Lematizacija in oblikoskladenjsko označevanje korpusa SentiCoref

Eva Pori,* Jaka Čibej,* Tina Munda,† Luka Terčon,† Špela Arhar Holdt*†

* Filozofska fakulteta, Univerza v Ljubljani, Aškerčeva 2, 1000 Ljubljana
eva.pori@ff.uni-lj.si; jaka.cibej@ff.uni-lj.si; spela.arharholdt@ff.uni-lj.si

† Fakulteta za računalništvo in informatiko, Univerza v Ljubljani, Večna pot 113, 1000 Ljubljana
tina.munda@fri.uni-lj.si; luka.tercon@fri.uni-lj.si

Povzetek

V prispevku predstavimo proces in rezultate ročnega pregledovanja lem in oblikoskladenjskih oznak MULTEXT-East v6 korpusa SentiCoref, ki bo pod okriljem projekta Razvoj slovenščine v digitalnem okolju (RSDO) vključen v novi učni korpus za slovenščino (trenutni ssj500k). Opišemo delotoke označevalne kampanje, ki je ena najobsežnejših tega tipa v našem prostoru, označevalne dileme, ki so razkrile določene vrzeli v referenčnih označevalnih smernicah, kot tudi rešitve in rezultate, ki smo jih oblikovali med delom in jih bo mogoče uporabiti v prihodnje.

Lemmatization and Morphosyntactic Annotation in the SentiCoref Corpus

The paper presents the process and the results of manual lemma and MULTEXT-East v6 morphosyntactic tag annotation in the SentiCoref corpus, which is planned to be included in the new Slovene training corpus (currently known as ssj500k) as part of the "Development of Slovene in a Digital Environment" project. The paper describes the workflows of the annotation campaign – which was among the most extensive campaigns of this type in Slovenia –, the annotation dilemmas that revealed gaps in previous versions of annotation guidelines, as well as the resulting solutions that will be useful in future annotation campaigns.

1. Uvod

Med leti 2020 in 2023 s podporo Ministrstva za kulturo Republike Slovenije in Evropskega sklada za regionalni razvoj poteka aplikativni projekt Razvoj slovenščine v digitalnem okolju (RSDO).[1] Med cilji projekta je infrastruktura za kontinuirano grajenje slovenskih korpusov: delotoki sprotnega zbiranja besedil, označevalni cevovod in dokumentacija za označevanje na različnih jezikovnih ravneh ter nekatera nova orodja za ročno označevanje ter pregledovanje korpusnih podatkov. Kot temeljna jezikovna vira za razvoj cevovoda za strojno označevanje sodobne slovenščine sta v nadgradnjo vključena tudi leksikon besednih oblik Sloleks (Dobrovoljc et al., 2019) in učni korpus ssj500k (Krek et al., 2020), s katerim se povezuje tudi pričujoči prispevek.

Učni korpus ssj500k v različici 2.3 (Krek et al., 2021) vsebuje 27.829 povedi oz. 500.295 besednih pojavnic, označenih na ravneh od stavčne segmentacije, tokenizacije, lematizacije, oblikoslovja in oblikoskladnje prek odvisnostne skladnje, imenskih entitet in večbesednih leksemov do udeleženskih vlog. Kot je značilno za učne korpuse, so jezikoslovne oznake ročno pregledane, s čimer je dosežena zanesljivost, ki jo potrebujemo za nadzorovano učenje strojnih postopkov. Na rezultate vplivata tudi obseg in zastopanost gradiva, zato je glavni cilj nadgradnje povečanje učnega korpusa na 1.000.000 besednih pojavnic. Na projektu bo za višje, kompleksnejše nivoje označevanja pripravljeno omejeno število novooznačenih povedi, osnovni nivoji pa bodo ročno pregledani za vse novo gradivo.

V prispevku predstavljamo označevalno kampanjo, v kateri smo ročno pregledali in popravili tokenizacijo, segmentacijo, leme in oblikoskladenjske oznake sistema MULTEXT-East (Erjavec, 2012) v korpusu SentiCoref 1.0 (Žitnik, 2019), ki predstavlja približno 76 % predvidene povečave učnega korpusa.[2] SentiCoref vsebuje besedila z novičarskih portalov, v katera so ročno vpisane oznake koreferenc in imenskih entitet, in odgovarja na potrebo, da se v učni korpus vključi gradivo, ki omogoča označevanje jezikovnih značilnosti prek meja povedi (Arhar Holdt in Čibej, 2021).

Namen prispevka je opisati delo, rezultate, zlasti pa označevalne dileme na ravni lem in oblikoskladnje, ki so razkrile določene vrzeli v referenčnih označevalnih smernicah (Holozan et al., 2008), kot tudi rešitve, ki smo jih oblikovali med delom in jih je mogoče uporabiti za prihodnje primerljive naloge. Novi učni korpus bo skupaj z nadgrajenimi označevalnimi smernicami ob zaključku projekta RSDO odprto na voljo na repozitoriju CLARIN.SI.

2. Preteklo in sorodno delo

Učni korpus ssj500k se kot referenčni vir za nadzorovano učenje strojnega jezikoslovnega označevanja sodobnih slovenskih pisnih besedil razvija že več kot desetletje (Krek et al., 2020). Do sedaj so bili na tem korpusu naučeni različni označevalniki, npr. Obeliks (Grčar et al., 2012), ReLDI (Ljubešić in Erjavec, 2016), nevronski označevalnik, ki ga je razvil Belej (2018), in CLASSLA StanfordNLP (Ljubešić in Dobrovoljc, 2019), ki se nadalje razvija tudi na projektu RSDO.

Začetki učnega korpusa segajo v čas projekta MULTEXT-East, ki je spodbudil razvoj sistema za oblikoskladenjsko označevanje (tudi) slovenščine (Dimitrova et al., 1998). Sistem oznak je bil revidiran in nadgrajen pod okriljem projekta Jezikoslovno označevanje slovenščine (JOS), v katerem je nastal korpus jos100k (Erjavec in Krek, 2008). Nato je bilo v projektu

[1] Spletna stran, ki predstavlja projektne cilje in sodelujoče partnerje: https://slovenscina.eu/.

[2] Za preostalih 24 % so v načrtu raznolike besedilne množice, ki bodo zagotovile (a) temelje za semantično označevanje, kot npr. slovenska različica vzporednega korpusa Elexis-WSD (Martelli et al., 2022), (b) izbrane nezastopane besedilne vrste, npr. tvite, ki predstavljajo uporabniško generirane spletne vsebine, (c) v rabi redkejše dvoumne besedne oblike: enakopisne zaimke, dvojinske oblike ipd. (Arhar Holdt in Čibej, 2021: 49–50).

Sporazumevanje v slovenskem jeziku pregledanih dodatnih 400.000 besed, pripravljene pa so bile tudi referenčne smernice za označevanje lem in oblikoskladnje po sistemu JOS oz. MULTEXT-East v4 (Holozan et al., 2008). Trenutna različica korpusa vsebuje oznake sistema MULTEXT-East v6, ki na sistemski ravni vsebuje 1.900 možnih oznak z informacijo besedne vrste in različnih slovarsko-slovničnih značilnosti, kot so npr. spol, sklon, število in lastnoimenskost pri samostalnikih.[3]

SentiCoref 1.0 (Žitnik, 2019) je korpus z 837 besedili oz. približno 433.000 pojavnicami, ki je bil vzorčen iz korpusa SentiNews 1.0 (Bučar, 2017). Čeprav SentiCoref 1.0 neposredno ne vsebuje enakih oznak sentimenta kot SentiNews 1.0, sta korpusa medsebojno povezljiva. SentiCoref 1.0 vsebuje tudi oznake imenskih entitet (oseb, organizacij in lokacij) ter koreferenc na imenske entitete skupaj s koreferenčnimi verigami, ki označujejo sentiment za vsako entiteto. SentiCoref 1.0 je odprto dostopen pod licenco CC BY 4.0 na repozitoriju CLARIN.SI, in sicer v tabelaričnem formatu TSV3, ki ga podpira označevalno orodje INCEpTION (Klie et al., 2018), naslednik orodja WebAnno (Eckart de Castilho et al., 2014).

3. Priprava na označevanje

3.1. Priprava podatkov

SentiCoref 1.0 je sicer tokeniziran, ne vsebuje pa lem in oblikoskladenjskih oznak. Kar zadeva delitev na pojavnice, SentiCoref 1.0 ni bil zasnovan z mislijo na potencialne dodatne jezikoslovne nivoje označevanja, zato v nekaterih primerih odstopa od tokenizacijskih pravil, ki jih pri označevanju korpusov trenutno uporabljamo v slovenskem prostoru (označevalnik *classla*[4] oz. vanj vključeni tokenizator *Obeliks*[5]), npr. pri deljenju kratic ("STA-jev" > "STA", "-", "jev") in številnikov ("2,356" > "2", ",", "356"). Prav tako delitev v SentiCorefu 1.0 ne vsebuje podatkov o presledkih. Pred strojnim oblikoskladenjskim označevanjem in ročnim popravljanjem oblikoskladenjskih oznak je bilo treba najprej popraviti tokenizacijo (vzporedno z njo tudi strojno lematizacijo) ter razdeliti besedilo na povedi (stavčna segmentacija). Za pregledovanje smo korpus pripravili v tabelaričnem formatu v okolju *Google Preglednice* (ang. *Google Sheets*), saj INCEpTION ne podpira spreminjanja tokenizacije. Tokenizacija je bila v celoti popravljena ročno, stavčna segmentacija pa je bila najprej strojno pripisana (na podlagi ločil), nato pa ročno pregledana in potrjena.

Pri pregledovanju segmentacije je bilo 17.095 strojno pripisanih koncev povedi ročno potrjenih kot ustreznih (z ujemanjem treh pregledovalcev in potrditvijo končnega razsojevalca oz. kuratorja). 2.528 koncev povedi so pregledovalci pripisali ročno: pri 2.151 koncih so se strinjali vsi pregledovalci (in kurator), pri 275 po dva, pri 156 pa je konec povedi označil le en pregledovalec. 2.992 koncev povedi je bilo potrjenih kot neustreznih; od tega jih je bilo 1.409 označenih avtomatsko, 940 ročno s popolnim ujemanjem med tremi pregledovalci, 167 ročno z ujemanjem dveh pregledovalcev, 476 ročno z oznako le enega pregledovalca. Pri večini primerov, v katerih je razsojevalec zavrnil odločitve pregledovalcev, gre za popravke tokenizacije in lem, ko so pregledovalci npr. kot konec stavka označili piko, ki je v resnici del okrajšave ("d.o.o", "." > "d.o.o.").

Na popravljenem in ustrezno segmentiranem korpusu smo leme in oblikoskladenjske oznake označili z označevalnikom CLASSLA StanfordNLP v0.0.11.[6]

3.2. Priprava smernic za označevanje

Pri pregledovanju oznak smo sledili smernicam za oblikoskladenjsko označevanje JOS (Holozan et al., 2008), ki vključujejo nabor oblikoskladenjskih oznak (MSD), splošna načela lematizacije ter natančnejše opredelitve posameznih označevalnih kategorij in podkategorij, ponazorjene z označenimi korpusnimi primeri. Smernice smo pripravili v okolju *Google Dokumenti* (ang. *Google Docs*), da smo jih lahko dopolnjevali na podlagi sprotne obravnave ključnih označevalskih dilem ter ponovnega pregleda in evalviranja problematičnih mest. Predvsem na te vidike smernic se bomo osredotočili tudi v nadaljevanju.

4. Pregledovanje oznak

4.1. Obseg in delotoki označevalne kampanje

Ročni pregled strojno označenega gradiva je potekal v okolju *Google Preglednice*. Podatki iz 837 besedil so bili pripravljeni v prav toliko datotekah. Vsaka datoteka je vsebovala metapodatke in za pregledovanje relevantne informacije: obliko pojavnice, lemo, strojno pripisano oblikoskladenjsko oznako (z možnostjo izbire popravka s spustnega seznama vseh obstoječih oznak, kar je olajšalo popravljanje in zmanjšalo možnost zatipka) in celico za morebiten komentar pregledovalca.

Podatke je pregledovalo 24 študentov jezikoslovnih smeri, razdeljenih v 3 skupine. Vsaka izmed teh skupin študentov je pregledovala iste datoteke; namen tega, da vsako pojavnico pregledajo 3 študenti, je bil doseči večjo zanesljivost odločitev. Vsakemu izmed 8 pregledovalcev v skupini je bila dodeljena besedna vrsta oz. več besednih vrst, pri čemer je dodelitev potekala na osnovi preferenc študentov, predhodno ugotovljenih v anketi. Glede na težavnost označevanja ter pogostost vsake besedne vrste v korpusu sta samostalnik pregledovala dva študenta; glagol, pridevnik in zaimek po en študent; za izbiro oznake preprostejše besedne vrste pa smo združili v skupine, pri čemer je en študent pregledoval po eno skupino: prislov in členek; predlog in veznik; številnik, okrajšava, medmet in "neuvščeno". Pred pričetkom pregledovanja so bile pregledovalcem predstavljene smernice (gl. 3.2) in demonstracija postopka v obliki videa. Pregledovanje je potekalo v dveh fazah.

4.1.1. Pregledovanje

Uvodni teden pregledovanja je bil namenjen poglobljeni seznanitvi s smernicami in razreševanju potencialnih nejasnosti, zato je bilo vsakemu pregledovalcu dodeljenih le 5 datotek. Število datotek se je postopoma zviševalo do 20 tedensko, hkrati pa smo okretnejšim ali bolj časovno razpoložljivim pregledovalcem omogočili večji obseg dela (individualizirani pristop). Analiza (ne)ujemanja med tremi vzporednimi pregledovalci je predstavljala izhodišče za 2. fazo – kuracijo.

[3] Označevalni sistem je opisan na spletni strani: http://nl.ijs.si/ME/V6/msd/.
[4] https://github.com/clarinsi/classla
[5] https://github.com/clarinsi/obeliks
[6] https://pypi.org/project/classla/0.0.11/

4.1.2. Kuracija

Posamezne odločitve pregledovalcev smo uredili v enotno tabelo, da so bile ob pojavnicah prikazane vse 3 odločitve, pri čemer so bile posebej označene tiste oznake, pri katerih je med pregledovalci prišlo do razhajanja. Naloga kuratorjev je bila pregledati prav te pojavnice in jim pripisati končno oznako. 7 kuratorjev je bilo izbranih iz vrst pregledovalcev, po eden za vsako besedno vrsto oz. skupino besednih vrst. Označevalna kampanja se je zaključila v 12 tednih, od katerih so bili štirje namenjeni samo kuraciji.

4.2. Označevalne dileme

Ob kuraciji smo identificirali dve vrsti označevalnih težav: (a) primeri, pri katerih so bile označevalne smernice jasne, a pri delu niso bile dosledno upoštevane in (b) primeri, ki so se pokazali kot zahtevnejši: slabše predstavljeni v smernicah in mestoma tudi nedosledno obravnavani v obstoječem ssj500k 2.3.[7]

Težave prvega tipa smo analizirali, odpravili nekonsistentnosti in jih označili v skladu z označevalnimi smernicami. Nekaj več informacij o tipičnih tovrstnih težavah povzemamo v poglavju 4.3. Posebno pozornost pa smo posvetili drugi skupini težav, ki smo jih identificirali kot bolj kompleksne in zahtevne, saj so njihove rešitve zahtevale premislek o odprtih vprašanjih na ravni lematizacije in oblikoskladnje (tudi) v korpusu ssj500k in posledično nadgradnjo označevalnih smernic. V nadaljevanju predstavimo te težave, v poglavju 5 pa predlagane spremembe smernic.

4.2.1. Občnoimenska prekrivnost v stvarnih lastnih imenih

Pregledovalcem je težave povzročalo pravilo, da je v stvarnih lastnih imenih, kjer je lastnoimenski samostalnik prekriven z občnoimenskim samostalnikom, tako lema kot oblikoskladenjska oznaka občnoimenska. Iz tega sledi, da je lematizacija slovenskih imen podjetij, časopisov, revij, knjig, tudi televizijskih oddaj, serij ali filmov ipd. z malo začetnico: npr. podjetje ***Iskra*** [iskra, Sozei]; časnik ***Delo*** [delo, Sosei]. Na iskanje prekrivnosti, ki zaradi pomenske oddaljenosti občnoimenske "ustreznice" pogosto ni enoznačno (gl. tudi 4.2.3), je bilo treba večkrat opozoriti, saj je bilo pregledovalcem bolj intuitivno ohraniti zapis leme z veliko začetnico. Opozarjati jih je bilo treba tudi, da načelo prekrivnosti dogovorno velja samo pri samostalnikih (stranka ***Zares*** [Zares, Slmei]). Manj težav smo zaznali pri pregledovanju tistih primerov stvarnih lastnih imen, ki niso imela prekrivne leme z občnim samostalnikom in smo jih lematizirali z veliko začetnico (podjetje ***Mercator*** [Mercator, Slmei]).

4.2.2. Izlastnoimenski svojilni pridevniki

Del pravila, da pri svojilnih pridevnikih, ki izvirajo iz osebnih ali zemljepisnih lastnih imen, ohranjamo lemo z veliko začetnico (***Aškerčeva*** *ulica* > lema: Aškerčev), je bil jasen, več dilem je bilo pri pregledovanju tistih izlastnoimenskih svojilnih pridevnikov, ki se v rabi pišejo z malo ali pa prehajajo v zapis z malo, ker niso v pomenu prave svojine (***Parkinsonova*** *bolezen* > lema: parkinsonov).

Pri lematizaciji izlastnoimenskih pridevnikov v stvarnih lastnih imenih je pregledovalce zmedla različna obravnava primerov v korpusu ssj500k (***Delova*** *dopisnica* > lema: Delov vs. ***Magov*** *novinar* > lema: magov), zato je bilo treba ta del pravila, ki v izhodiščnih smernicah ni bil pojasnjen, posebej razložiti.

4.2.3. Tuja stvarna lastna imena

Ker načelo prekrivnosti z občnoimenskimi samostalniki (gl. 4.2.1) velja primarno za slovenske samostalnike, se je pogosto pojavljalo vprašanje, katere besede obravnavati kot slovenske (prevzete besede, ki se pregibajo s slovensko morfologijo, vedno umeščamo med slovenske, če potrditve za pregibanje v rabi ni, pa se je treba odločiti na podlagi drugih kriterijev). Dileme so se nanašale predvsem na: (a) prevzete besede, ki pogosto nastopajo kot deli tujejezičnih imen sicer slovenskih podjetij (tip *leasing, holding*) ter (b) ostale občnoimenske besede v tujejezičnih zvezah, ki so prekrivne s slovenskimi občnoimenskimi samostalniki, pri čemer pa pogosto ne izpolnjujejo kriterija pomenske prekrivnosti (tip *trans, global*).

Podrobneje smo obravnavali tudi skupino stvarnih lastnih imen tipa *Zagrebačka banka*, *Večernji list*. Ker gre za imena v hrvaščini, ki zaradi sorodnosti s slovenščino mestoma prinašajo besedje, enako slovenskemu, so bili pregledovalci v dilemi, ali tako pridevnik kot samostalnik označiti kot slovensko besedo in pri tem pridevniku pripisati v slovenščini neobstoječo lemo, ali (vsaj) pridevnik umestiti med tujejezično besedišče.

4.2.4. Ločevanje pridevnikov od prislovov

Odločitve pregledovalcev so se pogosto razhajale pri primerih, ki so izkazovali tipično povedkovnodoločilno rabo pridevnikov oz. obravnavo pridevniških oblik, ki so se prekrivale z osnovno prislovno obliko. Smernice so že vsebovale splošno navodilo o označevanju pridevnikov, ki lahko nastopajo v prilastkovi ali povedkovi rabi (*Sledil je* ***prelomni*** *korak* > pridevnik kot levi prilastek; *uradno še ni* ***rehabilitiran*** > pridevnik kot povedkovo določilo), pa tudi pravilo za ločevanje pridevnikov od prislovov v primeru pridevniškega niza (***uradno*** *prečiščeno besedilo* > prislov). Niso pa naslovile razlike med pridevniško in prislovno lemo pri posameznih zahtevnejših primerih (npr. *smotrno, potrebno, mogoče, možno* v primerih kot npr. *bi bilo* ***smotrno****, da bi [...]*), ki so se tudi v korpusu ssj500k pokazali kot nekonsistentno označeni: pogosto smo zasledili prislovno lemo namesto dogovorno ustrezne pridevniške leme. Neskladja so predstavljala izhodišče za nadaljnje analize, ki so vključevale ponovni pregled vseh primerov oz. zgledov (v korpusu SentiCoref) s prekrivnimi pridevniškimi in prislovnimi oblikami ter oblikovanje dopolnjenega pravila za pripisovanje pridevniških in prislovnih lem.

4.2.5. Nesklonljivi prilastki (tip *bruto, solo*)

Pregledovalci so imeli težave z razumevanjem navodila v izhodiščnih smernicah, da tiste primere tipa bruto, solo (npr. ***solo*** *uspeh*, *rast* ***bruto*** *zadolževanja*, ***info*** *točka*), ki so sklonljivi, obravnavamo kot samostalnike, tiste, ki niso, pa kot pridevnike. Predvsem v navodilu ni jasno, kako preverjati (ne)sklonljivost in kaj je vodilo za odločitev (sistemska možnost, gradivo).

[7] Smernice Holozan et al. (2008) predstavljajo v slovenskem prostoru sprejet in široko apliciran označevalni standard, zato smo jim sledili v največji možni meri. Tudi dopolnitev smernic, ki smo jo pripravili na projektu RSDO, ostaja v zastavljenih konceptualnih okvirih. Morebitne korenitejše spremembe označevalnega sistema, kjer izstopa predvsem vprašanje lematiziranja (pravopisno, ne pa tudi oblikoslovno) različnih samostalnikov in tudi drugih besednih vrst z veliko ali malo začetnico, zahtevajo širši premislek, ki ga nakažemo v pogl. 6.

4.2.6. Prislovne zveze (tip *na novo*)

Težave so bile tudi z obravnavo t. i. prislovnih zvez oz. označevanjem nepredložnega dela teh zvez. Smernice posredno nakazujejo, naj označevanje teži k pridevniškim lemam (*na **drobno*** > lema: droben), se je pa recimo pri primeru *v živo* pokazalo, da so bili v korpusu ssj500k vsi takšni primeri označeni kot prislovni (*v **živo*** > lema: živo). Na osnovi tega neskladja smo naredili podrobnejšo analizo in odkrili več primerov neenotnega označevanja enakovrstnih primerov.

4.3. Pregledani podatki

Analiza popravkov po koncu pregledovanja in kuriranja kaže, da se delež vnesenih popravkov sklada s pričakovanim deležem napak pri avtomatskem označevanju slovenskih besedil z označevalnikom CLASSLA StanfordNLP (Ljubešič in Dobrovoljc 2019: 31–32). Na ravni lematizacije je bilo skupaj popravljenih 5.588 lem, kar je približno 1,3 % vseh pojavnic v korpusu, kar se sklada s približno 98-odstotno natančnostjo lematizacije. Na ravni oblikoskladenjskih oznak je bilo skupaj 12.586 popravkov, kar pomeni 2,9 % vseh oznak v korpusu (ob skoraj 97-odstotni natančnosti oblikoskladenjskega označevanja).

Pri popravkih lem so bili med najpogostejšimi lastnoimenskimi samostalniki, ki so prekrivni z občnoimenskimi (npr. ***Luka** Koper* > lema: luka), okrajšave, sestavljene iz ene ali dveh črk (npr. *dr.* > lema: dr.), pa tudi besede s prekrivnimi oblikami v oblikoskladenjski paradigmi (npr. *delo* in *del*). Pri popravkih oblikoskladenjskih oznak je šlo povečini za ločevanje med občnimi in lastnoimenskimi samostalniki (tip *Leasing – leasing*; 1538 popravkov oz. 12 %; v obratni smeri iz občnoimenskega v lastno je bilo popravkov manj: 235 oz. 1,8 %), med moškim in ženskim spolom (825 popravkov oz. 6,6 %; pri tem gre npr. za imena določenih strank, kot je *Desus*) ter med prekrivnimi oblikami v imenovalniku, tožilniku in rodilniku (skupaj 1.617 popravkov oz. 12,8 % pri samostalnikih; npr. neživi samostalniki moškega spola: *odbor*, *posel* v imenovalniku in tožilniku). Na ravni besednih vrst je šlo največkrat za težje ločevanje med prekrivnimi prislovi in prirednimi vezniki (npr. *tako*; 130 popravkov oz. 1,1 %), med lastnoimenskimi samostalniki in neuvrščenimi tujejezičnimi izrazi (npr. *Amnesty International*; 118 popravkov oz. 1,0 %) ter med členki in prirednimi vezniki (npr. *sicer*, *niti*, *ne*; 97 popravkov oz. 0,7 %). Ker je bila količina popravkov relativno majhna, bi se bilo v prihodnjih označevalnih kampanjah morda smiselno osredotočiti le na najpogostejše pričakovane napake. Kot vodilo lahko pri tem služijo v tem poglavju naštete najpogostejše dileme in težave.

5. Nadgradnja označevalnih smernic

Na podlagi analize najpogostejših označevalskih dilem in pregleda označevalnih odločitev v korpusu ssj500k smo pripravili rešitve glede (nadaljnjega) pregledovanja in dopolnitve smernic za problematične kategorije, naštete v poglavju 4.2. Nadgrajene smernice bodo objavljene ob koncu projekta RSDO.

I. **Občnoimenska prekrivnost v stvarnih lastnih imenih:** splošno načelo, da stvarna imena, prekrivna z občnim samostalnikom, označujemo kot občni samostalnik in lematiziramo z malo začetnico, ostala, ki prekrivnosti ne izkazujejo, pa z veliko začetnico, smo dopolnili s konkretnimi zgledi rabe. Izbrali smo kategorije, ki so povzročale največ težav (podjetja in časnike), npr. *O tem, da so bile v **Iskri*** [iskra, Somem] *potrebne spremembe, so čivkali že vrabci na veji.*; *Večino hrane kupimo v **Mercatorju*** [Mercator, Slmem] *ali **Intersparu*** [Interspar, Slmem].; *Kot smo poročali v prejšnji številki **Mladine*** [mladina, Sozer].

II. **Izlastnoimenski svojilni pridevniki:** v smernice smo dodali pravila za rabo velike in male začetnice s primeri:

(a) Pridevniki iz osebnih in zemljepisnih lastnih imen: načeloma ohranjamo lemo z veliko začetnico, tiste primere, ki se v rabi pišejo z malo ali so na prehodu v zapis z malo, ker niso v pomenu prave svojine, pa lematiziramo z malo, npr. *Celjska občina je prejšnji teden objavila razpis za najem vile v **Aškerčevi*** [Aškerčev, Psnzem] *7 v Celju.*; *Gre za zdravilo za zdravljenje **parkinsonove*** [parkinsonov, Psnzer] *bolezni.*

(b) Pridevniki iz stvarnih lastnih imen: dodatno smo opredelili načelo lematizacije primerov tipa ***Delova** dopisnica* > lema: Delov in ***Magov** novinar* > lema: magov. Pri primerih, kjer je bila prekrivnost sistemsko sicer možna, vendar v dejanski rabi neizkazana, smo ohranili veliko začetnico, npr. *S tega stališča je polemika z **Mladininim*** [Mladinin, Psnmeo] *doktorjem sociologije že skorajda na robu smiselnega* (občni samostalnik *mladina* sicer obstaja, vendar je svojilni pridevnik v rabi izredno redek, tj. ima eno samo pojavitev v referenčnem korpusu Gigafida 2.0). Nasprotno je v primerih, ki izkazujejo pogostejšo rabo svojilnega pridevnika, npr. *vsi pa občudujejo njegovi operi Jevgenij Onjegin in **Pikova*** [pikov, Psnzei] *dama.*

(c) Pridevniki na -ski, -ški kot del zemljepisnih lastnih imen: lematiziramo jih z malo začetnico, pri čemer je treba posebej izpostaviti razliko v odnosu do primerov tipa *Kranjska*, *Štajerska* ipd. Pri imenih regij gre za samostalnike in jih lematiziramo z veliko: *V Vinski kleti **Goriška*** [goriški, Ppnsmi] *Brda zadovoljni s poslovanjem v minulem letu*; *Črnivec je poleg prelaza Volovjek najsevernejši cestni prehod, ki povezuje **Kranjsko*** [Kranjska, Slzet] *in **Štajersko*** [Štajerska, Slzet].

(č) Splošni pridevniki kot del zemljepisnih lastnih imen: lematiziramo jih z malo (tip *nov*, *spodnji*), če v splošni rabi ne obstajajo, pa ohranimo veliko začetnico, npr. *Britanija, Avstralija in **Nova*** [nov, Ppnzei] *Zelandija*; *Mlekarna Celeia iz **Arje*** [Arji, Ppnzer] *vasi je namreč edina domača mlekarna v večinski lasti zadrug.*

III. **Tuja stvarna lastna imena:** po posvetu s širšo projektno ekipo smo se odločili, da bomo oblikovno prekrivne občnoimenske samostalnike "iskali" tudi v tujejezičnih večbesednih stvarnih lastnih imenih. Pri tem je treba upoštevali predvsem dve merili: pregibanje v rabi in prevzetost (prisotnost v referenčnih priročnikih, npr. *Hypo **Leasing*** [leasing, Somei], *Infond **Holding*** [holding, Somei]), ne pa nujno tudi merilo pomenske prekrivnosti – v nekaterih primerih lahko ima tuja beseda v stvarnem lastnem imenu podoben pomen, kot ga ima (prekrivna) slovenska beseda, v nekaterih pa ne. Primere, ki so bili oblikovno prekrivni, pomensko pa ne, smo zbrali na posebnem seznamu in po analizi sprejeli odločitev, da jih vse obravnavamo kot občne samostalnike, npr. *Trade **Trans*** [trans, Somei] *Invest*, *Prevent **Global*** [global, Somei].

V smernice smo dodali odločitev glede označevanja slovenščini sorodnih tujih primerov (tip *Večernji list*): pri

samostalnikih upoštevamo načelo prekrivnosti s slovenskim občnoimenskim besediščem, pridevnike obravnavamo kot tuje besedišče, pri katerem ostane lema enaka besedni obliki, npr. *Jutarnji* [Jutarnji, Nj] *list* [list, Somei], *Zagrebačka* [Zagrebačka, Nj] *banka* [banka, Sozei].

IV. **Ločevanje pridevnikov od prislovov:** pri opredelitvi razlike med pridevnikom in prislovom v vlogi povedkovega določila smo v smernicah izpostavili skladenjski vidik. Opredelitev razlike, da je beseda v vlogi povedkovega določila prislov, če je iz stavka izpustljiva, pridevnik pa, če je nepogrešljiva (obvezna), smo podkrepili s primeri, npr. *O tem ni ***(mogoče)*** [mogoč, Ppnsei] *sklepati.*; ***(Mogoče)** [mogoče, Rsn] ste ga vznemirili.*

V. **Nesklonljivi prilastki (tip *bruto, solo*):** pri obravnavi nesklonljivih prilastkov se je smiselno opreti na preverjanje njihove sklonljivosti v dejanski rabi. Oblikovali smo pravilo, da če najdemo potrditev v referenčnem korpusu, da se določen primer lahko pregiba kot samostalnik, potem to opcijo upoštevamo, če pa potrditve ne najdemo, primer dosledno obravnavamo kot pridevnik: *so se do konca leta povprečne **neto*** [neto, Ppnmein] *plače realno povečale za okoli 33 odstotkov.*

VI. **Prislovne zveze (tip *na novo*):** v smernice smo dodali eksplicitno pravilo, da primere tega tipa obravnavamo kot zveze predloga in pridevnika. Na primerih, ki so pregledovalcem predstavljali največ težav, smo ponazorili, da obravnavamo nepredložni del zveze torej kot pridevnik in ne kot prislov, npr. *Če bi se **na*** [na, Dt] ***hitro*** [hiter, Ppnset] *ozrl, bi videl, da ga zasledujejo.*

6. Zaključek in nadaljnje delo

Pregledovanje osnovnih označevalnih nivojev korpusa SentiCoref predstavlja eno najobsežnejših kampanj te vrste v našem prostoru in – ob kampanji, ki se je osredotočala na gradivo računalniško posredovane komunikacije Janes (Čibej et al., 2018) – tudi eno prvih priložnosti za ponovitev dela z uporabo metodologije, ki se je vzpostavila pri pripravi izhodiščne različice učnega korpusa.

Po opravljeni kuraciji, končni kontroli kvalitete označenega in statističnem pregledu dilem in popravkov je mogoče potegniti nekaj zaključkov. Pomembno je, da so se pomanjkljivosti označevalnih smernic kazale zlasti pri temah, povezanih z označevanjem lastnih imen (samostalnikov, izlastnoimenskih pridevnikov), še zlasti pri odločitvah, ki so povezane s presojanjem, ali je določena beseda slovenska ali tujejezična. Ker korpus SentiCoref vsebuje atipično visoko število raznovrstnih lastnih imen (tako je bil namreč zgrajen), smo pogosto srečevali težave, ki so bile pri pripravi ssj500k redkejše in za smernice manj relevantne.

Obstoj kategorije lastnoimenskosti na ravni oblikoskladnje in posledično lematiziranje ob iskanju prekrivnosti med občno- in lastnoimenskimi entitetami odpira konceptualne težave, ki bi jih kazalo v ponovno premisliti. Prva je, da je označevalno kategorijo najti samo pri samostalnikih, prekrivnost (po nekako drugačni logiki) iščemo tudi pri pridevnikih, ne pa pri ostalih besednih vrstah. Težava je tudi, da pri odločitvah glede zapisa leme z veliko ali malo začetnico na raven oblikoskladenjskega označevanja prenašamo vprašanja, ki se dotikajo pravopisa (oz. pravopisov, če upoštevamo, da se vse dileme preslikavajo in potencirajo pri srečevanju s tujejezičnimi elementi), pri čemer sistem sledi predpostavki, da avtorji besedil pravopisu vedno sledijo. Pri razlikovanju obravnave zemljepisnih/osebnih imen ter stvarnih imen se v sistem še dodatno vpletajo načela, ki so bolj kot na oblikoskladnjo vezane na (s semantiko in trenutnim pravopisom povezano) metajezikovno klasifikacijo referentov. Zdi se, da na ravni označevanja lem in oblikoskladnje sprejemamo odločitve, ki bi sodile na raven jezikovnega opisa in predpisa, ob čemer se opiramo na jezikovne vire, kjer prav te odločitve pogosto še niso sprejete.

Ker težave pripisovanja občno- oz. lastnoimenskosti samostalnikov prednjačijo v veliki sliki vseh opravljenih popravkov, obenem pa identifikacijo lastnoimenskih zvez v zadnjih letih uspešno opravljamo pri označevanju imenskih entitet, bi kazalo ponovno razmisliti o dodani vrednosti te kategorije na ravni oblikoskladnje. Če se izkaže, da je kljub vsemu koristna, bi se določene težave dalo odpraviti z radikalnejšim posegom v smernice, npr. z odpovedjo iskanja prekrivnih občnih in lastnih samostalnikov in sledenju rabi, kakršna se v besedilih pojavlja. Enako velja za obravnavo tujega besedišča, ki ga po trenutnem sistemu med slovenske samostalnike umeščamo precej popustljivo in obenem nedosledno. S širjenjem označevanja na besedilne vrste, kjer je tujejezičnih elementov več in v slovenščino prehajajo po manj predvidljivih vzorcih, bi bilo smiselno opredeliti jasen namen ločevanja po jezikih in oblikovati dosledne in pripisljive kriterije zanj. Problem bi bilo dobro nasloviti celovito in podati rešitve za vse relevantne označevalne ravni, ne le lematizacijo in oblikoskladnjo.

Druga večja skupina označevalnih težav je bila vezana na enakopisne oblike, pogosto pridevnike in prislove, pa tudi nekatere slovnične besedne vrste. Tudi tu je opaziti, da se v smernicah pojavljajo semantični (ne le oblikoslovni in skladenjski) kriteriji za presojanje, kar pa se je izkazalo za manj pereče od (po novem vsaj deloma naslovljenih, ne pa povsem odpravljenih) dilem glede uporabe referenčnih jezikovnih virov, npr. za opredeljevanje sklonljivosti. Pri tej skupini težav je ključna ugotovitev, da označevanje tudi v ssj500k ni potekalo povsem usklajeno, zato smo ob delu pripravili seznam težav, ki bi jih bilo v prihodnosti smiselno preveriti in ustrezno urediti za nazaj.

Pri vsem pa je treba upoštevati, da je strojno pripisovanje lem in oblikoskladenjskih oznak za slovenščino že doseglo raven, ko bi bilo celovite ročne preglede smotrno nadomestiti z delnimi, za katere pa bi bilo treba razviti (referenčne in dokumentirane) postopke za avtomatsko ali polavtomatsko identifikacijo problematičnih mest. Spoznanja, ki jih navajamo v prispevku, so lahko izhodišče za takšno nadaljevanje.

Pregledani in popravljeni SentiCoref bo v nadaljevanju projekta RSDO umeščen ob ostale besedilne množice, ki bodo sestavljale povečani učni korpus za slovenščino. V prihodnje bomo v celotnem učnem korpusu izvedli še serijo polavtomatskih popravkov (npr. ali so enobesedni vezniki, kot je "zato", vedno ustrezno označeni kot vezniki), s čimer bomo poskrbeli, da bodo enake dileme v celotnem učnem korpusu razrešene konsistentno. Na podoben način bomo učni korpus primerjali tudi s Slovenskim oblikoslovnim leksikonom Sloleks (Dobrovoljc et al., 2019), da npr. preverimo, ali se glagolski vid glagolov v učnem korpusu ujema s Sloleksom. V okviru projekta RSDO je istočasno z nadgradnjo učnega korpusa potekala tudi nadgradnja Sloleksa, zato smo nalogo prestavili na poznejši termin.

Učni korpus bo skupaj z nadgrajenimi označevalnimi smernicami in ostalo dokumentacijo ob zaključku projekta javnosti odprto na voljo na repozitoriju CLARIN.SI.

7. Zahvala

Projekt *Razvoj slovenščine v digitalnem okolju* (RSDO) sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru *Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014–2020*. Raziskovalna programa št. P6-0411 (Jezikovni viri in tehnologije za slovenski jezik) in št. P6-0215 (*Slovenski jezik - bazične, kontrastivne in aplikativne raziskave*) je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna. Avtorice in avtorji se sodelujočim v označevalni kampanji iskreno zahvaljujemo za vse delo, prav tako pa tudi recenzentoma za relevantne in konstruktivne komentarje.

8. Literatura

Špela Arhar Holdt in Jaka Čibej. 2021. Analize za nadgradnjo učnega korpusa ssj500k. V: Š. A. Holdt, ur., *Nova slovnica sodobne standardne slovenščine: viri in metode*, str. 15–53. Znanstvena založba Filozofske fakultete, Ljubljana. Zbirka Sporazumevanje. https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/325/477/7313-1.

Primož Belej. 2018. *Oblikoskladenjsko označevanje slovenskega jezika z globokimi nevronskimi mrežami*. Magistrsko delo, Fakulteta za računalništvo in informatiko, Univerza v Ljubljani.

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

Jaka Čibej, Špela Arhar Holdt, Tomaž Erjavec in Darja Fišer. 2018. Ročno označeni korpusi Janes za učenje jezikovnotehnoloških orodij in jezikoslovne raziskave. V: D. Fišer, ur., *Viri, orodja in metode za analizo spletne slovenščine*, str. 44–73. Znanstvena založba Filozofske fakultete, Ljubljana. Zbirka Prevodoslovje in uporabno jezikoslovje. https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/111/203/2416-1.

Ludmila Dimitrova, Tomaž Erjavec, Nancy Ide, Heiki Jaan Kaalep, Vladimir Petkevič in Dan Tufis. 1998. Multext-east: Parallel and comparable corpora and lexicons for six central and eastern European languages. V: *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, zvezek 1, str. 315–319, Montreal, Quebec, Kanada. Association for Computational Linguistics. https://aclanthology.org/P98-1050.pdf.

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

Kaja Dobrovoljc, Simon Krek, Peter Holozan, Tomaž Erjavec, Miro Romih, Špela Arhar Holdt, Jaka Čibej, Luka Krsnik in Marko Robnik-Šikonja. 2019. *Morphological lexicon Sloleks 2.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1230.

Richard Eckart de Castilho, Chris Biemann, Irina Gurevych in Seid Muhie Yimam. 2014. WebAnno: a flexible, web-based annotation tool for CLARIN. V: *Proceedings of the CLARIN Annual Conference (CAC) 2014*, Soesterberg, Nizozemska. https://www.clarin.eu/sites/default/files/cac2014_submission_6_0.pdf.

Tomaž Erjavec in Simon Krek. 2008. The JOS morphosyntactically tagged corpus of Slovene. V: *Proceedings. 6th International Conference on Language Resources and Evaluation (LREC 2008)*, str. 322–327, Marakeš, Maroko. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/pdf/89_paper.pdf.

Tomaž Erjavec. 2012. MULTEXT-East: Morphosyntactic resources for Central and Eastern European languages. *Language Resources and Evaluation*, 46(1):131–142.

Miha Grčar, Simon Krek in Kaja Dobrovoljc. 2012. Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik (Obeliks: statistical morphosyntactic tagger and lemmatizer for Slovene). V: *Proceedings of the 8th Language Technologies Conference*, zvezek C, str. 89–94, Ljubljana, Slovenija. IJS. http://nl.ijs.si/isjt12/proceedings/isjt2012_17.pdf.

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

Jan-Christoph Klie, Michael Bugert, Beto Bullosa, Richard Eckart de Castilho in Irina Gurevych. 2018. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. V: *Proceedings of System Demonstrations of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, ZDA. https://aclanthology.org/C18-2002.pdf.

Simon Krek, Kaja Dobrovoljc, Tomaž Erjavec, Sara Može, Nina Ledinek, Nanika Holz, Katja Zupan, Polona Gantar, Taja Kuzman, Jaka Čibej, Špela Arhar Holdt, Teja Kavčič, Iza Škrjanec, Dafne Marko, Lucija Jezeršek in Anja Zajc. 2021. *Training corpus ssj500k 2.3*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1434.

Simon Krek, Tomaž Erjavec, Kaja Dobrovoljc, Polona Gantar, Špela Arhar Holdt, Jaka Čibej in Janez Brank. 2020. 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, Slovenija. Inštitut za novejšo zgodovino. http://nl.ijs.si/jtdh20/pdf/JT-DH_2020_Krek-et-al_The-ssj500k-Training-Corpus-for-Slovene--Language-Processing.pdf.

Nikola Ljubešić in Kaja Dobrovoljc. 2019. What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian. V: *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, str. 29–34. Firence, Italija. The Association for Computational Linguistics, Stroudsburg. https://www.aclweb.org/anthology/W19-3704.

Nikola Ljubešić in Tomaž Erjavec. 2016. Corpus vs. Lexicon Supervision in Morphosyntactic Tagging: the Case of Slovene. V: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, str. 1527–1532, Pariz,

Francija. European Language Resources Association (ELRA). https://aclanthology.org/L16-1242.pdf.

Federico Martelli, Roberto Navigli, Simon Krek, Jelena Kallas, Polona Gantar, Svetla Koeva, Sanni Nimb, Bolette Sandford Pedersen, Sussi Olsen, Margit Langemets, Kristina Koppel, Tiiu Üksik, Kaja Dobrovoljc, Rafael Ureña-Ruiz, José-Luis Sancho-Sánchez, Veronika Lipp, Tamás Váradi, András Győrffy, Simon László, Valeria Quochi, Monica Monachini, Francesca Frontini, Carole Tiberius, Rob Tempelaars, Rute Costa, Ana Salgado, Jaka Čibej, in Tina Munda. 2022. *Parallel sense-annotated corpus ELEXIS-WSD 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1674.

Slavko Žitnik. 2019. *Slovene corpus for aspect-based sentiment analysis - SentiCoref 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1285.