# Analize za nadgradnjo učnega korpusa ssj500k

*Špela ARHAR HOLDT*
Fakulteta za računalništvo in informatiko Univerze v Ljubljani,
Filozofska fakulteta Univerze v Ljubljani,
spela.arharholdt@fri.uni-lj.si

*Jaka ČIBEJ*
Filozofska fakulteta Univerze v Ljubljani, Institut »Jožef Stefan«,
jaka.cibej@ff.uni-lj.si

Abstract

The paper presents a series of linguistic analyses aimed at improving the ssj500k Slovene training corpus and – to a lesser extent – the Sloleks morphological lexicon of Slovene. Both resources are vital in supervised machine learning of linguistic annotation for modern written Slovene as well as in other language-related tasks. First, the analysis focuses on the representation of morphosyntactic tags in the training corpus, resulting in suggestions on how to expand the corpus with unrepresented and ambiguous word forms and tags. The analysis reveals several shortcomings of the lexicon and inconsistencies within the MULTEXT-East v6 tagging scheme. These need to be addressed in the future. Second, the analysis categorizes sentences and texts containing non-standard and foreign-language elements as well as evaluates the adequacy of the ssj500k training corpus for the annotation of language elements on paragraph- and text-levels, resulting in suggestions on how to expand the training corpus with new texts. This will enable the corpus to be annotated on higher levels and new taggers to be trained by also taking into account language elements in non-standard Slovene.

**Ključne besede:** učni korpus, ssj500k, Sloleks, oblikoskladnja, oznake MSD
**Keywords:** training corpus, ssj500k, Sloleks, morphosyntax, MSD tags

# 1 Uvod

Učni korpusi so premišljeno grajene in zanesljivo (tipično ročno) označene podatkovne množice, ki se uporabljajo pri strojnem učenju postopkov za obdelavo naravnega jezika. Učni korpus ssj500k, ki je na repozitoriju CLARIN.SI raziskovalni skupnosti trenutno na voljo v različici 2.2 (Krek et al. 2019), je referenčni vir za nadzorovano učenje strojnega jezikoslovnega označevanja sodobnih slovenskih pisnih besedil. Kot tak predstavlja pomemben člen v verigi, ki prek učenja označevalnikov,[1] označevanja besedil in uporabe strojno pripisanih oznak za napredna podatkovna luščenja oz. korpusne poizvedbe vodi do jezikovnih podatkov, na osnovi katerih lahko nastane sodoben, empiričen, korpusno podprt slovnični opis ali katerikoli drug na označenih besedilih temelječi rezultat.

Korpus ssj500k se razvija že več kot desetletje, kar izčrpno predstavlja prispevek Krek et al. (2020b). V različici 2.2 vsebuje 27.829 povedi, označenih na različnih jezikovnih ravneh, od segmentacije, tokenizacije, lematizacije, oblikoslovja in oblikoskladnje prek odvisnostne skladnje, imenskih entitet in večbesednih leksemov do udeleženskih vlog. Osnovne ravni oznak so pripisane vsem povedim v korpusu, ostale pa zajemajo omejen nabor povedi.[2] Prva naloga za nadaljnji razvoj korpusa je zato označevanje dodatnih povedi na višjih označevalnih ravneh. Razen tega je treba premisliti o povečanju korpusnega obsega, dodajanju novih označevalnih ravni (tudi takšnih, ki posegajo prek meja posamezne povedi, npr. označevanje koreferenčnosti) in nenazadnje evalvirati in izboljšati označevanje na obstoječih ravneh.

1 Do sedaj so bili na tem korpusu naučeni označevalniki Obeliks (Grčar et al. 2012), ReLDI (Ljubešić in Erjavec 2016) in CLASSLA StanfordNLP (Ljubešić in Dobrovoljc 2019). Različni označevalniki svoj model znanja gradijo na različne načine, v porastu je tudi metodologija, ki se na jezikovne vire, kot sta učni korpus in leksikon oblik, zanaša v manjši meri. Zato je treba posebej poudariti, da se prispevek posveča izključno nadzorovanemu učenju in znotraj tega okvira temelji na predpostavki, da ciljna izboljšava virov (lahko) izboljša natančnost označevanja sodobne pisne slovenščine. Predpostavko je mogoče preveriti po nadgradnji virov z empiričnimi evalvacijami označevalne natančnosti za izbrana orodja.

2 Segmentacija, lematizacija, oblikoskladnja JOS ter UD so označene pri vseh 27.829 povedih, večbesedne enote pri 13.511 povedih, skladnja JOS pri 11.411 povedih, imenske entitete pri 9.488 povedih, skladnja UD pri 8.000 povedih in udeleženske vloge pri 5.501 povedih (Krek et al. 2020b: Tabela 1).

Poleg učnega korpusa se prispevek dotika tudi določenih pomanjkljivosti leksikona Sloleks, odprto dostopnega jezikovnega vira (Dobrovoljc et al. 2019b), ki v trenutni različici prinaša oblikoslovne informacije za 100.805 slovenskih besed različnih besednih vrst. Vsebino in prioritete za nadgradnjo leksikona pregledno predstavljajo Dobrovoljc et al. (2015). Pričujoči prispevek leksikon analizira predvsem kot referenčni vir za nabor (razpoložljivih oz. možnih) oblikoskladenjskih oznak za slovenščino in v tej luči na seznam razvojnih prioritet dodaja nekaj vsebinsko specifičnih novih točk. Seveda pa je namembnost leksikona, kot tudi učnega korpusa, širša od tematike, ki jo pokriva prispevek, zato imajo jezikoslovne evalvacije in izboljšave obeh virov lahko tudi širši pozitivni vpliv.

Delo, ki ga predstavljava, je nastalo pod okriljem projekta Nova slovnica sodobne standardne slovenščine: viri in metode.[3] Projekt je med drugim vseboval analize jezikoslovnega označevanja korpusov in izdelavo predloga izboljšav tabel oblikoskladenjskih oznak JOS, na osnovi katerih bo osnovan nadaljnji razvoj ssj500k, deloma pa tudi leksikona Sloleks. V središču raziskovalnega zanimanja so štiri teme: (a) zastopanost oznak MSD v učnem korpusu, (b) pojavnost povedi oz. besedil, ki vsebujejo nestandardne in tujejezične elemente, (c) identifikacija morebitnih nedoslednosti ali težav sistema za oblikoskladenjsko označevanje, (č) ustreznost korpusa ssj500k za označevanje jezikovnih značilnosti na odstavčni ali širši ravni. V prispevku predstaviva motivacijo za izbiro naštetih tem, podatke in izsledke posameznih analiz. Razpravo zaključi razdelek s strnjenimi ugotovitvami v obliki priporočil za nadgradnjo učnega korpusa ter leksikona.

## 2 Zastopanost oblikoskladenjskih oznak v učnem korpusu

Ker je slovenščina oblikoslovno bogat jezik, sodi med temeljne označevalne ravni poleg segmentacije, tokenizacije in lematizacije tudi

3 Projekt Nova slovnica sodobne standardne slovenščine: viri in metode (J6-8256) je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije med letoma 2017 in 2020. Projektna spletna stran: http://slovnica.ijs.si/.

oblikoskladenjsko označevanje, ki v korpusu ssj500k že od začetka njegovega razvoja temelji na sistemu oznak MULTEXT-East (Erjavec 2012).[4] Skupaj z učnim korpusom se je razvijal tudi označevalni sistem: pred različico 2.0 je bil korpus ssj500k označen po sistemu MULTEXT-East v4, ki je bil pripravljen v projektu JOS.[5] Ssj500k 2.0 je bil označen s sistemom MULTEXT-East v5, ki je ostal v delovni različici in ni bil posebej dokumentiran in popisan – je pa enak bolje dokumentiranemu MULTEXT-East v6, s katerim je označen korpus ssj500k 2.2.[6]

Z različnimi verzijami oznak so bile označene tudi različne edicije referenčnega pisnega korpusa Gigafida (Logar et al. 2012, Krek et al. 2020a),[7] kar je pri analizah, ki sledijo, treba upoštevati. K razlikam med verzijami oznak in vprašanjem, ki bi se jim bilo treba posvetiti pri nadaljnjem razvoju MULTEXT-East za slovenščino, se vračamo v razdelku 4 tega prispevka. V prispevku oblikoskladenjske oznake navajava brez razvezave oz. dodatnega opisa, npr. 'Gp-ppd'. Vse informacije, ki omogočajo interpretacijo oznak, so na voljo na spletni strani: http://nl.ijs.si/ME/V6/msd/html/msd-sl.html, tako za slovensko kot angleško različico oznak in s konkretnimi primeri označenih besed.[8]

Sistem MULTEXT-East v6, ki je uporabljen za ssj500k 2.2, vsebuje 1.900 oblikoskladenjskih oznak, v učnem korpusu pa se pojavi 1.304 od vseh možnih oznak, kar pomeni približno 80-odstotno zastopanost (Krek et al. 2020b: Tabela 3). Predvidevati je, da so manjkajoče oznake različnih vrst in različnega vpliva na učenje

4 Od različice 2.2 naprej so v ssj500k pripisane tudi oblikoskladenjske oznake sistema Universal Dependencies (UD) (Dobrovoljc et al. 2019a). Tem oznakam se v prispevku ne posvečamo, ker pa so v veliki meri strojno preslikane iz oznak MULTEXT-East, bodo izboljšave slednjih pozitivno vplivale tudi na oznake UD.

5 Jezikoslovno označevanje slovenščine, projektna stran: http://nl.ijs.si/jos/josMSD-sl.html. Na tej strani je natančneje predstavljeno tudi ozadje priprave označevalnega sistema oz. njegove prilagoditve za slovenščino, ki kot oblikoskladenjsko bogat jezik prinaša precej višje število oznak kot večina zahodnoevropskih jezikov (http://nl.ijs.si/jos/msd/html-sl/msd.background.html).

6 Nabor oznak in informacije o označevalnem sistemu MULTEXT-East v6 so na voljo na strani: http://nl.ijs.si/ME/V6/msd/html/msd-sl.html.

7 Verjetno tudi drugih korpusnih virov – v tem prispevku se osredotočamo samo na referenčni korpus.

8 Pri samostalnikih denimo beležimo: ali gre za lastno ali občno ime; spol; število; sklon ter (ne)živost pri samostalnikih moškega spola v tožilniku ednine.

označevanja, vendar natančnejša analiza stanja z identifikacijo akutnih mest do sedaj še ni bila opravljena. V nadaljevanju zato najprej predstavimo skupine oznak, ki v korpusu ssj500k 2.2 manjkajo, pri čemer za primerjavo uporabimo zastopanost teh oznak v referenčnem korpusu standardne pisne slovenščine Gigafida 2.0 (Krek et al. 2020a), nato se natančneje posvetimo oznakam za najbolj problematično skupino, zaimke. Kot je omenjeno v uvodnem delu prispevka, analiza temelji na predpostavki, da lahko dopolnitev učnega korpusa in leksikona izboljša strojno označevanje slovenščine, kar bo mogoče preveriti po sami nadgradnji, vendar pa so analize pomembne tudi za razvoj samega označevalnega sistema, torej nabora in vrednosti označevalnih kategorij.

## 2.1 Oblikoskladenjske oznake v korpusu ssj500k 2.2 in Gigafida 2.0

Tabela 1 prikazuje zastopanost oblikoskladenjskih oznak MULTEXT-East v6 v učnem korpusu ssj500k 2.2, primerjalno s korpusom Gigafida 2.0. Vrstice prikazujejo, koliko oznak za posamezno besedno vrsto se (ne) pojavlja v učnem korpusu, v stolpcih pa je informacija o referenčnem korpusu. Za samostalnik denimo v zadnjem stolpcu tabele vidimo, da zajema skupno 104 oznake (krepki tisk). V vrsticah spodaj sledi podatek, da se od tega nabora v ssj500k pojavlja skupno 95 oznak in da se jih 9 v ssj500k ne pojavlja. Na drugi strani v krepkem tisku 2. in 3. stolpca vidimo, da se v korpusu Gigafida pojavlja skupno 97 samostalniških oznak in 7 se jih ne pojavlja. Ostale celice po principu matrike pokažejo, da se 95 oznak pojavlja tako v korpusu ssj500k kot Gigafida; 2 oznaki se pojavita v Gigafidi, ne pa tudi v ssj500k; 7 oznak pa se ne pojavi ne v Gigafidi ne v ssj500k.

V središču zanimanja so oznake, ki se bodisi ne pojavijo v nobenem od korpusov, kar nakazuje morebitne težave na ravni označevalnega sistema, bodisi se v referenčnem korpusu pojavljajo, ne pojavijo pa se v učnem korpusu. Pri slednjih oblikujemo predlog, katere manjkajoče oznake bi bilo pri korpusni nadgradnji smiselno ciljno zagotoviti.

**Tabela 1:** Zastopanost oblikoskladenjskih oznak v ssj500k 2.2 in Gigafida 2.0 po besednih vrstah.

| Oznake MULTEXT-East v6 v korpusu ssj500k 2.2 | Se pojavlja v Gigafida 2.0 | Se ne pojavlja v Gigafida 2.0 | Skupna vsota |
|---|---|---|---|
| **Ločilo (U)** | **1** | | **1** |
| se pojavlja | 1 | | 1 |
| **Okrajšava (O)** | **1** | | **1** |
| se pojavlja | 1 | | 1 |
| **Medmet (M)** | **1** | | **1** |
| se pojavlja | 1 | | 1 |
| **Členek (L)** | **1** | | **1** |
| se pojavlja | 1 | | 1 |
| **Veznik (V)** | **2** | | **2** |
| se pojavlja | 2 | | 2 |
| **Prislov (R)** | **4** | | **4** |
| se pojavlja | 4 | | 4 |
| **Predlog (D)** | **5** | **1** | **6** |
| se pojavlja | 5 | | 5 |
| se ne pojavlja | | 1 | 1 |
| **Neuvrščeno (N)** | **4** | **4** | **8** |
| se pojavlja | 2 | | 2 |
| se ne pojavlja | 2 | 4 | 6 |
| **Samostalnik (S)** | **97** | **7** | **104** |
| se pojavlja | 95 | | 95 |
| se ne pojavlja | 2 | 7 | 9 |
| **Glagol (G)** | **145** | **11** | **156** |
| se pojavlja | 128 | 2 | 130 |
| se ne pojavlja | 17 | 9 | 26 |
| **Števnik (K)** | **152** | **63** | **215** |
| se pojavlja | 146 | 6 | 152 |
| se ne pojavlja | 6 | 57 | 63 |
| **Pridevnik (P)** | **243** | **36** | **279** |
| se pojavlja | 240 | 1 | 241 |
| se ne pojavlja | 3 | 35 | 38 |
| **Zaimek (Z)** | **683** | **439** | **1.122** |
| se pojavlja | 635 | 34 | 669 |
| se ne pojavlja | 48 | 405 | 453 |
| **Skupna vsota** | **1.339** | **561** | **1.900** |

Oznake za ločilo, okrajšavo, medmet, členek, veznik in prislov so glede zastopanosti v obeh korpusih neproblematične. Pri predlogu je potencialno problematična oznaka 'Di' (predlog, ki mu sledi imenovalnik), s katerim se označuje lema *via*.[9] Ker v označevalnem sistemu ta rešitev precej izstopa, bi bilo mogoče razmisliti o alternativah, npr. označevanju te besede kot prislov ali tožilniški predlog, odvisno od identificirane jezikovne rabe.[10]

V obeh korpusih se pojavljata oznaki 'N' za neuvrščene leme in 'Nj' za tujejezične. V referenčnem, ne pa tudi učnem korpusu najdemo oznake 'Ne' za emotikone ter 'Nw' za spletne strani. K vprašanju nabora oznak za neuvrščene leme se vračamo v razdelku 4.

V primerjavi z drugimi polnopomenskimi besednimi vrstami so samostalniške oznake relativno dobro zastopane. V obeh korpusih manjka 7 oznak, ki so lastnoimenske in povezane bodisi z dvojino bodisi s srednjim spolom: 'Slmdm', 'Slmdo', 'Slzdi', 'Slzdt', 'Slsmt', 'Slsmm', 'Slsmo'. V učnem korpusu manjkata poleg tega še oznaki 'Slzdd' in 'Slsmd'. Vrzel je posledica lastnoimenskih paradigem v leksikonu Sloleks, ki pogosto vključujejo samo edninske ali množinske podatke, čeprav je jezikovnosistemsko mogoče tudi lastnoimensko besedišče uporabljati v vseh slovničnih številih.[11] Parcialnost paradigem se prenaša v označevalni sistem MULTEXT-East v6, ki oznak za lastna imena srednjega spola v dvojini sploh ne vključuje; primerov tipa *(*dve*) *Sredozemlji* s samostalniškimi oznakami trenutno torej ni mogoče označiti. Za odpravo težave je treba dopolniti oblikoslovni leksikon, nato označevalni sistem.

Bolj zapletena je situacija z glagolskimi oznakami. V obeh korpusih manjkajo oznake 'Gp-pte-d', 'Gp-ppe-d', 'Gp-g---d', 'Gp--ppd', 'Gp-m', 'Gp-vpd', 'Gp-vdd'. Prvi trije primeri označujejo

9 V leksikonu Sloleks je to tudi edini primer leme, ki ustreza oznaki 'Di'.

10 V jezikovnih priročnikih se v rabi z imenovalnikom *via* interpretira kot prislov (zgled v SSKJ2: *potovati v Zagreb via Zidani Most*, https://www.fran.si/133/sskj2-slovar-slovenskega-knjiznega-jezika-2/4545803/via). V Pravopisu je enak primer označen kot kombinacija predloga s tožilnikom: https://fran.si/134/slovenski-pravopis/3807195/via). Pregled pojavitev leme *via* v korpusu Gigafida sicer pokaže, da gre pogosto za tujejezično rabo (npr. *via Mazzini*, *foglio del via*).

11 Na primer leme tipa *Slovenija*, ki vsebujejo samo ednino: https://viri.cjvt.si/sloleks/slv/headword/89554/Slovenija ali tipa *Jesenice*, ki vsebujejo samo množino: https://viri.cjvt.si/sloleks/slv/headword/62141/Jesenice.

nestandardno zapisan zanikani glagol *biti* (npr. *nebom*) in četrti je posledica nedoslednosti označevalnega sistema, ki trenutno navaja dve alternativni oznaki za enake primere, k čemur se vračamo v razdelku 4. Ostale manjkajoče oznake so posledica redkosti jezikovnih pojavov, kot je namenilnik glagola *biti* ali dvojinski velelnik *bodiva, bodita*. Ti obliki se v korpusu Gigafida 2.0 sicer pojavljata, vendar sta napačno lematizirani v *bosti.* Problem je mogoče nasloviti z vključitvijo korpusnega gradiva, ki bo podprlo lažje razdvoumljanje oblik.

Dve glagolski oznaki, ki se pojavita v učnem korpusu, ne pa tudi v Gigafidi, sta 'Gp-sdd-d' (*nista*) in 'Ggnsdd-n' (*imata, hočeta*). Manjkajoči oznaki razkrijeta še en problem označenosti referenčnega korpusa: dvojina se v rabi pojavlja,[12] vendar so oblike napačno označene kot tretja, ne druga oseba. Na drugi strani je najti 17 glagolskih oznak, ki se pojavljajo v korpusu Gigafida, ne pa tudi v učnem korpusu. 6 je namenjenih nestandardnim oblikam (npr. *sve, nebova*, glej razdelek 4), 8 je vezanih na prihodnjiške oblike glagola *iti* (npr. *pojde, pojdejo*) – ker ustrezajo eni sami lemi, jih ni treba razdvoumljati, zato z vidika strojnega označevanja niso problematične in jih v učni korpus ni treba dodajati. Podobno velja za oznake tipa 'Ggnspd-d', 'Ggvvpd' in 'Ggnvpd' (npr. *nimava, pomagajva, upajva*), ki so v referenčnem korpusu sicer pogoste, vendar v smislu oblikoskladenjskega označevanja nedvoumne.

Pridevniških oznak, ki se ne pojavljajo v nobenem od korpusov, je 35, od tega jih je 32 za dvojino predvsem srednjega in ženskega, v nekaj primerih tudi moškega spola (npr. *Zvonetovih, zrelejših, najzvestejšima*). Kot je razvidno iz primerov, je razlog za redkost poleg dvojine lahko tudi vrsta pridevnika (svojilni) ali stopnjevanost oblike; tudi preostale 3 oznake v tej skupini so presežniške (npr. *najzvestejšemu*). Razen naštetih oblik v učnem korpusu manjkajo 3 podobne, tj. dvojinske oznake, ki so v Gigafidi zajete ('Pppmdd', 'Ppsmdo', 'Psnmdd'), na drugi strani pa v Gigafidi manjka v učni korpus zajeta dvojinska oblika 'Pppmdt' (*zračnejša*). V učnem korpusu bi bilo

12 Ad hoc preverbo obstoja je mogoče opraviti z vključitvijo zaimka v iskalni pogoj, npr. »vidva nista«.

smiselno zagotoviti nekoliko višjo reprezentiranost dvojine, zlasti tistih primerov, ki so lahko za označevanje dvoumni.

Oznak za številnike, ki se ne pojavljajo v nobenem od korpusov, je 57 od 215 možnih (26,5-odstotna nezastopanost). V tej skupini oznak jih je kar 30 za vrsto 'drugo', kamor umeščamo primere tipa *trojni, tristoteri* itd.[13] Poleg tega manjka 12 oznak za zaimkovne številnike, v vseh primerih za srednji spol v dvojini (npr. *drugih, drugima*) – primeri tovrstne rabe se v korpusu Gigafida v resnici pojavljajo, vendar so napačno označeni kot množinski. Manjka 12 oznak za vrstilne številnike, prav tako večinoma v dvojini (npr. *tristotridesetih, tristotridesetima*). Zadnje 3 oznake so za glavne številnike: 'Kbgsdd', 'Kbgsmd' in 'Kbgzdd' (npr. *dvema, trem*), ki ponovno razkrijejo napačno označenost referenčnega korpusa. Številniških oznak, ki se ne pojavijo v učnem korpusu, v Gigafidi pa so prisotne, je 6. Prednjači vrsta 'drugo' (npr. *dvojnima*), najti je tudi po en primer zaimkovnega in vrstilnega številnika, primerljivega zgoraj naštetim primerom. Zelo podobna slika je pri 6-ih primerih, ki jih zajema učni korpus, ne pa tudi Gigafida.

Z vidika manjkajočih oznak najbolj izstopajo oznake za zaimke. V obeh korpusih manjka 405 od 1.122 zaimkovnih oznak, kar pomeni 36,1-odstotno nezastopanost. Zaimki tudi sicer izstopajo po razvejanosti sistema, saj je oznak zanje več kot za vse ostale besedne vrste skupaj, obenem pa so precej enoznačne: kar 70 % oznak v leksikonu Sloleks pokriva samo 1, 2 ali 3 leme. Pregled oznak, ki manjkajo v obeh korpusih, pokaže, da gre za precej različne primere na ravni vrste, slovničnega spola in števila, vendar ponovno prednjačijo oblike za dvojino in srednji spol. Med 34 primeri, ki so vključeni v učni korpus, ne pa v Gigafido, so najbolj relevantne oznake za osebne, povratne in svojilne zaimke (skupno 24 primerov, npr. *vama, svoje, najino*), ker nakazujejo mesta, kjer so v Gigafidi morda prisotne označevalne napake. 48 primerov, ki se pojavljajo v referenčnem, ne pa učnem korpusu, predstavlja Tabela 2.

13 V referenčnih priročnikih, kot je SSKJ, so tovrstni primeri sicer umeščeni med pridevnike in tudi pri morebitni optimizaciji označevalnega sistema bi bilo smiselno premisliti o njihovi prerazvrstitvi.

**Tabela 2:** Zaimkovne oznake, ki se pojavljajo v referenčnem, ne pa v učnem korpusu.

| Vrsta zaimka | Št. oznak | Oznake in pogostost v Gigafida 2.0 | Primeri pojavnic |
|---|---|---|---|
| Kazalni | 2 | Zk-sdd: 376, Zk-mdo: 172 | tistima, takšnima |
| Nedoločni | 3 | Zn-mdo: 370, Zn-mdd: 28, Zn-zdo: 2 | enakima, nekima, redkokaterima |
| Nikalni | 3 | Zl-mmo: 237, Zl-mdd: 3, Zl-smo: 2 | nobenimi, nobenima, nobenimi |
| Osebni | 3 | Zodmdi: 1.774, Zopzdi: 1.739, Zotzdi: 555 | vidva, midve, onidve |
| Oziralni | 7 | Zz-sdr: 3.516, Zz-mmo: 1.072, Zz-mdr: 924, Zz-mdo: 35, Zz: 16, Zz-zdo: 5, Zz-mdd: 1 | kakršnihkoli, kakršnimi, kolikršnih, kakršnima, čigarkoli, kolikršnima, kolikršnima |
| Svojilni | 29 | Zsdzete: 6.530, Zspmmoe: 4.454, Zspzerd: 4.143, Zspmemd: 2.534, Zstmddd: 1.463, Zspmeod: 1.327, Zsdzeid: 1.248, Zstmmod: 1.241, Zstmedd: 1.199, Zspmddm: 991, Zsdmede: 980, Zstmddem: 933, Zsdmmid: 924, Zsdmmoe: 701, Zstmddez: 691, Zsdmemd: 656, Zsdmerd: 601, Zsdmmrd: 506, Zspmedd: 456, Zspmdde: 402, Zspmmod: 393, Zspzdod: 306, Zsdmeod: 140, Zstmddm: 129, Zsdmedd: 96, Zsdmddm: 85, Zsdmdoe: 34, Zsdzdod: 28, Zsdmddd: 6 | tvojo, mojimi, najine, najinem, njunima, najinim, vajina, njunimi, njunemu, našima, tvojemu, njegovima, vajini, tvojimi, njenima, vajinem, vajinega, vajinih, najinemu, mojima, najinimi, najinima, vajinim, njihovima, njihnjima, njihnima, vajinemu, vašima, tvojima, vajinima, vajinima |
| Vprašalni | 1 | Zv-mdd: 48 | kolikšnima |

Kot je razvidno iz podatkov, so za dodatno vključitev v učni korpus najbolj relevantni svojilni in osebni zaimki, zlasti najpogostejše oblike, kot npr. *tvojo, mojimi, najine, njunima, vidva* itd. Vključiti je možno tudi manjkajoče oznake za oziralne zaimke, druge skupine pa se zdijo opcijske. V nadaljevanju zaimkovne oblike preučimo še z vidika njihove enakopisnosti z drugimi besednimi vrstami in med različnimi vrstami zaimka.

## 2.2 Enakopisnost zaimkov z drugimi besednimi vrstami ali drugimi vrstami zaimkov

V Tabeli 3 navajamo besedne oblike, ki jih je glede na leksikon Sloleks mogoče pripisati zaimenski lemi, obenem pa tudi lemi kake druge besedne vrste. Pri vsaki dvoumni obliki opredelimo vrsto problema in število pojavitev z določeno oznako v korpusu ssj500k 2.0. V analizo ne zajemamo enakopisnih oblik, ki se v učnem korpusu že pojavljajo z vsemi možnimi besednimi vrstami,[14] tudi če v korpusu niso reprezentirane vse oblikoskladenjske lastnosti (npr. *mene* se pojavi kot 'Sozer' in 'Sozmi', ne pa kot 'Sozmt'). Prav tako niso vključena lastna imena, ki so v zapisu z malimi črkami enakopisna zaimkom (npr. *vanje – Vanje*).

**Tabela 3:** Zaimenske oblike, ki so enakopisne polnopomenskim – vrzeli v učnem korpusu.

| Dvoumna oblika | Vrsta problema | POS s frekvenco | Oblikoskladenjske oznake s frekvenco |
|---|---|---|---|
| isti | enakopisni glagol *istiti* | G: 0<br>Z: 42 | Ggvste: 0, Ggvvde: 0<br>Zn-mei: 7, Zn-met: 10, Zn-mmi: 3,<br>Zn-sdi: 0, Zn-sdt: 0, Zn-zdi: 0, Zn-zdt: 0,<br>Zn-zed: 2, Zn-zem: 20 |
| istim | enakopisni glagol *istiti* | G: 0<br>Z: 6 | Ggvspe: 0<br>Zn-meo: 0, Zn-mmd: 1, Zn-seo: 5,<br>Zn-smd: 0, Zn-zmd: 0 |
| jaz | enakopisni samostalnik *jaz* | S: 0<br>Z: 118 | Somei: 0, Sometn: 0<br>Zop-ei: 118 |
| jest | nestandardna oblika za zaimek *jaz* | G: 0<br>Z: 7 | Ggnm: 0<br>Zop-ei: 7 |
| kaj | enakopisni samostalnik *kaja* + enakopisni prislov | R: 92<br>S: 0<br>Z: 582 | Rsn: 92<br>Sozdr: 0, Sozmr: 0<br>Zv-sei: 274, Zv-set: 307, Zv-ser: 1 |
| kaka | enakopisni glagol *kakati* | G: 0<br>Z: 7 | Ggnste: 0<br>Zv-mdi: 0, Zv-mdt: 0, Zv-smi: 0, Zv-smt: 0,<br>Zv-zei: 7 |

14 *Enako, kako, kar, nekaj, nekak, vse, čemu* (prislov, zaimek); *vas, tema, temi, mene* (samostalnik, zaimek); *nič, tem* (prislov, samostalnik, zaimek); *tako, čim* (prislov, veznik, zaimek); *si* (glagol, zaimek), *meni* (glagol, samostalnik, zaimek); *tak* (medmet, zaimek); *vi* (številnik, zaimek). Od izpuščenih primerov gre izpostaviti obliki *kva* (prislov, zaimek) in *neki* (členek, zaimek), kjer gre pri eni ali obeh besednih vrstah za nestandardno obliko, na drugi strani pa primere *me, mi, ti, mu, je, one, ta, to,* kjer je pogosto rabljena zaimenska oblika enakopisna s tujejezično obliko, označeno kot 'Nj'.

| Dvoumna oblika | Vrsta problema | POS s frekvenco | Oblikoskladenjske oznake s frekvenco |
|---|---|---|---|
| kaki | enakopisni samostalnik *kaki* | S: 0<br>Z: 2 | Somei: 0, Sometn: 0<br>Zv-mmi: 0, Zv-sdi: 0, Zv-sdt: 0, Zv-zdi: 0, Zv-zdt: 0, Zv-zed: 2, Zv-zem: 0 |
| koje | tujejezična oblika v paru z arhaično slovensko | N: 1<br>Z: 0 | Nj: 1<br>Zv-mmt: 0, Zv-sei: 0, Zv-set: 0, Zv-zer: 0, Zv-zmi: 0, Zv-zmt: 0 |
| koji | tujejezična oblika v paru z arhaično slovensko | N: 1<br>Z: 0 | Nj: 1<br>Zv-mei: 0, Zv-met: 0, Zv-mmi: 0, Zv-sdi: 0, Zv-sdt: 0, Zv-zdi: 0, Zv-zdt: 0, Zv-zed : 0, Zv-zem: 0 |
| kolik | enakopisni samostalnik *kolika* | S: 2<br>Z: 0 | Sozdr: 0, Sozmr: 2<br>Zv-mei: 0, Zv-met: 0 |
| kolike | enakopisni samostalnik *kolika* | S: 2<br>Z: 0 | Sozer: 0, Sozmi: 1, Sozmt: 1<br>Zv-mmt: 0, Zv-zer: 0, Zv-zmi: 0, Zv-zmt: 0 |
| koliko | enakopisni samostalnik *kolika* | R: 90<br>S: 0<br>Z: 0 | Rsn: 90<br>Sozeo: 0, Sozet: 0<br>Zv-sei: 0, Zv-set: 0, Zv-zeo: 0, Zv-zet: 0 |
| kom | enakopisni samostalnik *koma*, trenutno okrajšava za *komad* v nestandardnem zapisu brez pike | S: 1<br>Z: 8 | Somei: 1<br>Sozdr: 0, Sozmr: 0<br>Zv-mem: 2, Zv-meo: 6 |
| mano | enakopisni samostalnik *mana* | S: 0<br>Z: 14 | Sozeo: 0, Sozet: 0<br>Zop-eo: 14 |
| moj | enakopisni samostalnik *moa* | S: 0<br>Z: 76 | Sozdr: 0, Sozmr: 0<br>Zspmeie: 62, Zspmete: 14 |
| mnogo | enakopisni prislov *mnogo* (in zaimek s./ž. spol) | R: 45<br>Z: 0 | Rsn: 45<br>Zn-sei: 0, Zn-set: 0, Zn-zeo: 0, Zn-zet: 0 |
| nekako | enakopisni prislov *nekako* (in zaimek s./ž. spol) | R: 51<br>Z: 0 | Rsn: 51<br>Zn-sei: 0, Zn-set: 0, Zn-zeo: 0, Zn-zet: 0 |
| nekoliko | enakopisni prislov *nekoliko* (in zaimek s./ž. spol) | R: 159<br>Z: 0 | Rsn: 159<br>Zn-sei: 0, Zn-set: 0, Zn-zeo: 0, Zn-zet: 0 |
| njem | nestandardna oblika (*n*, ki se sklanja brez vezaja) | S: 0<br>Z: 123 | Someo: 0, Sommd: 0<br>Zotmem: 106, Zotsem: 17 |

| Dvoumna oblika | Vrsta problema | POS s frekvenco | Oblikoskladenjske oznake s frekvenco |
|---|---|---|---|
| nje | nestandardna oblika (*n*, ki se sklanja brez vezaja) | S: 0<br>Z: 45 | Sommt: 0<br>Zotmmt: 1, Zotsmt: 0, Zotzer: 44, Zotzmt: 0 |
| nji | nestandardna oblika (*n*, ki se sklanja brez vezaja) | S: 0<br>Z: 1 | Sommi: 0, Sommo: 0<br>Zotzed: 0, Zotzem: 1 |
| njih | nestandardna oblika (*n*, ki se sklanja brez vezaja) | S: 0<br>Z: 156 | Somdm: 0, Sommm: 0<br>Zotmdr: 0, Zotmmm: 44, Zotmmr: 44, Zotmmt: 10, Zotsmm: 5, Zotsmr: 6, Zotsmt: 0, Zotzmm: 23, Zotzmr: 23, Zotzmt: 1 |
| oboje | enakopisni samostalnik *oboj* | S: 0<br>Z: 11 | Sommt: 0<br>Zc-mmt: 1, Zc-sei: 7, Zc-set: 3, Zc-zer: 0, Zc-zmi: 0, Zc-zmt: 0 |
| oboji | enakopisni samostalnik *oboj* | S: 0<br>Z: 3 | Sommi: 0, Sommo: 0<br>Zc-mmi: 3, Zc-sdi: 0, Zc-sdt: 0, Zc-zdi: 0, Zc-zdt: 0, Zc-zed: 0, Zc-zem: 0 |
| obojih | enakopisni samostalnik *obo*j | S: 0<br>Z: 1 | Somdm: 0, Sommm: 0<br>Zc-mdm: 1, Zc-mdr: 0, Zc-mmm: 0, Zc-mmr: 0, Zc-sdm: 0, Zc-sdr: 0, Zc-smm: 0, Zc-smr: 0, Zc-zdm: 0, Zc-zdr: 0, Zc-zmm: 0, Zc-zmr: 0 |
| prednjo | enakopisni pridevnik *prednji* (in zaimek *predme*) | P: 0<br>Z: 1 | Ppnzeo: 0, Ppnzet: 0<br>Zotzet--z: 1 |
| premnogo | enakopisni prislov *premnogo* (in zaimek s./ž. spol) | R: 1<br>Z: 0 | Rsn: 1,<br>Zn-sei: 0, Zn-set: 0, Zn-zeo: 0, Zn-zet: 0 |
| takole | enakopisni prislov *takole* (in zaimek s./ž. spol) | R: 29<br>Z: 0 | Rsn: 29<br>Zk-sei: 0, Zk-set: 0, Zk-zeo: 0, Zk-zet: 0 |
| tele | enakopisni samostalnik *tele* | S: 0<br>Z: 2 | Sosei: 0, Soset: 0<br>Zk-mmt: 0, Zk-sdi: 0, Zk-sdt: 0, Zk-zdi: 0, Zk-zdt: 0, Zk-zer: 1, Zk-zmi: 1, Zk-zmt: 0 |
| toliko | enakopisni prislov *toliko* (in zaimek s./ž. spol) | R: 163<br>Z: 0 | Rsn: 163<br>Zk-sei: 0, Zk-set: 0, Zk-zeo: 0, Zk-zet: 0 |
| vate | enakopisni samostalnik *vata* | S: 2<br>Z: 0 | Sommt: 0, Sozer: 2, Sozmi: 0, Sozmt: 0<br>Zod-et--z: 0 |
| ve | enakopisni glagol *vedeti* | G: 76<br>Z: 0 | Ggnste: 76<br>Zodzmi: 0 |

| Dvoumna oblika | Vrsta problema | POS s frekvenco | Oblikoskladenjske oznake s frekvenco |
|---|---|---|---|
| ves | enakopisni samostalnik *vesa* | S: 0<br>Z: 102 | Sozdr: 0, Sozmr: 0<br>Zc-mei: 22, Zc-met: 80 |
| vsej | enakopisni glagol *vsejati* + nestandardna oblika za členek *vsej* | G: 0<br>Z: 49<br>L: 1 | Ggdvde: 0<br>Zc-zed: 5, Zc-zem: 44<br>L: 1 |

Rezultate je mogoče razdeliti v več skupin. Na eni strani je najti enakopisnost z nestandardnimi in tujejezičnimi oblikami (glede posebne obravnavane teh glej razdelek 3). Izjema je oblika *kom,* kjer je poved z nestandardno okrajšavo (*kom* namesto *kom.*) treba zamenjati s povedjo, ki vsebuje ustrezno obliko samostalnika *koma*. Z vidika zaznamovanosti dodatno izstopata primera *koje, koji*, ki bosta kot 'Nj' posebej obravnavana (razdelek 3), vprašanje pa je, ali sta kot vprašalni zaimek v sodobni standardni slovenščini res še prisotna.

Za nadgradnjo učnega korpusa so relevantni predvsem primeri, kjer je zaimenska oblika enakopisna s polnopomensko besedo: vključiti želimo tako povedi, ki vsebujejo zaimensko obliko, kot povedi z enakopisnim samostalnikom, glagolom, pridevnikom ali prislovom. Glavno metodološko vprašanje je, ali oz. kdaj zaradi redkosti oblik v jezikovni rabi tovrstno (umetno) vključevanje postane kontraproduktivno. V pomoč je lahko frekvenca leme v referenčnem korpusu, pri čemer je treba upoštevati, da ravno pri obravnavanih primerih oblike in torej tudi leme niso natančno označene.

V prvi skupini navajamo primere, kjer bi bilo treba dodati povedi z nezaimensko obliko. V oklepaju je navedena frekvenca v korpusu Gigafida 2.0:[15] (a) glagoli *kakati* (484), *vsejati* (57) in *istiti* (57); (b) samostalniki *tele* (8.244), *jaz* (6.251), *oboj* (5.822, velika količina napačno lematiziranih), *mana* (3.463, veliko napačno lematiziranih), *kaki* (2.982, precej primerov pridevniške vrste),

15 Dostop prek platforme noSketch Engine (https://www.clarin.si/noske/), poizvedbe maj 2020, korpus Gigafida 2.0 (referenčni, dedupliciran, objavljen 11. 4. 2019). Pri izdelavi iskalnega pogoja so upoštevane oblike in besednovrstne oznake.

*vesa* (879, veliko lastnih imen), *moa* (785, veliko napačno lematiziranih) in *kaja* (314, veliko lastnih imen); (c) pridevnik *prednji* (30.325). Glede na podatke bi bilo v učni korpus dobro dodati primere *prednji, tele, jaz* in *kaki.* Redke leme *vesa, kakati, vsejati* in *istiti* se zdijo manj ključne, ne bi pa njihova vključitev škodovala, saj je zastopanost enakopisnih zaimkov pri vseh naštetih primerih zadostna. Kontraproduktivna bi lahko bila vključitev v rabi redke in najbrž arhaične leme *kaja*, prav tako vključitev oblik za leme *oboj*, *moa* in *mana*, saj trenutna lematizacija že sedaj napačno prepoznava zaimke kot samostalnike.

V drugi skupini so primeri, kjer bi bilo treba v korpus dodati zaimensko obliko: (a) glagol *vedeti* (869.903), (b) samostalnika *vata* (3.861) in *kolika* (1.019); (c) prislovi *mnogo* (89.142), *premnogo* (125), *nekako* (87.711), *nekoliko* (302.982), *takole* (40.483), *toliko* (324.313). Medtem ko so polnopomenske oblike v rabi dobro zastopane, se nekateri enakopisni zaimki pojavljajo relativno redko. Glede na podatke o pogostosti bi bilo v učni korpus smiselno dodati povedi z manjkajočimi osebnimi zaimki *ve* in *vate.* Pri ostalih zaimkih je možno upoštevati (zaradi napak sicer manj zanesljivo) frekvenco oblike: *nekako* (132), *takole* (261), *nekoliko* (74), *toliko* (44), *mnogo* (26), *koliko* (13), *kolik* (10), *kolike* (3), *premnogo* (2).

Po načinu gornje analize smo obravnavali tudi enakopisnost oblik pri zaimkih različne vrste. Za dvoumne se izkažejo le nekateri pari osebnih in kazalnih zaimkov: oblike *oni, one*, *ona, te* in *ti* so v učnem korpusu že reprezentirani tako kot osebni kot kazalni zaimki. Oblika *ono* je vključena le kot kazalni zaimek, ne pa tudi osebni, kar bi bilo mogoče dopolniti. Nabor vseh do sedaj opredeljenih dopolnitev in sprememb strnjujemo v razdelku 6.

## 3 Obravnava nestandardnih in tujejezičnih besedil

Učni korpus ssj500k je bil v začetku razvoja zasnovan kot splošni učni korpus za slovenščino, zato so bila vanj poleg standardnih pisnih vključena tudi nestandardna in transkribirana govorjena besedila. V vmesnem času so bili pod okriljem projekta Jezikoslovna

analiza nestandardne slovenščine[16] (Fišer et al. 2018) zgrajeni novi korpusni viri, v prvi vrsti korpus spletne slovenščine Janes 1.0, ki vsebuje slovenske tvite, forumska sporočila, blogovske zapise in komentarje na novice (skoraj 253 milijonov pojavnic), poleg tega pa nabor učnih korpusov, specializiranih za označevanje nestandardne spletne slovenščine Janes-Norm, Janes-Tag, Janes-Syn (Čibej et al. 2016, Erjavec et al. 2016, Arhar Holdt et al. 2016).

Skladno s tem razvojem je bil referenčni korpus Gigafida na prehodu v različico 2.0 posodobljen iz korpusa pisne v korpus pisne standardne slovenščine – iz njega so bila odstranjena besedila, za katera je bilo znano in predvideno, da so v njih prisotne nestandardne jezikovne prvine (odkloni od standardne slovenščine na ravni zapisa, besedišča, skladnje in sloga). To je zajemalo zlasti uporabniške spletne vsebine, npr. komentarje, forumska sporočila in druga spletna besedila, v nekaterih primerih pa tudi leposlovje (npr. prevod romana Trainspotting) in časopise (npr. lokalne medije, (delno) pisane v regionalni jezikovni različici slovenščine, kot je glasilo zamejskih Slovencev v Italiji – Novi Matajur). Postopek odstranjevanja nestandardnih besedil iz korpusa Gigafida je opisan v Krek et al. (2020a).

Iz podobnih razlogov in na primerljiv način bi bilo smiselno posodobiti tudi učni korpus ssj500k in v njem ustrezno označiti besedila, ki vsebujejo veliko nestandardnih jezikovnih prvin, oz. besedila, ki so v celoti oz. v večji meri iz tujejezičnih elementov. Razvojna skupnost na ta način lahko izbere in uporabi tiste dele korpusa, ki so optimalni za določeno nalogo. V nadaljevanju so predstavljene analize, na osnovi katerih je mogoče zasnovati tovrstno nadgradnjo.

## 3.1 Nestandardna besedila

V korpus ssj500k 2.2 so vključeni vzorci 1.655 besedil s 414 različnimi naslovi (pri nekaterih besedilih je naslov neznan, nekateri

16 Projekt Jezikoslovna analiza nestandardne slovenščine (J6-6842) je sofinancirala ARRS med letoma 2014 in 2018. Projektna spletna stran: http://nl.ijs.si/janes/.

vzorci pa so vzeti iz istega besedila in imajo zato enak naslov). Glede na zvrst (Slika 1) je večina besedil neumetnostnih (94 %), le manjši del pa umetnostnih (6 %).[17] Na tej točki je treba omeniti, da delitev na besedilne zvrsti, ki je uporabljena v korpusu ssj500k 2.2, ni skladna z delitvijo v Gigafidi 1.0 in 2.0, kjer so besedila razdeljena na spletna in tiskana, tiskana pa na periodična (časopisi in revije), knjižna (strokovna in leposlovna) in druga. Metapodatke v učnem korpusu bi bilo zato treba posodobiti v skladu z novo tipologijo, uporabljeno v Gigafidi, in obenem poskrbeti tudi za reprezentativnost korpusnega gradiva.

- **Vsa besedila** [1.655]
  - **Neumetnostna** [1.559]
    - **Neumetnostna** [8]
    - **Nestrokovna** [1.192]
    - **Strokovna** [359]
      - **Strokovna** [26]
      - **Naravoslovna in tehnična** [182]
      - **Humanistična in družboslovna** [151]
  - **Umetnostna** [96]
    - **Umetnostna** [4]
    - **Prozna** [88]
    - **Pesniška** [2]
    - **Dramska** [2]

**Slika 1:** Tipologija besedil iz korpusa ssj500k 2.2 glede na besedilne zvrsti.

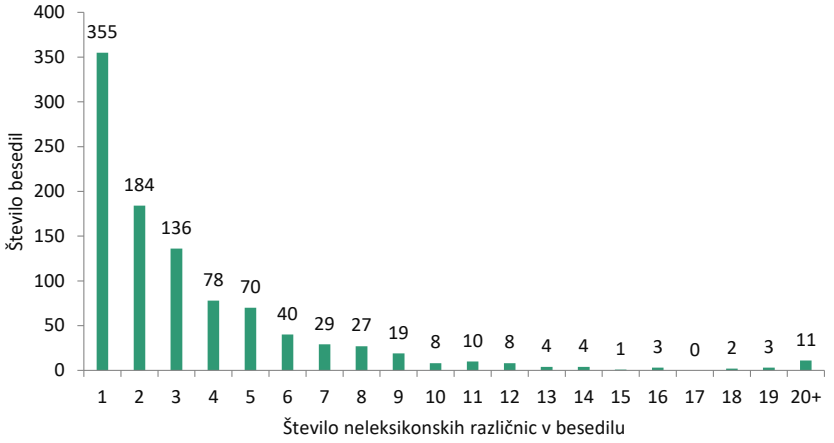
Preverili smo, katera besedila, ki so vključena v učni korpus, izstopajo od standardne jezikovne rabe na nivoju besedišča, tako da smo izvozili besedila glede na vsebnost različnih neleksikonskih različnic, tj. oblik, katerih kombinacija oblike, leme in oblikoskladenjske oznake ni zabeležena v oblikoslovnem leksikonu Sloleks in ki obenem niso ločila (‘U’), lastna imena (‘Sl.*’), svojilni pridevniki (‘Ps.*’), tujejezični elementi (‘Nj’) ali glavni števniki (‘Kag’). Če namreč oblike, ki pripadajo tem kategorijam, niso zabeležene v Sloleksu, je mnogo

17 Nekaj nekonsistentnosti se pokaže pri označenosti besedil z metapodatki, saj se npr. pod oznako neumetnostnih besedil pojavi 8 besedil, ki so prav tako označena zgolj kot neumetnostna (ne pa kot bodisi strokovna bodisi nestrokovna). Podobno je pri umetnostnih besedilih, kjer so 4 besedila označena le kot umetnostna, ne pa tudi kot bodisi prozna, pesniška ali dramska.

verjetneje, da oblika v leksikon (še) ni vključena, kot pa da gre za pravo nestandardno obliko.[18]

Vsaj eno neleksikonsko različnico smo zabeležili v 992 besedilih (60 % vseh besedil v učnem korpusu), a so v več kot 68 % teh besedil prisotne le tri tovrstne oblike ali manj. Razporeditev oblik po besedilih kaže Slika 2 (v graf so vključena le besedila, ki vsebujejo vsaj eno neleksikonsko različnico).

Po številu neleksikonskih različnic izstopata besedili z identifikacijskama številkama ssj369 in ssj370, ki sta vzeti iz slovenskega prevoda romana Trainspotting. V besedilih najdemo 165 neleksikonskih različnic (ssj370) oz. 57 (ssj369). Pregled neleksikonskih različnih pokaže veliko nestandardnih jezikovnih prvin, npr. nestandardne zapise (*tko, omenu*, *tistmu*, *blo*) in nestandardno besedišče (*falit*, *prbasan*, *štengah*). Besedili je ob nadgradnji učnega korpusa torej treba ustrezno označiti.




**Slika 2:** Razporeditev neleksikonskih različnic v besedilih korpusa ssj500k 2.2.

V preostalih besedilih tovrstnih prvin v tolikšni meri nismo zasledili oz. so bile prisotne le izjemoma (npr. kot občasne zatipkane

18 Pri izvozu neleksikonskih različnic se razkrijejo tudi napake pri ročnem označevanju in lematizaciji učnega korpusa (npr. oblika *devizah*, ki je lematizirana v *devize* namesto v *deviza*) oz. pomanjkljivosti v Sloleksu (npr. oblika *vneto*, ki je v Sloleksu 2.0 zabeležena samo kot pridevnik *vnet*, ne pa kot prislov *vneto*). S tako zaznanimi popravki je torej mogoče izboljšati tako učni korpus kot oblikoslovni leksikon.

oblike). Naslednja tri besedila z največ neleksikonskimi različnicami so npr. vzeta iz poljudnoznanstvene revije Življenje in tehnika (ssj1378, 32 različnic, in ssj744, 24 različnic) ter iz Družinske enciklopedije zdravil (ssj1663, 30 različnic). Veliko zabeleženih oblik je v teh primerih iz kategorije specializiranega besedišča (npr. *feromagnete*, *panemon*, *hematopoetske*, *vazodilatator*), zato besedila z vidika nestandardnosti niso problematična. Podobno velja tudi za naslednjih 41 besedil, ki vsebujejo več kot 10 neleksikonskih različnic – večina jih je bodisi specializiranih (*laminacijo*, *razhroščevanja, flavonolni, oligomerni*) oz. še nevključenih v leksikon (*štajerščine*, *identificirajoča*).

Potencialno problematični sta še besedili ssj1505 (Access za Windows 95 v uporabi; 23 različnic, npr. izseki kode in računalniških ukazov, *HTML*, *source*, *if*, *arguments;* zatipkani izrazi, *konkurečen*) in ssj384 (brez naslova; 21 različnic, prav tako s področja računalništva, *submit*, *IMG*, *border*, *explorer*, *px*).

Iz besedil smo izvozili tudi oblike, ki so v Sloleksu označene kot nestandardne. Vsebovane so v 33 besedilih (2 % vseh besedil v korpusu), a je v 29 besedilih prisotna le ena takšna oblika. Vseh povedi, ki vsebujejo vsaj eno nestandardno obliko, je 47. Ponovno izstopata besedili ssj369 in ssj370 iz romana Trainspotting, ki v 13 povedih vsebujeta skupno 5 različnih nestandardnih oblik: *kva*, *vseen*, *reku*, *jest*, *mal*. Preostalih 34 povedi vsebuje npr. nestandardne oblike *otroci* (kot orodnik množine), *Sydneya*, *LCDjev* in *prizadane*. Tudi tem povedim je torej treba pripisati ustrezno oznako.

## 3.2 Povedi s pojavnicami, označenimi kot Neuvrščeno

Korpus ssj500k 2.2 vsebuje 27.829 povedi, od teh jih 457 vsebuje vsaj eno pojavnico, ki je označena z oblikoskladenjsko oznako neuvrščeno (‘N’) oz. tujejezično (‘Nj’). 16 od teh povedi je iz prevoda romana Trainspotting, v katerem so bili z oblikoskladenjsko oznako ‘N’ označeni skupaj pisani besedni nizi (*čeuva*, *navjo*, *dab*, *bga*, *tlelevš*). Preostalih 441 povedi smo ročno pregledali in glede na vsebino označili kot problematične oz. neproblematične. Kot problematičnih

je bilo označenih 205 povedi (približno 45 % vseh povedi, ki vsebujejo ‘N’ oz. ‘Nj’). 138 povedi je bilo problematičnih zaradi prevelike vsebnosti oznak ‘N’, 58 pa zaradi prevelike vsebnosti oznak ‘Nj’.

V prvi skupini pogosto najdemo povedi z inicialkami avtorjev (zgled 1), spletnimi naslovi in številkami (2), računalniškimi ukazi in izseki kode (3), navedbami del (4) oziroma nestandardnimi jezikovnimi prvinami (5).

[1] *(pk, dm)*, ssj715.3575.12701
[2] *Več na www.pohodafestival.sk*, ssj1415.6902.23897
[3] *- rw-r--r--1 root root 3315 Jun 2 1997 CHARSETS*, ssj480.2595.9255
[4] *Psychology, Harper& Row, New York, 1987.*, ssj570.2945.10468
[5] *Gospa Stepperjeva in jest mor’va skuhat’ kosilo.*, ssj75.471.176

V drugi skupini so povedi, ki so v celoti (zgled 6) ali v veliki večini v tujem jeziku (7).

[6] *»Pa pukla je Avstrija – preskupo je, ljudi nemaju para za Prater i gađanje.«*, ssj90.598.2258
[7] *Madame ne mangera pas de marrons glacés? se je zarežal le petit.*, ssj75.479.1799

Med 138 povedi, ki so vsebovale pojavnice ‘N’, je bilo 18 povedi označenih kot delno problematičnih – neuvrščene pojavnice so sicer v manjšini, a so nastale zaradi napačne tokenizacije (zgleda 8 in 9; problematične pojavnice so podčrtane).

[8] *Sprememba drugega< HEAD> v</ HEAD> in</ H3> v</ H1> odpravi tudi ti dve napaki.*, ssj385.2212.7838
[9] *Marsikdo ne ve, da je leta 1981 g (dč) a Hiteova izdala podobno študijo o moških.*, ssj860.4223.14915

Od 58 povedi, ki so vsebovale pojavnice ‘Nj’, je bilo 22 označenih kot delno problematičnih – gre npr. za povedi, v katerih se pojavljajo citati v tujem jeziku (zgled 10) oz. v katerih tujejezične prvine zajemajo večji del povedi, ki pa je kljub temu legitimna (11).

[10] *»Con permiso!« je zavpil bolniški strežnik.*, ssj594.3044.10788
[11] *(NEMŠKO: SCHLAG; FRANCOSKO: COUP; ČEŠKO: RÁNA; SLOVAŠKO: RANA; HRVAŠKO: UDARAC; ŠPANSKO: GOLPE; DANSKO: SLAG; ŠVEDSKO: SLAG) Udarec je gibanje palice navzdol, ki ga igralec naredi z namenom, da bi udaril po žogici in jo premaknil.*, ssj557.2904.10287

## 3.3 Druge problematične povedi

Med analizo je bilo označenih tudi 125 povedi, ki ne vsebujejo niti neuvrščenih niti tujejezičnih pojavnic, a so kljub temu problematične. Gre predvsem za zelo kratke povedi, ki vsebujejo navedbe avtorjev fotografij (63 povedi; zgled 12), vzorce za navajanje literature in telefonske številke (24 povedi, zgled 13).

[12] *(Foto: T. G.)*, ssj55.352.1422
[13] *V: Geschichtliche Grundbegriffe, Stuttgart 1975, 2. zv., 647 nn. /32*, ssj112.705.2653

Posebna kategorija so povedi, ki so vzete iz prepisov sej Državnega zbora kot vzorec govorjenih besedil. Tovrstnih problematičnih povedi je bilo označenih 22, večinoma pa vsebujejo le opombe o številu prisotnih poslancev (zgled 14), nekaj pa je izjav, ki so tudi napačno segmentirane (zgleda 15 in 16 bi denimo morala biti združena v isto poved, kar nakazuje, da je na določenih mestih v korpusu treba popraviti stavčno segmentacijo).

[14] *(71 prisotnih)*, ssj142.936.3584
[15] *Nadaljevali bomo s 3.*, ssj142.916.3502
[16] *TOČKO DNEVNEGA REDA – PREDLOG ZAKONA O GOZDOVIH, z razpravo, h kateri imamo še nekaj prijavljenih.*, ssj142.916.3503

Preostanek sestoji iz naslednjega: dve povedi, ki izvirata iz besedil, ki niso bila uradno objavljena; 9 povedi, pri katerih je v samo poved zajeta tudi številka strani (zgled 17); fragmentirane povedi s *press*, ki vsebujejo tudi nestandardne zapise besed (4 povedi; zgled 18) in ena prazna poved.

[17] *17 Ti odgovori seveda zrcalijo zgolj predstave vprašanih pred njihovo lastno zakonsko izkušnjo*.
[18] *(slavnostni govornik bo nikola keramičar press)*, ssj818.4049.14316

Smernice za označevanje nestandardnih in tujejezičnih besedil, ki so osnovane na predstavljenih podatkih, navajamo v razdelku 6.

# 4 Posodobitev označevalnega sistema MULTEXT-East in prilagoditev označevanja

Označevalni sistem MULTEXT-East v6, ki je bil uporabljen za označevanje Gigafide 2.0 in ssj500k 2.2, trenutno vključuje tudi oznake, ki se nanašajo specifično na nestandardne jezikovne prvine. Vsebinska posodobitev obeh korpusov z vidika standardnosti besedil, ki je bila predstavljena v razdelku 3, zahteva premislek, ali je smiselno skladno z novostmi urediti tudi nabor oblikoskladenjskih oznak. Pri tem gre dodati, da nekatere od teh oznak tudi za označevanje nestandardnega jezika v resnici niso v rabi: v okviru gradnje korpusa nestandardne spletne slovenščine Janes (Fišer et al. 2018) je princip označevanja temeljil na normalizaciji nestandardnih oblik v standardne (npr. *nebom* → *ne bom*), normalizirane oblike pa so bile nato označene z oblikoskladenjskimi oznakami, ki se nanašajo na standardne oblike. V nadaljevanju predstavljamo podroben pregled problematičnih oblikoskladenjskih oznak in podamo predloge za spremembo označevalnega sistema.

## 4.1 Oznake za nestandardne oblike pomožnega glagola *biti* in drugih glagolov

V označevalni shemi so problematične predvsem oznake za nestandardne oblike pomožnega glagola *biti*, ki v primerjavi s standardno različico izražajo dodatne slovnične lastnosti (npr. *sve*, ki izraža dvojino in ženski spol).

Sedem oznak, ki jih prikazuje Tabela 4, ni dokumentiranih v specifikacijah označevalnih sistemov MULTEXT-East v4 in v6, a so kljub

temu prisotne v korpusih Gigafida 1.0 in Gigafida 2.0 (ne pa tudi v učnem korpusu ssj500k 2.2). Vse navedene oznake so zelo redke in v večjem delu primerov pripisane napačno. Oznaka 'Gp-ppdzd' (*nebove*) je denimo pripisana zemljepisnemu lastnemu imenu *Nebove,* oznaka 'Gp-ppdzn' (*bove*) se večinoma pojavlja pri osebnem in zemljepisnem imenu (*José Bove*, dolina *Bove*), podobno je pri 'Gp-ppmzn' *(bome: Med pripravami je bilo delavcem naročeno, naj vrata *bome* zapahnejo.)* in pri oznaki 'Gp-spmzd' (*nisme*), kjer gre večinoma za zatipkane oblike *nisem*.

**Tabela 4:** Nedokumentirane oznake za nestandardne oblike glagola *biti*.

| Oznaka | Značilnosti | Primer pojavnice | Pogostost (GF1.0) | Pogostost (GF2.0) | Pogostost (ssj500k 2.2) |
|---|---|---|---|---|---|
| Gp-pdm-d | glagol vrsta=pomožni oblika=prihodnjik oseba=druga število=množina nikalnost=zanikani | neboste | 253 | 0 | 0 |
| Gp-ppdzd | glagol vrsta=pomožni oblika=prihodnjik oseba=prva število=dvojina spol=ženski nikalnost=zanikani | nebove | 17 | 12 | 0 |
| Gp-ppdzn | glagol vrsta=pomožni oblika=prihodnjik oseba=prva število=dvojina spol=ženski nikalnost=nezanikani | bove | 69 | 44 | 0 |
| Gp-ppmzn | glagol vrsta=pomožni oblika=prihodnjik oseba=prva število=množina spol=ženski nikalnost=nezanikani | bome | 25 | 13 | 0 |
| Gp-ptd-d | glagol vrsta=pomožni oblika=prihodnjik oseba=tretja število=dvojina nikalnost=zanikani | nebosta | 37 | 0 | 0 |
| Gp-spdzd | glagol vrsta=pomožni oblika=sedanjik oseba=prva število=dvojina spol=ženski nikalnost=zanikani | nisve | 1 | 0 | 0 |
| Gp-spmzd | glagol vrsta=pomožni oblika=sedanjik oseba=prva število=množina spol=ženski nikalnost=zanikani | nisme | 66 | 4 | 0 |

Oznake, ki jih prikazuje Tabela 5, še vedno ostajajo v specifikacijah MULTEXT-East v6. Nekatere se nanašajo zgolj na nestandardne oblike, nekatere pa so problematične, ker se nanašajo na standardne oblike, a so zasnovane po sistemu, ki dopušča tudi oznake za nestandardne oblike – predvsem v primerih, ko oznaka opredeljuje (ne) zanikanost, standardna pa je le nezanikana oblika (npr. *neboš*/*boš*).

**Tabela 5:** Oznake za nestandardne oblike glagolov in z njimi povezane oznake v specifikacijah MULTEXT-East v6.

| Oznaka | Značilnosti | Primer pojavnice | Pogostost (GF1.0) | Pogostost (GF2.0) | Pogostost (ssj500k 2.2) |
|---|---|---|---|---|---|
| Gp-spdzn | glagol vrsta=pomožni oblika=sedanjik oseba=prva število=dvojina spol=ženski nikalnost=nezanikani | sve | 3.027 | 1.369 | 0 |
| Gp-g---d | glagol vrsta=pomožni oblika=pogojnik nikalnost=zanikani | nebi | 0 | 0 | 0 |
| Ggvspdz | glagol vrsta=glavni vid=dvovidski oblika=sedanjik oseba=prva število=dvojina spol=ženski | greve | 49 | 46 | 0 |
| Gp-ppe-n | glagol vrsta=pomožni oblika=prihodnjik oseba=prva število=ednina nikalnost=nezanikani | bom | 438.560 | 366.379 | 172 |
| Gp-ppe-d | glagol vrsta=pomožni oblika=prihodnjik oseba=prva število=ednina nikalnost=zanikani | nebom | 0 | 0 | 0 |
| Gp-ppm-n | glagol vrsta=pomožni oblika=prihodnjik oseba=prva število=množina nikalnost=nezanikani | bomo | 729.526 | 651.941 | 290 |
| Gp-ppm-d | glagol vrsta=pomožni oblika=prihodnjik oseba=prva število=množina nikalnost=zanikani | nebomo | 479 | 4 | 0 |
| Gp-ppd | glagol vrsta=pomožni oblika=prihodnjik oseba=prva število=dvojina | bova | 0 | 0 | 0 |

| Oznaka | Značilnosti | Primer pojavnice | Pogostost (GF1.0) | Pogostost (GF2.0) | Pogostost (ssj500k 2.2) |
|---|---|---|---|---|---|
| Gp-ppd-n | glagol vrsta=pomožni oblika=prihodnjik oseba=prva število=dvojina nikalnost=nezanikani | bova / boma | 33.715 | 33.295 | 8 |
| Gp-ppd-d | glagol vrsta=pomožni oblika=prihodnjik oseba=prva število=dvojina nikalnost=zanikani | nebova | 24 | 18 | 0 |
| Gp-pde-n | glagol vrsta=pomožni oblika=prihodnjik oseba=druga število=ednina nikalnost=nezanikani | boš | 112.838 | 81.996 | 33 |
| Gp-pde-d | glagol vrsta=pomožni oblika=prihodnjik oseba=druga število=ednina nikalnost=zanikani | neboš | 447 | 7 | 0 |
| Gp-pte-n | glagol vrsta=pomožni oblika=prihodnjik oseba=tretja število=ednina nikalnost=nezanikani | bo | 5.859.226 | 5.992.004 | 2.283 |
| Gp-pte-d | glagol vrsta=pomožni oblika=prihodnjik oseba=tretja število=ednina nikalnost=zanikani | nebo | 0 | 0 | 0 |
| Gp-ptm-n | glagol vrsta=pomožni oblika=prihodnjik oseba=tretja število=množina nikalnost=nezanikani | bodo, bojo | 2.498.082 | 2.661.368 | 1.037 |
| Gp-ptm-d | glagol vrsta=pomožni oblika=prihodnjik oseba=tretja število=množina nikalnost=zanikani | nebodo, nebojo | 1.163 | 11 | 0 |

Treh oznak – ‘Gp-g---d’ (*nebi*), ‘Gp-ppe-d’ (*nebom*) in ‘Gp-pte-d’ (*nebo*) ni mogoče najti v nobenem korpusu, kar je indikator, da najverjetneje oznake nikoli niso pravilno pripisane. Pri nekaterih drugih oznakah, ki jih ni v učnem korpusu ssj500k 2.2, a so prisotne v obeh različicah Gigafide, je mogoče opaziti zelo strm upad pojavljanja ob

prehodu z Gigafide 1.0 na Gigafido 2.0, npr. ‘Gp-spdzn’ (*sve*; 55-odstotni upad), ‘Gp-ppm-d’ (*nebomo*; več kot 99-odstotni upad), ‘Gp--pde-d’ (*neboš*; več kot 98-odstotni upad) in ‘Gp-ptm-d’ (*nebodo*, *nebojo*; več kot 99-odstotni upad). Pri obliki *sve*, pri kateri upad ni tako strm, najdemo veliko napačno označenih primerov, v katerih gre v resnici za tujejezične zapise v srbščini, hrvaščini, bosanščini itn. Podobno je z oznako ‘Ggvspdz’ (*greve*), ki ima v Gigafidi 2.0 le 46 zadetkov, večina je označenih napačno, saj gre za lastno ime *Greve* (kot priimek).

Oznake ‘Gp-ppd’ (*bova*), ‘Gp-ppd-n’ (*bova/boma*) in ‘Gp-ppd-d’ (*nebova*) izkazujejo nekonsistentnost v označevalnem sistemu, saj sta za isto obliko (*bova*) na voljo dve konkurenčni oznaki. Pri tem se najbolj splošna oznaka ‘Gp-ppd’, ki ne vsebuje kategorije zanikanosti, v korpusih sploh ne pojavi. Zanikanost je kategorija, ki jo je smiselno imeti označeno pri glagolih *imeti* in *hoteti*, saj imata v standardni slovenščini tako zanikane (*nimam, nočem*) kot nezanikane oblike (*imam, hočem*). V primerih, ko gre za par oznak, ki označujeta zanikanost in nezanikanost, zanikanost pa je prisotna samo v nestandardnih oblikah (*bomo*/**nebomo*, *boš*/**neboš*), bi bilo smiselno odstraniti obe oznaki in ju nadomestiti z eno, ki ne vsebuje kategorije zanikanosti.

Tudi oznako ‘Gp-g---d’ (*nebi*) bi bilo smiselno odstraniti, ni pa treba popravljati njene sorodne oblike ‘Gp-g’ (*bi*). Ta ne vsebuje nikalnosti (kar pa je prav tako nekonsistentno s trenutnim sistemom, po katerem so poimenovane druge oznake, ki izražajo zanikanost ali nezanikanost).

## 4.2 Druge oznake

Tabela 6 prikazuje zaimkovne oznake, ki so bile odstranjene v specifikacijah MULTEXT-East v6. Ta sprememba ni pojasnjena oziroma dokumentirana. Nobena od oznak se ne pojavi v obravnavanih korpusih, a ostaja vprašanje, ali ne gre morda za podoben problem kot pri zaimkih *ve*/*me*, ki se ne pojavljata, ker ju ni v učnem korpusu, in ali ni bila ukinitev preuranjena. Zaimek ‘Zopsmi’ (npr. *me [dekleta]*)

bi bilo npr. smiselno pričakovati v besedilih, četudi redko. Odstranjene oznake bi bilo torej smiselno dodati in podrobneje preučiti, kako funkcionalne so v označevalnem sistemu.

**Tabela 6:** Oznake zaimkov, odstranjene v različici MULTEXT-East v6.

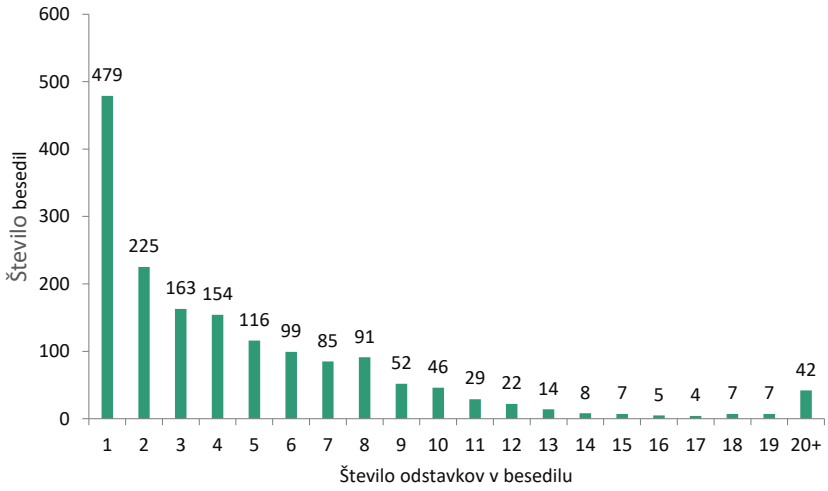
| Oznaka | Značilnosti | Primer pojavnice | Pogostost (GF1.0) | Pogostost (GF2.0) | Pogostost (ssj500k 2.2) |
|---|---|---|---|---|---|
| Zopsdi | zaimek vrsta=osebni oseba=prva spol=srednji število=dvojina sklon=imenovalnik | medve, midve | 0 | 0 | 0 |
| Zopsmi | zaimek vrsta=osebni oseba=prva spol=srednji število=množina sklon=imenovalnik | me | 0 | 0 | 0 |
| Zv----em | zaimek vrsta=vprašalni število_svojine=ednina spol_svojine=moški | katerega | 0 | 0 | 0 |
| Zv----ez | zaimek vrsta=vprašalni število_svojine=ednina spol_svojine=ženski | katere | 0 | 0 | 0 |
| Zv----es | zaimek vrsta=vprašalni število_svojine=ednina spol_svojine=srednji | katerega | 0 | 0 | 0 |
| Zv----d | zaimek vrsta=vprašalni število_svojine=dvojina | katerih | 0 | 0 | 0 |
| Zv----m | zaimek vrsta=vprašalni število_svojine=množina | katerih | 0 | 0 | 0 |

V korpusih manjkajo tudi nekatere oznake za podskupine kategorije Neuvrščeno, npr. napaka tokenizacije (‘Nt’), napaka programa (‘Np’). Uporabljene pa so oznake ‘Nj’ (tujejezično) ter oznake, ki so bile uvedene za označevanje elementov, ki jih najdemo v spletnih besedilih, npr. omembe uporabnikov (‘Na’), URL-naslovi (‘Nw’), ključniki (‘Nh’) ter emotikoni in emodžiji (‘Ne’). Omeniti je sicer treba, da so prav pri teh oznakah v korpusih trenutno določena neskladja, ki po vsej verjetnosti izhajajo iz rabe različnih označevalnikov (oz. njihovih različic). V prejšnjih različicah označevalnega sistema ločila niso imela pripisane oblikoskladenjske oznake, na

prehodu iz različice 4 v različico 6 pa je bila zaradi konsistentnosti dodana oznaka za ločila (‘U’).

## 5 Gradivna razdrobljenost in reprezentativnost

Ob načrtih za nadgradnjo in širitev učnega korpusa ssj500k smo preverili tudi, kako gradivno razdrobljen je učni korpus ssj500k 2.2 oz. kako obsežni so segmenti, ki so vzeti iz istega izvornega besedila. Ta podatek je pomemben za načrtovanje označevalnih nivojev, ki jih učni korpus še ne vsebuje in segajo preko meja povedi ali odstavka, npr. za označevanje koreferenčnosti in podobnih jezikovnih značilnosti. Slika 3 predstavlja razporeditev odstavkov po besedilih v korpusu ssj500k 2.2. Dobra polovica besedil (52 %) vsebuje tri odstavke ali manj, le 11 % besedil pa vsebuje 10 odstavkov ali več. V povprečju en odstavek vsebuje 3,42 povedi, poved pa v povprečju 18,83 pojavnice.




**Slika 3:** Razporeditev odstavkov po besedilih v korpusu ssj500k 2.2.

Za oceno trenutnega stanja se v prispevku osredotočamo na primernost korpusa za označevanje koreferenčnosti. Za slovenščino sta bila s koreferencami že označena korpusa coref149 (Žitnik 2018), ki zajema del besedil iz učnega korpusa ssj500k 1.4, in SentiCoref 1.0

(Žitnik 2019), ki vsebuje besedila iz korpusa SentiNews 1.0 (Bučar 2017) in ni prekriven s korpusom ssj500k. Coref149 vsebuje 149 odstavkov iz korpusa ssj500k, ki vsebujejo vsaj 100 besed in najmanj 6 imenskih entitet. To predstavlja le 2 % od 8.137 odstavkov uporabljene različice učnega korpusa.

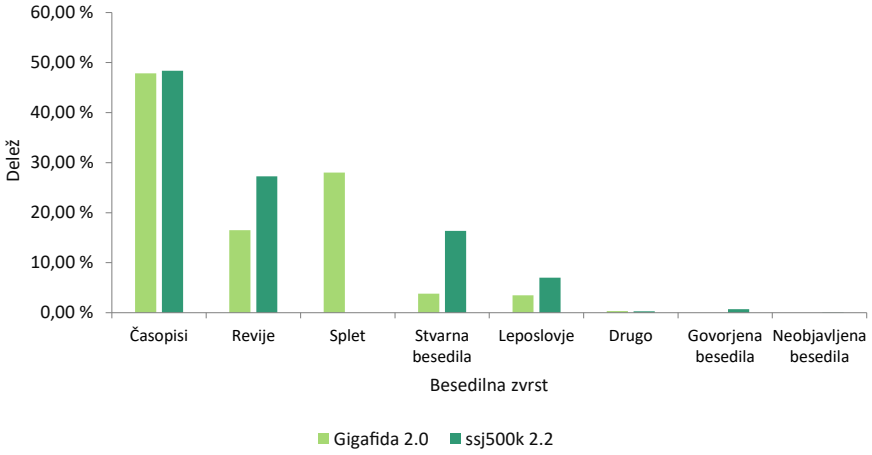
Če z vidika naštetih dveh kriterijev pogledamo besedila v korpusu ssj500k 2.2, ugotovimo, da njegova uporabnost ni dosti višja: kriterijem ustreza 193 (2,3 %) odstavkov. Še 151 odstavkov (1,9 %) z najmanj 100 besedami vsebuje od 2 do 5 imenskih entitet, 145 (1,7 %) pa je odstavkov s 50–100 besedami, ki vsebujejo vsaj 6 imenskih entitet. Ostaja še 1.015 odstavkov (12,5 %), ki vsebujejo najmanj 100 besed, a (zaenkrat) ne vsebujejo nobenih oznak za imenske entitete.

Kot je bilo omenjeno v Uvodu, je z imenskimi entitetami v različici ssj500k 2.2 označenih 9.488 povedi oz. 498 besedil, kar pomeni 30 % celotnega učnega korpusa (Krek et al. 2020b). Glede na kriterije gradnje korpusa coref149 je torej v ssj500k 2.2 za označevanje imenskih entitet in posledično koreferenc na voljo še nekaj gradiva, a bi tudi v primeru, da so vse omenjene kategorije odstavkov relevantne, to predstavljalo le 1.504 odstavke oz. dobrih 18 % celotnega korpusa. Po vseh ocenah je torej razdrobljenost korpusa ssj500k previsoka, da bi lahko služil kot učni korpus za označevanje koreferenčnosti. Pri njegovi nadaljnji širitvi je torej poleg vseh do sedaj naštetih želja treba upoštevati tudi to, da morajo biti besedila ustrezne dolžine.

Pri širjenju učnega korpusa je treba paziti tudi, da razširjena različica ostane karseda reprezentativen vzorec korpusa pisne standardne slovenščine Gigafida tako po časovni kot po besedilnozvrstni sestavi. Slika 4 prikazuje razporeditev besed po besedilnih zvrsteh v korpusih Gigafida 2.0 in ssj500k 2.2.[19] Največja razlika med korpusoma se pokaže pri spletnih besedilih, ki jih v učnem korpusu

19 Tipologija besedilnih zvrsti in prenosnika v ssj500k se razlikuje od tiste v Gigafidi. V tem prispevku smo za namene primerjave metapodatke iz ssj500k preslikali na metapodatke v Gigafidi, kar pa v nekaterih primerih ni povsem natančno, saj bi bil potreben natančnejši pregled besedil po naslovih (nekatera besedila glede na metapodatke v ssj500k lahko po tipologiji v Gigafidi 2.0 npr. sodijo bodisi pod revije bodisi pod leposlovje).

ssj500k ni, čeprav v Gigafidi zajemajo precejšen delež.[20] Razlog za to je po vsej verjetnosti to, da spletna besedila niso bila vključena v korpus FidaPlus, od koder je bil vzorčen korpus JOS1M, iz katerega je bil nato vzorčen ssj500k. Po drugi strani v Gigafidi ni neobjavljenih oz. govorjenih besedil, manjši delež pa je tudi leposlovja in stvarnih besedil.




**Slika 4:** Razporeditev besed po besedilnih zvrsteh v Gigafidi 2.0 in ssj500k 2.2.

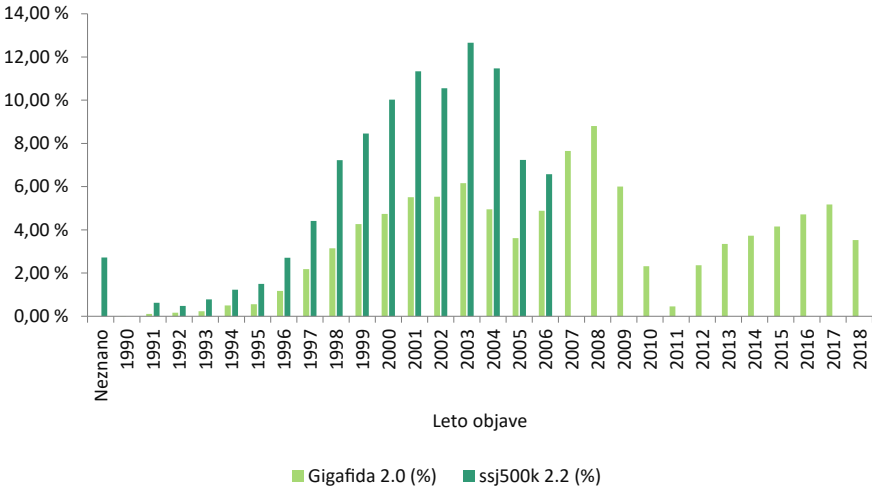
Tabela 7 prikazuje deleže besed v Gigafidi 2.0 in sskj500k 2.2 po besedilnih zvrsteh. V predzadnjem stolpcu je navedena razporeditev besed, če bi ssj500k razširili na milijon besed in ob tem upoštevali enako porazdelitev besedilnih zvrsti kot v Gigafidi 2.0. V zadnjem stolpcu je navedeno, koliko besed bi bilo potrebno dodati oziroma odvzeti, da bi dosegli takšno stanje. Odstraniti bi bilo treba govorjena in neobjavljena besedila, dodati pa predvsem spletna besedila (280.198 besed) in časopise (236.231 besed), manjši del pa tudi revij (28.840 besed) in drugih besedil (1.981 besed). Ker je delež stvarnih besedil v ssj500k 2.2 precej višji od tistega v Gigafidi 2.0, bi bilo ob upoštevanju nove porazdelitve treba iz učnega korpusa

20 V Gigafido 2.0 so bila v kategorijo spletnih besedil vključena tudi časopisna besedila, ki izhajajo na spletu in so bila v korpus vključena z zbiralnikom IJS Newsfeed (Krek et al. 2020a).

izločiti tudi 43.888 besed iz stvarnih besedil, a je glede na to, da je ssj500k označen ročno in na več ravneh (kar je časovno zamudno), te podatke smiselno obdržati kljub morebitnemu odstopanju od idealne porazdelitve.

**Tabela 7:** Primerjava razporeditve besed po besedilnih zvrsteh v korpusih Gigafida 2.0 in ssj500k 2.2.

| Zvrst | Gigafida 2.0 | Gigafida 2.0 (%) | ssj500k 2.2 | ssj500k 2.2 (%) | Razširjeni ssj500k | Sprememba v številu besed ob širitvi |
|---|---|---|---|---|---|---|
| Časopisi | 542.721.362 | 47,83 | 242.067 | 48,38 | 478.298 | +236.231 |
| Revije | 187.417.840 | 16,52 | 136.330 | 27,25 | 165.170 | +28.840 |
| Splet | 317.938.703 | 28,02 | 0 | 0 | 280.198 | +280.198 |
| Stvarna besedila | 42.944.398 | 3,78 | 81.735 | 16,34 | 37.847 | -43.888 |
| Leposlovje | 39.715.765 | 3,50 | 35.064 | 7,01 | 35.001 | -63 |
| Drugo | 3.955.865 | 0,35 | 1.505 | 0,30 | 3.486 | +1.981 |
| Govorjena besedila | 0 | 0 | 3.459 | 0,69 | 0 | -3459 |
| Neobjavljena besedila | 0 | 0 | 135 | 0,03 | 0 | -135 |
| Skupaj | 1.134.693.933 | 100,00 | 500.295 | 100,00 | 1.000.000 | +499.705 |




**Slika 5:** Razporeditev besed v korpusih Gigafida 2.0 in ssj500k 2.2 po letih objave besedil.

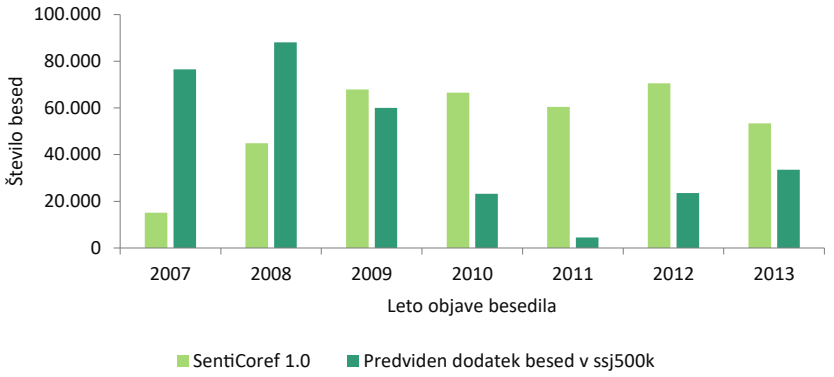
Slika 5 prikazuje razporeditev besed v korpusih Gigafida 2.0 in ssj500k 2.2 po letih objave besedil. Razvidno je, da v korpusu ssj500k primanjkuje predvsem novejših besedil iz let 2007–2018, saj vsebinsko že dalj časa ni bil posodobljen. Problematična so tudi besedila, pri katerih je metapodatek o letu objave neznan – pri teh bi bilo dobro metapodatke dopolniti, če jih je mogoče ugotoviti z natančnejšim pregledom besedil. Za učenje označevalnikov to načeloma nima posledic, je pa kljub temu dobro, da je korpus čimbolj reprezentativen in bogato označen z metapodatki, da ga je mogoče poljubno filtrirati.

Tabela 8 prikazuje deleže besed v Gigafidi 2.0 in sskj500k 2.2 po letih objave besedil, skupaj s predvidenimi dodatki h korpusu ssj500k, če bi bil razširjen na milijon besed in če bi pri tem ohranil enako porazdelitev besed po letih kot Gigafida 2.0. Iz tabele je razvidno, da so v učnem korpusu besedila iz večine let med 1991 in 2005 v primerjavi z Gigafido 2.0 nekoliko nadreprezentirana, kar pa odraža predvsem dejstvo, da so bila v Gigafido ob posodobitvah vključena predvsem novejša besedila, ni pa bilo dodanih novih besedil iz zgodnejših let. Glede na opravljeni razrez bi bilo v učni korpus največ besedil treba dodati za leta 2008 (88.086 besed), 2007 (76.492 besed) in 2009 (60.017 besed), manjše količine pa za ostala leta med 2007 in 2018 ter za leta 1990 (zanemarljiva količina), 1999 in 2002.

Preverili smo še, v kolikšni meri bi bilo za dopolnitev učnega korpusa mogoče uporabiti gradivo korpusa SentiCoref 1.0, ki je že označeno s koreferencami in imenskimi entitetami. V projekciji je predvidena širitev ssj500k na milijon besed, torej približno enkratna povečava njegovega trenutnega obsega. SentiCoref 1.0 vsebuje 837 besedil z novičarskih portalov rtvslo.si, 24ur.com, dnevnik.si, finance.si in zurnal24.com. Glede na tipologijo besedilnih zvrsti v Gigafidi 2.0 torej besedila sodijo med spletna. SentiCoref 1.0 zajema skupno približno 379.000 besed, kar je skoraj 76 % predvidenega povečanja ssj500k. Podrobnejši razrez korpusa SentiCoref 1.0 po letu objave besedila v primerjavi s predvidenimi dodatki h korpusu ssj500k glede na posamezno leto (glej Tabelo 8) prikazuje Slika 6.

**Tabela 8:** Deleži besed v Gigafidi 2.0 in sskj500k 2.2 po letih objave besedil.

| Leto objave | Gigafida 2.0 | Gigafida 2.0 (%) | ssj500k 2.2 | ssj500k 2.2 (%) | Razširjeni ssj500k | Sprememba v številu besed ob širitvi |
|---|---|---|---|---|---|---|
| Neznano | 0 | 0 | 13.584 | 2,72 | 0 | -13.584 |
| 1990 | 87.366 | 0,01 | 0 | 0 | 77 | +77 |
| 1991 | 1.225.109 | 0,11 | 3.127 | 0,63 | 1.080 | -2.047 |
| 1992 | 1.883.601 | 0,17 | 2.387 | 0,48 | 1.660 | -727 |
| 1993 | 2.670.988 | 0,24 | 3.933 | 0,79 | 2.354 | -1.579 |
| 1994 | 5.735.339 | 0,51 | 6.133 | 1,23 | 5.055 | -1.078 |
| 1995 | 6.311.833 | 0,56 | 7.489 | 1,50 | 5.563 | -1.926 |
| 1996 | 13.268.443 | 1,17 | 13.531 | 2,70 | 11.693 | -1.838 |
| 1997 | 24.745.780 | 2,18 | 22.088 | 4,41 | 21.808 | -280 |
| 1998 | 35.657.270 | 3,14 | 36.141 | 7,22 | 31.425 | -4.716 |
| 1999 | 48.421.615 | 4,27 | 42.318 | 8,46 | 42.674 | +356 |
| 2000 | 53.749.946 | 4,74 | 50.190 | 10,03 | 47.370 | -2.820 |
| 2001 | 62.566.212 | 5,51 | 56.732 | 11,34 | 55.139 | -1.593 |
| 2002 | 62.822.765 | 5,54 | 52.819 | 10,56 | 55.365 | +2.546 |
| 2003 | 69.916.212 | 6,16 | 63.341 | 12,66 | 61.617 | -1.724 |
| 2004 | 56.195.504 | 4,95 | 57.378 | 11,47 | 49.525 | -7.853 |
| 2005 | 41.105.613 | 3,62 | 36.232 | 7,24 | 36.226 | -6 |
| 2006 | 55.400.787 | 4,88 | 32.872 | 6,57 | 48.824 | +15.952 |
| 2007 | 86.795.219 | 7,65 | 0 | 0 | 76.492 | +76.492 |
| 2008 | 99.950.427 | 8,81 | 0 | 0 | 88.086 | +88.086 |
| 2009 | 68.100.586 | 6,00 | 0 | 0 | 60.017 | +60.017 |
| 2010 | 26.352.060 | 2,32 | 0 | 0 | 23.224 | +23.224 |
| 2011 | 5.155.242 | 0,45 | 0 | 0 | 4.543 | +4.543 |
| 2012 | 26.736.600 | 2,36 | 0 | 0 | 23.563 | +23.563 |
| 2013 | 38.002.753 | 3,35 | 0 | 0 | 33.492 | +33.492 |
| 2014 | 42.320.908 | 3,73 | 0 | 0 | 37.297 | +37.297 |
| 2015 | 47.152.788 | 4,16 | 0 | 0 | 41.556 | +41.556 |
| 2016 | 53.564.921 | 4,72 | 0 | 0 | 47.206 | +47.206 |
| 2017 | 58.709.992 | 5,17 | 0 | 0 | 51.741 | +51.741 |
| 2018 | 40.088.054 | 3,53 | 0 | 0 | 35.329 | +35.329 |




**Slika 6:** Primerjava korpusa SentiCoref 1.0 in predvidenega dodatka besed v korpus ssj500k po letu objave besedil.

Po scenariju enkratne povečave je za leta med 2009 in 2013 SentiCoref 1.0 preobsežen, zlasti če upoštevamo, da je treba v ssj500k poleg spletnih besedil dodati tudi časopise, revije in drugo, zapolniti pa je treba tudi vrzel za leta med 2014 in 2018. Določiti je torej treba smiselno kompromisno rešitev, ki obenem ohrani čim več podatkov iz korpusa SentiCoref 1.0 in v čim večji meri upošteva kriterije reprezentativnosti. Spoznanja predstavljenih analiz povzemamo v sledečem razdelku.

## 6 Smernice za nadgradnjo učnega korpusa ssj500k in leksikona Sloleks

Analiza je razkrila šibka mesta učnega korpusa ssj500k in identificirala možnosti za nadgradnjo tako kot korpusa kot tudi označevalnega sistema MULTEXT-East in oblikoslovnega leksikona Sloleks.

**Sistem oznak MULTEXT-East** je treba urediti predvsem na ravni vsebnosti oznak, ki so namenjene označevanju nestandardnih jezikovnih prvin. Glede na rezultate analiz (razdelek 4) predlagamo odstranitev oznak za nestandardni zapis pomožnega glagola *biti* (za npr. *nebom, greve*) in posodobitev parov, ki opredeljujejo (ne)zanikanost pri glagolih, kjer je standardna samo nezanikana različica zapisa (**neboš/boš*). Tako v označevalnem sistemu prisotne kot trenutno

nedokumentirane oznake za nestandardne oblike, ki se kljub temu pojavljajo v referenčnem korpusu (za npr. *bove, bome*), je pri označevanju prihodnjih različic smiselno nadomestiti z najbližjimi standardnimi ustreznicami, npr. *greve* ('Ggvspdz') označimo z enako oznako kot *greva* ('Ggvspd'). Na drugi strani bi bilo v sistem treba dodati manjkajoče oznake za zaimke srednjega spola (za npr. *medve, me*) in v povezavi z dopolnitvami leksikona tudi oznake za dvojino srednjega spola lastnih imen (**Sredozemlji*). Od sprememb, ki so nastale med različicama v4 in v6, kaže obdržati oznako 'U' za ločila ter nabor oznak za elemente, značilne za elemente iz spletnih besedil ('Na', 'Nh', 'Ne', 'Nw'). Odprto ostaja še vprašanje potencialno problematične oznake za predloge z imenovalnikom ('Di'), ki trenutno izstopa v sistemu, tudi glede na aktualne jezikovne priročnike.

Kot predpogoj za navedene spremembe označevalnega sistema je treba zagotoviti nadgradnjo **oblikoslovnega leksikona Sloleks**. Kot omenjeno, je treba dopolniti pomanjkljive lastnoimenske paradigme in preveriti vsebnost neželenih nestandardnih oblik. Rezultati analize (razdelek 2) pričajo tudi o potrebi po uvedbi oznak za arhaične oblike oz. leksikonske enote, ki bi pomagale pri razdvoumljanju enakopisnih besednih oblik (npr. vprašalni zaimek *koji,* samostalnik *kaja*).

Skladno z razvojem referenčnih korpusnih virov za slovenščino predlagamo **označitev nestandardnih delov učnega korpusa,** kar glede na analize (razdelek 3) obsega 291 problematičnih stavkov v skupnem obsegu 1.872 pojavnic. To znaša približno 0,4 odstotka celotnega korpusa, a lahko metapodatki o nestandardnih besedilih npr. omogočijo naprednejše in raznovrstne evalvacije označevalnikov in drugih orodij. Treba pa je zasnovati tipologijo oznak, saj pri vseh besedilih, ki so bila zaznana kot potencialno problematična, ne gre nujno samo za nestandardne jezikovne prvine, temveč za zelo specifične jezikovne elemente (npr. izseki računalniške kode).

Na drugi strani je treba zagotoviti **dopolnitev učnega korpusa za boljšo zastopanost dvoumnih oblikoskladenjskih oznak.** Kot kažejo rezultati (razdelek 2), v učnem korpusu manjkajo oznake, ki pokrivajo dvojinske oblike, kar vodi v napačno označevanje referenčnega korpusa z enakopisnimi oblikami v množini ali v neustrezni glagolski

osebi. V učni korpus bi bilo zato smiselno dodati nabor (približno 50 do 100) povedi, ki bi ciljno pokrile dvojinske oznake različnih besednih vrst (npr. za *bodiva, nista, imata, drugima*). Za nadgradnjo učnega korpusa so relevantni tudi primeri, kjer je zaimenska oblika enakopisna s polnopomensko besedo, npr. *prednji, tele, jaz, kaki, ve, vate*. Za vključitev so relevantne tudi zaimenske oblike, ki so pogoste v referenčnem in neobstoječe v učnem korpusu, npr. *tvojo, mojimi, najine, njunima*. Na drugi strani je iz korpusa treba odstraniti povedi, ki vsebujejo nestandardne in tujejezične enakopisne oblike, npr. *kva, neki, jest* ter *me, one, to*.

Glede na ocene (razdelek 5) učni korpus v trenutni različici ni primeren za označevanje jezikovnih prvin na odstavčni ravni. **Pri gradivnem širjenju korpusa** je zato nujno zagotoviti, da bodo besedila ustrezne dolžine (npr. vsaj 100 besed in vsaj 6 imenskih entitet) in zaključena – to v veliki meri razrešuje predlagani dodatek iz korpusa SentiCoref 1.0. Poskrbeti je treba tudi za uravnoteženost glede na besedilno vrsto ter leto izida: (a) odstraniti bi bilo treba govorjena in neobjavljena besedila, dodati pa predvsem spletna besedila (280.198 besed) in časopise (236.231 besed), manjši del pa tudi revij (28.840 besed) in drugih besedil (1.981 besed) in (b) največ besedil bi bilo treba dodati za leta 2008 (88.086 besed), 2007 (76.492 besed) in 2009 (60.017 besed), manjše količine pa za ostala leta med 2007 in 2018 ter za leta 1990 (zanemarljiva količina), 1999 in 2002.

Ob nadgradnji učnega korpusa je treba nenazadnje **posodobiti metapodatke o besedilni zvrsti**, da bodo skladni s tipologijo iz korpusa Gigafida. Problematična so tudi besedila, pri katerih je metapodatek o letu objave neznan – pri teh bi bilo dobro metapodatke dopolniti, če jih je mogoče ugotoviti z natančnejšim pregledom besedil.

Priložnost za nadgradnjo ponuja projekt Razvoj slovenščine v digitalni dobi, ki bo potekal med letoma 2020 in 2022 s finančno podporo Ministrstva za kulturo Republike Slovenije. Razvoj učnega korpusa ssj500k bo temeljil na predstavljenih analizah in bo zagotovil povečavo korpusa, dodatno označevanje na različnih ravneh in odpravo identificiranih pomanjkljivosti.

*Zahvala*

Prispevek je nastal s financiranjem Agencije za raziskovalno dejavnost Republike Slovenije, in sicer raziskovalnega projekta Nova slovnica sodobne standardne slovenščine: viri in metode (J6-8256) ter programske skupine Jezikovni viri in tehnologije za slovenski jezik (P6-0411). Avtorja se zahvaljujeva Dafne Marko za pomoč pri analizi tujejezičnih pojavnic v učnem korpusu in dr. Kaji Dobrovoljc za preliminarne analize nestandardnih prvin v korpusu. Zahvaljujeva se tudi obema recenzentoma za natančno branje in koristne predloge.

## Reference

Arhar Holdt, Š., Fišer, D., Erjavec, T. in Krek, S. (2016). Syntactic annotation of Slovene CMC: first steps. V D. Fišer in M. Beißwenger (ur.), *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities* (str. 3–6). Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: http://nl.ijs.si/janes/cmc-corpora2016/proceedings.

Bučar, J. (2017). Manually sentiment annotated Slovenian news corpus SentiNews 1.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1110.

Čibej, J., Arhar Holdt, Š., Erjavec, T. in Fišer, D. (2016). Razvoj učne množice za izboljšano označevanje spletnih besedil. V T. Erjavec in D. Fišer (ur.), *Zbornik konference Jezikovne tehnologije in digitalna humanistika* (str. 40–46). Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: http://www.sdjt.si/wp/wp-content/uploads/2016/09/JTDH-2016_Cibej-et-al_Razvoj-ucne-mnozice.pdf.

Dobrovoljc, K., Krek, S. in Erjavec, T. (2015). Leksikon besednih oblik Sloleks in smernice njegovega razvoja. V V. Gorjanc, Gantar, P., Kosem, I. in Krek, S. (ur.), *Slovar sodobne slovenščine: problemi in rešitve* (str. 80–105). Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/15/47/489-1.

Dobrovoljc, K., Erjavec, T. in Ljubešić, N. (2019a). Improving UD processing via satellite resources for morphology. V A. Rademaker in F. Tyers (ur.), *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)* (str. 24–34). Stroudsburg: Association for

Computational Linguistics. Dostopno prek: https://www.aclweb.org/anthology/W19-80.pdf.

Dobrovoljc, K., Krek, S., Holozan, P., Erjavec, T., Romih, M., Arhar Holdt, Š., Čibej, J., Krsnik, L. in Robnik-Šikonja, M. (2019b). Morphological lexicon Sloleks 2.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1230.

Erjavec, T. (2012). MULTEXT-East: morphosyntactic resources for Central and Eastern European languages. *Language Resources and Evaluation*, 46 (1), 131–142. https://doi.org/10.1007/s10579-011-9174-8.

Erjavec, T., Čibej, J., Arhar Holdt, Š., Ljubešić, N. in Fišer, D. (2016). Gold-standard datasets for annotation of Slovene computer-mediated communication. V A. Horák et al. (ur.), *RASLAN 2016: Recent Advances in Slavonic Natural Language Processing: proceedings* (str. 29–40). Brno: Tribun EU. Dostopno prek: https://nlp.fi.muni.cz/raslan/raslan16.pdf.

Fišer, D., Ljubešić, N. in Erjavec, T. (2018). The Janes project: language resources and tools for Slovene user generated content. *Language Resources and Evaluation*, 54 (1), 223–246. https://doi.org/10.1007/s10579-018-9425-z.

Grčar, M., Krek, S. in Dobrovoljc, K. (2012). Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. V T. Erjavec in J. Žganec Gros (ur.), *Zbornik Osme konference Jezikovne tehnologije* (str. 89–94). Ljubljana: Institut Jožef Stefan. Dostopno prek: http://nl.ijs.si/isjt12/proceedings/isjt2012_17.pdf.

Krek, S., Dobrovoljc, K., Erjavec, T., Može, S., Ledinek, N., Holz, N., Zupan, K., Gantar, P., Kuzman, T., Čibej, J., Arhar Holdt, Š., Kavčič, T., Škrjanec, I., Marko, D., Jezeršek, L. in Zajc, A. (2019). Training corpus ssj500k 2.2, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1210.

Krek, S., Arhar Holdt, Š., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešić, N., Kosem, I. in Dobrovoljc, K. (2020a). Gigafida 2.0: the reference corpus of written standard Slovene. V N. Calzolari (ur.), *LREC 2020: Twelfth International Conference on Language Resources and Evaluation: conference proceedings* (str. 3340–3345). Pariz: European Language Resources Association. Dostopno prek: http://www.lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf.

Krek, S., Erjavec, T., Dobrovoljc, K., Gantar, P., Arhar Holdt, Š., Čibej, J. in Brank, J. (2020b). The ssj500k Training Corpus for Slovene Language

Processing. V D. Fišer in T. Erjavec (ur.), *Jezikovne tehnologije in digitalna humanistika: zbornik konference* (str. 24–33). Ljubljana: Inštitut za novejšo zgodovino. Dostopno prek: http://nl.ijs.si/jtdh20/pdf/JT-DH_2020_Krek-et-al_The-ssj500k-Training-Corpus-for-Slovene--Language-Processing.pdf.

Ljubešić, N. in Erjavec, T. (2016). Corpus vs. Lexicon Supervision in Morphosyntactic Tagging: The Case of Slovene. V N. Calzolari (ur.), *LREC 2016: Tenth International Conference on Language Resources and Evaluation: proceedings* (str. 1527–1531). Pariz: European Language Resources Association. Dostopno prek: http://www.lrec-conf.org/proceedings/lrec2016/pdf/811_Paper.pdf.

Ljubešić, N. in Dobrovoljc, K. (2019). What does neural bring? Analysing improvements in morphosyntactic annotation and lemmatisation of Slovenian, Croatian and Serbian. V *BSNLP 2019: Proceedings of the workshop, The 7th Workshop on Balto-Slavic Natural Language Processing* (str. 29–34). Dostopno prek: https://www.aclweb.org/anthology/W19-3704.

Logar, N., Grčar, M., Brakuš, M., Erjavec, T., Arhar Holdt, Š. in Krek, S. (2012). *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba.* Ljubljana: Trojina, zavod za uporabno slovenistiko: Fakulteta za družbene vede. E-izdaja (2020). Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/233/333/5394-1.

Žitnik, S. (2018). Slovene coreference resolution corpus coref149, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1182.

Žitnik, S. (2019). Slovene corpus for aspect-based sentiment analysis – SentiCoref 1.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1285.