Nadgradnja učnega korpusa ssj550k v SUK 1.0

Špela ARHAR HOLDT
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
Univerza v Ljubljani, Filozofska fakulteta

Jaka ČIBEJ
Univerza v Ljubljani, Filozofska fakulteta

Kaja DOBROVOLJC
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
Institut »Jožef Stefan«

Tomaž ERJAVEC
Institut »Jožef Stefan«

Polona GANTAR
Univerza v Ljubljani, Filozofska fakulteta

Simon KREK
Univerza v Ljubljani, Filozofska fakulteta
Institut »Jožef Stefan«

Tina MUNDA
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
Institut »Jožef Stefan«

Nejc ROBIDA
Univerza v Ljubljani, Filozofska fakulteta
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko

Luka TERČON
Univerza v Ljubljani, Filozofska fakulteta

Slavko ŽITNIK
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko

Povzetek

V prispevku predstavljamo nadgradnjo učnega korpusa, ki je namenjen učenju strojnih postopkov za jezikoslovno označevanje besedil v sodobni standardni slovenščini. Nova različica korpusa ssj500k, ki smo ga preimenovali v SUK (*slovenski učni korpus*), prinaša nova besedila in nove ročno pregledane jezikoslovne oznake različnih vrst. Korpus smo povečali s petsto tisoč na več kot milijon pojavnic z vključitvijo treh odprto dostopnih jezikovnih virov, ki vsak na svoj način odpravljajo predhodno identificirane pomanjkljivosti ssj500k: SentiCoref 1.0, ELEXIS-WSD za slovenščino in iz korpusa Gigafida 2.0 pripravljena množica Ambiga. Pregledovanje jezikoslovnih oznak opišemo po ravneh: tokenizacija, stavčna segmentacija, lematizacija, oblikoskladnja MULTEXT-East, oblikoslovje ter skladnja Universal Dependencies, skladnja JOS-SYN, udeleženske vloge, imenske entitete in koreference. Za vse ravni smo posodobili označevalne smernice, ki so pregledno zbrane in na voljo za nadaljnje delo. Na podatkih korpusa SUK smo naučili novo različico strojnega označevalnika CLASSLA-Stanza, ki dosega presežne vrednosti za vse evalvirane ravni. Z bogatim naborom ročno pregledanih jezikoslovnih oznak predstavlja učni korpus SUK enega temeljnih jezikovnih virov za sodobno slovenščino, zato zahteva neprestano posodabljanje in nadgrajevanje, kar predstavimo v zaključnem poglavju s smernicami za nadaljnji razvoj.

Ključne besede: učni korpus, ssj500k, SUK, jezikoslovno označevanje, označevalne smernice

Abstract

In this paper, we present an upgrade to the training corpus for linguistic annotation of modern standard Slovene. The new version of the ssj500k corpus, renamed to SUK, introduces both new texts and new manually reviewed linguistic tags of various types. The corpus has been expanded from 500,000 to over a million tokens by incorporating three openly accessible language resources, each addressing the previously identified shortcomings of ssj500k: SentiCoref 1.0, ELEXIS-WSD for Slovene, and a dataset prepared from the Gigafida 2.0 corpus called Ambiga. We describe the linguistic annotation process at various levels: tokenization, segmentation, lemmatization, MULTEXT-East morphology, Universal Dependencies

syntax, JOS-SYN syntax, semantic role labelling, named entity recognition, and coreference resolution. We have updated annotation guidelines, which are systematically compiled and available for further work. Using the SUK corpus data, we trained a new version of the automatic tagger CLASSLA-Stanza, which achieves outstanding results for all evaluated levels. With its manually-reviewed linguistic tags, the SUK corpus is foundational for modern Slovene, requiring ongoing improvements, which we detail in the final section with future development guidelines.

Keywords: training corpus, ssj500k, SUK, linguistic annotation, annotation guidelines

1 Uvod

Učni korpusi (ang. *training corpora*) so premišljeno grajene besedilne množice z zanesljivimi (tipično ročno pripisanimi ali pregledanimi) dodatnimi informacijami, ki se uporabljajo pri nadzorovanem strojnem učenju postopkov za obdelavo naravnega jezika. Ti postopki so lahko različni, med najbolj ključnimi za nadaljnje delo z jezikovnimi podatki pa je jezikoslovno označevanje: delitev besedila na gradnike (besede oz. pojavnice, večbesedne enote, povedi) in pripis jezikoslovnih informacij tem gradnikom. Učni korpusi za jezikoslovno označevanje zato spadajo v temeljno digitalno infrastrukturo določenega jezika in kot taki zahtevajo kontinuiran razvoj in nadgrajevanje.

Za nadzorovano učenje strojnega jezikoslovnega označevanja besedil v sodobni standardni slovenščini[1] se v našem prostoru že več kot desetletje razvija učni korpus, ki je bil do nedavnega poimenovan ssj500k (Krek idr., 2020a). Ta je vseboval 27.829 povedi (oz. približno 500.000 pojavnic, ki so korpusu dale ime), označenih na različnih jezikovnih ravneh, od segmentacije, tokenizacije, lematizacije, oblikoslovja in oblikoskladnje prek odvisnostne skladnje,

1 Za označevanje nestandardne slovenščine so na voljo učni korpusi iz zbirke Janes (Čibej idr., 2018); nedavna nadgradnja množic Janes-Tag in Janes-Norm je predstavljena v poročilu Arhar Holdt idr. (2023). Za označevanje starejše slovenščine pa je na voljo učni korpus goo300k (Erjavec, 2015).

imenskih entitet in večbesednih enot do udeleženskih vlog. Pod okriljem projekta Razvoj slovenščine v digitalnem okolju (RSDO)[2] je bil učni korpus nadgrajen z novimi besedili in oznakami, zaradi spremembe obsega pa smo ga preimenovali v SUK, *slovenski učni korpus*.

Nadgradnja korpusa predstavlja pomemben razvojni korak ne le v smislu prenove jezikovnega vira, pač pa tudi z vidika metodologije označevanja. Za vse ravni jezikovnih oznak, ki smo jih pripisovali in pregledovali, so bile posodobljene označevalne smernice, ki so po koncu projekta urejeno zbrane ter objavljene in tako na voljo za nadaljnje nadgradnje.[3] SUK 1.0, ki je pod odprto licenco na voljo na repozitoriju CLARIN.SI (Arhar Holdt idr., 2022), je bil že pod okriljem projekta uporabljen za izboljšavo strojnega označevalnika za slovenščino.

Pripravo korpusa smo z vidika projektnih ciljev predstavili v poročilu (Arhar Holdt idr., 2023), dela na posameznih označevalnih ravninah se dotikajo tudi nekateri prispevki, ki jih navajamo v nadaljevanju. V tem prispevku želimo raziskovalno-razvojni skupnosti jedrnato in celovito predstaviti nadgradnjo učnega korpusa ssj500k v SUK in s tem omogočiti njegovo usklajeno nadaljnje nadgrajevanje. Najprej predstavimo nabor besedilnih množic, s katerimi smo korpus nadgradili, sledi opis ročnega označevanja oz. pregledovanja oznak po jezikovnih ravninah in primerjava predhodne korpusne sestave z novo. Prispevek zaključimo s podatki o izboljšavah strojnega označevalnika, ki služijo kot ocena korpusne nadgradnje, ter smernicami za nadaljnje delo.

2 Metodologija

2.1 Povečanje korpusnega obsega

Korpus ssj500k (v različici 2.3: Krek idr., 2021) obsega 27.829 povedi in je v celoti ročno pregledan na ravni tokenizacije, stavčne segmentacije, oblikoskladenjskih oznak in lem. Večbesedne enote so

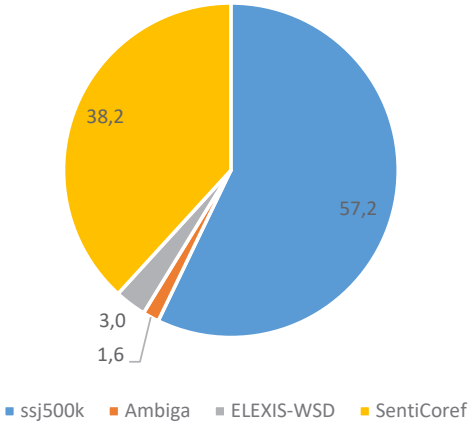
2 Spletna stran projekta z dostopom do rezultatov: https://slovenscina.eu/.
3 Smernice so dostopne na https://wiki.cjvt.si/shelves/jezikoslovno-oznacevanje-korpusov.

označene in pregledane pri 13.511 povedih, skladnja JOS-SYN pri 11.411 povedih, imenske entitete pri 9.488 povedih, skladnja UD pri 8.000 povedih in udeleženske vloge (SRL) pri 5.501 povedi (Krek idr., 2020a, Tabela 1). Ena od prioritet nadaljnjega razvoja je bilo povečanje razpoložljivega gradiva za višje označevalne ravni, v analizah korpusne sestave pa sta bili identificirani tudi potreba po povečanju s korpusnimi besedili, ki omogočajo označevanje prek meja povedi, ter dopolnitvi korpusa za boljšo zastopanost oblikoskladenjskih oznak in dvoumnih besednih oblik (Arhar Holdt in Čibej, 2021, 49–50).

Z upoštevanjem identificiranih potreb in v želji po učinkoviti izrabi že obstoječega gradiva smo za povečanje korpusa izbrali tri odprto dostopne jezikovne vire: (a) **SentiCoref 1.0** (Žitnik idr., 2022) je korpus besedil s slovenskih novičarskih portalov, ki je za namene analize sentimenta opremljen z oznakami imenskih entitet in koreferenc. Korpus odgovarja na potrebe po vključitvi gradiva za označevanje prek meja povedi, prinaša pa tudi vključitev novega označevalnega nivoja, ki spada na področje semantike – koreferenc. (b) **ELEXIS-WSD za slovenščino** (Martelli idr., 2021) je slovenski del 10-jezičnega vzporednega korpusa, ki vsebuje 2.024 povedi iz Wikipedijinih člankov. Korpus vsebuje ročno pripisane oznake za razdvoumljanje pomenov (ang. *word-sense disambiguation*) in kot tak ob korpusu SentiCoref predstavlja drugo izhodišče za strojno učenje na semantični ravni. (c) Iz korpusa **Gigafida 2.0** (Krek idr., 2020b) je bila pripravljena množica **Ambiga,** nabor 603 povedi, ki vsebujejo v predhodnem učnem korpusu nezastopane oblikoskladenjske oznake in pojavnice, identificirane kot problematične za strojno označevanje, npr. enakopisne zaimke, redke dvojinske oblike in podobno.

Novi učni korpus SUK tako sestavljajo množice ssj500k 2.3 (586.187 pojavnic oz. 57,2 %), SentiCoref 1.0 (391.962 pojavnic oz. 38,2 %), ELEXIS-WSD (31.233 pojavnic oz. 3 %) in Ambiga (16.257 pojavnic oz. 1,6 %), kot predstavlja Graf 1. Ko so bile množice za povečanje korpusnega obsega določene, je sledil strojni pripis jezikovnih oznak in njihov celoviti ročni pregled na ravni tokenizacije, stavčne segmentacije, lem in oblikoskladenjskih oznak, za izbrane dele korpusa pa še pripis in urejanje oznak na višjih označevalnih

ravneh. V nadaljevanju predstavljamo delo z oznakami, in sicer ločeno po označevalnih ravninah.



Graf 1: Besedilna sestava učnega korpusa SUK 1.0.

2.2 Segmentacija, tokenizacija, lematizacija, oblikoskladnja MULTEXT-East

Osnovni nivoji korpusnega označevanja: segmentacija, tokenizacija, lematizacija in oblikoskladnja po sistemu MULTEXT-East (žargonsko tudi MSD; ang. *morpho-syntactic description*) so bili ročno pregledani na celotnem gradivu, ki predstavlja nadgradnjo učnega korpusa (512.588 besednih pojavnic).

SentiCoref 1.0, največja izmed novih množic nadgrajenega učnega korpusa, je bil označen po fazah: (a) tokenizacija, lematizacija in segmentacija na povedi z orodjem CLASSLA-Stanza[4] (verzija 0.0.11), (b) ročni pregled teh treh ravni, (c) strojno oblikoskladenjsko označevanje po sistemu MULTEXT-East v6 z istim orodjem, (č) ročni pregled oblikoskladenjskih oznak. Ročni pregled je temeljil na uveljavljenih smernicah[5] in je potekal v spletnem okolju *Google Preglednice* (ang. *Google Sheets*). Tokenizacijo, lematizacijo in

4 https://github.com/clarinsi/classla
5 https://wiki.cjvt.si/books/04-oblikoskladnja-multext-east/page/oznacevalne-smernice, gl. Različica 1.0.

segmentacijo je pregledovalo 9 študentov, medtem ko je pri ročnem pregledu MSD-oznak sodelovalo 24 študentov jezikoslovnih smeri v razponu približno štirih mesecev, kar predstavlja eno najobširnejših tovrstnih označevalnih akcij v našem prostoru. Pregledovanje je potekalo po principu trojnega ujemanja: vsako pojavnico so neodvisno drug od drugega pregledali 3 študenti – oznake, ki so jih enotno izbrali vsi trije označevalci, so bile sprejete, oznake, pri katerih je prišlo do neujemanja, pa so bile znova pregledane v fazi kuracije (za natančnejši popis metodologije gl. Pori idr., 2022). Pri pregledu MSD-jev se je množica ELEXIS-WSD pridružila SentiCorefu (ostale ravni so bile pregledane predhodno), pri Ambigi pa je označevanje vseh štirih ravni zaradi omejenega obsega poteklo v enem koraku.

V nadaljevanju predstavljamo izzive, ki smo jih identificirali v označevalni kampanji, in rešitve, ki so vključene v nadgrajene smernice.[6] Gre za težje in mejne primere, ki so bili v predhodnih označevalnih smernicah slabše zastopani ali pa sploh niso bili, ali pa je pri pregledovanju teh pogosto prišlo do neupoštevanja smernic in s tem nedoslednosti. Dileme smo analizirali, tudi s pomočjo že označenih podatkov v ssj500k, in jih po kuraciji, kolikor je bilo mogoče, uskladili.

Prekrivnost samostalnikov v slovenskih stvarnih lastnih imenih z občnoimenskimi: Pravilo, da samostalnikom, ki so del stvarnih lastnih imen in so prekrivni z občnoimenskimi samostalniki, pripišemo občnoimenskost in jih lematiziramo z malo začetnico, je bilo v obstoječih smernicah sicer obravnavano, a označevalcem ni bilo intuitivno. Gre za primere tipa podjetje *Iskra* (lema: iskra, MSD: Sozei), časnik *Delo* (lema: delo, MSD: Sosei). Vendar to pravilo velja le za samostalnike, ne pa tudi za druge besedne vrste in ne za primere, kjer nesamostalniška besedna vrsta nastopa kot samostalnik, npr. stranka *Zares* (MSD: Slzei, lema: Zares).

Pridevniki iz osebnih in zemljepisnih lastnih imen: Pri izlastnoimenskih svojilnih pridevnikih, ki zaznamujejo vrsto in ne prave svojine ter tudi že prehajajo v zapis z malo začetnico, se je pri določanju leme

6 https://wiki.cjvt.si/books/04-oblikoskladnja-multext-east/page/oznacevalne-smernice, gl. Različica 2.0.

izkazala za težjo odločitev med malo in veliko začetnico. V obstoječih smernicah ni bilo jasnega razlikovanja med to kategorijo pridevnikov in pravimi svojilnimi pridevniki. Tako smo v nadgrajenih smernicah dodatno pojasnili obravnavo izlastnoimenskih pridevnikov: (a) pri pridevnikih iz osebnih lastnih imen imamo poleg teh, ki izražajo pravo svojino (*Pahorjeva* (lema: Pahorjev, MSD: Psnzei) *[mlada struja]*) še tiste, ki zaznamujejo vrsto in jih v rabi pogosto najdemo zapisane z malo začetnico; te primere lematiziramo z malo začetnico (*[zdravljenje] parkinsonove* (lema: parkinsonov) *[bolezni]*); (b) pri pridevnikih iz stvarnih lastnih imen (*Magov [novinar]*, *Delova [dopisnica]*) smo opredelili načelo lematizacije, in sicer z malo začetnico lematiziramo tiste, ki v referenčnem korpusu Gigafida 2.0 izkazujejo svojilno rabo (*Magov* [novinar]; lema: magov (prek mag = čarovnik), medtem ko primere, kjer je svojina konceptualno sicer možna, vendar v rabi ni izkazana, lematiziramo z veliko začetnico (*Delova [dopisnica]*; lema: Delov).

Tuja stvarna lastna imena: Tu so izziv predstavljali primeri dveh tipov: (a) tuja stvarna lastna imena iz slovenščini sorodnih jezikov, ki se v slovenskih besedilih zaradi morfološke podobnosti pregibajo po slovenskih vzorcih (npr. hrvaška imena: *Zagrebačka banka*, *Večernji list*) in (b) deli tujih stvarnih lastnih imen, ki so prevzeti v slovenščino in so pomensko prekrivni z izvorno tujo besedo (npr. *leasing*, *holding*) ali pa so oblikovno prekrivni s slovenskimi samostalniki, a si s tujo besedo ne delijo pomena, pa tudi besedni vrsti v obeh jezikih nista nujno isti (npr. *trans*, *global*). Odločili smo se, da bomo tako v primerih tipa (a) kot (b) upoštevali prekrivnost s slovenskim občnim samostalnikom, če je zadovoljeno vsaj enemu izmed dveh meril: 1) potencialno prekriven samostalnik kot del tujega lastnega imena se v rabi pregiba; 2) tuj samostalnik je prevzet, kar potrjujejo referenčni priročniki za slovenščino (npr. *[Hypo] Leasing*; lema: leasing, MSD: Somei; *[Infond] Holding*; lema: holding, MSD: Somei). Pomenska prekrivnost besede v enem in drugem jeziku ni bila nujen pogoj za uvrstitev tovrstnih primerov med občnoimenske samostalnike (*[Trade] Trans [Invest]*; lema: trans, MSD: Somei; *[Prevent] Global*; lema: global, MSD: Somei). Kot velja pri obravnavi (delov) stvarnih imen, ki jih sestavljajo neizpodbitno slovenske besede, tudi

v tujih stvarnih lastnih imenih prekrivnost iščemo le pri samostalnikih. To velja posebej izpostaviti, saj so v jezikih, sorodnih slovenščini, lahko tudi nesamostalniške besedne vrste oblikovno podobne slovenskim in se kot take lahko tudi pregibajo. Pri teh besedah je lema enaka obliki, MSD-oznaka pa 'neuvrščeno' (*Večernji* (lema: Večernji, MSD: Nj) *list* (lema: list, MSD: Somei), *Zagrebačka* (lema: Zagrebačka, MSD: Nj) *banka* (lema: banka, MSD: Sozei).

Ločevanje pridevnikov od prislovov: Obravnavali smo vprašanje, kateri besedni vrsti pripada oblika besede, ki je enaka prislovu in pridevniku, ko je ta beseda (a) v vlogi povedkovega določila (npr. *[... bi bilo] smotrno, [da bi ...]*) ali (b) v strukturi z nedoločnikom (npr. *[O tem ni] mogoče [sklepati.]*). Predhodne smernice tega niso obravnavale, kar se je odražalo tudi v korpusu ssj500k, kjer tovrstni primeri niso bili enotno označeni. Po pregledu in analizi pojavitev tovrstnih primerov v korpusu SentiCoref smo oblikovali pravilo, da besedi v obeh naštetih skladenjskih vlogah pripišemo pridevniško lemo in MSD-oznako, če v stavku ni izpustljiva (je obvezna, da je stavek koherenten), in nasprotno – prislovno lemo in oznako, če je stavek koherenten tudi brez nje (npr. *[O tem ni] mogoče* (lema: mogoče, MSD: Ppnsei) *[sklepati.]* > *O tem ni sklepati.**; *Mogoče* (lema: mogoče, MSD: Rsn) *[ste ga vznemirili.]* > *Vznemirili ste ga.*).

Predložne prislovne zveze: Podobno kot pri prejšnji dilemi je bila težava pri razlikovanju med pridevnikom in prislovom v prislovnih zvezah s predlogom (npr. *na novo*, *v živo*). Tudi tovrstni primeri so bili v korpusu ssj500k označeni neenotno in po analizi smo določili, da nepredložnemu delu v predložnih prislovnih zvezah pripišemo pridevniško lemo in MSD-oznako (*[na] novo* (lema: nov, MSD: Ppnset)).

Nesklonljivi prilastki: V obstoječih smernicah je bilo pravilo, da nesklonljive prilastke (npr. *solo*, *neto*, *bruto*) označimo kot samostalnike, kadar so sklonljivi, in kot pridevnike, kadar niso, vendar kriterij sklonljivosti ni bil jasno opredeljen. Tako smo oblikovali pravilo, da določen primer označimo kot samostalnik, če v referenčnem korpusu najdemo potrditev, da se lahko pregiba kot samostalnik (npr. *pop*, *elektro*), in kot pridevnik, če te potrditve ni (npr. *neto*, *repro*).

2.3 Oblikoslovje in skladnja po sistemu UD

Universal Dependencies (UD) je označevalna shema, ki si prizadeva za mednarodno oz. medjezično usklajeno slovnično označevanje besedil na oblikoslovni in skladenjski ravni, da bi pospešila razvoj večjezičnih jezikovnih tehnologij na eni strani in kontrastivnih jezikoslovnih analiz na drugi (de Marneffe idr., 2021). V zbirko več sto korpusov, označenih s to shemo, je bila leta 2015 priključena tudi univerzalna odvisnostna drevesnica za pisno slovenščino, drevesnica SSJ (Dobrovoljc idr., 2017), ki je ob prvi objavi vsebovala 8.000 razčlenjenih povedi korpusa ssj500k (primer na Sliki 1), v projektu RSDO pa smo jo bistveno nadgradili tako z vidika obsega kot z vidika dokumentiranosti smernic in infrastrukturne podpore za njeno nadaljnjo analizo (Dobrovoljc in Ljubešić, 2022; Dobrovoljc idr., 2023).

Jedrne smernice sheme UD, kakršne so dokumentirane na uradni spletni strani projekta,[7] za vsako izmed predlaganih »univerzalnih« oznak (17 besednih vrst, 24 oblikoskladenjskih lastnosti, 37 odvisnostnih skladenjskih relacij) podajajo razmeroma splošno opredelitev s ponazoritvami na nekaj izbranih primerih v različnih jezikih, način prenosa teh smernic na svoje konkretne jezikovne podatke pa je prepuščen avtorjem drevesnic za posamezne jezike. Ker za slovenščino ob nastanku prvotne drevesnice SSJ te smernice niso bile sistematično dokumentirane, je bil prvi korak znotraj projekta RSDO zato namenjen izčrpnemu popisu smernic UD za slovenščino, tako na spletni strani projekta (v angleščini) kot v obliki samostojnega priročnika v slovenščini.[8] Slednji poleg velikega števila ponazoritev prototipičnih in mejnih primerov vsake oznake vsebuje tudi ločeno poglavje s smernicami za označevanje kompleksnejših skladenjskih struktur (npr. elipse, primerjave, poudarjalni členki, besedilni povezovalci ...). Pri tem smo poleg opisa prvotnih smernic uvedli tudi nekaj manjših izboljšav na mestih, kjer je bila prvotna označenost korpusa SSJ-UD nedosledna ali

7 https://universaldependencies.org/

8 https://wiki.cjvt.si/books/07-universal-dependencies/page/oznacevalne-smernice, gl. Različica 1.0.

neustrezna glede na splošne, jezikovno univerzalne smernice. To pa ne velja za vse identificirane neskladnosti, saj nekatere predstavljajo precejšen odmik od doslej uveljavljenih označevalnih praks v slovenskem prostoru in bi jih bilo zato smiselno najprej nasloviti s širšo strokovno diskusijo. Tovrstna mesta smo popisali v ločeni prilogi[9] h krovnim smernicam.

Ker sta si označevalna sistema JOS in UD na ravni pripisovanja besednih vrst in drugih oblikoslovnih lastnosti precej podobna, so bila že ob nastanku prvotne odvisnostne drevesnice UD za slovenščino izdelana podrobna pravila za preslikavo oblikoskladenjskih oznak JOS v besedne vrste in oblikoskladenjske lastnosti sistema UD,[10] s katerimi je bil v celoti pretvorjen tudi učni korpus ssj500k. Na enak način smo z avtomatsko pretvorbo v univerzalne oblikoslovne oznake (besedne vrste in druge oblikoskladenjske lastnosti) pretvorili tudi novi učni korpus SUK z ročno pripisanimi oblikoskladenjskimi oznakami JOS. Ker se pretvorbena pravila v času od nastanka prejšnjih različic korpusov niso spremenila, smo v okviru projekta RSDO pretvorbo opravili zgolj na novo dodanih besedilih korpusa SUK in opravili ustaljeni ročni pregled povedi z glagolom *biti* za razdvoumljanje med pojavitvami pomožnega in glavnega glagola (po en označevalec na primer).

Poleg zgoraj opisanega označevanja celotnega korpusa SUK na oblikoslovni ravni smo prvotni korpus ssj500k oz. SSJ v obsegu 8.000 povedi dodatno povečali še za 5.435 novih ročno razčlenjenih povedi v obliki dvofazne označevalne kampanje. V prvi fazi razširitve so označevalci ročno pregledali 3.411 polpretvorjenih povedi korpusa ssj500k, ki zaradi omejene natančnosti pretvorbenih pravil v času nastanka prvotnega korpusa SSJ-UD niso bile javno objavljene, pri čemer so se označevalci osredotočili predvsem na pripisovanje novih oz. manjkajočih povezav (22.377 oz. 23,5 % vseh pojavnic). V drugi fazi širitve je bil skladenjsko razčlenjen še podkorpus ELEXIS-WSD, ki vsebuje 2.024 povedi, in sicer z ročnim pregledom vseh strojno

9 https://wiki.cjvt.si/books/07-universal-dependencies/page/oznacevalne-smernice, gl. Različica 1.0 – Priloga.

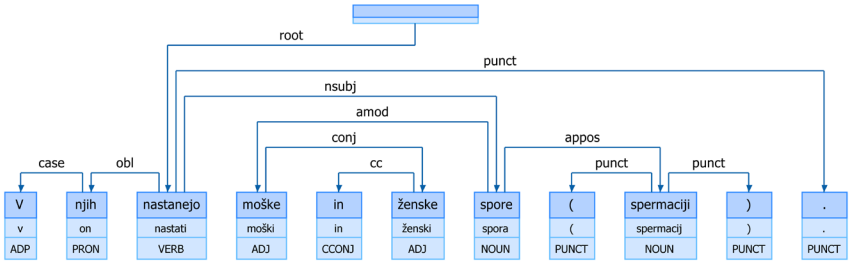
10 https://github.com/clarinsi/jos2ud

pripisanih razčlemb orodja CLASSLA-Stanza. V obeh fazah so vsako poved pregledali 2–3 neodvisni označevalci in končni kurator, pri čemer smo za označevanje uporabili orodje Q-CAT (Brank, 2022), ki odslej podpira tudi uvoz datotek v formatu CoNLL-U, za kuracijo pa spletno platformo WebAnno, ki jo vzdržuje CLARIN.SI. Pred objavo je bila glede na nekoliko spremenjene izhodiščne smernice in druge identificirane nedoslednosti s hevrističnimi poizvedbami izboljšana tudi označenost prvotne drevesnice SSJ.

Rezultati vseh zgoraj opisanih aktivnosti so objavljeni kot del novega referenčnega učnega korpusa za slovenščino SUK 1.0, s čimer se je količina učnih podatkov tako na oblikoslovni kot skladenjski ravni skoraj podvojila (gl. Tabelo 1). Univerzalno skladenjsko razčlenjeni del korpusa SUK je bil po standardni delitvi na učno, validacijsko in testno množico obenem objavljen tudi kot del skupne mednarodne zbirke drevesnic UD v2.10 – kot nova, razširjena in izboljšana različica drevesnice SSJ. Nova različica SSJ v primerjavi s prvotno vsebuje skoraj enkrat večje število pojavnic (126.427, +89,9 %), s čimer se korpus SSJ po številu pojavnic danes umešča v zgornjo osmino vseh drevesnic UD po svetu. Z razširitvijo je drevesnica SSJ postala tudi bolj raznolika, saj se vsi trije podkorpusi (izvorne povedi iz ssj500k, nove povedi iz ssj500k, nove povedi iz ELEXIS-WSD) med seboj razlikujejo tako z vidika vrste vsebovanih besedil kot njihove skladenjske kompleksnosti.

Drevesnica SSJ, tj. univerzalno oblikoskladenjsko razčlenjeni podkorpus korpusa SUK, je bila kot samostojna podatkovna množica že integrirana v številna orodja in spletne portale po svetu,[11] po njej pa je mogoče brskati tudi s pomočjo lokalno razvitega orodja Q-CAT (Slika 1) in spletnega vmesnika Drevesnik, ki sicer trenutno omogočata zgolj prikaz univerzalnih besednih vrst in odvisnostnih skladenjskih relacij, ne pa oblikoslovnih lastnosti tipa Case=Nom.[12]

11 https://universaldependencies.org/tools.html
12 https://orodja.cjvt.si/drevesnik/



Slika 1: Primer označene povedi po shemi Universal Dependencies v orodju Q-CAT.

2.4 Skladnja po sistemu JOS-SYN

Sistem JOS-SYN, ki je bil zasnovan v projektu Jezikoslovno označevanje slovenščine (Erjavec idr., 2010), sledi spoznanjem slovenskega jezikoslovja (zlasti slovnici Toporišič, 2004), obenem pa temeljnim idejam, ki jih zarisujejo obstoječi uveljavljeni sistemi odvisnostnega označevanja. Ključna lastnost sistema je, da upošteva informacije, ki jih prinašajo oblikoskladenjske oznake JOS oz. njihova sodobna različica MULTEXT-East v6 (Erjavec, 2012). Na skladenjski ravni tako dodajamo samo informacije, ki jih še ni pokrila oblikoskladnja, kar omogoči robusten, intuitiven in hitro razložljiv označevalni sistem.[13]

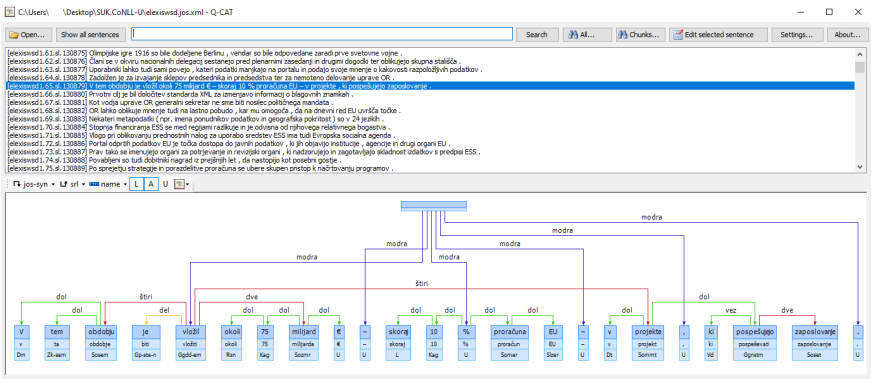
Skladenjska raven JOS-SYN je bila dobro zastopana že v prejšnji različici učnega korpusa: v ssj500k je bilo s tem sistemom označenih 11.411 povedi v 617 besedilih s skupnim obsegom 235.864 pojavnic (Krek idr., 2020a, 25–26). Na teh podatkih je že bil naučen skladenjski razčlenjevalnik za slovenščino, ki je dosegal 90,43 % za pravilno določeno mesto povezave oz. 87,52 % za pravilno določena mesto in tip povezave (Dobrovoljc idr., 2012). Cilj nove označevalne kampanje je bil označiti 2.024 novih povedi ELEXIS-WSD, s tem povečati obseg učnega gradiva, pri tem pa natančneje oceniti ter nadgraditi označevalne smernice.[14] Kampanja je trajala približno šti-

13 Sistem oznak je predstavljen na strani https://wiki.cjvt.si/books/06-odvisnostna-skladnja-jos-syn/page/predstavitev-oznak.

14 https://wiki.cjvt.si/books/06-odvisnostna-skladnja-jos-syn/page/oznacevalne-smernice, gl. Različica 1.0.

ri mesece, dva meseca za intenzivno označevanje in dva meseca za pripravo analiz in nadgradnjo smernic.

Povedi ELEXIS-WSD, v katerih je že bila ročno popravljena tokenizacija, segmentacija, lematizacija ter oblikoskladnja MULTEXT-East, smo najprej strojno skladenjsko označili z orodjem CLASSLA-Stanza (verzija 1.1.0), nato pa sta dva jezikoslovca s pomočjo orodja Q-CAT (Brank, 2022) ročno pregledala vsako od povedi in popravila strojno pripisane skladenjske oznake (Slika 2). Nejasnosti in neskladja v označevalnih rešitvah smo beležili in naslavljali sproti ob delu. Težja mesta označevanja, ki so izvirala iz nejasnosti označevalnih smernic ali novoodkritih označevalnih zadreg, smo jezikoslovno analizirali, poiskali rešitve in posodobili smernice. Poleg vrzeli v smernicah smo med delom identificirali tudi mesta, kjer so podatki v ssj500k označeni neskladno. Za določene vrste težav, ki jih navajamo v nadaljevanju, smo skladenjske oznake v podatkih ssj500k posodobili, nekaj usklajevanja bo treba še opraviti v prihodnje, nekatere težave pa se propagirajo z nižjih ravni, čemur se bo treba posvetiti v nadaljnjih projektih.



Slika 2: Označevanje odvisnostne skladnje JOS-SYN v programu Q-CAT.

Da bi označevalne smernice postale preproste za nadaljnje nadgrajevanje in uporabo, smo jih oblikovno in vsebinsko poenostavili, strukturo nadgradili in zagotovili dodatne zglede označevanja (več o tem v Arhar Holdt idr., 2023). V smernice smo dodali nova poglavja,

ki natančneje pojasnjujejo označevanje izbranih pojavov. Nova je denimo obravnava simbolov in ločil, ki nadomeščajo besede (npr. *% ° $* za besede *odstotek, stopinja, dolar*), znake *+ & /* - v pomenu veznikov 'in', 'ali' (npr. *srčno-žilna bolezen*), znak / v pomenu 'na' (*6 mg/kg*), znaka - in – v pomenu 'od'–'do', 'proti' (*v sezoni 2006–07*) ter znak - pri povezovanju kratic in števil v podredne zveze (*16-tonski*). Ti elementi pri predhodnem označevanju niso bili vpeti v besednozvezno skladnjo, zaradi česar je razpadla drevesnica vseh povedi, ki so jih vsebovale. Nove smernice, ki ločujejo besednozvezno povezljive znake od nepovezljivih, za povezljive pa jasno prikažejo načine povezovanja, so skladnejše s primerljivimi sistemi, tudi UD za slovenščino. Ker gre za večjo spremembo sistema označevanja, smo pregledali in uskladili obravnavo tovrstnih elementov tudi v ssj500k.

Obširnejša dopolnitev smernic je bila pripravljena tudi za obravnavo lastnih imen in tujejezičnih elementov. Problematiko smo strukturirali na ožje vsebinske sklope, za vsakega pripravili opis, temelječ na analizah predhodnega označevanja, pa tudi posebna opozorila, kjer je v preteklosti prihajalo do zmede. Navodila za označevanje lastnih imen so bila predhodno precej skopa, posledično pa je v ssj500k opaziti velike neskladnosti označevanja, tako pri določanju, ali zvezo obravnavati kot slovensko ali kot fragment v tujem jeziku (glede na smernice se fragmenti v tujem jeziku tipično ne povezujejo v drevesnico), kot tudi odločanje, kaj je jedro pri zvezah, ki prihajajo iz tujega jezika. Precej nedoslednosti je najti pri povezovanju tujejezičnih členov tipa *de, la, the*, za katera velja posebna obravnava, vendar jih označevalci težko prepoznavajo, kadar gre za manj znane tuje jezike. Najtrši oreh pri označevanju pa so tujejezična stvarna lastna imena, kjer naj bi označevanje sledilo odločitvam na ravni oblikoskladnje, vendar tudi tam smernice niso optimalne (Pori idr., 2022).

Od sprememb je možno izpostaviti še nekaj takšnih, ki so vezane na označevanje specifičnih struktur (za referenco gl. nove smernice[15]). V poglavje, ki se posveča označevanju struktur tipa *nujno je*,

15 https://wiki.cjvt.si/books/06-odvisnostna-skladnja-jos-syn/page/oznacevalne-smernice, gl. Različica 2.0

smo dodali obravnavo struktur *treba je,* saj je za označevalce koristno na enem mestu videti, da so pridevniki v takšnih primerih z glagola *biti* vezani s povezavo DOL, prislovi pa s TRI. Na podoben način smo v poglavje *Polstavčni desni prilastki*, ki se je predhodno osredotočalo na pridevniške, deležniške in nedoločniške polstavke (povezava DOL), dodali še primer obravnave deležijskih desnih prilastkov (povezava TRI). Nadgradili smo poglavje o prilastkovih odvisnikih, ki zdaj vsebuje tudi navodila za označevanje t. i. nepravih odvisnikov, prilastkovih odvisnikov v povedih s pristavki ter primerov tipa *dovolj star, da*. Nenazadnje, pojasnili smo navodila za označevanje osebka pri pasivnih strukturah s *se* (npr. *v hudih primerih se daje adrenalin*).

Posodobili smo dve mesti smernic, kjer se nahajajo vnaprej pripravljeni (zaključeni) seznami besed, ki jih označujemo po določenih pravilih, in sicer informativni seznam zvez, ki jih povezujemo s povezavo SKUP,[16] ter seznam členkov, ki jih povezujemo v besedne zveze kot določujoči element. Oba seznama smo posodobili na osnovi analiz predhodnega označevanja in pojavnosti obravnavanih jezikovnih elementov v referenčnem korpusu, upoštevali pa smo tudi označevalne prakse pri skladenjskem sistemu UD za slovenščino.

Pri preverbah že označenega gradiva smo identificirali tudi nedoslednosti, ki ne izvirajo nujno iz nejasnosti smernic in bi jih bilo treba v nadaljnjih projektih sistematično nasloviti in odpraviti. Poleg že omenjene težave z označevanjem (zlasti tujejezičnih stvarnih) lastnih imen so se kazala neujemanja pri povezovanju členkov in prislovov (npr. *vsaj, izključno*) in slovničnih besed, ki lahko nastopajo kot različne besedne vrste (npr. *niti, razen*), povezovanju pridevnikov, kadar modificirajo številnike (npr. *dodatnih 400 milijonov*), označevanju pridevniške in samostalniške vezljivosti, latinskih poimenovanj, citatov in drugih fragmentov (npr. pri zvezah s *pa tudi*), ločevanjem med osebkom in povedkovim določilom; predmetom in prislovnim določilom; oznakami TRI in ŠTIRI ter prilastkovimi in drugimi odvisniki (npr. v stavkih s *ko, preden, dokler*). Za urejanje doslednosti so ključni zlasti problemi, ki se lahko propagirajo na višje

16 Besede, ki imajo variantni zapis skupaj ali narazen, večbesedne veznike in podobne večbesedne enote.

ravni (udeleženske vloge) ali so posledica nerešenih vprašanj na nižjih ravneh (oblikoskladnja).

2.5 Udeleženske vloge po sistemu SRL

Označevanje korpusa s semantičnimi kategorijami izhaja iz potrebe po strojnem procesiranju jezikovnih podatkov, ki so semantične narave, in zadeva različne možnosti njihove uporabe, kot je razvoj sistemov za luščenje informacij, sistemov za odgovarjanje na vprašanja, izboljšava delovanja skladenjskih razčlenjevalnikov, strojnih prevajalnikov ipd.

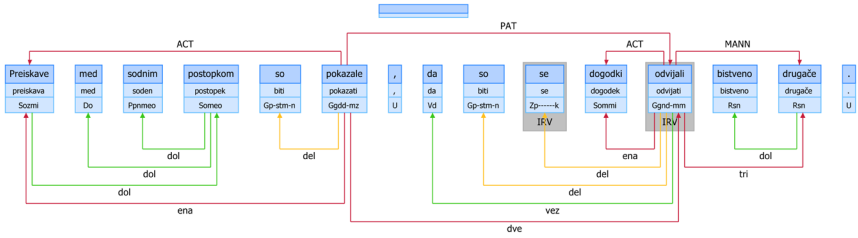
Celotni del semantično označenega korpusa SUK predstavlja **podkorpus SRL**, ki vsebuje dva dela. Korpus **SRL-ssj500k** vsebuje 9.724 ročno označenih povedi iz priprave predhodne različice učnega korpusa (ssj500k 2.3) in povedi, ki so bile v ssj500k 2.3 označene na morfološki in skladenjski ravni, niso pa bile označene na semantični ravni. **SRL-WSD** predstavlja korpus ELEXIS-WSD, ki vsebuje 2.024 povedi. Razen že predhodno ročno pregledanih povedi sta bila korpusa najprej avtomatsko označena na semantični ravni s pomočjo SRL parserja (Björkelund idr., 2009), korpus SRL-WSD pa tudi na morfološki in skladenjski ravni z orodjem CLASSLA-Stanza po sistemu JOS in UD. V označevalno kampanjo na semantični ravni je bilo skupaj vključenih 11.748 povedi, od tega je bilo 5.501 povedi ponovno pregledanih, 6.247 povedi pa je bilo najprej avtomatsko označenih, nato pa ročno pregledanih. Odločitve so bile na koncu usklajene v celotnem podkorpusu SRL učnega korpusa SUK.

Pri snovanju slovenskega modela za semantično označevanje (Krek idr., 2016) smo se glede na analizo označevalnih sistemov odločili, da bomo izhajali iz funkcijskega generativnega pristopa Praške odvisnostne drevesnice (ang. *Prague Dependency Treebank*, PDT; Mikulová idr., 2006).[17] Z vidika optimizacije pomenske razdrobljenosti, upoštevanja slovenskih specifik in prekrivnosti oznak med posameznimi sistemi smo nabor ustrezno zreducirali, kot je opisano v

17 Strnjen pregled vseh semantičnih oznak, njihov opis in zgledi je na voljo na tej povezavi: https://wiki.cjvt.si/books/10-udelezenske-vloge-srl/page/predstavitev-oznak.

Arhar Holdt idr. (2023). Podroben opis semantičnih oznak in pravila za njihovo uporabo vsebujejo Smernice za semantično označevanje učnega korpusa,[18] ki so bile v okviru projekta RSDO nadgrajene in posodobljene glede na vsebinske analize.

Vse povedi v na novo označenem in popravljenem predhodno že označenem korpusu je označila podiplomska študentka slovenistike na podlagi prve različice smernic in na podlagi sprotnih konzultacij in navodil. Celoten na novo označen korpus je nato pregledala soavtorica tega prispevka s pomočjo sistematičnih in ciljnih preverjanj. Za označevanje je bilo uporabljeno orodje Q-CAT (Brank, 2022), kot prikazuje Slika 3.



Slika 3: Prikaz semantične označevalne ravni v orodju Q-CAT.

Izhodišče semantičnega označevanja je predstavljal posamezni glagol v vseh svojih pojavitvah znotraj vnaprej določenih pomenskih skupin, npr. glagoli govorjenja, premikanja, kognitivnih procesov ipd., kar je omogočilo prepoznavanje tipičnih udeleženskih vlog, ki se povezujejo s posameznimi pomeni glagolov znotraj skupnega pomenskega polja. Z označevanjem smo začeli pri pogostejših glagolih (*biti, imeti, morati, iti, začeti, vedeti*) ter nadaljevali z upoštevanjem sorodnih pomenskih skupin, npr. glagolov rekanja (*povedati, reči, praviti, govoriti*). Na koncu smo označili glagole z zgolj eno pojavitvijo v povedi (pribl. 1200). Na ta način smo v največji možni meri zajeli povedi, za katere je bilo mogoče izpeljati čim bolj sistematične in usklajene jezikovne rešitve.

18 https://wiki.cjvt.si/books/10-udelezenske-vloge-srl/page/oznacevalne-smernice, gl. Različica 1.0.

V procesu označevanja je bil korpus nadgrajen tudi z vsebinskega vidika, pri čemer dodana vrednost temelji na jezikoslovnem premisleku že obstoječih odločitev v skladu z novimi spoznanji pri izdelavi semantičnih virov, analize vezljivostnih vzorcev pri izdelavi Vezljivostnega leksikona (Gantar, 2021; Gantar, 2023) in na upoštevanju potreb jezikovnotehnološke skupnosti.

V približno 75 % korpusa so popravljena in poenotena razmerja med udeleženci pri glagolih rekanja po načelu: REC = naslovnik glagolskega dejanja, RESLT = konkretni končni rezultat ali "izdelek" glagolskega dejanja (npr. izjava sama, ki jo največkrat uvaja odvisni stavek), PAT = vsebina ali tema glagolskega dejanja.

Druge pomembne vsebinske izboljšave korpusa temeljijo na analizi nekaterih problematičnih skladenjskih struktur in poenotenju odločitev v povezavi z označevanjem skladenjskega nivoja. Sem sodi poenotenje in uskladitev opredeljevanja razmerja med udeleženci v skladenjsko enakovrednih povedih tipa: *kdo ali kaj je kdo ali kaj*. Na podlagi smernic predhodnega semantičnega označevanja učnega korpusa ssj500k smo z udeleženske vlogo ACT, ki v splošnem zajema vršilce in pobudnike dejanja, označevali samostalnike v imenovalniku, ki nastopajo kot osebki glagola *biti*; samostalniška povedkova določila ob glagolu *biti* pa kot prizadeto (PAT): *območje medenice(ACT)* je središče telesa*(PAT)*; *problem begunc*ev*(ACT)* je stvar države*(PAT)*. Glede na omenjena izhodišča smo že na ravni prvotnega označevanja tu predvidevali največ neenotnosti na pomenskem nivoju in odstopanja med skladenjskim in pomenskim nivojem, predvsem zaradi težav pri odločanju o izhodišču in določilu stavka na pomenski ravni in o polno- oz. nepolnopomenski vlogi glagola *biti*, ki odloča med osebkovo in povedkovodoločilno vlogo na skladenjski ravni. V zvezi s tem smo pri nadgradnji korpusa sprejeli odločitev, da v skladenjsko enakovrednih povedih pomenska interpretacija sledi pravilu: kar izvem novega = prizadeti (PAT) udeleženec, o komer ali čemer izvem kaj novega = nosilni udeleženec (ACT). To v veliki meri ustreza označevanju na skladenjski ravni, kjer se temu, kar je na pomenski ravni aktant, pripisuje odvisni del povedka (povezava *dol*), temu, kar na

pomenskem nivoju opredeljujemo kot prizadeto, pa je na skladenjski ravni tipično pripisan osebek (povezava *ena*):

Dogodek v Ankaranu(dol-ACT) je bila dramatična nesreča(ena-PAT).
Gostja večera(dol-ACT) bo Desa Muck(ena-PAT).
Večina potnikov(dol-ACT) so bile ženske(ena-PAT).

Označevanje je v skladu z zgornjimi odločitvami dosledno izpeljano na pribl. 90 % povedi združenega korpusa SRL, medtem ko je smiselnost poenotenja z razmerij na skladenjski ravni (tj. ena-ACT; dva-PAT) eden od jezikoslovnih premislekov, ki terjajo širši jezikovni konsenz.

Z omenjeno vsebinsko nadgradnjo so povezane tudi odločitve pri drugih udeleženskih vlogah glagola *biti* po sistemu: *biti* + samostalnik = PAT: *dogodek(ACT) je bil nesreča(PAT)*; biti + pridevnik = RESLT: *je osamljena(RESLT)*; biti + prislov = MANN: *bo toplo(MANN)*. Popravki so bili izvedeni tudi na predhodno ročno že označenih povedih, s čimer smo želeli doseči enotnost označevanja pri nekaterih najpogostejših semantičnih vzorcih.

Prav tako so bile deloma poenotene odločitve, aplicirane na korpusne povedi v približno 80 %, pri razumevanju agentnih in deagentnih rab. Pri označevanju smo sledili pomenski interpretaciji izhodiščnega udeleženca kot vršilca dejanja (ACT), ki mu praviloma ni mogoče dodati še enega vršilca, ne da bi se pri tem spremenil pomen: *dogodki(ACT) so se odvijali bistveno drugače – *ACT je odvijal dogodke …*, in pravilu, da morajo ostati udeleženske vloge v agentnih in deagentnih strukturah nespremenjene, kjer prihaja do diskrepance med skladenjskim in pomenskim nivojem: *stvar(PAT-ena) je malce bolj zapletena – zgodbo(PAT-dve) sta sami(ACT-tri) zapletli.* Pri nadaljnji nadgradnji učnega korpusa bi bilo smiselno upoštevati tudi neenotnosti v pomenski interpretaciji, ki niso bile sistematično odpravljene, npr. *med njimi so se širile govorice(ACT)* : *potem je začela širiti govorice(PAT)*.

Nadaljnje izboljšave korpusa vidimo na več ravneh: z aktualizacijo semantičnih oznak glede na označevalni sistem PDT (opisano v

Arhar Holdt idr. (2023, 44–45)); z nadgradnjo korpusa z naborom semantičnih oznak glede na jezikoslovne analize, ki zahtevajo konsenz tudi na drugih označevalnih ravneh; ter z nadgradnjo korpusa s semantičnimi kategorijami, ki se oblikujejo znotraj pobud za povezovanje konceptov na medjezikovni ravni (npr. UniDive,[19] ELEXIS[20]).

2.6 Imenske entitete

Imenske entitete (ang. *named entities*; NE) so samostalniki in samostalniške besedne zveze, ki identificirajo neko osebo (oznaka PER), lokacijo (oznaka LOC), organizacijo (oznaka ORG) ali drug edinstven objekt v realnem prostoru in času (oznaka MISC). Tem standardnim oznakam se pridružuje še kategorija svojilnih pridevnikov, izpeljanih iz osebnega lastnega imena (oznaka DERIV-PER), npr. [*Obamova*] DERIV-PER *izvolitev*), ki se je kot odgovor na potrebo po celovitejši anonimizaciji osebnih podatkov pokazala kot nepogrešljiva. Imenske entitete so na ortografski ravni pogosto izražene z veliko začetnico (npr. *Slovenska tiskovna agencija*) ali kratico (npr. *STA*), vendar pa velika začetnica in kratica ne označujeta samo imenskih entitet (npr. *BDP*). Identifikacija imenskih entitet v besedilu je pomembna za odkrivanje koreferenčnosti, analiziranje sentimenta, ekstrakcijo informacij, povezav in dogodkov ter druge naloge, povezane s procesiranjem naravnega jezika.

V projektu RSDO so bile imenske entitete ročno pregledane v korpusih SentiCoref 1.0 in ELEXIS-WSD, tj. v 20.166 povedih oz. 96,31 % novega gradiva. SentiCoref je že vseboval strojno pripisane oznake, entitete, ki se pojavljajo v koreferenčnih verigah, pa so bile tudi ročno pregledane, medtem ko je bil ELEXIS-WSD predoznačen v projektu, z orodjem CLASSLA-Stanza. Pri ročnem pregledu obeh korpusov smo sledili predhodno uveljavljenim označevalnim smernicam.[21] Kampanja pregledovanja je potekala v spletnem orodju INCEpTION (Klie idr., 2018), ki je preprosto za uporabo in nudi

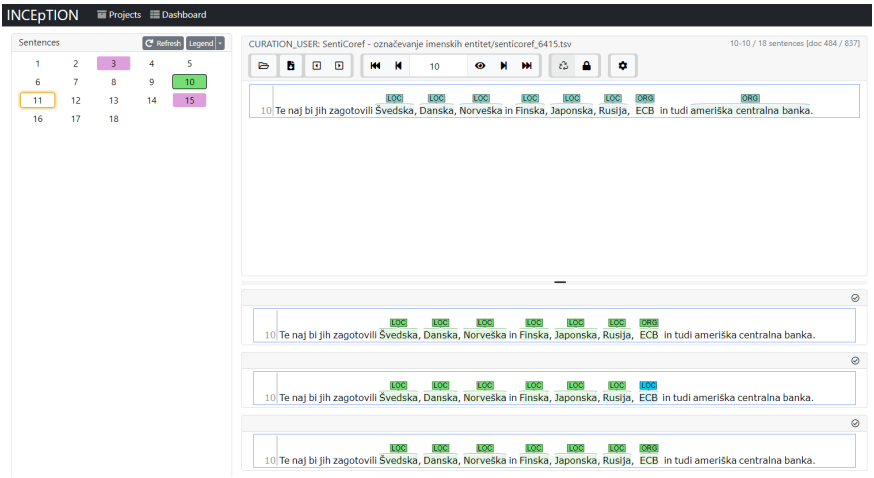
19 https://www.cost.eu/actions/CA21167/
20 https://elex.is/
21 https://wiki.cjvt.si/books/08-imenske-entitete/page/oznacevalne-smernice, gl. Različica 1.1.

dober pregled nad že opravljenim delom. Gradivo so pod vodstvom koordinatorja pregledovale tri študentke jezikoslovnih smeri. Vsako poved so pregledale vse tri študentke, neujemanja med pripisanimi oznakami pa je v fazi kuracije posebej obravnaval koordinator in tem primerom tudi pripisal končne oznake (Slika 4).

Pri označevanju se je pojavil pomislek glede obravnave ženskih oblik priimkov, ki so tvorjeni iz moških priimkov in so z oblikovnega vidika svojilni pridevniki (npr. *Kresalova*). Po tem kriteriju bi jim morali prisoditi oznako DERIV-PER, a smo tovrstnim primerom pripisali oznaka PER, saj pomensko delujejo kot osebno lastno ime, poleg tega pa so oblikoskladenjske lastnosti zabeležene na nivoju oblikoskladnje.

Kot problematično se je izkazalo tudi določanje začetka imenske entitete v primerih, ko je prvi del uradnega imena organizacije zapisan z malo začetnico, ker ga pisec besedila dojema kot vrstno poimenovanje (npr. *občina Gornja Radgona*). Obveljalo je splošno pravilo, po katerem je glavni kazalnik, da celotno enoto označimo kot imensko entiteto, velika začetnica ([*Občina Gornja Radgona*] ORG). V določenih primerih pa lahko kot imensko entiteto obravnavamo tudi primere, ki so zapisani z malo začetnico, a vsebujejo vse sestavine uradnega imena te institucije. Tak primer je [*ameriška centralna banka*]ORG, uradno slovensko poimenovanje pa je *Ameriška centralna banka*. Če je institucija zapisana kot parafraza uradnega imena, ne glede na to, ali je zapisana z malo ali veliko začetnico, je ne označimo kot imensko entiteto, npr. *Karavanški predor*, saj je uradno ime *predor Karavanke*. Posebna problematika označevalnega sistema je predvsem predpostavka, da avtorji besedil vedno upoštevajo pravopisna pravila in se tudi pozanimajo o uradni obliki imena določene institucije.

Pri obravnavi dilem se je tudi izkazalo, da bi poleg obstoječih oznak potrebovali še oznako za pridevnike iz stvarnih lastnih imen (npr. *Mercatorjev*), za katere bi po vzoru DERIV-PER lahko uvedli oznako DERIV-ORG. Enako velja za svojilne pridevnike iz entitet z oznako LOC (npr. *Lunin*), ki bi jim lahko pripisali oznako DERIV-LOC. Uvedba novih kategorij bi bil radikalnejši poseg v obstoječe smernice, kar bi bilo v prihodnje smiselno temeljiteje premisliti.



Slika 4: Prikaz faze kuracije v orodju INCEpTION: v spodnjih treh vrsticah vidimo odločitve pregledovalk, v zgornji vrstici pa je prikazana končna odločitev kuratorja.

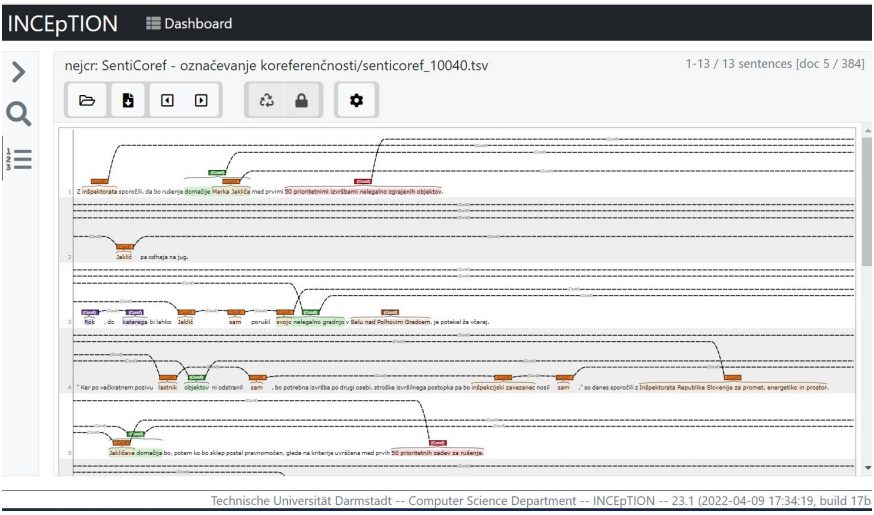
2.7 Koreference

»Odkrivanje koreferenčnosti je ena izmed treh ključnih nalog ekstrakcije informacij iz besedil, kamor spadata še prepoznavanje imenskih entitet in ekstrakcija povezav« (Žitnik in Bajec, 2018). V okviru projekta RSDO smo jo iskali v 837 besedilih korpusa SentiCoref 1.0. Besedila so obsegala 18.142 povedi oz. 391.962 pojavnic. Za iskanje koreferenc so najprimernejša krajša zaključena besedila, zato smo za to nalogo izbrali množico SentiCoref, drugo gradivo v učnem korpusu je namreč razdeljeno na odstavke ali še krajše enote.

V izbranih besedilih so predhodno že označili koreference (Žitnik in Bajec, 2018), vendar se je izkazala potreba po nadgradnji označevalnega sistema za slovanske jezike, saj ti referenčnost pogosto izražajo tudi morfemsko. Kot osnovo novega označevalnega sistema smo uporabili označevalne smernice ReLDI: Uputstvo za anotiranje koreferenci (interni dokument za projektno rabo), ki so v sklopu iniciative ReLDI 2008 nastale za potrebe srbskega jezika. Smernice smo prevedli v slovenščino, jih uredili in prilagodili, pri čemer je bila najpomembnejša odločitev označevalcev, da za razliko od srbske kampanje na ravni koreferenc ne označujemo skladenjskih značilnosti – za

slovenščino so te v korpusu SUK namreč dosledneje in celoviteje določene pri oblikoskladenjskih in skladenjskih oznakah. Smernice smo skupno pripravljali in dopolnjevali v spletnem urejevalniku *Google Dokumenti*, končna različica pa je na voljo na portalu Wiki CJVT.[22]

Kampanja označevanja koreferenčnosti je bila, kot kampanja označevanja imenskih entitet, opravljena na platformi INCEpTION (Slika 5). Gradivo sta pregledala dva raziskovalca, eden pa je kampanjo tudi koordiniral. Osnovne dileme so bile večinoma razrešene v uvajalni fazi, nekatere tudi pozneje med sprotno komunikacijo ob problemih pri označevanju samih besedil. Pomemben del uvajalne faze je bilo na primer poenotenje in določitev jasnejše terminologije v smernicah. Določili smo razmerja med termini *entiteta*, *koreferenčnost*, *koreferenca* in *omenitev*. Prav tako smo iz izvornih smernic odstranili del gradiva, ki je primerjalo označevalni sistem z alternativnimi pogledi na koreferenčnost, in poskrbeli za natančno členjenost poglavij ter pravilno označenost zgledov. Poenostavljene in eksplicitne smernice so izboljšale komunikacijo med označevalcema in sam označevalni sistem.



Slika 5: Označevanje koreferenc v orodju INCEpTION.

22 https://wiki.cjvt.si/books/09-odkrivanje-koreferencnosti/page/oznacevalne-smernice, gl. Različica 1.6.

Ob nadaljnjem označevanju se je v praksi izrazila pomembna konceptualna dilema označevanja koreferenčnosti, kadar so povezave med posameznimi omenitvami v besedilu zanikane ali pa je o povezavi posamezne omenitve z antecendentom pisec besedila izrazil dvom. Take primere najdemo predvsem pri novinarskih prispevkih, katerih temelj je naklonski členek *naj*, saj s pogojnikom (kondicionalom) izražajo konstanten dvom o resničnosti povezave med vršilcem dejanja in samim dejanjem. Pravilo, da se koreferenčnosti ob dvomu o povezavi znotraj samega besedila ali njenem zanikanju ne označuje, se je izkazalo za neizvedljivo in je posledično povzročalo precej težav. Pri nadaljnjem razvoju označevalnega sistema bo ta izziv treba upoštevati in sprejeti drugačne smernice ali pa te natančneje določiti s primeri konkretnih besedil, ne samo posameznih povedi. Nekaj težav je povzročal tudi vrstni red označevanja posameznih omenitev, saj je lahko v enem stavku samo ena koreferenca na posameznega referenta, zato smo določili, da imajo samostalniki prednost pred na primer zaimki. Pri veriženju vseh izpeljanih pridevnikov in lastnoimenskih samostalnikov pa je prihajalo do obsežnega kopičenja koreferenčnih povezav in s tem drobljenja, ki bi lahko povzročilo poznejšo slabšo ekstrakcijo informacij naučenih modelov.

Dele izvornih smernic, ki za slovenščino niso bili relevantni, smo umestili na konec dokumenta in zapisali posebno opozorilo, da v naši označevalni kampanji niso bili upoštevani. Ta del navodil smo v dokumentu vseeno ohranili, saj je v njih mnogo primerov povedi, označenih s koreferencami.

V nadaljnjih kampanjah bi bilo smiselno evalvirati uspešnost in ustreznost posameznih označevalnih odločitev in smernice še enkrat posodobiti. Trenutne zglede v smernicah je treba nadomestiti in dopolniti z realnimi zgledi iz korpusnih besedil, saj označevalna praksa razkrije številne izzive, na katere teoretične smernice ne dajejo natančnih odgovorov.

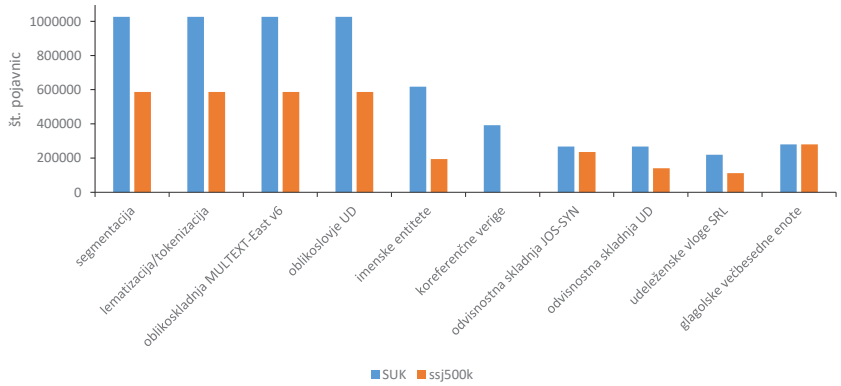
3 Kvantitativni pregled korpusa

Vseh 1.025.639 pojavnic novega učnega korpusa SUK je označenega in ročno pregledanega na ravni stavčne segmentacije, tokenizacije, lematizacije in oblikoskladenjskih oznak. Skoraj dve tretjini korpusa vsebujeta oznake imenskih entitet, dobrih 38 % celotnega korpusa pa je označenega na nivoju koreferenc. Približno četrtina korpusa vsebuje oznake odvisnostne skladnje po sistemih JOS-SYN in UD, približno petina pa oznake udeleženskih vlog SRL. Z oznakami glagolskih večbesednih enot je označenih 27 % gradiva, vse še iz ssj500k. Natančni podatki po označevalnih nivojih so prikazani v Tabeli 1.

Tabela 1: Količina pregledanega gradiva v SUK po označevalnih nivojih.

Označevalni nivo	Pojavnice	Povedi	Besedila	% celotnega SUK
Segmentacija	1.025.639	48.594	2.908	100
Lematizacija/tokenizacija	1.025.639	48.594	2.908	100
Oblikoskladnja MULTEXT-East v6	1.025.639	48.594	2.908	100
Oblikoslovje UD	1.025.639	48.594	2.908	100
Imenske entitete	617.832	29.654	1.336	60,24
Koreferenčne verige	391.962	18.142	837	38,22
Odvisnostna skladnja JOS-SYN	267.097	13.435	618	26,04
Odvisnostna skladnja UD	267.097	13.435	618	26,04
Udeleženske vloge SRL	219.216	11.748	598	21,37
Glagolske večbesedne enote	280.522	13.511	754	27,35

Označenost SUK-a v primerjavi s ssj500k je predstavljena v Grafu 2.



Graf 2: Primerjava označenega gradiva v ssj500k in SUK po označevalnih nivojih.

4 Kodiranje korpusa

Tako kot ssj500k je tudi SUK kodiran v formatu XML s shemo, ki sledi priporočilom TEI,[23] vendar po nadgrajeni kodirni shemi, ki jo priporoča CLARIN.SI.[24] Ker je SUK sestavljen iz več podkorpusov, ki imajo različne metapodatke o besedilih in ravni označevanja, je korpus oblikovan kot krovna datoteka TEI s kolofonom in povezavami na posamezne datoteke podkorpusov. Vsak podkorpus nato vsebuje razdelke z označenimi besedili.

Slika 6 prikazuje začetek podkorpusa SentiCoref, kjer vrhnji zamejuje podkorpus, gnezdeni pa prvo besedilo. V elementu <bibl> so podani metapodatki besedila, nato pa sledi začetek prvega odstavka in nato prve povedi. Prvi dve besedi sestavljata imensko entiteto tipa organizacija (<seg type="name" subtype="org">), besedi pa sta označeni z MULTEXT-East oblikoskladenjskimi oznakami in oblikoslovnimi lastnostmi po sistemu Universal Dependencies ter s svojo lemo.

23 https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html
24 https://github.com/clarinsi/TEI-schema, gl. tudi https://www.clarin.si/repository/xmlui/page/data#tei

```
<div xmlns="http://www.tei-c.org/ns/1.0" xml:id="senticoref" xml:lang="sl">
   <div xml:id="senticoref1" xml:lang="sl">
      <bibl corresp="#senticoref1">
         <title>Zakaj se hrana draži?</title>
         <note type="handle">http://hdl.handle.net/11356/1285</note>
         <note type="source">SentiCoref 1.0</note>
         <note type="url">http://www.24ur.com/novice/gospodarstvo/zakaj-se-hrana-drazi.html</note>
         <note type="main_url">www.24ur.com</note>
         <note type="keywords">iztok, jarc, michel, barnier, hrana, podražitev</note>
         <note type="author">STA / V.L.</note>
         <date>2007-09-03</date>
      </bibl>
      <p xml:id="senticoref1.1">
         <s xml:id="senticoref1.1.1">
            <seg type="name" subtype="org" xml:id="senticoref1.1.1.ne1">
               <w ana="mte:Ppnzei" msd="UPosTag=ADJ|Case=Nom|Degree=Pos|Gender=Fem|Number=Sing"
                  lemma="evropski" xml:id="senticoref1.1.1.t1">Evropska</w>
               <w ana="mte:Sozei" msd="UPosTag=NOUN|Case=Nom|Gender=Fem|Number=Sing"
                  lemma="komisija" xml:id="senticoref1.1.1.t2">komisija</w>
            </seg>
```

Slika 6: Primer zapisa korpusa v TEI.

Kompleksnejše označevalne ravni, kot sta skladnja ali koreferenčnost, so kodirane v elementih <linkGrp> znotraj svoje povedi, ta element pa vsebuje skupino povezav, ki med seboj povežejo ustrezne elemente, lahko pa tudi podajo funkcijo povezave, kar se uporablja pri skladenjskih povezavah.

Kot ilustrira Slika 7 na primeru označevanja koreferenc, lahko poved vsebuje tudi segmente, ki združujejo večbesedne izraze (<seg type="coref">), ti segmenti, ali pa segmenti za imenske entitete, pa so nato prek povezav združeni v koreferenčno verigo.

Zapis XML oz. TEI je zelo ekspresiven, datoteke je tudi mogoče formalno validirati, vendar pa je za učinkovito uporabo takšnega kodiranja potrebna ustrezna programska oprema in poznavanje standarda XML, pa tudi zapis TEI je kompleksen in zahteva privajanje. V računalniškem jezikoslovju se je v zadnjih letih uveljavil bistveno bolj enostaven zapis CoNLL-U, ki je bil razvit v sklopu projekta Universal Dependencies, zato je korpus dostopen tudi v takšnem zapisu, ki pa sicer ne zajema kompleksnejših vrst oznak, kot so koreference.

CoNLL-U je enostaven tabelaričen format, namenjen zapisu oznak jezikoslovnih ravni do odvisnostne skladnje, dopušča pa tudi pripisovanje enostavnih metapodatkov povedim, odstavkom in besedilom. V stolpcu ‘ostalo’ je mogoče dopisati poljubne atribute

```
   <seg type="coref" xml:id="senticoref1.1.1.phr9-1">
      <w ana="mte:Ppnmeid"
         msd="UPosTag=ADJ|Case=Nom|Definite=Def|Degree=Pos|Gender=Masc|Number=Sing"
         lemma="francoski" xml:id="senticoref1.1.1.t17">francoski</w>
      <w ana="mte:Somei" msd="UPosTag=NOUN|Case=Nom|Gender=Masc|Number=Sing"
         lemma="kolega" xml:id="senticoref1.1.1.t18">kolega</w>
   </seg>
</seg>
<pc ana="mte:U" msd="UPosTag=PUNCT" lemma="." xml:id="senticoref1.1.1.t19">.</pc>
<linkGrp type="COREF">
   <link target="#senticoref1.1.1.nel #senticoref1.1.2.nel #senticoref1.1.3.nel"/>
   <link target="#senticoref1.1.1.phr52-1 #senticoref1.1.3.phr52-2 #senticoref1.1.11.phr52-3"/>
```

Slika 7: Primer zapisa koreferenčnih verig v TEI.

posameznim pojavnicam, z uporabo zapisa IOB pa tudi lastnosti niza pojavnic, kar je uporabljeno za označevanje imenskih entitet. Slika 8 ilustrira, ravno tako na primeru začetka korpusa SentiCoref, zapis oznak v formatu CoNLL-U, pri čemer smo izpustili nekaj za ilustracijo nepomembnih stolpcev.

```
# newdoc id = senticoref1
# title = Zakaj se hrana draži?
# handle = http://hdl.handle.net/11356/1285
# source = SentiCoref 1.0
# url = http://www.24ur.com/novice/gospodarstvo/zakaj-se-hrana-drazi.html
# main_url = www.24ur.com
# keywords = iztok, jarc, michel, barnier, hrana, podražitev
# author = STA / V.L.
# date = 2007-09-03
# newpar id = senticoref1.1
# sent_id = senticoref1.1.1
# text = Evropska komisija mora narediti analizo vzrokov rasti cen hrane , menita kmetijski minister Jarc in njegov francoski kolega .
1       Evropska        evropski        ADJ     Agpfsn  Case=Nom|Degree=Pos|Gender=Fem|Number=Sing_     NER=B-org
2       komisija        komisija        NOUN    Ncfsn   Case=Nom|Gender=Fem|Number=Sing                  NER=I-org
3       mora            morati          VERB    Vmpr3s  Aspect=Imp|Mood=Ind|Number=Sing|Person=3...     NER=O
```

Slika 8: Zapis korpusa v formatu CoNLL-U.

Ker ima SUK dva zapisa skladnje (UD in JOS-SYN), ima vsak skladenjsko označeni podkorpus dve različici CoNLL-U datotek, eno s skladnjo UD in angleškimi oznakami MULTEXT-East, drugo pa s skladnjo JOS in slovenskimi oznakami MULTEXT-East.

5 Dostopnost korpusa

Korpus SUK je dostopen za prevzem na repozitoriju jezikovnih virov raziskovalne infrastrukture CLARIN.SI (Arhar Holdt idr., 2022) pod licenco CC BY-SA 4.0, ki dovoljuje uporabo za poljubne namene, vključno s komercialnimi, vendar pod pogojem, da se prizna avtorstvo korpusa in, če se korpus nadgradi, da je nadgrajeni korpus na voljo pod enakimi pogoji kot izvirnik.

Vnos v repozitoriju vsebuje dve stisnjeni datoteki, in sicer SUK.TEI.zip (20 MB, ki se razširi v 198 MB) s korpusom, kodiranim v TEI, in SUK.CoNLL-U.zip (23 MB, razširjen v 169 MB) s korpusom v formatu CoNLL-U.

Korpus je za pregledovanje in analizo dostopen tudi prek konkordančnikov CLARIN.SI, tj. noSketch Engine (Rychlý, 2007) in KonText (Machálek, 2020), pri čemer so povezave do konkordančnikov na voljo prek vnosa v repozitoriju.

6 Ocena uspešnosti označevanja in novi označevalni modeli

Učna množica SUK je bila v okviru projekta RSDO uporabljena za učenje novih jezikovnih modelov za označevanje besedil. Pri tem smo uporabili označevalno orodje CLASSLA-Stanza,[25] ki je bilo v pretekli različici že naučeno na učnem korpusu ssj500k (Ljubešić in Dobrovoljc, 2019; Terčon in Ljubešic, 2023).

Kot pripravo na učenje modelov smo najprej izvedli delitev korpusa SUK na učno, validacijsko in testno podmnožico v razmerju 8 : 1 : 1.[26] Z uporabo učne in validacijske množice smo naučili modele za štiri ravni slovničnega označevanja: oblikoskladenjsko označevanje, lematizacija, skladenjsko razčlenjevanje (tako po sistemu JOS-SYN kot tudi po sistemu Universal Dependencies) in označevanje udeleženskih vlog. Naučene modele smo nato ovrednotili na testni množici. Rezultati evalvacije modelov so prikazani v Tabeli 2.[27]

Za vsak model so podane vrednosti izražene z oceno F1 v obliki odstotka, pri čemer je za oblikoskladenjsko označevanje prikazana ocena F1 za oznake vseh treh sistemov (MULTEXT-East v6, UD

25 https://pypi.org/project/classla/

26 Izvorna koda za proces delitve in vse nastale podmnožice so dostopne na https://github.com/clarinsi/suk-split.

27 Tabela prikazuje rezultate modelov, ki so pri učenju in evalvaciji za napovedovanje oznak uporabljali tudi novo različico slovenskega oblikoslovnega leksikona Sloleks 3.0 (Čibej idr., 2022). Celoten proces učenja in evalvacije modelov, vključno z uspešnostjo modelov, ki niso uporabljali leksikona, je podrobneje opisan na GitHub repozitoriju https://github.com/clarinsi/classla-training.

besedne vrste in UD lastnosti), za skladenjsko razčlenjevanje pa je prikazan F1 za splošno uveljavljeno oceno LAS (ang. *Labeled Attachment Score*), ki se pogosto uporablja za vrednotenje uspešnosti odvisnostnega označevanja (Nivre in Fang, 2017).

Tabela 2: Uspešnost modelov za vsako označevalno raven.

Označevalna raven	F1
Oblikoskladenjsko označevanje	97,08
Lematizacija	98,97
UD-skladnja	90,57
Skladnja JOS-SYN	93,89
Udeleženske vloge	76,24

Novi označevalni modeli večinoma dosegajo F1 vrednosti nad 90, z izjemo modela za označevanje udeleženskih vlog, za katerega je bilo v učni množici na voljo najmanj učnih podatkov od vseh zgoraj omenjenih označevalnih ravni. Po pregledu uspešnosti modela pri napovedovanju posameznih udeleženskih vlog se je izkazalo, da lahko pričakujemo precej višjo natančnost napovedovanja pri vlogah, ki so bistveno pogostejše (npr. ACT – F1 87,76 in PAT – F1 86,37), medtem ko pri redkejših model dosega manjše vrednosti (npr. ACMP – F1 36,36). Uspešnost in analiza najpogostejših napak modela za skladenjsko razčlenjevanje po sistemu UD sta podrobneje predstavljena v Dobrovoljc idr. (2023), podatki za skladnjo JOS-SYN pa v Arhar Holdt idr. (2023, 28).

Vsi omenjeni jezikovni modeli so že vključeni v najnovejšo različico orodja CLASSLA-Stanza 2.1 kot privzeti jezikovni modeli za označevanje standardne slovenščine.

7 Sklep in nadaljnje delo

V prispevku smo predstavili nadgradnjo učnega korpusa ssj500k v SUK 1.0. Korpus je temelj za učenje jezikoslovnega označevanja sodobne slovenščine, zato ga je nujno kontinuirano izboljševati in razvijati metodologijo njegove priprave. Specifične identificirane težave

in prioritete za vsako posamezno označevalno ravnijo so popisane v Arhar Holdt idr. (2023), tu pa navajamo splošne smernice nadaljnjega razvoja učnega korpusa SUK.

Prva razvojna prioriteta je **izboljševanje kakovosti korpusa in označevalnih smernic za vse vključene jezikovne ravnine.** Pri analizah označenosti korpusa ssj500k in delu z novimi podatki so se razkrile označevalne nedoslednosti, posredno pa tudi šibka mesta označevalnih smernic. Deloma so bile težave odpravljene, mestoma pa zahtevajo dodatno delo in širši strokovni konsenz o pripisovanju določenih jezikoslovnih kategorij na posameznih ravninah. Premišljeno usklajevanje potrebujemo tako znotraj korpusa, kjer se določene težave propagirajo med jezikovnimi ravninami, kakor tudi med korpusom SUK in drugimi jezikovnimi viri (leksikonom Sloleks, drugimi učnimi korpusi ipd.).

Druga prioriteta je **povečevanje korpusa**. Po gradnji korpusa SUK se potrjuje, da sta strojna lematizacija in pripisovanje oblikoskladenjskih oznak že dovolj natančna, da celostni ročni pregledi oznak niso več smiselni. V prihodnje bi se bilo bolje omejiti na ročno preverbo pri težavnih kategorijah, ki jih je mogoče predhodno (pol) avtomatsko identificirati. Za višje ravni je treba stremeti k povečanju deleža ročno označenega gradiva, za vse ravni pa zagotoviti ciljne in celostne analize mest, kjer se strojni označevalnik moti, in v primeru redkosti ali razpršenosti jezikovnih pojavov dodati ustrezno izbrane dodatne povedi ali besedila.

V povezavi s prejšnjo točko je treba izboljšati **žanrsko reprezentativnost korpusa** oz. vključiti domene sodobne standardne slovenščine, ki se glede na raziskave jezikovnih značilnosti pomembneje razlikujejo od splošne rabe, npr. pravna in uradovalna besedila, znanstveni jezik in pisanje učečih se. Kot omenjeno, trenutno poteka razvoj učnih korpusov za označevanje nestandardne slovenščine, govorjenega jezika in starejših besedil ločeno, zato je treba pri gradnji zagotoviti metodološko skladnost, npr. vključenost primerljivih oznak in pregledno, transparentno žanrsko specifično nadgrajevanje smernic. Na drugi strani je zlasti za semantično in diskurzno raven mogoče izbrati in dodati **nove vrste oznak**, saj so poleg koreferenčnosti relevantni tudi podatki za pomensko razdvoumljanje,

detekcijo metafor, ugotavljanje mnenj, detekcijo sovražnega govora in podobno.

Nujno je zagotoviti kontinuiran razvoj **orodij in spletnih servisov za označevanje, analizo in vizualizacijo** jezikoslovno označenih podatkov. Podatke višjih označevalnih ravni (skladnja, SRL) je trenutno mogoče vizualizirati v programu Q-CAT, ki omogoča tudi napredno iskanje; prav tako je mogoče po oznakah iskati v konkordančnikih, ki jih ponuja infrastruktura CLARIN.SI. Vendar sta vizualizacija in izvoz označenih podatkov v obeh orodjih omejena, zato niso enostavno, pregledno dostopni. Poseben izziv so oznake na ravneh, ki presegajo meje povedi, kot so koreferenčne verige. V nadaljevanju je treba razviti možnosti za uporabniku prijazno sočasno pregledovanje in izvažanje bogato označenih podatkov v berljivem formatu za jezikoslovne analize in druge nadaljnje rabe.

Učne množice in označevalne sheme je potrebno usklajevati s standardizacijskimi pobudami v **mednarodnem prostoru** ter sodelovati pri njihovem nastajanju. Mednarodno standardizirane učne množice omogočajo, da se za slovenščino razvijajo orodja v okviru mednarodnih konzorcijev in da je slovenščina del večjezikovnih učnih in evalvacijskih množic.

Glavni namen korpusa je razvoj strojnih označevalnikov, pri čemer je treba zagotoviti tudi njihovo **kakovostno in transparentno vrednotenje.** Na to potrebo odgovarja portal SloBENCH,[28] ki vsebuje evalvacijske množice za različne naloge procesiranja naravnega jezika. Ogrodje omogoča, da vsakdo gradi lastne modele in preko portala odda napovedi, kjer se avtomatsko izvede vrednotenje in objavi v izbrani lestvici. Trenutno je v ogrodju SloBENCH na voljo 8 javnih lestvic. Za naloge prepoznavanja imenskih entitet, odkrivanja koreferenčnosti in razčlenjevanja po sistemu UD so bile že pripravljene evalvacijske množice, ki so skladne z metodologijo priprave korpusa SUK. Za nadaljnji razvoj bi bilo potrebno v avtomatske sisteme vrednotenja vključiti tudi ostale ravni, npr. JOS-SYN in udeleženske vloge. Zaradi zagotavljanja čimbolj transparentnega vrednotenja je potrebno dodatno pripraviti označene korpuse, ki ne bodo javno

28 https://slobench.cjvt.si/

dostopni in bodo namenjeni izključno vrednotenju.

Nenazadnje je mogoče omeniti tudi potrebo po **diseminaciji korpusa SUK** ne le v domači in mednarodni razvojni skupnosti, ampak tudi v drugih strokah, zlasti jezikoslovju. Ročno pregledane oznake takšnega obsega in raznolikosti, kot jih prinaša SUK, so redek in trenutno neizkoriščen potencial za kvantitativne in kvalitativne empirične analize jezikovnih pojavov, s tem pa za posodobitev jezikovnega opisa, ki ga v prostoru nujno potrebujemo in bi imel pozitiven povratni vpliv tudi na jezikoslovno označevanje sodobne slovenščine.

Zahvala

Projekt Razvoj slovenščine v digitalnem okolju sta med leti 2020 in 2023 sofinancirali Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj (operacija se je izvajala v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014–2020). Program Jezikovni viri in tehnologije za slovenski jezik (P6-0411) sofinancira Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije iz državnega proračuna.

Literatura

Arhar Holdt, Š., Bordon, D., Čibej, J., Dobrovoljc, K., Gantar, P., Lenardič, J., Munda, T., Pori, E., Robida, N., Terčon, L., in Žitnik, S. (2023). *Slovenski učni korpus: Množici SUK 1.0 in Janes-Tag 3.0: Poročilo projekta Razvoj slovenščine v digitalnem okolju*.

Arhar Holdt, Š., in Čibej, J. (2021). Analize za nadgradnjo učnega korpusa ssj500k. V Š. Arhar Holdt (Ur.), *Nova slovnica sodobne standardne slovenščine: viri in metode* (1. izd.). Znanstvena založba Filozofske fakultete.

Arhar Holdt, Š., Krek, S., Dobrovoljc, K., Erjavec, T., Gantar, P., Čibej, J., Pori, E., Terčon, L., Munda, T., Žitnik, S., Robida, N., Blagus, N., Može, S., Ledinek, N., Holz, N., Zupan, K., Kuzman, T., Kavčič, T., Škrjanec, I., … Zajc, A. (2022). *Training corpus SUK 1.0*. http://hdl.handle.net/11356/1747

Björkelund, A., Hafdell, L., in Nugues, P. (2009). Multilingual Semantic Role Labeling. V J. Hajič (Ur.), *Proceedings of the Thirteenth Conference*

on Computational Natural Language Learning (CoNLL 2009): Shared Task (str. 43–48). Association for Computational Linguistics. https://aclanthology.org/W09-1206

Brank, J. (2022). *Q-CAT Corpus Annotation Tool 1.4*. http://hdl.handle.net/11356/1684

Čibej, J., Gantar, K., Dobrovoljc, K., Krek, S., Holozan, P., Erjavec, T., Romih, M., Arhar Holdt, Š., Krsnik, L., in Robnik-Šikonja, M. (2022). *Morphological lexicon Sloleks 3.0*. http://hdl.handle.net/11356/1745

Čibej, J., Holdt, Š. A., Erjavec, T., in Fišer, D. (2018). *Ročno označeni korpusi Janes za učenje jezikovnotehnoloških orodij in jezikoslovne raziskave*. https://api.semanticscholar.org/CorpusID:165686166

de Marneffe, M.-C., Manning, C. D., Nivre, J., in Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, *47*(2), 255–308. https://doi.org/10.1162/coli_a_00402

Dobrovoljc, K., Erjavec, T., in Krek, S. (2017). The Universal Dependencies Treebank for Slovenian. V T. Erjavec, J. Piskorski, L. Pivovarova, J. Šnajder, J. Steinberger, in R. Yangarber (Ur.), *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing* (str. 33–38). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1406

Dobrovoljc, K., Krek, S., in Rupnik, J. (2012). Skladenjski razčlenjevalnik za slovenščino. V T. Erjavec in J. Žganec Gros (Ur.), *Zbornik Osme konference jezikovne tehnologije: Zvezek C* (str. 42–47). Institut »Jožef Stefan«.

Dobrovoljc, K., in Ljubešić, N. (2022). Extending the SSJ Universal Dependencies Treebank for Slovenian: Was It Worth It? V S. Pradhan in S. Kuebler (Ur.), *Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) within LREC2022* (str. 15–22). European Language Resources Association. https://aclanthology.org/2022.law-1.3

Dobrovoljc, K., Terčon, L., in Ljubešić, N. (2023). Universal Dependencies za slovenščino: Nove smernice, ročno označeni podatki in razčlenjevalni model. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave*, *11*(1), 218–246. https://doi.org/10.4312/slo2.0.2023.1.218-246

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. (2015). The IMP historical Slovene language resources. *Language Resources and Evaluation*, *49*(3), 753–775. https://doi.org/10.1007/s10579-015-9294-7

Erjavec, T., Fišer, D., Krek, S., in Ledinek, N. (2010). The JOS Linguistically Tagged Corpus of Slovene. V N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, in D. Tapias (Ur.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2010/pdf/139_Paper.pdf

Gantar, P. (2021). Strojno berljiv Vezljivostni leksikon slovenskih glagolov. V Š. Arhar Holdt (Ur.), *Nova slovnica sodobne standardne slovenščine: viri in metode* (1. izd., str. 259–297). Znanstvena založba Filozofske fakultete.

Gantar, P. (2023). Analiza udeleženskih vlog s skladenjskega, pomenskega in leksikalnega vidika. V M. Smolej in M. Schlamberger Brezar (Ur.), *Prispevki k preučevanju slovenske skladnje* (1. izd., str. 77–97). Založba Univerze v Ljubljani. https://doi.org/10.4312/9789612970987

Klie, J.-C., Bugert, M., Boullosa, B., de Castilho, R., in Gurevych, I. (2018). The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. V D. Zhao (Ur.), *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations* (str. 5–9). Association for Computational Linguistics. https://aclanthology.org/C18-2002

Krek, S., Arhar Holdt, Š., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešić, N., Kosem, I., in Dobrovoljc, K. (2020). Gigafida 2.0: the reference corpus of written standard Slovene. V N. Calzolari (Ur.), *LREC 2020 : Twelfth International Conference on Language Resources and Evaluation* (str. 3340–3345). ELRA - European Language Resources Association.

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. (2021). *Training corpus ssj500k 2.3*. http://hdl.handle.net/11356/1434

Krek, S., Erjavec, T., Dobrovoljc, K., Gantar, P., Arhar Holdt, Š., Čibej, J., in Brank, J. (2020). The ssj500k training corpus for Slovene language processing. V D. Fišer in T. Erjavec (Ur.), *Jezikovne tehnologije in digitalna humanistika: zbornik konference* (str. 23–33). Inštitut za novejšo zgodovino.

Krek, S., Gantar, P., Dobrovoljc, K., in Škrjanec, I. (2016). Označevanje udeleženskih vlog v učnem korpusu za slovenščino. V T. Erjavec in D. Fišer (Ur.), *Zbornik konference Jezikovne tehnologije in digitalna humanistika* (str. 106–110). Znanstvena založba Filozofske fakultete.

Ljubešić, N., in Dobrovoljc, K. (2019). What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian. V T. Erjavec, M. Marcińczuk, P. Nakov, J. Piskorski, L. Pivovarova, J. Šnajder, J. Steinberger, in R. Yangarber (Ur.), *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing* (str. 29–34). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3704

Machálek, T. (2020). KonText: Advanced and Flexible Corpus Query Interface. V N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, in S. Piperidis (Ur.), *Proceedings of the Twelfth Language Resources and Evaluation Conference* (str. 7003–7008). European Language Resources Association. https://aclanthology.org/2020.lrec-1.865

Martelli, F., Navigli, R., Krek, S., Tiberius, C., Kallas, J., Gantar, P., Koeva, S., Nimb, S., Pedersen Sandford, B., Olsen, S., Langements, M., Koppel, K., Üksik, T., Dobrovoljc, K., Ureña-Ruiz, R.-J., Sancho-Sánchez, J.-L., Lipp, V., Váradi, T., Györffy, A., … Munda, T. (2021). Designing the ELEXIS Parallel Sense-Annotated Dataset in 10 European Languages. V I. Kosem in M. Cukr (Ur.), *eLex 2021 Proceedings: Proceedings of the eLex 2021 conference* (str. 377–395). Lexical Computing CZ, s.r.o.

Mikulová, M., Bémová, A., Hajič, J., Hajičová, E., Havelka, J., Kolářová, V., Kučová, L., Lopatková, M., Pajas, P., Panevová, J., Razímová, M., Sgall, P., Štěpánek, J., Urešová, Z., Veselá, K., in Žabokrtský, Z. (2006). *Annotation on the tectogrammatical layer in the Prague Dependency Treebank: Annotation manual*.

Nivre, J., in Fang, C.-T. (2017). Universal Dependency Evaluation. V M.-C. de Marneffe, J. Nivre, in S. Schuster (Ur.), *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)* (str. 86–95). Association for Computational Linguistics. https://aclanthology.org/W17-0411

Pori, E., Čibej, J., Munda, T., Terčon, L., in Arhar Holdt, Š. (2022). Lematizacija in oblikoskladenjsko označevanje korpusa SentiCoref. V D. Fišer in T. Erjavec (Ur.), *Jezikovne tehnologije in digitalna humanistika: zbornik konference*. Inštitut za novejšo zgodovino.

Rychlý, P. (2007). Manatee/Bonito-A Modular Corpus Manager. V P. Sojka in A. Horák (Ur.), *Recent Advances in Slavonic Natural Language Processing, RASLAN* (str. 65–70). Masaryk University.

Tercon, L., in Ljubesic, N. (2023). CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages. *ArXiv*, *abs/2308.04255*. https://api.semanticscholar.org/CorpusID:261881905

Toporišič, J. (2004). *Slovenska slovnica*. Obzorja.

Žitnik, S., in Bajec, M. (2018). Odkrivanje koreferenčnosti v slovenskem jeziku na označenih besedilih iz coref149. *Slovenščina 2.0: Empirične, Aplikativne in Interdisciplinarne Raziskave*, *6*(1), 37–67.

Žitnik, S., Blagus, N., in Bajec, M. (2022). Target-level sentiment analysis for news articles. *Knowledge-Based Systems*, *249*, 108939. https://doi.org/https://doi.org/10.1016/j.knosys.2022.108939