Ročno označeni korpusi Janes za učenje jezikovnotehnoloških orodij in jezikoslovne raziskave

Jaka Čibej, Špela Arhar Holdt, Tomaž Erjavec, Darja Fišer

Izvleček

V tem poglavju najprej predstavimo splošni postopek in delotok izdelave ročno označenih korpusov (od priprave podatkov, izdelovanja smernic za označevanje, dela z označevalno platformo in poteka označevalne kampanje do pretvorbe v končni format ter objave in distribucije), pri čemer se podrobneje posvetimo največjima tako nastalima korpusoma Janes-Norm (približno 185.000 pojavnic) in Janes-Tag (približno 75.000 pojavnic), katerih glavni namen je izboljšava jezikovnotehnoloških orodij za tokenizacijo, stavčno segmentacijo, normalizacijo, lematizacijo in oblikoskladenjsko označevanje. Drugi del poglavja poda pregled vseh ročno označenih korpusov Janes: poleg že omenjenih Janes-Norm in Janes-Tag še Janes-Syn (skladnja v RPK), Janes-Kratko (pojavi krajšanja v RPK), Janes-Vejica (raba vejice v RPK), Janes-Preklop (preklapljanje koda v RPK) in Janes-Geo (raba nestandardnih jezikovnih prvin v RPK v odvisnosti od regionalnega izvora uporabnikov). V njem na kratko predstavimo vsebino in strukturo vsakega korpusa ter opišemo njegov predvideni namen.

Ključne besede: slovenščina, računalniško posredovana komunikacija, lematizacija, normalizacija, oblikoskladenjsko označevanje, odprti podatki, Text Encoding Initiative, CLARIN.SI

1 UVOD

Ročno označene podatkovne množice so v okviru proučevanja računalniško posredovane komunikacije (RPK) in razvoja jezikovnotehnoloških orodij za računalniško obdelavo naravnega jezika ključnega pomena. Z vidika jezikoslovnih raziskav omogočajo sistematičen in natančen kvantitativni in kvalitativni vpogled v proučevane pojave, z vidika jezikovnih tehnologij (kot je to podrobneje opisano v Ljubešić et al. (2018)) pa služijo kot podlaga za učenje in izboljšavo orodij, ki olajšajo nadaljnje analize na večjih korpusih, avtomatizirajo označevalne postopke (npr. prepoznavanje imenskih entitet in pripisovanje metapodatkov besedilom ali uporabnikom) in izboljšajo natančnost jezikoslovnega označevanja besedil (npr. normalizacija, lematizacija in oblikoskladenjsko označevanje).

Čeprav je slovenska RPK (kot to velja za druge jezike) zbir različnih komunikacijskih praks oz. besedilnih žanrov, je mogoče v njej v splošnem opaziti jezikovne značilnosti, ki se razlikujejo od standardne pisne slovenščine, kakršna je zbrana in reprezentirana tudi v referenčnih korpusih za slovenščino. Orodja za jeziko(slo)vno označevanje, ki so bila razvita na osnovi standardnega gradiva, zato pri jeziku RPK izkazujejo slabšo natančnost. Težave se pojavljajo na vseh označevalnih ravneh, od tokenizacije, stavčne segmentacije, oblikoskladenjskega označevanja in lematizacije do višjih označevalnih ravni, kot je npr. skladnja. Dodaten izziv za optimizacijo označevalnih postopkov predstavlja dejstvo, da številne značilnosti nestandardne slovenščine tudi v jezikoslovnem smislu še niso dobro raziskane, ne same na sebi ne v razmerju do standardnega jezika. Rešitev, ki jo ponuja projekt JANES, je zato razvoj ročno označenih korpusov ciljno za namene učenja jezikovnotehnoloških orodij in proučevanja jezikovnih pojavov v slovenski RPK. S tem se slovenščina pridružuje jezikom, pri katerih so bile potrebe po prilagoditvi označevalnih orodij že identificirane in rešitve uspešno uporabljene v praksi. Med sorodnimi raziskavami za tuje jezike velja npr. izpostaviti izboljšanje oblikoskladenjskega označevanja nemških tvitov (Rehbein et al. 2013) ter avtomatske normalizacije nemških družbenomedijskih besedil (Laarmann-Quante in Dipper 2016, Ueberwasser 2013), ročno označeni korpusi pa so bili uporabljeni tudi za izboljšanje razreševanja anaforičnih sklicev v angleških besedilih (Poesio et al. 2017) in prepoznavanja imenskih entitet (Bontcheva et al. 2017, Benikova et al. 2014).

V poglavju predstavljamo ročno označene korpuse, ki so bili izdelani v okviru projekta. Vsa besedila, ki so vključena vanje, so bila vzorčena iz korpusa Janes, ki je podrobneje predstavljen v Erjavec et al. (2018). V 2. razdelku opišemo splošni postopek različnih stopenj izdelave korpusov, posebno pozornost pa posvetimo največjima ročno označenima korpusoma Janes-Norm in Janes-Tag (Erjavec et al. 2016c), ki služita kot ponazoritvena primera. Temu sledita še kratek pregled

poglavitnih značilnosti vseh tako nastalih korpusov (3. razdelek) ter sklep, v katerem nakažemo možnosti za izboljšave in prihodnje delo.

2 Izdelava ročno označenih korpusov

V tem razdelku predstavljamo postopek izdelave ročno označenih korpusov od priprave podatkov in označevanja do končne pretvorbe v format TEI (*Text Encoding Initiative*).[1] Postopek se je od korpusa do korpusa nekoliko razlikoval glede na kompleksnost problema in število vpletenih označevalcev/razsodnikov, a je v vseh primerih sledil naslednjim stopnjam:

1. priprava podatkov,
2. izdelava tipologije in smernic za označevanje,
3. označevanje oz. označevalna kampanja z razsojanjem,
4. končni izvoz in pretvorba podatkov.

V naslednjih podrazdelkih ta postopek podrobneje predstavimo na primeru označevalne kampanje, katere cilj je bila izdelava ročno označenega korpusa za izboljšanje avtomatskega označevanja slovenske RPK na petih ravneh: tokenizacija, stavčna segmentacija in normalizacija (korpus Janes-Norm) ter dodatno lematizacija in oblikoskladnja (korpus Janes-Tag).

2.1 Priprava podatkov

Besedila, ki sestavljajo ročno označene korpuse, predstavljene v tem prispevku, so bila vzorčena iz korpusa Janes. V prvi fazi izdelave korpusov Janes-Norm in Janes-Tag sta bila izdelana dva vzorca:

1. Kons1, ki ga sestavljajo tviti, in
2. Kons2, ki sestoji iz forumskih sporočil in komentarjev na blogovske zapise ter novice.

Kons1 vsebuje 4.000 tvitov, ki so bili vzorčeni naključno, a z upoštevanjem določenih omejitev: nismo upoštevali tvitov, ki so bili daljši od 120 znakov,[2] saj so ti pogosto okrajšani oz. odrezani pri koncu. Prav tako nismo upoštevali tvitov, ki so jih objavili računi nezasebnih uporabnikov (npr. organizacije,

1 http://www.tei-c.org/index.xml

2 Tviti so bili zajeti v času, ko je bila dolžina tvita omejena na 140 znakov.

agencije, podjetja), saj ti največkrat ne izkazujejo tipičnih značilnosti jezika RPK. Obenem smo poskrbeli, da je vzorec vseboval tako relativno standarden jezik (s čimer smo želeli v korpus vključiti standardne, a kljub temu za RPK značilne jezikovne prvine) kot zelo nestandardnega: v vzorec smo zato dodali po 1.000 tvitov z različnimi kategorijami jezikovne (L1–L3) in tehnične (T1–T3) nestandardnosti, avtomatsko pripisane po metodi, opisani v Ljubešić et al. (2018). Pri tem 1 označuje visoko stopnjo standardnosti, 3 pa visoko stopnjo nestandardnosti. V Kons1 smo vključili štiri kategorije, ki so prispevale po 1000 tvitov: prve tri kategorije (T1L3, T3L1 in T3L3) vsebujejo tvite z najvišjo stopnjo nestandardnosti (tehnične, jezikovne ali obeh), zadnja (T1L1) pa tvite, ki ne kažejo znakov nestandardnosti.

Tudi Kons2 vsebuje 4.000 besedil, vzorčen pa je bil po enakih kriterijih kot Kons1. Ker za razliko od tvitov forumska sporočila in komentarji na novice in blogovske zapise niso omejeni z dolžino, smo upoštevali samo besedila dolžine 20-280 znakov, da bi zagotovili medsebojno primerljivost vzorcev Kons1 in Kons2 glede na dolžino.

Pred ročnim označevanjem sta bila vzorca z obstoječimi orodji (Erjavec 2011; Ljubešić et al. 2014) avtomatsko označena na vseh petih obravnavanih ravneh jezikoslovnih oznak (tokenizacija, stavčna segmentacija, normalizacija, lematizacija in oblikoskladnja).

2.2 Označevalne smernice

Na podlagi ročne analize manjšega podkorpusa z 200 naključno vzorčenimi tviti iz vseh štirih kategorij smo pripravili označevalne smernice, ki obravnavajo tehnične in jezikoslovne vidike označevalnega postopka.

Tehnične smernice (Erjavec et al. 2016a) vsebujejo celotno označevalno shemo v WebAnnu (Eckart de Castilho et al. 2014; platforma je podrobneje predstavljena v razdelku 2.3.1) in predstavljajo splošna navodila za delo s platformo (npr. kako združujemo ali ločujemo pojavnice, kako brišemo nerelevantne ali avtomatsko generirane tvite, kako ravnamo s kompleksnejšimi večplastnimi oznakami), jezikoslovne smernice (Čibej et al. 2016c) pa pojasnjujejo merila, ki jih mora označevalec upoštevati med označevanjem. Smernice za označevanje RPK so v glavnem sledile smernicam za označevanje standardnih (Holozan et al. 2008) in historičnih (Erjavec 2015) slovenskih besedil, a z nekaterimi prilagoditvami, saj gre za drugačen medij. Vidike smernic, ki so najbolj značilni za RPK, na kratko povzemamo v nadaljevanju.

2.2.1 Stavčna segmentacija

Stavčna segmentacija je delitev besedila na stavke.[3] Pri segmentiranju tvitov na stavke smo kot glavni kriterij upoštevali končna ločila (pike, klicaje, vprašaje, tri- ali večpičje ter narekovaje). Tviti pa vsebujejo tudi nekatere druge prvine, ki se lahko sopojavljajo s končnimi ločili oziroma, če končnih ločil ni, same delujejo na podoben način. Te prvine so emotikoni in emojiji (;-), =D, 😊, 😈), ključniki (#justsayin), sklici na uporabniška imena (@avtor) in URL-naslovi (http://t.co/fqVqV92mzc). Ob odsotnosti končnih stavčnih ločil lahko te prvine prevzamejo njihovo vlogo in končajo stavek. Ker se lahko stavek konča tudi s serijo več tovrstnih prvin, kot konec stavka obravnavamo zadnjo prvino:

> Liverpool zasluženo owna Twitter, ampak na vrhu je pa fucking Iago Aspas hahaha :) #nogomet #LFC #SOULIV http://t.co/LCyEvyoVD7 ¶

Če se prvine pojavljajo skupaj z ločilom, jih obravnavamo kot ločen stavek:

> Življenje Je Cirkus. js sm pa cefur. Luka Stigl js sm se poscal v hlace k sm se vidu. bolano. ¶ :) ... ¶ http://t.co/QyzKRZqZnS ¶

2.2.2 Tokenizacija

Pri avtomatski tokenizaciji (tj. delitvi besed na pojavnice) so se pojavile napake, ki jih je bilo treba popraviti ročno. Med najpogostejšimi napakami so bile okrajšave, emotikoni, obrazila in besede, ki so vsebovale ločila. Pri okrajšavah (npr. slov. za »slovenski«) je tokenizator piko pogosto obravnaval kot končno stavčno ločilo in kot posamezno pojavnico. V takšnih primerih je bilo piko treba združiti z okrajšavo.

Emotikoni so se pogosto pojavljali v serijah in brez vmesnih presledkov, zaradi česar jih je tokenizator obravnaval kot ločila in jih delil. V tovrstnih primerih serije nismo delili na posamezne emotikone, temveč smo jih združili v eno pojavnico:

> :) ¶ :) ¶ : ¶ * ¶ * → :):):**
>
> \ ¶ m ¶ / ¶ (¶ - ¶ _ ¶ - ¶) → \m/(-_-)

3 Izraz »stavek« uporabljamo v širšem pomenu, ki zajema logično strukturno celoto. V nekaterih primerih je to poved, v drugih zgolj ena sama (ne)jezikovna prvina (npr. emotikon ali URL-naslov). V primerih konec stavka (oziroma meje med pojavnicami v primerih, ki prikazujejo popravke tokenizacije) označujemo s simbolom ¶.

Podoben pristop smo ubrali pri ločenih obrazilih in besedah, ki so vsebovale ločila:

TV ¶ - ¶ ja → TV-ja

sms ¶ - ¶ i → sms-i

žen ¶ (¶ sk ¶) ¶ am → žen(sk)am

politik ¶ (¶ e ¶ / ¶ o ¶) → politik(e/o)

2.2.3 Normalizacija

Pojem *normalizacija* v našem kontekstu pomeni pripisovanje oblike, ki je po zapisu čim bolj prilagojena standardni. Pojavnice smo normalizirali samo na nivoju zapisa, ne pa denimo na nivoju besedišča (*farbat – farbati*, ne **barvati*) ali skladnje (*nisem bral knjigo – nisem bral knjigo*, ne **knjige*). Na ravni normalizacije sta se za najbolj problematični izkazali dve kategoriji besed, in sicer:

1. nestandardne besede z več različicami zapisa in brez neposredne standardne ustreznice, in
2. tujejezične jezikovne prvine z različnimi stopnjami prevzetosti.

Besede iz prve kategorije (npr. *orng*, *ornk*, *oreng*, *orenk*; *fovš*, *favš*, *fouš*, *fauš*, *fowš*) se najpogosteje pojavljajo le v govorjenem jeziku in v standardni slovenščini nimajo neposredne ustreznice, zato je določanje normalizirane oblike težavno. Normalizirano obliko smo v teh primerih določili s pomočjo dodatnega merila: označevalec je v podkorpusu tvitov korpusa Janes z regularnimi izrazi poiskal vse prisotne različice zapisa, kot normalizirano obliko pa je izbral najpogostejšo (v zgornjih primerih sta to *ornk* in *fouš*).

Na podoben način je bila problematična normalizacija besed iz druge kategorije, ki je vsebovala tujejezične prvine z različnimi stopnjami prilagoditve slovenskemu zapisu in oblikoslovju (npr. *updateati*, *updajtati*, *updejtati*, *apdejtati*). Normalizacija z izvirnimi/citatnimi oblikami (npr. *po-update-ati*) bi v mnogih primerih v korpus uvedla umetne oblike, ki jih v realni jezikovni rabi ni, zato smo pri označevanju tujejezičnih prvin upoštevali naslednji merili:

a) če je bil zapis docela fonetiziran (npr. *danke schön* → *dankešn*, *appreciate* → *apršiejt*), smo besedo obravnavali na enak način kot slovensko nestandardno besedo z več različicami zapisa (glej primer *ornk* zgoraj);

b) če je beseda izkazovala kakršnekoli tujejezične značilnosti (npr. neslovenske črke ali tujejezični zapis), smo normalizirano obliko določili tako, da

smo iz podkorpusa tvitov Janes izbrali najpogostejši zapis med tistimi, ki so še vedno izkazovali tujejezične značilnosti (npr. *updateati*, *updajtati*, *updejtati* → *updejtati*).

Nekaterih prvin, ki so značilne za tvite in RPK (npr. sklici na uporabniška imena, ključniki, URL-naslovi, emotikoni in emojiji), nismo normalizirali, ne glede na to, ali so bile njihove oblike v skladu s standardno ali ne. Pri normalizaciji prav tako nismo popravljali skladnje (npr. nepravilne rabe sklonov ali napak pri ujemanju – niti naključnih), pogostih leksikalnih napak (*moči* – *morati*) ali napak v slogu ali registru (*rabiti* – *potrebovati*).

2.2.4 Lematizacija

Pripisovanje lem je v največji možni meri sledilo smernicam za označevanje korpusa ssj500k (Holozan et al. 2008), ki je v vmesniku SketchEngine služil tudi kot referenčni vir za označevalce. Razlike ali dopolnitve označevalnega sistema zadevajo žanrske specifike označevanih besedil, pri čemer gre izpostaviti tujejezične prvine in raznovrstne kratice, ki se v spletni slovenščini pojavljajo mnogo pogosteje in oblikovno bolj raznorodno kot v standardnem jeziku.

Podobno kot pri normalizaciji je tudi pri lematizaciji med večjimi izzivi označevanja določanje meje med tujejezičnim in slovenskim besediščem. V tvitih se tujejezične prvine pojavljajo kot posamezne besede različnih besednih vrst in variant zapisa (*share*, *shareati*, *share-ati*, *šerati*), kot besedne zveze ali daljši segmenti. Zadnje smo označevali kot niz pojavitev v tujem jeziku, pri čemer so leme enake oblikam, oblikoskladenjska oznaka pa je *Nj*. Podobno velja za občnoimenske besedne zveze (*bonus score*, *sugar rush*) in posamezne besede, ki so v besedilu zapisane brez jasno razvidnih prilagoditev slovenskemu zapisu oz. pregibanju (*jailbreak*, *hrvatskog*).

Pri besedah, ki prilagoditev izražajo, smo lemo določili v skladu s slovenskimi oblikoslovnimi načeli (*benchmarki* → *benchmark*, *chatala* → *chatati*). Pri odločanju, ali besedo obravnavati kot tujejezično ali prevzeto, so bili uporabljeni tudi referenčni leksikalni viri, predvsem SSKJ[4] in SNB[5] ter leksikon besednih oblik Sloleks. Vprašanja uvrščanja kratičnih poimenovanj med kratice in okrajšave na eni strani ter občna (*lol*, *drž.*) in lastna imena (*Sds*, *Slo.*) na drugi so bila razrešena že na ravni normalizacije, označevalci pa so v teh primerih pri pripisovanju lem (in oblikoskladenjskih oznak) sledili normaliziranim oblikam.

4 http://www.fran.si/130/sskj-slovar-slovenskega-knjiznega-jezika

5 http://www.fran.si/131/snb-slovar-novejsega-besedja

Projektnospecifična je še odločitev, da se URL-naslovi lematizirajo v domeno (*http://t.co/ZaVQdnaN5p → t.co*), s čimer omogočimo preglednejše prikazovanje korpusnih podatkov v vmesniku. Pri ostalih tviterskih prvinah (uporabniška imena, ključniki, emotikoni) je lema enaka obliki.

2.2.5 Oblikoskladnja

Tudi na oblikoskladenjski ravni so bile osnovno izhodišče za označevanje smernice korpusa ssj500k po sistemu za oblikoskladenjsko označevanje JOS. Med razlikami gre v prvi vrsti omeniti širitev sistema z naslednjimi novimi oznakami: *Nh* za ključnike; *Nw* za URL- in e-naslove; *Na* za sklice na uporabniška imena; in *Ne* za emotikone in emojije. Z naštetimi oznakami in načelom lematizacije, pri katerem lema sledi izvorni obliki, smo na enostaven in sledljiv način rešili vprašanje označevanja tvitersko specifičnih prvin. Glede na sistem JOS smo uvedli še eno pomembno novo oznako, in sicer za ločila: v sistemu JOS za ločila ni bila predpisana specifična oznaka (v formatu XML/TEI so bila identificirana s posebnim elementom, ne pa tudi z posebno oznako), v projektu JANES pa smo za to uporabili oznako *U*. Vse uporabljene oznake sledijo sistemu MULTEXT-East V5 (v izdelavi), dostopne pa so na http://nl.ijs.si/ME/V5/msd/.

Pri ročnem označevanju nista bili uporabljeni oznaki *Nt* (zatipkana beseda) ter *Np* (tokenizacijska napaka), saj so bile tovrstne težave ročno odpravljene že na ravni normalizacije in tokenizacije.

Zaradi specifik tviterske komunikacije se je pri označevanju pojavljalo večje število pomensko nejasnih oz. dvoumnih primerov (npr. *dobr* kot pridevnik ali prislov). Kot je to veljalo za označevanje ssj500k, so označevalci take primere interpretirali in označili po principu najverjetnejše možnosti. Podobno načelo je veljalo za označevanje samocenzuriranih besed (*v p***i → Sozem*). V primeru odstopov od norme na skladenjski ravni so bile oznake pripisane skladno z dejansko (in ne pričakovano) pojavitvijo. Tipični tovrstni primeri so na ravni rabe sklonov (*nisem oblikovala intergalaktično brisačo → Sozet*, ne *Sozer*), števila (*Z Martino smo se tekmovale → Ggnd-mz*, ne *Ggnd-dz*) in rabe kategorije živosti (*jaz vem za kvalitetnega centra z nba izkušnjami → Sometd*, ne *Sometn*).

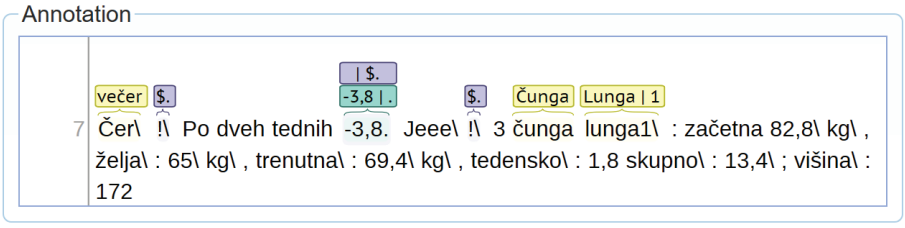
Nazadnje je treba omeniti še označevanje zaprtih besednih vrst, ki je skladno z načeli ssj500k v izhodišču potekalo leksikonsko pogojeno, a z možnostjo dodajanja nestandardnega besedišča. Kategorija, ki je na ta način dobila največ novih elementov, je členek (npr. *eto*, *evo*, *ajde*, *naka*, *kao*, *glih*/*lih* in *ta* v primerih tipa *ta star*). Pri drugih kategorijah se potreba po dopolnitvi pojavlja redkeje, npr. z veznikom *samo* (kot v primeru *Nism še vidu, sam so rekl da je dobr*).

2.3 Postopek ročnega označevanja

Ker so bila obstoječa orodja za obdelavo besedil naučena na standardni slovenščini in bi se na nestandardni odrezala bistveno slabše, je bilo za večjo natančnost pri lematizaciji in oblikoskladenjskem označevanju besedil ključno, da v vzorcih Kons1 in Kons2 najprej ročno popravimo napake v tokenizaciji in stavčni segmentaciji ter obenem normaliziramo besede z nestandardnim zapisom. Pri označevanju smo zato v prvi fazi ročno popravili avtomatsko označene nivoje tokenizacije, stavčne segmentacije in normalizacije, zatem pa smo ročno popravljeni podmnožici ponovno uvozili v označevalno orodje kot vzorca Kons1-MSD in Kons2-MSD ter ročno popravili še napake na ravneh lematizacije in oblikoskladenjskih oznak. V podmnožice, ki smo jih izbrali za drugo fazo označevalne kampanje, smo vključili več nestandardnih besedil kot standardnih, s čimer smo v korpusu želeli zagotoviti čim več prvin, ki so značilne za RPK.

2.3.1 Označevalna platforma

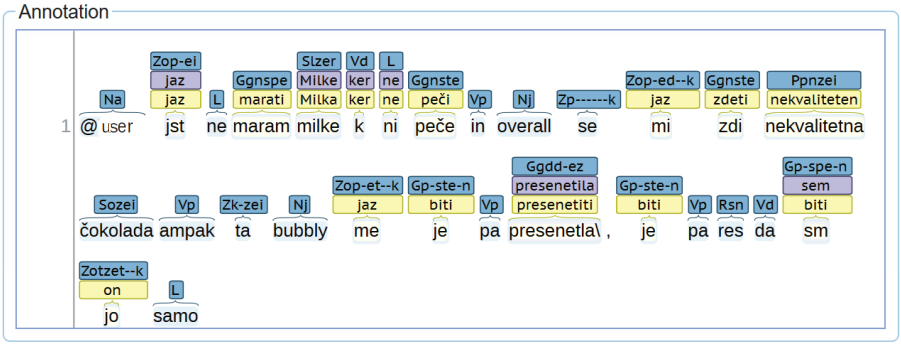
Označevanje je potekalo v spletni označevalni platformi WebAnno (Eckart de Castilho et al. 2014), ki med drugim omogoča večplastno označevanje, vsaki plasti pa lahko pripišemo tudi več kot eno vrednost. Slabost orodja je ta, da ni namenjeno popravljanju tokenizacijskih napak, zato je to precej zapleteno. Iz tega razloga smo poleg tokenizacijske plasti z več vrednostmi v orodje uvedli tudi nabor posebnih simbolov, s pomočjo katerih smo lahko ločevali in združevali pojavnice ter določali meje med stavki. Primer prikazuje Slika 1: tokenizator je v tem primeru niz »-3,8.« napačno obravnaval kot eno pojavnico, označevalec pa jo je ločil na dve in dodal stavčno mejo na drugo pojavnico (pika). S poševnicami smo zaznamovali, da v izvirnem besedilu med pojavnicama ni bilo presledka.[6]



Slika 1: Popravljanje mej med pojavnicami in stavki v WebAnnu (oznake za normalizacijo so označene z rumeno, za tokenizacijo z zeleno in za stavčno segmentacijo z vijolično).

6 Če je bila poševnica prisotna tudi v izvirnem besedilu, je bila v označevalni platformi prikazana kot *\$0*. Na ta način smo ločili prave poševnice od tistih, s katerimi smo zaznamovali pomanjkanje presledkov med pojavnicami.

Slika 2 prikazuje primer z oznakami za oblikoskladnjo (modra), lematizacijo (rumena) in normalizacijo (vijolična):



Slika 2: Označevanje lem in oblikoskladenjskih oznak v WebAnnu.

Platforma omogoča tudi t. i. razsojanje, pri katerem razsodnik sprejme dokončno odločitev v primerih, kjer prihaja do razhajanja med različnimi označevalci; primer je podan v Sliki 3.

Slika 3: Razsojanje v WebAnnu.

2.3.2 Označevalna kampanja

Označevalne kampanje vseh korpusov so se razlikovale v številu označevalcev in obsegu označevanega gradiva, a so potekale na podoben način, in sicer v več fazah. V nadaljevanju predstavljamo pregled in opis označevalne kampanje za korpusa Janes-Norm in Janes-Tag, ki je zajemala tri stopnje:

a) NTS-Kons1 – normalizacijo, tokenizacijo in stavčno segmentacijo vzorca Kons1 (od decembra 2015 do marca 2016);

b) NTS-Kons2 – normalizacijo, tokenizacijo in stavčno segmentacijo vzorca Kons2 (od marca 2016 do maja 2016); in

c) LO-Kons1&2 – lematizacijo in oblikoskladenjsko označevanje vzorcev Kons1 in Kons2 (od marca 2016 do oktobra 2016).

Ob začetku prvega dela označevalne kampanje (NTS-Kons1) smo priredili dvodnevno delavnico, na kateri so se označevalci seznanili z delom v WebAnnu in z označevalnimi smernicami. Na delavnici je sodelovalo 11 študentov jezikoslovnih smeri na magistrski stopnji.

Teoretičnemu uvodu v WebAnno s praktičnim delom in predstavitvi smernic je sledila uvajalna označevalna faza, med katero so udeleženci označili manjše število tvitov. Cilji označevanja so bili naslednji:

a) vsak tvit mora biti pravilno razdeljen na stavke;

b) vsak tvit mora biti pravilno razdeljen na pojavnice; in

c) vse pojavnice morajo imeti pripisano normalizirano obliko; dvoumne pojavnice ohranijo izvirno, nenormalizirano obliko.

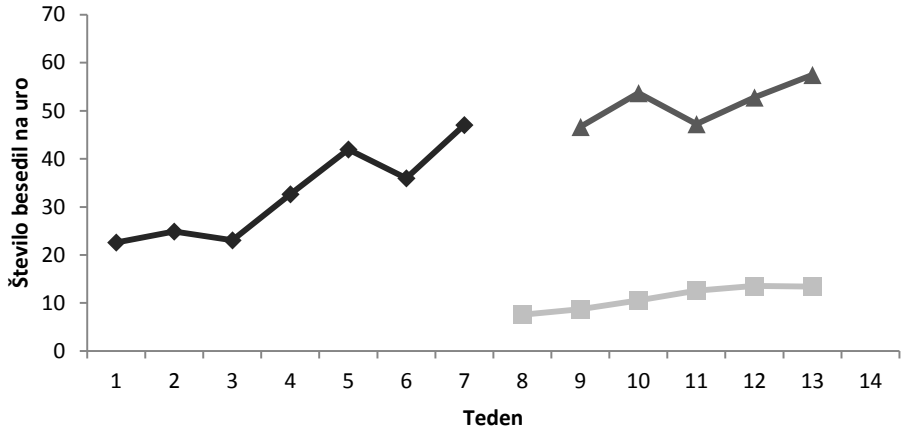
Uvajalni označevalni fazi je sledila diskusija, na kateri smo z označevalci razpravljali o njihovih odločitvah in razhajanjih med njihovimi oznakami, podali pa smo tudi pravilne rešitve in razloge zanje, da bi čim bolj uskladili odločitve označevalcev in izboljšali njihovo ujemanje. V drugem delu kampanje (LO-Kons1) smo na enodnevni delavnici označevalce seznanili s konceptom oblikoskladenjskih oznak in lem ter jim predstavili smernice. Tudi tej delavnici je sledila uvajalna faza, cilj pa je bil tokrat vsaki pojavnici v tvitu (z izjemo ločil) pripisati ustrezno lemo in oblikoskladenjsko oznako. Odločitve smo skupaj prediskutirali in utemeljili z načeli iz smernic.

Obema delavnicama je sledila preizkusna faza. V delu NTS-Kons1 smo označevalce razdelili v dve skupini po 5 oz. 6 označevalcev, vsaki skupini pa smo dodelili 100 tvitov iz preizkusne množice, pri katerih so morali popraviti avtomatsko pripisane oznake in dodati nove, kjer je bilo to potrebno. Pri oblikoskladenjskih oznakah in lematizaciji smo označevalce razdelili v pare, vsak par pa je označil po 50 tvitov iz preizkusne množice. Oznake sta nato ročno preverila razsodnika, ki sta ocenila tudi natančnost označevalcev. Na podlagi rezultatov sta bila v delu NTS-Kons1 iz kampanje izključena dva nezanesljiva označevalca, v obeh delih pa sta razsodnika po začetni evalvaciji dopolnila smernice za označevanje še s primeri, ki so se v preizkusni seji izkazali za problematične.

Delotok označevanja je vključeval skupino označevalcev in dva razsodnika z dobrim poznavanjem smernic za označevanje. Razsodnika, ki sta bila zadolžena tudi za vodenje označevalne kampanje, sta v tedenskih fazah posamezni skupini

označevalcev[7] dodelila določeno število datotek, po koncu vsake faze pa sta oznake ročno preverila in, če je bilo potrebno, označevalcem podala konstruktivno povratno informacijo ter na ta način odstranila najpogostejše oz. najresnejše napake. Če so označevalci med delom naleteli na posebno problematično dilemo, so bile z njo dopolnjene tudi smernice za označevanje. Ustvarjen je bil tudi e-poštni seznam, na katerem so lahko označevalci razsodnikoma zastavljali vprašanja in razreševali problematične ali dvoumne primere, ki niso bili vključeni v smernice.

Med delom smo spremljali učinkovitost označevalcev, tako da smo v vsaki fazi merili razmerje med časom označevanja in številom označenih besedil (glej Sliko 4).



Slika 4: Učinkovitost označevalcev pri označevanju vzorcev Kons1 in Kons2. Črni del predstavlja NTS-Kons1, temno sivi NTS-Kons2 in svetlo sivi LO-Kons1&2.

Z grafa je razvidno, da normalizacija, tokenizacija in stavčna segmentacija potekajo mnogo hitreje od lematizacije in oblikoskladenjskega označevanja. Dobro usposobljeni označevalci lahko v eni uri normalizirajo med 45 in 55 besedil, lematizirajo in oblikoskladenjsko označijo pa le nekaj nad 10 besedil. Na tej točki je treba znova poudariti, da so označevalci preverjali in popravljali napačne avtomatsko pripisane oznake. Označevanje brez predhodnih avtomatskih postopkov bi po vsej verjetnosti vzelo bistveno več časa.

7 Na začetku so bili označevalci razdeljeni v skupine po 3, pozneje pa v pare. Zelo natančni označevalci so v nekaterih fazah označevali tudi posamezno.

2.4 Izvoz in končni zapis označenega korpusa

Posebno pozornost smo namenili formatu podatkov, da bi vse ročno preverjene oznake združili v enovit zapis. Pri zapisu korpusa Janes uporabljamo priporočila za kodiranje besedil TEI, ki so v uporabi tudi pri večini obstoječih slovenskih korpusov. Ker WebAnno formata TEI ne podpira, smo med razpoložljivimi formati izbrali tabelarični format TSV, v katerem je vsaka pojavnica skupaj z identifikatorjem in vsemi oznakami zapisana v svoji vrstici.

Izdelali smo pretvornik med izvornim formatom TEI in formatom TSV, ki omogoča ciklično izvažanje in uvažanje ter združevanje oznak v skupni TEI. Rezultat tako vsebuje vse oznake izvornega TEI, dopolnjene s popravki ročnega označevanja, pretvornik pa je uporaben tudi za druge označevalne kampanje z drugim naborom oznak.

Končne različice ročno označenih korpusov so torej kodirane kot datoteke XML v formatu TEI P5 (TEI Consortium), ki vključujejo kolofon TEI z metapodatki o korpusu ter telo, ki ga sestavljajo anonimni bloki (<ab>), od katerih vsak vsebuje po eno besedilo. Poleg tega vsak dokument vsebuje tudi oblikoskladenjske specifikacije, ki so kodirane kot knjižnica struktur lastnosti TEI. To omogoča, da oblikoskladenjsko oznako razgradimo na posamezne sestavne dele (pare atributov in vrednosti) oz. da jo lokaliziramo v slovenščino.

```
<ab xml:id="janes.blog.publishwall.4264.3" type="blog" subtype="T1L3">
   <s>
      <w lemma="kaj" ana="#Rgp">Kaj</w><c> </c>
      <w lemma="biti" ana="#Va-r3s-y">ni</w><c> </c>
      <w lemma="ta" ana="#Pd-nsn">to</w><c> </c>
      <choice>
         <orig><w>tazadnje</w></orig>
         <reg>
            <w lemma="ta" ana="#Q">ta</w><c> </c>
            <w lemma="zadnji" ana="#Agpnsn">zadnje</w>
         </reg>
      </choice><c> </c>
      <choice>
         <orig><w>AAjevska</w></orig>
         <reg><w lemma="aa-jevski" ana="#Agpfsn">AA-jevska</w></reg>
      </choice><c> </c>
      <w lemma="molitev" ana="#Ncfsn">molitev</w>
      <pc ana="#Z">?</pc>
   </s>
</ab>
```

Slika 5: Izsek iz XML TEI 5.

Kot prikazuje Slika 5, je vsak element <ab> (tj. besedilo) označen s svojo identifikacijsko oznako/kodo iz korpusa Janes, z vrsto vira (tviti, komentarji na novice, forumska sporočila ali blogovski zapisi) in s kategorijo standardnosti (T1L1, T1L3, T3L1 ali T3L3), vsak blok pa vsebuje zaporedne stavke (<s>) iz besedila. Pojavnice so kodirane kot besede (<w>) ali ločila (<pc>), izvirni »jezikovni« presledki pa so ohranjeni z elementom TEI za znak (»character«, <c>). Pojavnice so označene z oblikoskladenjskimi oznakami, ki so kazalci na svoje definicije v knjižnici struktur lastnosti, besede pa so označene tudi z lemami.

Za kodiranje standardne oblike besed z nestandardnim zapisom smo uporabili element TEI <choice> z dvema podrednima elementoma za izvirno (<orig>) in normalizirano obliko (<reg>). Ta pristop ima to prednost, da omogoča večbesedne preslikave in razlikuje med jezikoslovnimi oznakami izvirnika in normalizirane oblike. Trenutno označujemo samo normalizirane oblike.

Kodiranje TEI smo nato prevedli v vertikalni format CQP, ki ga uporablja Sketch Engine (Rychlý 2007), korpus pa namestili na NoSketch Engine (instalacija CLARIN.SI).

2.5 Varovanje osebnih podatkov in avtorskih pravic

Vsi ročno označeni korpusi so po obsegu majhni in ne vsebujejo občutljivih osebnih podatkov, zato jih ne dojemamo kot problematične z vidika varovanja osebnih podatkov ali avtorskih pravic (Erjavec et al. 2016b). Besedil, ki so vključena v korpuse, nismo anonimizirali, v malo verjetnem primeru pritožb pa bomo posamezna problematična besedila odstranili iz javno dostopnih korpusov.

3 ROČNO OZNAČENI KORPUSI

V nadaljevanju predstavljamo vse ročno označene korpuse, ki so bili izdelani v okviru projekta JANES. Delimo jih na dve kategoriji:

a) korpuse za učenje jezikovnotehnoloških orodij in

b) korpuse za jezikoslovne raziskave.

Poleg korpusov Janes-Norm in Janes-Tag, ki smo se jima posvetili že ob opisu postopka izdelave ročno označenih korpusov, v kategorijo korpusov za učenje jezikovnotehnoloških orodij uvrščamo še skladenjsko označeni Janes-Syn, med korpuse za jezikoslovne raziskave pa štejemo Janes-Kratko, Janes-Vejica, Janes-Preklop in Janes-Geo.

Vsi korpusi so prosto dostopni pod licenco CC BY-SA 4.0 na repozitoriju CLARIN.SI. Povezave so navedene v ustreznih podrazdelkih.

3.1 Korpusi za učenje jezikovnotehnoloških orodij

V tem podrazdelku podrobneje predstavimo korpuse za učenje jezikovnotehnoloških orodij Janes-Norm, Janes-Tag in Janes-Syn. Prvi dve sta bili v okviru projekta za te namene že uporabljeni, kar je podrobneje opisano v Ljubešić et al. (2018).

3.1.1 Janes-Norm

Janes-Norm vsebuje besedila iz vzorcev Kons1 in Kons2, njegova glavna vloga pa je učenje in preizkušanje orodij za tokenizacijo, stavčno segmentacijo in normalizacijo slovenske RPK. Tabela 1 prikazuje velikost korpusa oz. njegovih delov glede na različne stopnje standardnosti in besedilne žanre.

Tabela 1: Velikost in sestava korpusa Janes-Norm.

Janes-Norm												
	Besedila	%	Pojavnice	%	Besede	%	Norm.	%	Prave norm.	%	Večbesedne norm.	%
Vse	7.816	100,0	184.755	100,0	142.848	100,0	39.304	27,5	16.604	42,2	815	4,9
T1L1	1.979	25,3	48.437	26,2	37.666	26,4	7.883	20,9	795	10,1	78	9,8
T1L3	1.936	24,8	47.426	25,7	34.861	24,4	12.609	36,2	6.566	52,1	239	3,6
T3L1	1.954	25	41.472	22,4	33.071	23,2	6.458	19,5	1.018	15,8	153	15
T3L3	1.947	24,9	47.420	25,7	37.250	26,1	12.354	33,2	8.225	66,6	345	4,2
Blogi	1.159	14,8	20.981	11,4	16.258	11,4	3.567	21,9	1.621	45,4	87	5,4
Forumi	1.572	20,1	37.647	20,4	30.960	21,7	7.556	24,4	3.796	50,2	214	5,6
Komentarji	1.145	14,6	23.489	12,7	19.083	13,4	4.628	24,3	1.880	40,6	92	4,9
Tviti	3.940	50,4	102.638	55,6	76.547	53,6	23.553	30,8	9.307	39,5	422	4,5

V celoti korpus vsebuje 7.816 besedil, ki so relativno enakomerno razporejena po štirih vključenih kategorijah (ne)standardnosti. Omeniti je treba, da je vrstni red besedil tako v korpusu Janes-Norm kot v korpusu Janes-Tag naključen, saj je korpus tako lažje razdeliti na učno in testno množico ter obenem zajeti vse besedilne tipe. Korpus ne vsebuje vseh 8.000 vzorčenih besedil (4.000 iz Kons1 in 4.000 iz Kons2) oz. 2.000 besedil za vsako od kategorij standardnosti, saj so označevalci imeli možnost, da posamezna besedila označijo kot nerelevantna (npr. če so bila avtomatsko generirana, povsem nerazumljiva, brez kakršnihkoli jezikovnih prvin ali v celoti v tujem jeziku). Teh besedil v končno različico korpusa nismo vključili.

Skupaj besedila vsebujejo skoraj 185.000 pojavnic oz. 144.000 besed (pri čemer kot besedo štejemo vse pojavnice razen ločil, številk in specifičnih elementov RPK, kot so npr. e-naslovi, URL-naslovi, ključniki, sklici na uporabniška imena ter emojiji in emotikoni). Iz Tabele 1 je razvidno, da so deleži med kategorijami standardnosti večinoma ohranjeni tudi pri pojavnicah in besedah. Glede na besedilne tipe približno polovica besedil, pojavnic in besed izvira iz tvitov, po okrog 12 % besedil pa iz blogovskih zapisov in komentarjev na novice.

Stolpec Norm. prikazuje delež normaliziranih pojavnic glede na skupno število besed. Stolpec Prave norm. podaja delež jezikovno kompleksnejših normalizacij glede na vse normalizirane besede. Med jezikovno kompleksnejše normalizacije štejemo vse popravke, ki ne vključujejo zgolj kapitalizacije (*slovenija* → *Slovenija*) ali rediakritizacije (*macka* → *mačka*). Kot je razvidno, je bila normalizirana več kot četrtina besed (27,5 %), 42 % od teh pa je vključevalo kompleksnejše normalizacije. Kot lahko pričakujemo, standardna besedila vsebujejo mnogo manj normaliziranih besed, stopnja jezikovne standardnosti (L) pa korelira s potrebo po normalizaciji. Glede na žanr so v korpusu Janes-Norm na splošno najbolj standardni komentarji na bloge (21,9 %), sledijo pa jim forumska sporočila in komentarji na novice. Največji delež besed, ki jih je bilo treba normalizirati, izkazujejo tviti (30,8 %).

Slika je nekoliko drugačna, če opazujemo samo jezikovno kompleksnejše normalizacije, saj je v tvitih tovrstnih normalizacij le 39,2 %, v komentarjih na novice in bloge 40,6 % in 45,4 %, v forumskih sporočilih pa več kot pol (50,1 %). Iz tega sklepamo, da uporabniki najbolj upoštevajo rabo diakritičnih znamenj na forumih, najmanj pa v tvitih. Vzrok za to je najverjetneje v napravah, s katerih uporabniki objavljajo: forumska sporočila pišejo na računalnikih, tvite pa na telefonih.

Zadnji stolpec podaja število in delež primerov (glede na vse jezikoslovno kompleksnejše normalizacije), pri katerih je normalizacija vključevala ločevanje ali združevanje besed. Kot smo že omenili, je tovrstne normalizacije še posebej težavno modelirati, a rezultati kažejo, da ne gre za pogost pojav, saj zajema le okrog

5 % jezikovnih normalizacij. Povedano drugače, četudi teh primerov sploh ne bi obravnavali, bi končni upad natančnosti ne bil znaten.

Omeniti je treba tudi, da so v Janes-Norm vključene tudi oblikoskladenjske oznake in leme, a so bile za del, ki ni prekriven z Janes-Tag, pripisane le avtomatsko in zato vsebujejo napake.

Janes-Norm je kot podatkovna množica na voljo na repozitoriju CLARIN.SI (Erjavec et al. 2016), iskanje po korpusu pa je omogočeno v konkordančnikih CLARIN.SI, in sicer KonText in NoSketch Engine. Slika 6 prikazuje primer iskanja vseh pojavnic z normalizirano obliko *koliko*.


Query koliko 125 (676.6 per million)

Page 1 of 7 Go Next | Last

janes.forum.medovernet.5809700	moraš uporabit primerno silo Pozdravljeni!	Koliko	dni oziroma mesecev je potrebno jemati
tid.5457101769225175O4	@tomltoml @vinkovasle1 ka pa pol tolk pizdite	kolk	so jih za glavo skrajšali, ..-vaših.. @surfon
tid.5027441568765706 24	kdo ima daljšega v Piranu oz. kdo bo zidal	koliko	in kje. Hot VS Pope let the games begin
tid.5401280977195499 52	komunajzarju", z veseljem ti bom prisluhnil.	Kolko	je blo onih infomatikov v JU, @petrasovdat
janes.blog.publishwall.3702.6	posojajo, pa jim ne bomo nikoli vrnili,...	koliko	smo že dolžni ?? 30 miljard, zato pa moramo
tid.5064982280835235 84	Polom. Fak. Polom. Ja tle sm. Ne. Ne vem	kolk	časa bom tle. Kva ti? Aja. A gre pingvin
janes.forum.avtomobilizem.18.75263.1530748	tut da se avti prodajo tut če jih nevem	kok	spluvamo @chatek hmmm ... florentinca san
janes.forum.medovernet.9184124	fast...tebe bi pa jaz koj mela hahahaha	kok	bi se midva nasmejala Evo, še ena slikca
janes.news.rtvslo.266505.14	sociedad je kr en vaški klub.. a ne? '..	kolk	je že blo? je pa itak "obvezna sestavina
tid.4231098289074995 21	in ste že začeli z antipropagando? jao,	kolk	ste poceni.. @Sraqa Sorry :) We need to
tid.3897948595487617 024	@PrimozP @PEroCaks pojma nimam. Toliko da lahko	kolko	tolko normalno oddajajo. Videl sem razmere
tid.3862089329951498 25	potem si Janko ne bo mogel več zmišljevati,	koliko	so vredni slabi krediti, ki bodo prenešeni
janes.news.mladina.163356.9	prebrati več kot komentar pod člankom in tvit?	Koliko	strani že ima ta popravek zgodovine? Več
janes.news.rtvslo.301991.1317	http://t.co/lxLrPc87pp gapponJa vidimo,	kolko	jih podpira 3500, 3500 - mogoče 5000 jih
tid.5046017830341877 76	in vodi. @m4tija cuj nism :-) ne vem od	kok	jim je uspel... sm pa ze zamenju geslo
tid.3869063368079114 24	sem bil v paru z Mertljem, ne glede nato,	koliko	mislijo, da ne moreva skupaj igrat. " @NK_MARIBOR
janes.news.rtvslo.306503.3	mogoče da je res okusno... sam me zanima	kolk	porcij bi mogu pojest en delavc k pride
janes.news.rtvslo.344929.19	gledam studio na planet tv pa ne morem verjet	kolk	se je dani bavec postaral... sploh ne zgleda

Slika 6: Iskanje pojavnic z normaliziranо obliko *koliko* v korpusu Janes-Norm preko vmesnika NoSketch Engine.

3.1.2 Janes-Tag

Janes-Tag je podmnožica korpusa Janes-Norm in vsebuje vzorca Kons1-MSD in Kons2-MSD, zasnovan pa je bil kot zlati standard za lematizacijo in označevanje oblikoskladenjskih oznak oz. za učenje in preizkušanje slovenskih oblikoskladenjskih označevalnikov in lematizatorjev (več o tem v poglavju Ljubešić et al. (2018)). Različica 2.0 poleg oblikoskladenjskih oznak in lem vsebuje še oznake lastnih imen, zato se jo lahko uporabi tudi kot učno množico za prepoznavalnike imenskih entitet v slovenski RPK.

Tabela 2 prikazuje velikost celotnega korpusa Janes-Tag in njegovih podkategorij glede na stopnjo standardnosti in vključene žanre.

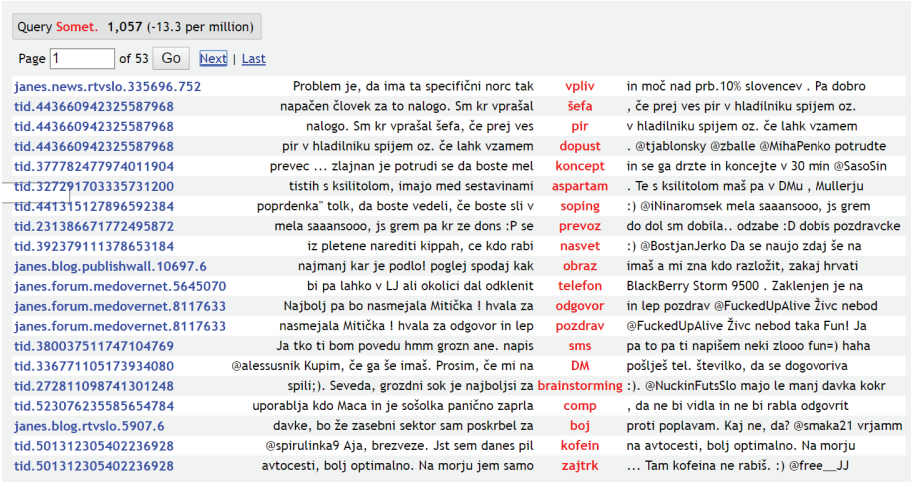
Tabela 2: Velikost in sestava korpusa Janes-Tag.

Janes-Tag														
	Besedila	%	Imena	%	Pojavnice	%	Besede	%	Norm.	%	Prave norm.	%	Večbesedne norm.	%
Vse	2.958	100,0	4.780	100,0	75.276	100,0	56.555	100,0	18.829	33,3	10.103	53,7	379	3,8
T1L1	275	9,3	254	5,3	6.695	8,9	5.400	9,5	954	17,7	77	8,1	11	14,3
T1L3	1.219	41,2	2.040	42,7	32.329	42,9	23.159	40,9	8.762	37,8	4.447	50,8	150	3,4
T3L1	245	8,3	159	3,3	4.559	6,1	3.788	6,7	589	15,5	126	21,4	12	9,5
T3L3	1.219	41,2	2.327	48,7	31.693	42,1	24.208	42,8	8.524	35,2	5.453	64,0	206	3,8
Blogi	269	9,1	180	3,8	5.046	6,7	3.952	7,0	848	21,5	370	43,6	24	6,5
Forumi	403	13,6	211	4,4	9.445	12,5	7.761	13,7	1.894	24,4	935	49,4	46	4,9
Komentarji	303	10,2	339	7,1	6.097	8,1	4.801	8,5	1.250	26	522	41,8	20	3,8
Tviti	1.983	67,0	4.050	84,7	54.688	72,6	40.041	70,8	14.837	37,1	8.276	55,8	289	3,5

Korpus v celoti vsebuje nekaj manj kot 3.000 besedil in dobrih 75.000 pojavnic oz. 56.000 besed, torej je približno pol manjši od korpusa Janes-Norm. Korpus kot učna množica torej ni velik (za primerjavo: zajema približno desetino velikosti ročno označenega učnega korpusa ssj500k (Krek et al. 2015)), a je kljub temu znatno pripomogel k izboljšanju oblikoskladenjskega označevanja in lematizacije slovenske RPK. Poleg tega kot zlati standard omogoča tudi preizkušanje oblikoskladenjskih označevalnikov in lematizatorjev za slovensko RPK.

Zaradi kriterijev vzorčenja Kons1-MSD in Kons2-MSD so deleži besedil, pojavnic in besed glede na različne stopnje standardnosti precej drugačni kot pri korpusu Janes-Norm, saj smo se pri korpusu Janes-Tag osredotočili predvsem na besedila v kategoriji L3, ki zajemajo več kot 80 % celotnega korpusa. Z vidika žanrov večina besedil (67 %) in še večji delež pojavnic (72,8 %) izvira iz tvitov, kar odseva dinamiko označevalne kampanje. Deleži normaliziranih besed v Tabeli 2 so podobni tistim iz korpusa Janes-Norm, razlike pa lahko verjetno pripišemo naključnim dejavnikom vzorčenja.

Janes-Tag je kot podatkovna množica na voljo na repozitoriju CLARIN.SI (Erjavec et al. 2016), iskanje po korpusu pa omogočeno v konkordančnikih CLARIN.SI. Slika 7 npr. prikazuje iskanje vseh samostalnikov v tožilniški obliki v noSketch Engine (v vrstico CQL vpišemo *[tag="Somet."]*).



Query Somet. 1,057 (-13.3 per million)

Page 1 of 53 Go Next | Last

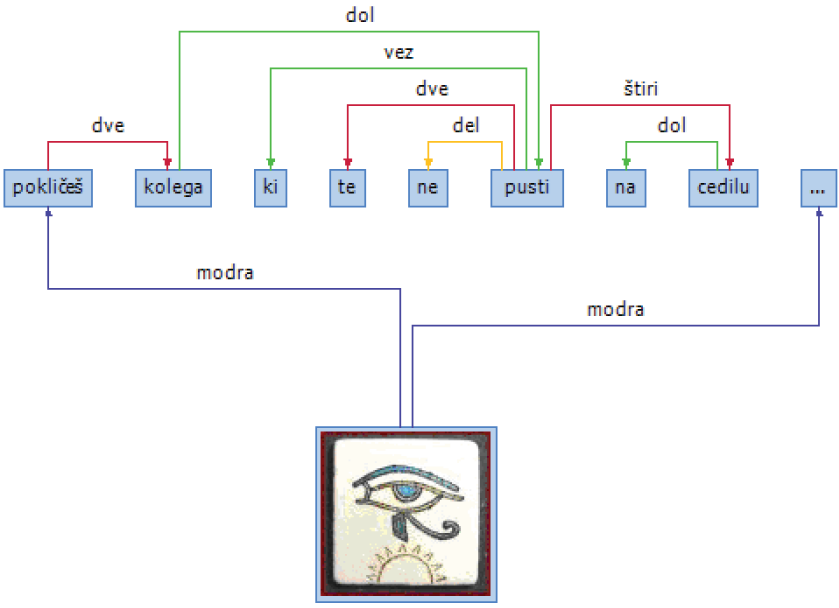
janes.news.rtvslo.335696.752	Problem je, da ima ta specifični norc tak	vpliv	in moč nad prb.10% slovencev . Pa dobro
tid.443660942325587968	napačen človek za to nalogo. Sm kr vprašal	šefa	, če prej ves pir v hladilniku spijem oz.
tid.443660942325587968	nalogo. Sm kr vprašal šefa, če prej ves	pir	v hladilniku spijem oz. če lahk vzamem
tid.443660942325587968	pir v hladilniku spijem oz. če lahk vzamem	dopust	. @tjablonsky @zballe @MihaPenko potrudte
tid.377782477974011904	prevec ... zlajnan je potrudi se da boste mel	koncept	in se ga drzte in koncejte v 30 min @SasoSin
tid.327291703335731200	tistih s ksilitolom, imajo med sestavinami	aspartam	. Te s ksilitolom maš pa v DMu , Mullerju
tid.441315127896592384	poprdenka" tolk, da boste vedeli, če boste sli v	soping	:) @iNinaromsek mela saaansooo, js grem
tid.231386671772495872	mela saaansooo, js grem pa kr ze dons :P se	prevoz	do dol sm dobila.. odzabe :D dobis pozdravcke
tid.392379111378653184	iz pletene narediti kippah, ce kdo rabi	nasvet	:) @BostjanJerko Da se naujo zdaj še na
janes.blog.publishwall.10697.6	najmanj kar je podlo! poglej spodaj kak	obraz	imaš a mi zna kdo razložit, zakaj hrvati
janes.forum.medovernet.5645070	bi pa lahko v LJ ali okolici dal odklenit	telefon	BlackBerry Storm 9500 . Zaklenjen je na
janes.forum.medovernet.8117633	Najbolj pa bo nasmejala Mitička ! hvala za	odgovor	in lep pozdrav @FuckedUpAlive Živc nebod
janes.forum.medovernet.8117633	nasmejala Mitička ! hvala za odgovor in lep	pozdrav	@FuckedUpAlive Živc nebod taka Fun! Ja
tid.380037511747104769	Ja tko ti bom povedu hmm grozn ane. napis	sms	pa to pa ti napišem neki zlooo fun=) haha
tid.336771105173934080	@alessusnik Kupim, če ga še imaš. Prosim, če mi na	DM	pošlješ tel. številko, da se dogoviva
tid.272811098741301248	spili;). Seveda, grozdni sok je najboljsi za	brainstorming	:). @NuckinFutsSlo majo le manj davka kokr
tid.523076235585654784	uporablja kdo Maca in je sošolka panično zaprla	comp	, da ne bi vidla in ne bi rabla odgovrit
janes.blog.rtvslo.5907.6	davke, bo že zasebni sektor sam poskrbel za	boj	proti poplavam. Kaj ne, da? @smaka21 vrjamm
tid.501312305402236928	@spirulinka9 Aja, brezveze. Jst sem danes pil	kofein	na avtocesti, bolj optimalno. Na morju
tid.501312305402236928	avtocesti, bolj optimalno. Na morju jem samo	zajtrk	... Tam kofeina ne rabiš. :) @free__JJ

Slika 7: Iskanje samostalnikov v tožilniški obliki v korpusu Janes-Tag preko vmesnika NoSketch Engine.

3.1.3 Janes-Syn

Kot pilotski vir za učenje skladenjskega razčlenjevanja slovenske RPK je bil pripravljen korpus Janes-Syn, ki je vzorčen iz Janes-Tag in obsega 200 tvitov (475 stavkov). Tviti so izbrani iz besedil zasebnih uporabnikov (ne pa tudi korporativnega tvitanja, ki je glede na preliminarna opažanja jezikovno bolj v skladu s standardom), vključujejo pa primere, ki so daljši od 120 znakov.

Janes-Syn je bil avtomatsko označen s skladenjskim razčlenjevalnikom SSJ (Dobrovoljc et al. 2012) in uvožen v program za pregledovanje drevesnic SSJ (avtor J. Brank, glej Sliko 8). Avtomatsko pripisane skladenjske povezave so bile ročno popravljene skladno z označevalnimi smernicami (Holozan et al. 2008), pri čemer smo sistem označevanja nadgradili z rešitvami za specifike nestandardnega jezika (Arhar Holdt 2016), in sicer z novostmi pri označevanju žanrsko specifičnih elementov, rabi tujejezičnih prvin, obravnavi eliptičnosti in fragmentarnosti jezika ter nestandardni rabi ločil. Več o pripravi Janes-Syn je mogoče prebrati v Arhar Holdt et al. (2016) in v Arhar Holdt (2018).



Slika 8: Primer označenega tvita v Označevalniku SSJ.

Janes-Syn je prosto dostopen za uporabo 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), iskanje po korpusu pa je omogočeno v konkordančnikih CLARIN.SI. Korpus je v konkordančnik umeščen tako, da je mogoče iskati tudi po skladenjskih oznakah: Slika 9 npr. prikazuje iskanje povezave 'ena', ki načeloma označuje stavčne osebke, v polje CQL vpišemo *[deprel="ena"]*.

Query ena 217 (49,498.2 per million)

Page 1 of 11 Go Next | Last

tid.266122536432062464	videti vsak tweet kandidatov, bi enostavno	sledili	njim?#predsednik12@petrasovdat V priemerjavi
tid.311838678605512704	priemerjavi s svojim predhodnikom nedvomno.	Okoliščine	in pogoji dela pa so mu bili vse prej kot
tid.342718195666399232	mu bili vse prej kot naklonjeni.Čeferin:	sodišče	podlega javnemu mnenju.Ni samo obsodilna
tid.342718195666399232	podlega javnemu mnenju.Ni samo obsodilna	sodba	tista, ki kaže na to, da pravna država
tid.342718195666399232	obsodilna sodba tista, ki kaže na to, da pravna	država	funkcionira.#poglediSlovenijeLažji med
tid.352691378020548608	#Glave s pregledom 3. sezone #GoT.Gobcamo	@anzet	@BokiNachbar @WiC_HmR @matevzluzar http://t.co/LWHogSK9nj
tid.361498211262791681	Blagovici je ustavljen lažno okvarjen romunski	kombi	; ustavljajo nič hudega sluteče naivne voznike
tid.366156416806944768	na SD kartico sem probala že stokrat.Tudi	telefon	to ponudi kot možnost.Rezultat?"Ni predmetov
tid.374883326189764608	@kricac; Bom ugibala - pripadnost?Današnja	mladina	tako zelo hlepi po tem.Mi pa tudi verjetno
tid.374883326189764608	Današnja mladina tako zelo hlepi po tem.	Mi	pa tudi verjetno nismo bili tako zelo drugačni
tid.381732876586217473	Čeprav ti letos ni šlo brez tebe ne bi bili	#junaki	.Rečem ti lahko le SREČNO.Šele zdaj videl
tid.390441098822561792	ti lahko le SREČNO.Šele zdaj videl kakšna	drama	je bila v CONCACAF,ko so ZDA v zadnjih
tid.390441098822561792	videl kakšna drama je bila v CONCACAF,ko so	ZDA	v zadnjih minutah priigrale Mehiki dvoboj
tid.397719565875953665	izkušnjah se da vzgajati brez nasilja.In tudi	sam	nisem bil nikoli tepen, za kar sem staršem
tid.412943107395973120	hvaležen.Ce vprasate mene (in mislim, da se	@anakobal_kobe	strinja), Tini do res odlicnega rezultata
tid.412943107395973120	do res odlicnega rezultata manjka le malo	teze	na spodnji smucki.:)Pa ti @stanka_d, si
tid.429342290239582209	manjka le malo teze na spodnji smucki.:)Pa	ti	@stanka_d, si čisto prestreljena.Od branjenja
tid.429342290239582209	zločincev se ti že pošteno blede.Si tudi	ti	del te "slavne" mreže #UDBA?@RomanLeljak
tid.439337873708679168	#novaplata http://t.co/mytlEbvTUzSuspendirana	tozilka	ne more vec preganjati novinarke.Odgovornost
tid.439337873708679168	vec preganjati novinarke.Odgovornost nosi	tozilec	, ki je prevzel zadevo.In njegov sef.http://t.co/4TRmccuocN

Slika 9: Primer iskanja skladenjskih oznak po Janes-Syn v NoSketchEnginu.

3.2 Korpusi za jezikoslovne raziskave

Z ročno označenimi korpusi za raziskave smo proučili štiri zanimive vidike slovenske računalniško posredovane komunikacije: načine krajšanja besedila, kodno preklapljanje, rabo vejice in rabo regionalnih jezikovnih različic na Twitterju.

3.2.1 Janes-Kratko

Janes-Kratko je korpus tvitov, ki je ročno označen z načini krajšanja po izdelani tipologiji (Goli et al. 2016a), ki pojave krajšanja uvršča na tri ravni: zapis (npr. krajšanje z ločili; *Slovenija - Slo.*), leksika (npr. nadomeščanje s kraticami; bruto domači proizvod – BDP) in skladnja (npr. izpust glagola; *Bi blo treba [∅] tiralico?*). Vsi nivoji so razdeljeni na skupno 32 podkategorij, ki med drugim na nivoju zapisa zajemajo npr. izpuščanje črk, opuščanje presledkov in ločil ter nadomeščanje s krajšimi nizi (npr. s številčnimi homofoni ali s tujejezičnimi črkami), na leksikalnem nivoju pa npr. zapis s kraticami in ustaljenimi okrajšavami s piko.

Glavni namen korpusa je ponuditi kvantitativen pregled nad vrsto in pogostostjo načinov krajšanja v slovenskih tvitih, zajema 777 tvitov (okoli 20.000 pojavnic), ki so bili vzorčeni iz korpusa Janes-Norm. V njem je zabeleženih skupno 3.464 pojavov krajšanja. Od tega je 87 % krajšanja na nivoju zapisa, na ostalih nivojih pa je krajšanja občutno manj: približno 12 % na leksikalnem in le dober 1 % na skladenjskem.

Query LN.* 90 (4,433.7 per million)

Page 1 of 5 Go Next | Last

tid.567399946710966272	drugih zornih kotov + počasni posnetki +	stat	! http://t.co/HheCGiojpc @MatevzNovak @cashkee
tid.617733529640824832	Glede na kratek čas od razpisa do izvedbe	ref.	v Grčiji, se sprašujem, kako uspešno je
tid.617733529640824832	Grčiji, se sprašujem, kako uspešno je bilo	info.	volivcev, ki je ključno za legitimnost.
tid.508689157519310848	moje je dobro, da gate trga. *in prestavi	tevejko	* Resna zadeva. V Mb med starejšimi pustoši
tid.598512578357231616	. *in prestavi tevejko * Resna zadeva. V	Mb	med starejšimi pustoši virusna pljučnica
tid.590962437379198976	npapi. chrome://flags/#enable-npapivpišeš v	nasl.	vrstico in klikneš "enable". @freeeky ok
tid.289085142964793346	#danasnjidan pred 10 leti je vlada razrešila	gen.	dir. Vursa Zorana Kovača, ker je na nov.
tid.289085142964793346	#danasnjidan pred 10 leti je vlada razrešila gen.	dir.	Vursa Zorana Kovača, ker je na nov. konf.
tid.289085142964793346	gen. dir. Vursa Zorana Kovača, ker je na	nov.	konf. kritiziral prenizka proračunska sredstva
tid.289085142964793346	dir. Vursa Zorana Kovača, ker je na nov.	konf.	kritiziral prenizka proračunska sredstva
tid.622036427992367104	segrevanje. Predstavljajte si zdej vse na	minimalcu	. Photo: V dogovoru s Sunčanom Stoneom
tid.605739619754364928	#lokalnevolitve2014 Mandarič, ki ga je Tito izgnal kot	kapital.	izdajalca, prihaja v deželo kjer prirejajo
tid.512893001082101760	letnika '96? Dajte se pripravit tudi na to. /	cc	@VitezDobTeka @BMWSlovenija @JelenaJal
tid.494794705264455680	@IgorGaberc @IgorLuksicSD @strankaSDS V Slo ni	polit.	higiene, kvečjemu politični waterboarding
tid.420282861996875776	sledila na TW, pa je tip vse preveč sral po	TL	, sem ga odsledil. Ku se pa celotna generacija
tid.519363999776133121	again. #matura Kakšno zgražanje zaradi "	nepreprič.	in sploš" odg. AB ; da "smo" izvolili PV
tid.519363999776133121	nepreprič. in sploš" odg. AB ; da "smo" izvolili	PV	, ki v celi kampanji ni dal enega konkr.
tid.519363999776133121	izvolili PV , ki v celi kampanji ni dal enega	konkr.	odg. nikogar ne moti! Hm. Al se on to poskuša
tid.380965979821318144	modre črte, zdej mi interactions ne folgajo,	app	ima krizo identitete in ne prikzuje fotk
tid.394481745154043905	nogavic pa ni #slabo #facepalm Setamo po	Lj	in pride Zoki mim, se ustavi pa da Emanuelu

Slika 10: Iskanje neustaljenih krajšav v korpusu Janes-Kratko preko vmesnika NoSketch Engine.

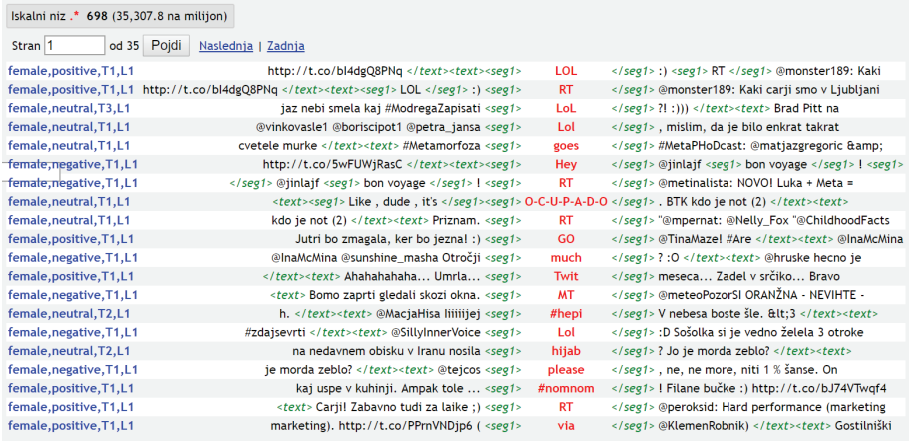
Rezultati raziskave na korpusu so podrobneje predstavljeni v Goli et al. (2016b), korpus pa je na voljo na repozitoriju CLARIN.SI (Goli et al. 2017) in v konkordančnikih CLARIN.SI. Slika 10 prikazuje iskanje vseh neustaljenih krajšav v korpusu (v polje CQL vpišemo *[seg="LN.*"]*).

3.2.2 Janes-Preklop

Korpus Janes-Preklop (glej Reher in Fišer 2018) vsebuje 1.104 tvite (19.769 pojavnic) in je namenjen proučevanju preklapljanja med jeziki v slovenskih tvitih. Preklopi so označeni na več nivojih: jezik preklopa, tip preklopa (medstavčno, zunajstavčno), stopnja prilagojenosti slovenskemu zapisu, stopnja razvidnosti oblikoslovne prilagojenosti (razvidnost iz obrazil in končnic) in vrsta besedne zveze preklopa (samostalniška besedna zveza ipd.).

V korpusu je označenih približno 1.400 preklopov, od tega sta približno dve tretjini znotrajstavčnih, tretjina pa medstavčnih. Jezikov, v katere preklapljajo uporabniki v korpusu, je skupno 9, najpogostejši pa so angleščina (90 % preklopov), hrvaščina/bosanščina/srbščina (4,5 %) in nemščina (3,5 %).

Podrobnejši rezultati raziskave kodnega preklapljanja so predstavljeni v Reher (2017), korpus pa je na voljo na repozitoriju CLARIN.SI (Reher et al. 2017) in v konkordačnikih CLARIN.SI. Slika 11 prikazuje iskanje vseh enobesednih preklopov v korpusu (v polje CQL vpišemo *<seg1>".*"</seg1>*).



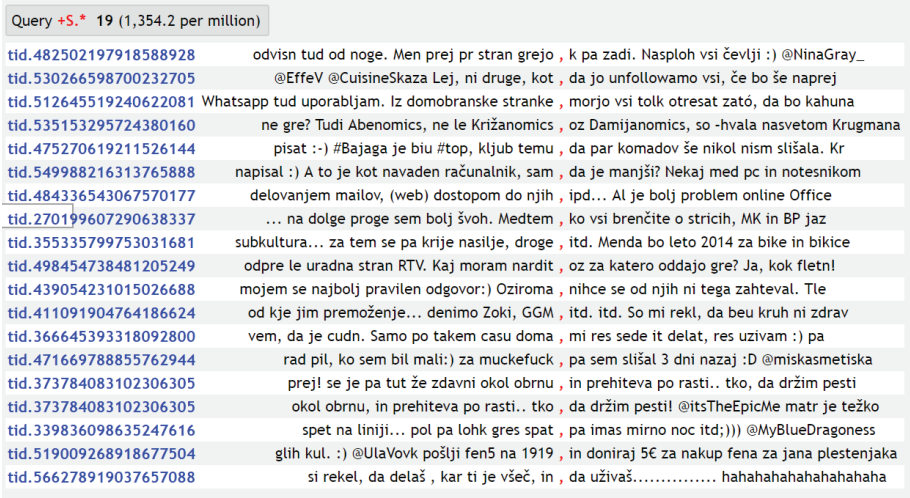
Slika 11: Iskanje enobesednih preklopov po korpusu Janes-Preklop v vmesniku NoSketch Engine.

3.2.3 Janes-Vejica

Janes-Vejica (Popič et al. 2017) je korpus tvitov, v katerih je v skladu z izdelano tipologijo ročno označena nestandardna (ne)raba vejice. Korpus vsebuje naključen vzorec 495 tvitov iz korpusa Janes v0.4, natančneje po 250 iz kategorij z visoko jezikovno nestandardnostjo (T1L3 in T3L3), pri čemer je bilo 5 tvitov izločenih iz prvotnega vzorca, ker so bili nerelevantni za jezikoslovne raziskave. V okviru raziskave na korpusu je bila razvita tudi sistemsko osnovana tipologija za opis nestandardne stave vejice (Popič et al. 2016a).

Glavni namen raziskave na korpusu Janes-Vejica je bil začrtati nadaljnje raziskave stave vejice v slovenščini, zlasti v primerjavi s standardnojezikovnim gradivom, ter določiti, v kolikšni meri raba vejice na Twitterju odstopa od jezikovnega standarda. Rezultati izpostavljajo nekaj novih težišč pri tej problematiki (npr. da je nestandardna raba vejice na Twitterju vezana predvsem na skladenjsko rabo). Izsledke so podrobneje predstavili Popič et al. (2016b).

Korpus Janes-Vejica je prosto in odprto dostopen na repozitoriju CLARIN.SI in njegovih konkordančnikih. Slika 12 prikazuje primer iskanja odvečnih vejic (v polje CQL vpišemo *[seg="\+S.*"]*).


Query +S.* 19 (1,354.2 per million)

tid.482502197918588928	odvisn tud od noge. Men prej pr stran grejo ,	k pa zadi. Nasploh vsi čevlji :) @NinaGray_
tid.530266598700232705	@EffeV @CuisineSkaza Lej, ni druge, kot ,	da jo unfollowamo vsi, če bo še naprej
tid.512645519240622081	Whatsapp tud uporabljam. Iz domobranske stranke ,	morjo vsi tolk otresat zató, da bo kahuna
tid.535153295724380160	ne gre? Tudi Abenomics, ne le Križanomics ,	oz Damijanomics, so -hvala nasvetom Krugmana
tid.475270619211526144	pisat :-) #Bajaga je biu #top, kljub temu ,	da par komadov še nikol nism slišala. Kr
tid.549988216313765888	napisal :) A to je kot navaden računalnik, sam ,	da je manjši? Nekaj med pc in notesnikom
tid.484336543067570177	delovanjem mailov, (web) dostopom do njih ,	ipd... Al je bolj problem online Office
tid.270199607290638337	... na dolge proge sem bolj švoh. Medtem ,	ko vsi brenčite o stricih, MK in BP jaz
tid.355335799753031681	subkultura... za tem se pa krije nasilje, droge ,	itd. Menda bo leto 2014 za bike in bikice
tid.498454738481205249	odpre le uradna stran RTV. Kaj moram nardit ,	oz za katero oddajo gre? Ja, kok fletn!
tid.439054231015026688	mojem se najbolj pravilen odgovor:) Oziroma ,	nihce se od njih ni tega zahteval. Tle
tid.411091904764186624	od kje jim premoženje... denimo Zoki, GGM ,	itd. itd. So mi rekl, da beu kruh ni zdrav
tid.366645393318092800	vem, da je cudn. Samo po takem casu doma ,	mi res sede it delat, res uzivam :) pa
tid.471669788855762944	rad pil, ko sem bil mali:) za muckefuck ,	pa sem slišal 3 dni nazaj :D @miskasmetiska
tid.373784083102306305	prej! se je pa tut že zdavni okol obrnu ,	in prehiteva po rasti.. tko, da držim pesti
tid.373784083102306305	okol obrnu, in prehiteva po rasti.. tko ,	da držim pesti! @itsTheEpicMe matr je težko
tid.339836098635247616	spet na liniji... pol pa lohk gres spat ,	pa imas mirno noc itd;))) @MyBlueDragoness
tid.519009268918677504	glih kul. :) @UlaVovk pošlji fen5 na 1919 ,	in doniraj 5€ za nakup fena za jana plestenjaka
tid.566278919037657088	si rekel, da delaš , kar ti je všeč, in ,	da uživaš..... hahahahahahahahahaha

Slika 12: Iskanje odvečnih vejic v korpusu Janes-Vejica preko vmesnika NoSketch Engine.

3.2.4 Janes-Geo

Janes-Geo (Čibej 2018) je korpus tvitov, ki so bili vzorčeni iz podkorpusa tvitov Janes-Tweet v0.3 glede na metapodatke o regionalni pripadnosti uporabnikov, avtomatsko pripisane s pomočjo geolokacije njihovih tvitov (Čibej in Ljubešić 2015, Čibej 2016). Skupno 321 uporabnikov, ki so vključeni v korpus, je bilo razporejenih v 9 regij: Ljubljano, Maribor in 7 regij, ki predstavljajo glavne narečne skupine (Ramovš 1931): Primorska, Rovtarska, Gorenjska, Dolenjska, Štajerska, Koroška in Panonska. Iz vsake regije je bilo vzorčenih po 500 tvitov (z izjemo Rovtarske, Koroške in Maribora, ki so prispevali 400, 260 in 330 tvitov) s kategorijo nestandardnosti L3 (v nekaterih manjših regijah zaradi pomanjkanja podatkov tudi L2), zato korpus vsebuje skupno nekaj manj kot 4000 tvitov oz. 64.000 pojavnic.

Glavni namen korpusa Janes-Geo je proučevanje medregionalnih jezikovnih razlik v slovenski RPK. Da bi ugotovili, ali se pogostost določenih nestandardnih jezikovnih pojavov razlikuje med uporabniki iz različnih regij, so bili tviti ročno označeni v skladu z za ta namen izdelano tipologijo nestandardnih jezikovnih prvin v slovenski RPK (z nekaterimi izjemami, kot sta raba ločil in skladnja). Oznake delimo na 6 glavnih kategorij: izpusti, transformacije, nestandardno besedje, nestandardno oblikoslovje, variantne različice pogostih besed in drugo.

Podrobnejši izsledki raziskave so predstavljeni v Čibej (2018), korpus pa je na voljo na repozitoriju CLARIN.SI (Čibej et al. 2018).

4 SKLEP

V poglavju smo predstavili postopek izdelave in ročnega označevanja korpusov v okviru projekta JANES s posebnim poudarkom na izdelavi korpusov Janes-Norm in Janes-Tag, ki služita kot zlata standarda za učenje in preizkušanje orodij za tokenizacijo, stavčno segmentacijo in normalizacijo na eni strani ter lematizacijo in oblikoskladenjsko označevanje RPK na drugi. Za odvisnostno označevanje skladenjske ravni jezika je bil pripravljen korpus Janes-Syn. Korpusa Janes-Norm in Janes-Tag sta bila že uporabljena za učenje normalizatorjev in oblikoskladenjskih označevalnikov, prilagojenih za nestandardni jezik (Ljubešić et al. 2018) kot tudi za označevanje celotnega korpusa Janes 1.0 (Erjavec et al. 2018).

Na drugi strani predstavljajo ročno označeni korpusi Janes-Kratko, Janes-Preklop, Janes-Vejica in Janes-Geo podlago za empirične analize jezikovnih prvin, ki vstopajo v središče raziskovalnega interesa s pojavom RPK: (ne)standardna stava vejice, raba tujejezičnih elementov in preklapljanje med jeziki, načini in

pogostost krajšanja jezikovnih elementov v besedilih ter regionalno specifične jezikovne prilagoditve pisne komunikacije.

Poleg ročno označenih korpusov velja kot pomemben projektni doprinos omeniti dobro dokumentirane in prosto dostopne označevalne tipologije ter smernice, ki bodo v korist nadaljnjim raziskavam na področju računalniško posredovane komunikacije. Predlagane rešitve je mogoče prilagoditi tudi za označevanje sorodnih jezikov, kar se je že izkazalo kot uspešno v okviru projekta ReLDI,[8] ki je organiziral označevalno kampanjo za normalizacijo, lematizacijo in oblikoskladenjsko označevanje besedil v okviru razvoja orodij za obdelavo hrvaške in srbske RPK (Miličević et al. 2016), učni množici, izdelani po vzoru Janes-Tag, pa tudi objavil v repozitoriju CLARIN.SI (Ljubešić et al. 2017a, Ljubešić et al. 2017b).

Prosta in odprta dostopnost rezultatov je bila med glavnimi vodili projekta in vsi korpusi so na voljo v repozitoriju CLARIN.SI in v konkordančniku NoSketch Engine. Pri prenosu v slednjega so bile dodatno upoštevane specifike vključenega gradiva, zaradi česar je mogoče s kombinacijo iskanja po korpusnospecifičnih oznakah in njihovega prikaza v konkordančnem nizu dobiti kvalitetnejši vpogled v jezikovne podatke. Nekaj možnosti za iskanje je opisanih na spletni strani projekta JANES, kjer so javno dostopne tudi vse navedene označevalne smernice, vključene pa so tudi v vnose v repozitoriju CLARIN.SI. Vse gradivo je za razliko od dosedanjih praks pri distribuciji korpusov RPK (npr. Frey et al. 2015, Chiari in Canzonetti 2014) na voljo pod zelo liberalno licenco Creative Commons Priznanje Avtorstva, s čimer so viri na voljo za nadaljnje raziskave in za razvoj komercialnih produktov tudi izven okvirov projekta JANES, licenca pa omogoča tudi morebitne izboljšave označevalnih tipologij, smernic in korpusov ter njihovo redistribucijo. Pomembna naloga za nadaljnje delo pa vsekakor ostaja nadaljnja nadgradnja označevalnih orodij, preizkus njihove točnosti na različnih vrstah jezikovnega gradiva ter nadaljnji koraki v smeri njihove optimizacije.

Zahvala

Avtorji prispevka se najlepše zahvaljujejo Kaji Dobrovoljc, Simonu Kreku in Katji Zupan za konstruktivne pripombe pri izdelavi smernic za označevanje korpusov Janes-Norm in Janes-Tag. Posebna zahvala gre vsem označevalcem, ki so sodelovali v označevalnih kampanjah korpusov: Teji Goli, Melaniji Kožar, Vesni Koželj, Poloni Logar, Klari Lubej, Dafne Marko, Barbari Omahen, Eneji Osrajnik, Predragu Petroviću, Poloni Polc, Aleksandri Rajković, Špeli Reher, Izi Škrjanec in Katji Zupan.

8 Regional Linguistic Data Initiative: https://reldi.spur.uzh.ch/.

Literatura

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

Arhar Holdt, Špela, Darja Fišer, Tomaž Erjavec in Simon Krek, 2016: Syntactic annotation of Slovene CMC: first steps. Fišer, Darja in Michael Beißwenger (ur.). *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities, 27-28 September 2016, Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia*. 1st ed. Ljubljana: Znanstvena založba Filozofske fakultete. 3–6.

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

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

Benikova, Darina, Chris Biemann in Marc Reznicek, 2014: NoSta-D Named Entity Annotation for German: Guidelines and Dataset. LREC 2014.

Kalina Bontcheva, Leon Derczynski in Ian Roberts, 2017: Crowdsourcing Named Entity Recognition and Entity Linking Corpora. Ide, Nancy in James Pustejovsky (ur.): *Handbook of Linguistic Annotation*. Dordrecht: Springer. 875–892.

Chiari, Isabella in Alessio Canzonetti, 2014: Le forme della comunicazione mediata dal computer: generi, tipi e standard di annotazione. Garavelli, Enrico in Elina Suomela-Härmä (ur.): *Dal manoscritto al web: canali e modalità di trasmissione dell'italiano. Tecniche, materiali e usi nella storia della lingua*. Firence: Franco Cesati Editore. 595–606.

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

Čibej, Jaka, 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.

Čibej, Jaka, Špela Arhar Holdt, Tomaž Erjavec, Darja Fišer in Katja Zupan, 2016c: *Smernice za označevanje računalniško posredovane komunikacije: tokenizacija, stavčna segmentacija, normalizacija, lematizacija in oblikoskladenjsko označevanje (v1.0)*. http://nl.ijs.si/janes/viri/

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

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

Čibej, Jaka, 2018: Regionalne jezikovne različice v slovenski računalniško posredovani komunikaciji: korpusni pristop z ročno označenim korpusom Janes-Geo. Fišer, Darja (ur.): *Viri, orodja in metode za analizo spletne slovenščine*. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani. 160–197.

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.

Eckart de Castilho, Richard, Chris Biemann, Iryna Gurevych in Sied Muhie Yimam. 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

Erjavec, Tomaž, 2011: Automatic linguistic annotation of historical language: ToTrTaLe and XIX century Slovene. *5th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2011)*. Portland: Association for Computational Linguistics. 33–38. http://aclweb.org/anthology-new/W/W11/W11-1505.pdf

Erjavec, Tomaž, 2015: The IMP historical Slovene language resources. Language Resources and *Evaluation* 49/3. 753–775.

Erjavec, Tomaž, Cyprian Laskowski, Jaka Čibej, Darja Fišer in Kaja Dobrovoljc, 2016a: *Navodila za označevanje računalniško posredovane komunikacije v WebAnno (v1.0)*. http://nl.ijs.si/janes/viri/

Erjavec, Tomaž, Jaka Čibej in Darja Fišer, 2016b: Omogočanje dostopa do korpusov slovenskih spletnih besedil v luči pravnih omejitev. *Slovenščina 2.0* 4/2. 189–219.

Erjavec, Tomaž, Darja Fišer, Jaka Čibej, Špela Arhar Holdt, Nikola Ljubešić in Katja Zupan, 2017: *CMC training corpus Janes-Tag 2.0*. Slovenian Language Resource Repository CLARIN.SI. http://hdl.handle.net/11356/1123

Erjavec, Tomaž, Jaka Čibej, Špela Arhar Holdt, Nikola Ljubešić in Darja Fišer, 2016c: Gold-Standard Datasets for Annotation of Slovene Computer-Mediated Communication. *Proceedings of the Tenth Workshop on Recent Advances in Slavonic Natural Languages Processing*, *RASLAN 2016*. Brno, Češka.

Erjavec, Tomaž, Darja Fišer, Jaka Čibej in Špela Arhar Holdt, 2016d*: CMC training corpus Janes-Norm 1.2*. Slovenian Language Resource Repository CLARIN.SI. http://hdl.handle.net/11356/1084

Erjavec, Tomaž, Nikola Ljubešić in Darja Fišer, 2017: Adapting a state-of-the-art tagger for South Slavic languages to non-standard text. *Proceedings of the EACL workshop*. The 6th Workshop on Balto-Slavic Natural Language Processing, April 4, 2017 Valencia, Spain. Stroudsburg: The Association for Computational Linguistics. 60–68. http://bsnlp-2017.cs.helsinki.fi/bsnlp2017-book.pdf

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, 2016: JANES v0. 4: Korpus slovenskih spletnih uporabniških vsebin. *Slovenščina 2.0* 4/2. 67–99.

Frey, Jennifer-Carmen, Aivars Glaznieks in Egon Stemle, 2015: The DiDi Corpus of South Tyrolean CMC Data. *Proceedings of the 2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media*. GSCL2015 (NLP4CMC2015). 1–6.

Goli, Teja, Eneja Osrajnik in Darja Fišer, 2016a: *Strategije krajšanja tvitov: tipologija oznak, v1.0*. http://nl.ijs.si/janes/viri/#Janes-Kratko

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

Goli, Teja, Eneja Osrajnik in Darja Fišer, 2017: *CMC shortening corpus Janes--Kratko 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1087

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

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

Laarmann-Quante, Ronja in Stefanie Dipper, 2016: An Annotation Scheme for the Comparison of Different Genres of Social Media with a Focus on Normalization. *Normalisation and Analysis of Social Media Texts (NormSoMe) Workshop*. 23–30. https://www.linguistics.ruhr-uni-bochum.de/~dipper/pub/normsome16_webVersion.pdf

Ljubešić, Nikola, Katja Zupan, Darja Fišer in Tomaž Erjavec, 2016: Normalising Slovene data: historical texts vs. user-generated content. Dipper, Stefanie, Friedrich Neubarth in Heike Zinsmeister (ur.): *Proceedings of the 13th Conference*

on Natural Language Processing (KONVENS), September 19-21, 2016, Bochum, Germany. 146–155. https://www.linguistics.rub.de/konvens16/pub/19_konvensproc.pdf

Ljubešić, Nikola, Daša Farkaš, Filip Klubička, Tomaž Erjavec, Maja Miličević, Mateja Filko, Denis Kranjčić in Barbara Dujmić, 2017: *Croatian Twitter training corpus ReLDI-NormTag-hr 1.1*. Slovenian Language Resource Repository CLARIN.SI. http://hdl.handle.net/11356/1121

Ljubešić, Nikola, Daša Farkaš, Filip Klubička, Tomaž Erjavec, Maja Miličević in Teodora Vuković, 2017: *Serbian Twitter training corpus ReLDI-NormTag-sr 1.1*. Slovenian Language Resource Repository CLARIN.SI. http://hdl.handle.net/11356/1120

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, Bulgaria. 371–378.

Ljubešić, Nikola, Tomaž Erjavec in Darja Fišer. 2014. Standardizing tweets with character-level machine translation. *Computational linguistics and intelligent text processing: 15th International Conference, CICLing 2014, Kathmandu, Nepal, April 6–12, 2014: proceedings: part II, (Lecture notes in computer science, 8404)*. Heidelberg: Springer. 164–175.

Ljubešić, Nikola, Tomaž Erjavec in Darja Fišer, 2018: Orodja za procesiranje nestandardne slovenščine. Fišer, Darja (ur.): *Viri, orodja in metode za analizo spletne slovenščine*. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani. 74–99.

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

Poesio, Massimo, Jon Chamberlain in Udo Kruschwitz, 2017: Phrase Detectives. Ide, Nancy in James Pustejovsky (ur): *Handbook of Linguistic Annotation*. Dordrecht: Springer. 1149–1176.

Popič, Damjan, Darja Fišer, Katja Zupan in Polona Logar, 2016b: Raba vejice v uporabniških spletnih vsebinah. *Proceedings of the Conference on Language Technologies & Digital Humanities, September 29th – October 1st, 2016 Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia, 2016*. 106–110.

Popič, Damjan, Katja Zupan in Darja Fišer, 2016a: *Smernice za označevanje nestandardne rabe vejice v uporabniških spletnih vsebinah*. http://nl.ijs.si/janes/viri/#Janes-Vejica

Popič, Damjan, Katja Zupan, Polona Logar, Teja Kavčič, Tomaž Erjavec in Darja Fišer, 2017: *Tweet comma corpus Janes-Vejica 1.0*. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1088

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

Rehbein, Ines, Emiel Visser in Nadine Lestmann, 2013: Discussing best practices for the annotation of Twitter microtext. *Proceedings of the Third Workshop on Annotation of Corpora for Research in the Humanities* (ACRH-3). 73.

Reher, Špela, 2017: *Slovenščina na prepihu: kodno preklapljanje v objavah slovenskih uporabnikov Twitterja. Kvalitativna in kvantitativna analiza tvitov iz korpusa nestandardne slovenščine Janes.* Magistrsko delo. Ljubljana: Filozofska fakulteta.

Reher, Špela, Tomaž Erjavec in Darja Fišer, 2017: *Tweet code-switching corpus Janes-Preklop 1.0*, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1154

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.

Rychlý, Pavel, 2007: Manatee/Bonito - A Modular Corpus Manager. *1st Workshop on Recent Advances in Slavonic Natural Language Processing*. Brno: Masarykova univerzita. 65–70.

TEI Consortium (ur.): *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/

Ueberwasser, Simone, 2013: Non-standard data in Swiss text messages with a special focus on dialectal forms. Zampieri, Marcos in Sascha Diwersy (ur.): *Non-standard Data Sources in Corpus-based Research*. Aachen: Shaker Verlag. 7–24. https://ueberwasser.eu/UeFiles/uni/Tagungen/2012Koeln/ueberwasser.pdf