Filozofske fakultete. 40–46.

Dobrovoljc Helena, 2008: Jezik v e-poštnih sporočilih in vprašanja sodobne normativistike. Košuta, Miran (ur.): *Slovenščina med kulturami, Zbornik Slavističnega društva Slovenije 19.* Celovec: Slavistično društvo Slovenije. 295–314.

Dobrovoljc, Kaja in Joakim Nivre, 2016: The Universal Dependencies Treebank of Spoken Slovenian. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC '16).* Portorož. 1566–1573.

Dobrovoljc, Kaja, Tomaž Erjavec in Simon Krek, 2016: Pretvorba korpusa ssj500k v Univerzalno odvisnostno drevesnico za slovenščino. *Proceedings of the Conference on Language Technologies and Digital Humanities.* Erjavec, Tomaž in Darja Fišer (ur.): *Proceedings of the Conference on Language Technologies and Digital Humanities.* Ljubljana: Znanstvena založba Filozofske fakultete. 190–192.

Dobrovoljc, Kaja, Simon Krek in Jan Rupnik, 2012: Skladenjski razčlenjevalnik za slovenščino. Erjavec, Tomaž in Jerneja Žganec Gros (ur.): *Zbornik Osme konference Jezikovne tehnologije.* Ljubljana: Institut Jožef Stefan. 42–47.

Erjavec, Tomaž, Darja Fišer, Simon Krek in Nina Ledinek, 2010: The JOS linguistically tagged corpus of Slovene. *LREC 2010, 7th International Conference on Language Resources and Evaluations*. Valletta. 1806–1809.

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.

Fišer, Darja, Tomaž Erjavec in Nikola Ljubešić, 2017: The compilation, processing and analysis of the Janes corpus of Slovene user-generated content. Wigham, Ciara R. in Gudrun Ledegen (ur.): *Corpus de communication médiée par les réseaux: construction, structuration, analyse*. Collection Humanités Numériques. Paris: L'Harmattan. V tisku.

Foster, Jennifer, Ozlem Cetinoglu, Joachim Wagner, Joseph Le Roux, Stephen Hogan, Joakim Nivre, Deirdre Hogan in Josef van Genabith, 2011: #hardtoparse: Pos tagging and parsing the twitterverse. Analyzing Microtext. *Papers from the 2011 AAAI Workshop.* 20–25.

Gantar, Polona, 2011: Slovnični in pomenski opisi v leksikalni bazi za slovenščino. Marušič, Franc in Rok Žaucer (ur.): *Zbornik prispevkov s simpozija 2011.* Nova Gorica: Univerza. 17–27.

Gantar, Polona, 2015: *Leksikografski opis slovenščine v digitalnem okolju.* Ljubljana: Znanstvena založba Filozofske fakultete.

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

Holozan, Peter, Simon Krek, Matej Pivec, Simon Rigač, Simon Rozman in Aleš Velušček, 2008: *Specifikacije za učni korpus.* Kamnik: Projekt »Sporazumevanje v slovenskem jeziku« ESS in MŠŠ. http://www.slovenscina.eu/Vsebine/Sl/Kazalniki/K2.aspx

Holozan, Peter. 2013: Uporaba strojnega učenja za postavljanje vejic v slovenščini. *Uporabna informatika* 21/4. 196–209.

Jakop, Nataša, 2008: Pravopis in spletni forumi – kva dogaja? Košuta, Miran (ur.): *Slovenščina med kulturami, Zbornik Slavističnega društva Slovenije 19.* Celovec: Slavistično društvo Slovenije. 315–327.

Jug Kranjec, Hermina, 1981: O pomenski in stilni vlogi besednega reda pri oblikovanju sporočilne perspektive povedi. *Jezik in slovstvo* 42/2-3. 37–42.

Kalin Golob, Monika, 2008: SMS-sporočila treh generacij. Košuta, Miran (ur.): *Slovenščina med kulturami, Zbornik Slavističnega društva Slovenije 19.* Celovec: Slavistično društvo Slovenije. 283–294.

Kaufmann, Max in Jugal Kalita, 2010: Syntactic normalization of twitter messages. *International conference on natural language processing, Kharagpur, India.*

Kranjc, Anja in Marko Robnik Šikonja, 2015: Postavljanje vejic v slovenščini s pomočjo strojnega učenja in izboljšanega korpusa Šolar. Fišer, Darja (ur.): *Zbornik konference Slovenščina na spletu in v novih medijih.* Ljubljana: Znanstvena založba Filozofske fakultete. 38–43.

Kranjc, Simona, 2003: Skladenjska analiza besedil, ki nastajajo v računalniških klepetih. Požgaj Hadži, Vesna (ur.): *Zbornik referatov z Drugega slovensko-hrvaškega slavističnega srečanja.* Ljubljana: Oddelek za slavistiko, Filozofska fakulteta. 69–82.

Kong, Lingpeng, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer in Noah A. Smith, 2014: A dependency parser for tweets. *Proc. of EMNLP*. Doha, Qatar. 1001–1012.

Krek, Simon, Kaja Dobrovoljc, Tomaž Erjavec, Sara Može, Nina Ledinek in Nanika Holz, 2015: *Training corpus ssj500k 1.4.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1052

Krek, Simon, 2015: Standardni in knjižni jezik – drugi poskus. Smolej, Mojca (ur.): *Slovnica in slovar – aktualni jezikovni opis (Obdobja 34)*. Ljubljana: Znanstvena založba Filozofske fakultete. 401–407.

Ledinek, Nina, 2014: *Slovenska skladnja v oblikoskladenjsko in skladenjsko označenih korpusih slovenščine*. Ljubljana: Založba ZRC.

Ljubešić, Nikola, Darja Fišer, Tomaž Erjavec, Jaka Čibej, Dafne Marko, Senja Pollak in Iza Škrjanec, 2015: Predicting the level of text standardness in user-generated content. *10th International Conference on Recent Advances in Natural Language Processing: Proceedings of RANLP 2015 Conference.* Hissar. 371–378.

Može, Sara, 2013: *FrameNet in večjezičnost: kontrastivna analiza glagolov premikanja v slovenščini in angleščini.* Doktorska disertacija. Ljubljana: Filozofska fakulteta.

Michelizza, Mija, 2015: *Spletna besedila in jezik na spletu.* Ljubljana: Založba ZRC.

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

Popič, Damjan, Darja Fišer, Katja Zupan in Polona Logar, 2016. Raba vejice v uporabniških spletnih vsebinah. Erjavec, Tomaž in Darja Fišer (ur.): *Proceedings of the Conference on Language Technologies and Digital Humanities.* Ljubljana, Slovenia. 149–153.

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

Schneider, Nathan, 2015: What I've learned about annotating informal text (and why you shouldn't take my word for it). *The 9th Linguistic Annotation Workshop held in conjuncion with NAACL 2015*. 152–157.

Stabej, Marko, Helena Dobrovoljc, Simon Krek, Polona Gantar, Damjan Popič, Špela Arhar Holdt, Darja Fišer in Marko Robnik Šikonja, 2016: Slovenščina Janes: pogovorna, nestandardna, spletna ali spretna? *Slovenščina 2.0* 4/2. 100–126.

Storrer, Angelika, 2013: Sprachverfall durch internetbasierte Kommunikation? Linguistische Erklärungsansätze – empirische Befunde. *Sprachverfall? Dynamik – Wandel – Variation. Jahrbuch des Instituts für Deutsche Sprache 2013*. De Gruyter Mouton. 171–196.

Toporišič, Jože, 1967: Besedni red v slovenskem knjižnem jeziku. *Slavistična revija* 15/1-2. 251–274.

Toporišič, Jože, 2004: *Slovenska slovnica (SS). Četrta, prenovljena in razširjena izdaja. 2. natis.* Maribor: Založba Obzorja. 667–678.

Toporišič, Jože, 2008: *Stilnost in zvrstnost.* Ljubljana: Založba ZRC.

Vidovič Muha, Ada, 2000: *Slovensko leksikalno pomenoslovje: govorica slovarja.* Ljubljana: Znanstvena založba Filozofske fakultete.

Zupančič, Nataša, 2009: *Korpusna analiza slovenskega jezika na spletnih forumih.* Magistrsko delo. Ljubljana: Filozofska fakulteta.