REPUBLIKA SLOVENIJA
MINISTRSTVO ZA KULTURO

EVROPSKA UNIJA
EVROPSKI SKLAD ZA REGIONALNI RAZVOJ
NALOŽBA V VAŠO PRIHODNOST

Slovenski učni korpus: Množici SUK 1.0 in Janes-Tag 3.0

Poročilo projekta Razvoj slovenščine v digitalnem okolju

Aktivnost DS1.2

Avtorji: Špela Arhar Holdt[○,□], David Bordon[□], Jaka Čibej[○,□], Kaja Dobrovoljc[○,□], Polona Gantar[□], Jakob Lenardič[◊], Tina Munda[○], Eva Pori[□], Nejc Robida[○,□], Luka Terčon[○], Slavko Žitnik[○]

[○] Fakulteta za računalništvo in informatiko Univerze v Ljubljani
[□] Filozofska fakulteta Univerze v Ljubljani
[◊] Inštitut za novejšo zgodovino

Ljubljana: Center za jezikovne vire in tehnologije, Univerza v Ljubljani, 2023

Operacijo Razvoj slovenščine v digitalnem okolju 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.

Vsebina

Operacijo Razvoj slovenščine v digitalnem okolju 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.

Operacijo Razvoj slovenščine v digitalnem okolju 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.

1 Špela Arhar Holdt: Projektni okvir

Poročilo oz. kazalnik *Slovenski učni korpus: Množici SUK 1.0 in Janes Tag 3.0* je nastalo pod okriljem projekta Razvoj slovenščine v digitalnem okolju, ki sta ga 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). Cilj projekta je bil zadovoljiti potrebe po računalniških izdelkih in storitvah s področja jezikovnih tehnologij za slovenski jezik za raziskovalne organizacije, podjetja in širšo javnost. Spletna stran projekta z dostopom do rezultatov: https://slovenscina.eu/.

Kazalnik se umešča v prvi projektni delovni sklop z naslovom *Jezikovni viri*. Namen delovnega sklopa je bil nadgraditi slovenske besedilne korpuse in leksikon besednih oblik. Prenovili smo učne množice in postopke za strojno označevanje sodobne slovenščine. Rezultat so osveženi in povečani jezikovni viri, ki so na voljo tako uporabniški skupnosti kot za strojno rabo. Z razvitimi postopki in orodji bo posodabljanje slovenskih korpusov v prihodnosti hitrejše in preprostejše.

Med področja, ki smo se jim na projektu posebej posvetili, sodi tudi nadgradnja učnega korpusa za slovenščino, ki naslavlja tako potrebe po strojnem procesiranju standardne kot nestandarne slovenščine: v tem kazalniku opisujemo nadgradnjo množic ssj500k ter Janes Tag. Rezultati, ki so po projektu odprto dostopni za nadaljnjo rabo, predstavljajo osnovo za gradnjo jezikovnih priročnikov in tehnologij za slovenščino. Novi učni množici sta odprto dostopni za nadaljnjo rabo.

2 Špela Arhar Holdt, Tina Munda: Opis projektnih aktivnosti

Cilj aktivnosti 1.2 Učne množice za razumevanje naravnega jezika je bilo nadgraditi oz. razviti napredno označeni učni množici za procesiranje in razumevanje naravnega jezika: učni korpus ssj500k in zbirko nalog SuperGLUE.

V tem kazalniku se osredotočamo na nadgradnjo učnega korpusa ssj500k, ki je bil v času prijave projekta v celoti ročno označen in pregledan na ravni tokenizacije, stavčne segmentacije, oblikoskladenjskih oznak in lem. Približno polovica korpusa je bila označena na ravni odvisnostne

Operacijo Razvoj slovenščine v digitalnem okolju 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.

skladnje po sistemih JOS in Universal Dependencies, imenskih identitet (NE) in glagolskih besednih zvez. Približno četrtina korpusa je vsebovala oznake semantičnih vlog (SRL). Za boljšo uporabnost učnega korpusa je bilo treba zagotoviti njegovo povečanje: v okviru projekta smo dodali nova korpusna besedila, ki omogočajo tudi označevanje prek meja povedi, povečali smo delež označenosti na vseh obstoječih ravneh ter dodali nove ravni označevanja, ki omogočajo razvoj orodij in postopkov na semantični ravni.

Na projektni aktivnosti je sodelovalo več raziskovalcev in zunanjih sodelavcev, kot so študenti. Soavtorstvo je navedeno pri vsakem posameznem projektnem rezultatu.

Delo je potekalo prek celotnega časa trajanja projekta po naslednjih korakih:
Povečanje korpusa ssj500k na nekaj čez 1.000.000 pojavnic. Pri tem smo uporabili gradiva, dostopna na CLARIN.SI: korpusa Coref149 in SentiCoref 1.0, ki sta že označena na ravneh koreferenc in imenskih entitet, ter ELEXIS-WSD, ki je v rabi za učenje pomenskega razdvoumljanja. K temu smo dodali tudi novopripravljeno množico slovenskih tvitov in korpus Ambiga, ki vsebuje redke in dvoumne besedne oblike.

- Pripis oznak v povečanem korpusu in celovit ročni pregled novega gradiva na ravni tokenizacije, stavčne segmentacije, lem in oblikoskladenjskih oznak. Pregled je potekal s pomočjo kuriranega študentskega dela ob preverjanju ujemanja med 3 označevalci.
- Pripis oznak na višjih označevalnih ravneh, ki omogočajo skladenjsko in semantično učenje: imenske entitete, koreference, odvisnostni skladnji JOS-SYN in Universal Dependencies ter udeleženske vloge SRL. Predvsem za slednje ravni je pomembna posodobitev označevalnih smernic, ki so po koncu projekta urejeno zbrane ter objavljene in tako na voljo za nadaljnje nadgradnje. Relacij in večbesednih enot na projektu nismo na novo označevali, so pa posredno zajete v obstoječe označevalne ravni, zlasti obe odvisnostni skladnji.
- Odločili smo se, da novooznačena besedila v skupnem obsegu 512.588 novih besednih pojavic[1] na repozitoriju CLARIN.SI objavimo v dveh ločenih množicah, kar omogoča enostavnejšo izbiro gradiva za učenje standardnih modelov na eni in nestandadnih na drugi strani. Predhodni ssj500k smo preimenovali v SUK (Slovenski učni korpus) in ga povečali z besedili standardne

[1] V poročilu ločujemo med "besednimi pojavnicami" - številom besed oz. oblik, ki se pojavljajo v korpusnih besedilih ter "pojavnicami", kjer pri štetju upoštevamo tudi ločila in druge posebne znake.

Operacijo Razvoj slovenščine v digitalnem okolju 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.

slovenščine. Tvite, ki vsebujejo elemente nestandardnega jezika, smo vključili v množico Janes-Tag.

REZULTATI AKTIVNOSTI

- Korpus SUK 1.0, povečan s 500.295 besednih pojavnic (ssj500k 2.3) na 881.668 besednih pojavnic (SUK 1.0), ki so v različnem deležu označene in ročno pregledane na več označevalnih nivojih: http://hdl.handle.net/11356/1747.
- Korpus Janes-Tag 3.0, povečan z 59.053 besednih pojavnic (Janes-Tag 2.1) na 190.268 besednih pojavnic (Janes Tag 3.0), ki so v različnem deležu označene in ročno pregledane na več označevalnih nivojih: http://hdl.handle.net/11356/1732.
- Prosto dostopne nadgrajene smernice za različne označevalne nivoje: https://wiki.cjvt.si/shelves/jezikoslovno-oznacevanje-korpusov.

3 Špela Arhar Holdt, Tina Munda, Jakob Lenardič: Sestava novega učnega korpusa

3.1 Standardni jezik (od ssj500k 2.3 do SUK 1.0)

V času prijave projekta RSDO je bil na voljo učni korpus ssj500k 2.2 (Krek et al. 2019), ki je za tem z nekaj formalnimi izboljšavami prerasel v različico 2.3 (Krek et al. 2021). Obe različici sta vsebovali enako količino besedil, jezikoslovno označenih na nivojih od tokenizacije, stavčne segmentacije, lematizacije, oblikoslovja in oblikoskladnje do odvisnostne skladnje in udeleženskih vlog, pa tudi na nivojih imenskih entitet in večbesednih leksemov. Vse oznake so bile tudi ročno pregledane, kot je značilno za učne korpuse kot primarne vire za postopke nadzorovanega strojnega učenja. Različica 2.3 je v primerjavi s prejšnjo prinesla popravljena skladenjska razmerja po novejši različici 2.8 sistema Universal Dependencies[2] (v nadaljevanju: UD), nadgradnjo zapisa TEI-formata in dodane UD-oznake v format VERT.

[2] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3687

Operacijo Razvoj slovenščine v digitalnem okolju 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.

V projektu RSDO smo korpus ssj500k 2.3 povečali z besedili treh podmnožic, nato pa nova besedila strojno označili ter ročno pregledali ter popravili na različnih označevalnih nivojih. Pripravili smo tudi podmnožico slovenskih tvitov, ki jo predstavimo v poglavju 3.2. Podmnožice, vključene v SUK, so:

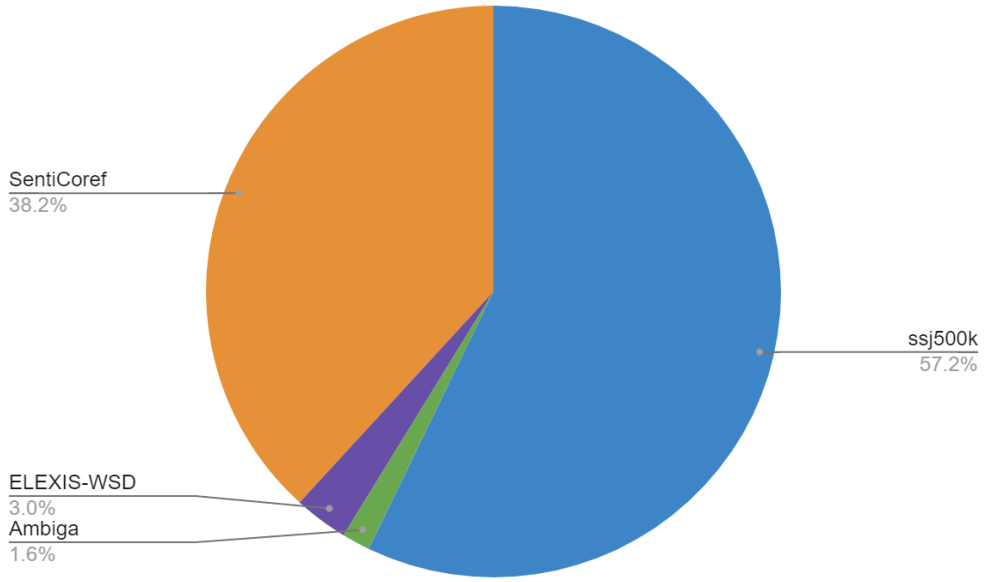
(a) korpus SentiCoref 1.0 (Žitnik 2019): korpus besedil s slovenskih novičarskih portalov, ki bil za namene analize sentimenta opremljen z oznakami imenskih entitet in koreferenc na imenske entitete skupaj s koreferenčnimi verigami, ki označujejo sentiment za vsako entiteto (Klemen in Žitnik 2022); predstavlja približno 89 % povečave;

(b) slovenska različica vzporednega korpusa ELEXIS-WSD (Martelli et al. 2022): slovenski del 10-jezičnega vzporednega korpusa, ki vsebuje 2024 povedi iz člankov na Wikipediji v vseh vključenih jezikovnih različicah. Za namene gradnje množice za nadzorovano učenje strojnega razdvoumljanja večpomenskih leksemov (angl. *word-sense disambiguation*, WSD) je bila množica opremljena z oblikoskladenjskimi oznakami po sistemu UD, vsem polnopomenskim leksemom pa so bili še pripisanimi pomeni; predstavlja 7 % povečave;

(c) množica Ambiga: nabor povedi iz korpusa Gigafida 2.0, ki vsebujejo dvoumne oblike, npr. enakopisne zaimke (*tvojima* ‘Zsdmdoe’ proti ‘Zsdmdde’), dvojinske oblike (*mlajšima* ‘Pppmdo’ proti ‘Pppmdd’ proti ‘Pppzdo’) idr., ki so bili identificirani z analizo pogostih napak pri strojnem označevanju osnovnih nivojev (Arhar Holdt in Čibej 2021, 49-59); predstavlja cca. 4 % povečave novega korpusa.

Novi učni korpus SUK tako sestavljajo množice ssj500k 2.3 (586.187 pojavnic oz. 57,2 %), SentiCoref (391.962 pojavnic oz. 38,2 %), ELEXIS-WSD (31.233 pojavnic oz. 3 %) in Ambiga (16.257 pojavnic oz. 1,6 %), kar predstavlja Graf 1.

Operacijo Razvoj slovenščine v digitalnem okolju 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.



Graf 1: Sestava novega učnega korpusa SUK 1.0

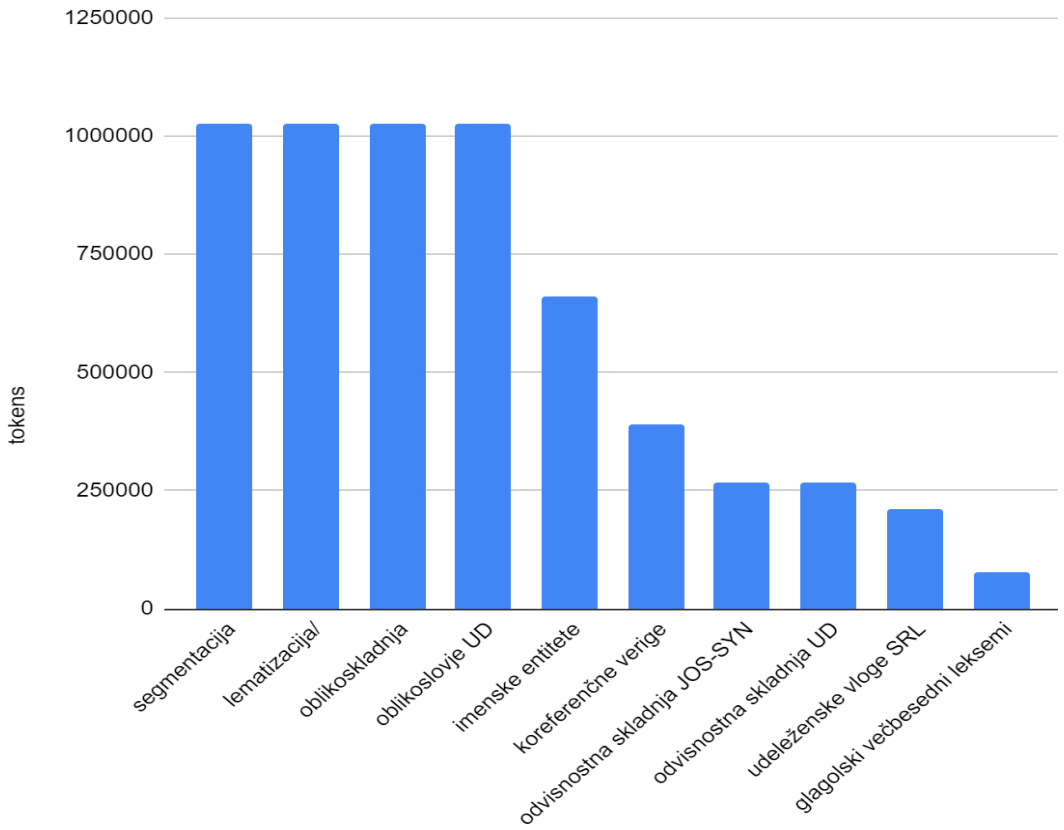
Vseh 1.025.639 pojavnic novega učnega korpusa SUK je označenega in ročno pregledanega na ravni stavčne segmentacije, tokenizacije, lematizacije in oblikoskladenjskih oznak, slednje po sistemih JOS oz. MULTEXT-East V6 (poglavje 4) in UD (poglavje 5). Skoraj dve tretjini (64,3 %) korpusa vsebujeta oznake imenskih entitet (Poglavje 8.1). Dobrih 38 % celotnega korpusa je označenih na ravni koreferenc, ki doslej v učnem korpusu še niso bile zastopane (poglavje 8.2). Približno četrtina korpusa vsebuje oznake odvisnostne skladnje po sistemih JOS (poglavje 6) in UD (poglavje 5), nekaj manj kot četrtina pa je označena na nivoju udeleženskih vlog SRL (poglavje 7). Z oznakami glagolskih večbesednih enot[3] je označenih 7 % gradiva, vse še iz ssj500k. Natančni podatki po označevalnih nivojih so prikazani v Tabeli 1 in Grafu 2.

[3] http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.1/

Operacijo Razvoj slovenščine v digitalnem okolju 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.

Tabela 1: Količina pregledanega gradiva v SUK po označevalnih nivojih

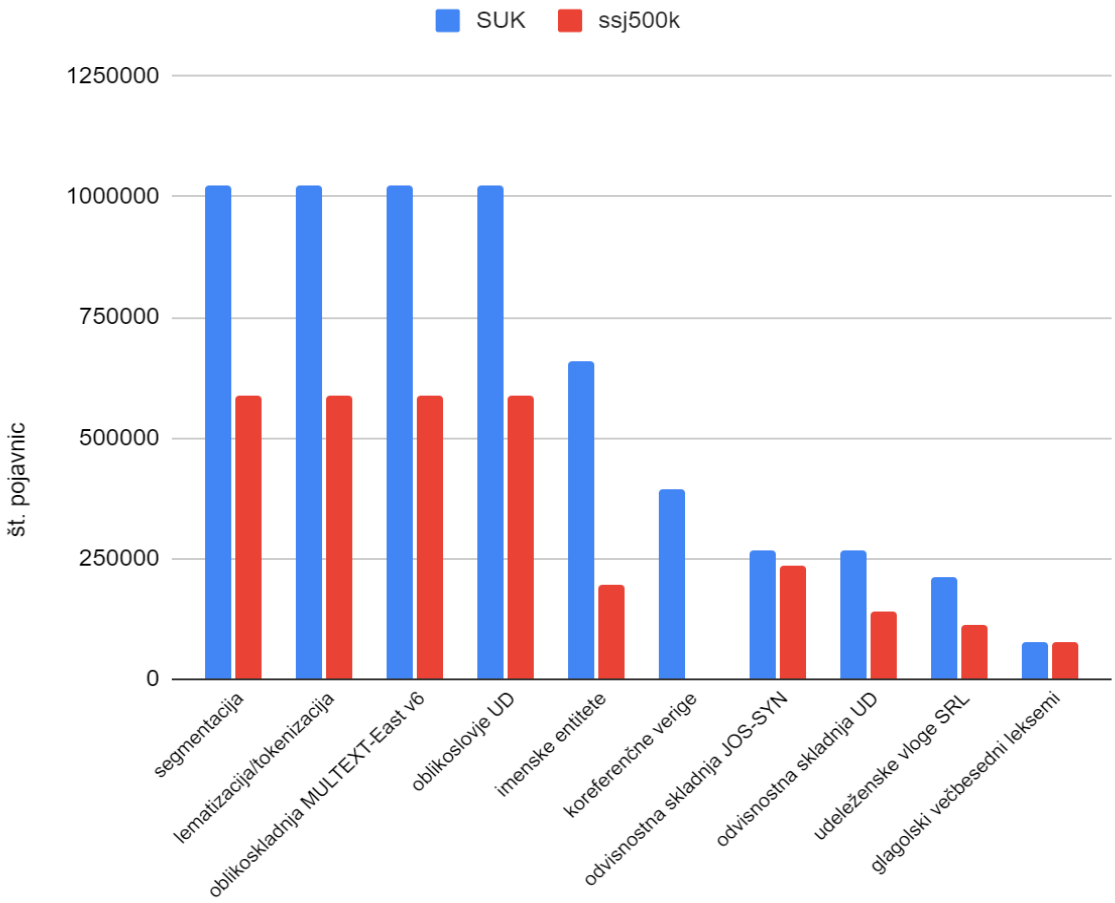
Označevalni nivo	Pojavnice	Povedi	Besedila	% celotnega SUK
Segmentacija	1.025.639	48.594	2.908	100
Lematizacija/tokenizacija	1.025.639	48.594	2.908	100
Oblikoskladnja MULTEXT-East v6	1.025.639	48.594	2.908	100
Oblikoslovje UD	1.025.639	48.594	2.908	100
Imenske entitete	659.059	31.577	1374	64,26
Koreferenčne verige	391.962	18.142	756	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	209.791	10.857	588	20,46
Glagolski večbesedni leksemi	76.126	2.933	586	7,42



Graf 2: Količina oznak po označevalnih nivojih v novem SUK

Operacijo Razvoj slovenščine v digitalnem okolju 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.

Doprinos SUK-a v primerjavi s prejšnjo različico učnega korpusa je predstavljen v Grafu 3. Na osnovnih nivojih (segmentacija, tokenizacija, lematizacija, oblikoskladnja po sistemih JOS in UD) je bilo označeno vse novo gradivo, kar predstavlja 75-odstotno povečanje korpusa ssj500k 2.3. Na nivoju skladnje JOS-SYN je novooznačenega gradiva za 13,24 % več, skladnje UD pa za skoraj 90 % več. Oznake za udeleženske vloge so v primerjavi s prejšnjim korpusom zrasle za dobrih 87 %, medtem ko je količina imenskih entitet bila povečana za več kot dvakrat (za 238,61 %). Koreferenčne verige predstavljajo nov, v učnem korpusu doslej še nezastopan označevalni nivo.



Graf 3: Primerjava označenega gradiva v ssj500k in SUK po označevalnih nivojih

Operacijo Razvoj slovenščine v digitalnem okolju 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.

3.2 Nestandardni jezik (od Janes-Tag 2.1 do Janes-Tag 3.0)

Za nadgradnjo korpusa ssj500k 2.3 v SUK je bila predvidena tudi vključitev korpusa računalniško posredovane komunikacije Janes-RSDO, ki je bil v okviru te aktivnosti tudi ročno pregledan na ravneh lematizacije in oblikoskladnje. Ker pa vsebuje nestandardna besedila (tvite), medtem ko vse ostale podmnožice, ki predstavljajo povečavo učnega korpusa v SUK, vsebujejo standardna besedila, je bil na koncu smiselneje vključen v nadgradnjo korpusa Janes-Tag iz različice 2.1[4] (Erjavec et al. 2019) v 3.0[5] (Lenardič et al. 2022).

Janes-RSDO 1.0 vsebuje 11.960 tvitov oz. 147.020 pojavnic. Normaliziran ter strojno označen in ročno pregledan na ravni tokenizacije in stavčne segmentacije je bil v okviru projektne aktivnosti 1.4 – Janes-Norm 3.0, medtem ko so leme in oblikoskladenjske oznake tam le strojno pripisane. Ročni pregled slednjih dveh ravni je potekal v okviru te projektne aktivnosti in v nadaljevanju predstavljamo njegov potek.

V tej aktivnosti je bilo 11.960 tvitov oz. 147.020 pojavnic, ki sestavljajo Janes-RSDO, ročno pregledanih na ravni lematizacije in oblikoskladnje po sistemu MULTEXT-East/JOS (Erjavec, 2012; Holozan et al. 2008). Kampanja je potekala v okolju *Google Preglednice* (angl. *Google Sheets*), kjer je bilo gradivo razdeljeno na 64 datotek. Pregled je opravljalo 6 študentov jezikoslovnih smeri, pri čemer sta eno datoteko neodvisno drug od drugega pregledala dva. Pregled je bil zastavljen v treh fazah. Najprej je potekalo ročno popravljanje MSD-oznak in lem, zaključni pregled slednjih odločitev pa je potekal dvostopenjsko: najprej sta (ne)ujemanje odločitev dveh pregledovalcev v vsaki datoteki pregledala dva študenta – kuratorja (iz vrst pregledovalcev), njune odločitve pa še vodja kampanje. Pri nalogi so bile uporabljene smernice za označevanje tokenizacije, lematizacije in oblikoskladnje, in sicer v okviru aktivnosti 1.2 nadgrajena verzija 2.0[6,7,8]. V nadaljevanju navajamo nekaj označevalnih dilem.

Povedkovnik: Z izjemo razdelka 5.6.3. smernice ne podajajo nobenih pravil za označevanje povedkovnika, ki v naboru oznak ni prepoznan kot svoja besedna vrsta. Problematičen je zlasti t. i.

[4] http://hdl.handle.net/11356/1238
[5] http://hdl.handle.net/11356/1732
[6] https://wiki.cjvt.si/books/tokenizacija/page/oznacevalne-smernice
[7] https://wiki.cjvt.si/books/lematizacija/page/oznacevalne-smernice
[8] https://wiki.cjvt.si/books/oblikoskladnja-multext-east/page/oznacevalne-smernice

Operacijo Razvoj slovenščine v digitalnem okolju 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.

»prislovni povedkovnik«, npr. *všeč* v *Meni je to všeč*. V ssj500k 2.3 je konsistentno označen kot prislov, tako da smo tovrstne zglede tudi mi označevali tako. Ampak to ni samoumevna razčlemba, saj v nasprotju s prototipskimi prislovi všeč nadzoruje skladenjske lastnosti drugih sestavnikov (konkretno se veže z dajalniškim in tožilniškim argumentom), to je pa kvečjemu lastnost pridevnikov v položaju povedkovega določila, npr. *O tem ni mogoče sklepati.* (gl. razdelek 5.5.2. v smernicah).

***Kaj* v položaju neujemalnega kvantifikatorja:** Gre za primere tipa *imaš kaj nalepk za menjati*; v ssj500k 2.3 so tovrstni zgledi označeni nekonsistentno ali kot prislovi ali kot zaimki; po analogiji s kvantifikacijskim *veliko*, ki je v ssj500k 2.3 vedno označen kot prislov, smo tovrstne zglede označili kot prislove, vendar to zopet ni samoumevno zaradi skladenjskega soizbora (rodilnik na dopolnilu).

***Kaj* v nestandardni členkovni rabi:** Po analogiji z nasprotjem v razdelku 6.3.2. v smernicah smo zglede tipa *kaj uro še zmerom prestavljamo jeseni ali smo končali s to rečjo?* obravnavali kot prislove. Razdelek 6.3.2. je bil pregledovalcem najbolj nerazumljiv – koristno bi bilo eksplicitno razložiti, zakaj je npr. *kaj* v *je neumno plačevati za kaj takšnega* obravnavan kot zaimek, v *Kako je kaj tam spodaj?* pa kot prislov. **Nestadardni *jap*:** Glede na smernice je nabor členkov fiksen, vendar smo sledili označevalni praksi v Janes-Tag, kjer je bil *jap* kot sopomenka za *ja/da* vedno označen kot členek.

4 Špela Arhar Holdt, Eva Pori, Tina Munda, Jaka Čibej: Segmentacija, tokenizacija, lematizacija, MSD

4.1 Obseg označevanja

Na projektu RSDO so bile na osnovnih označevalnih nivojih ročno pregledane pojavnice vseh vključenih podkorpusov, kar pomeni 586.472 pojavnic, od tega 512.588 besednih pojavnic.

Gradivo je bilo najprej tokenizirano, lematizirano in segmentirano na povedi z orodjem CLASSLA-Stanza[9], in sicer posamezno po množicah, razen če so množice te podatke že vsebovale. V primeru množice SentiCoref (za natančnejši popis metodologije gl. Pori et al. 2022), ki je največja med množicami za nadgradnjo, je omenjene tri ravni pregledalo devet študentov. Pri ostalih množicah je

[9] https://github.com/clarinsi/classla

Operacijo Razvoj slovenščine v digitalnem okolju 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.

pregled bodisi že bil opravljen v okviru drugih projektov (ELEXIS-WSD) bodisi je zaradi omejene velikosti potekal v nadaljevanju (Ambiga). Naslednji korak je bilo strojno pripisovanje oblikoskladenjskih oznak MULTEXT-East v6 (v nadaljevanju: MSD), in to z istim orodjem kot prejšnje ravni.

Sledil je ročni pregled MSD-oznak v vseh treh množicah za nadgradnjo učnega korpusa. Opravljalo ga je 24 študentov jezikoslovnih smeri 15 tednov. Študenti so bili razdeljeni v tri skupine po osem. Vsak izmed njih je bil zadolžen za pregled ene ali več besednih vrst v vseh datotekah. Vsakemu študentu smo določili, kaj bo pregledoval, na osnovi predhodno ugotovljenih preferenc. Besedne vrste, ki so lažje in torej hitrejše za pregled (njihova oblikoskladenjska oznaka vsebuje le malo opisnih kategorij) ali je njihova povprečna pojavnost na splošno nizka, smo smiselno združili in njihov pregled je prevzel po en pregledovalec iz vsake skupine (recimo številke, okrajšave, medmete in 'neuvrščeno' je pregledoval po en študent). Po drugi strani pa sta samostalnik zaradi pogoste pojavnosti pregledovala po dva študenta v vsaki skupini. Pred začetkom pregledovanja so bili študentom v obliki videoposnetka predstavljeni postopek označevanja in označevalne smernice (gl. razdelek 4.3), s katerimi so se – v delu, ki je relevanten za besedno vrsto, ki jo pregledujejo – v uvajalnem tednu dodobra spoznali.

Pregledovanje gradiva je potekalo v spletnem okolju *Google Preglednice* (angl. *Google Sheets*). Osnovna oblika preglednice za označevanje je bila (po stolpcih): pojavnica (besedna oblika), lema (osnovna oblika), strojno pripisana MSD-oznaka (ta celica je imela spustni seznam z vsemi možnimi MSD-ji, med katerimi so lahko pregledovalci tudi iskali oz. izbirali v primeru, da so morali oznako popraviti, hkrati pa so prednastavljene dovoljene vrednosti celic preprečevale zatipke), razvezava strojno pripisane oznake (besedni opis posameznih kategorij oznake) in komentar pregledovalca. Vsaka skupina študentov je pregledovala iste datoteke. S trojnim (ne)ujemanjem smo učinkovito odkrili pojavnice, katerih MSD-oznaka je težje določljiva oz. ni enoznačna. Takšne primere so v drugi fazi, t. i. fazi kuracije, znova pregledali t. i. kuratorji (iz vrst pregledovalcev). V tem koraku je bilo veliko komunikacije med kuratorji in projektno ekipo, pa tudi znotraj projektne ekipe, saj je bilo treba nekatere izpostavljene dileme premisliti in prediskutirati (gl. razdelek 4.2). S takšno organizacijo označevalne kampanje, ki je, ob označevalni kampanji računalniško posredovane komunikacije Janes (Čibej et al. 2018), ena najobsežnejših tovrstnih kampanj v našem prostoru, smo stremeli k najvišji natančnosti pripisanih oznak.

Operacijo Razvoj slovenščine v digitalnem okolju 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.

4.2 Opis izzivov in rešitev

V označevalni kampanji smo identificirali nekaj skupin težjih in mejnih primerov (tako pri lemah kot MSD-oznakah), ki so bile v obstoječih označevalnih smernicah (Holozan et al. 2008) slabše zastopane ali pa sploh niso bile, ali pa je pri pregledovanju pogosteje prišlo do neupoštevanja smernic in s tem nedoslednosti. Te primere smo s pomočjo strojnih metod ob koncu kuracije uskladili, težje primere pa posebej analizirali (tudi v oziru do označevalne prakse v ssj500k, kjer je bilo to mogoče) in v projektni skupini tudi prediskutirali. Prevladujoče izzive skupaj z rešitvami, ki so bile po kampanji vpeljane v nadgrajene smernice, predstavljamo v nadaljevanju.

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 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).

Operacijo Razvoj slovenščine v digitalnem okolju 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.

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) dele 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*). Po posvetu s širšo projektno ekipo smo se odločili, da bomo tako v primerih tipa (a) kot (b) upoštevali prekrivnost s slovenkim 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 morfološko 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). Opisane rešitve smo skupaj s primeri dodali v obstoječe smernice.

Ločevanje pridevnikov od prislovov: Velik pomislek je bil, 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.]*). Obstoječe smernice tega niso naslavljale, kar se je odražalo tudi v korpusu ssj500k, kjer tovrstni primeri niso enotno označeni. Po pregledu in analizi pojavitev tovrstnih primerov v največji podmnožici SentiCoref smo oblikovali pravilo, da besedi v obeh skladenjskih vlogah, tj. v (a) in (b), 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.*).

Operacijo Razvoj slovenščine v digitalnem okolju 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.

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 rabi najdemo potrditev, da se lahko pregiba kot samostalnik, in kot pridevnik, če v referenčnem korpusu ni primerov, kjer se pregiba kot samostalnik (npr. *[… so se do konca leta povprečne] neto* (lema: neto, MSD: Ppnmein) *[plače realno povečale …]*).

Predložne prislovne zveze: Podobno kot pri obravnavanju besedne oblike, ki je v vlogi povedkovega določila in v kombinaciji z nedoločnikom enaka prislovu in pridevniku, je bila težava pri razlikovanju med tema besednima vrstama tudi 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 načelo, da nepredložnemu delu v predložnih prislovnih zvezah pripišemo pridevniško lemo in MSD-oznako (*[na] novo* (lema: nov, MSD: Ppnset)).

V povezavi z orisanimi dilemami pri pregledovanju lematizacije in oblikoskladnje izpostavljamo še najpogostejše popravke. Analiza popravkov po koncu označevalne kampanje kaže, da se delež uveljavljenih 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 popravljenih 5.588 lem oz. približno 1,3 % vseh pojavnic v korpusu, kar se ujema s približno 98-odstotno natančnostjo strojne lematizacije. Na ravni oblikoskladenjskih oznak je bilo uvedenih 12.586 popravkov oz. 2,9 % vseh oznak v korpusu, kar odseva skoraj 97-odstotno natančnost strojnega oblikoskladenjskega označevanja. Popravki lem so najpogosteje zadevali lastnoimenskime samostalnike, ki so prekrivni z občnoimenskimi (npr. *Luka* [*Koper*] > lema: luka), okrajšave, sestavljene iz ene ali dveh črk (npr. *Dr.* > lema: dr.), in še prekrivne oblike v oblikoskladenjskih paradigmah različnih leksemov (npr. delo in del). Pri popravkih oblikoskladenjskih oznak je šlo večinoma za ločevanje med občnimi in lastnoimenskimi samostalniki (tip *Leasing* – *leasing*; 1.538 popravkov oz. 12 %; v obratni smeri, tj. iz občnoimenskega v lastnoimensko je bilo popravkov manj: 235 oz. 1,8 %), med moškim in ženskim spolom (825 popravkov oz. 6,6 %; npr. 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

Operacijo Razvoj slovenščine v digitalnem okolju 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.

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 %). Nedoslednosti v celotnem korpusu SUK smo analizirali in odpravili s serijo (pol)avtomatskih pregledov. (Pori et al. 2022)

4.3 Smernice in priporočila za nadaljnje delo

V okviru projekta RSDO smo nadgradili smernice za oblikoskladenjsko označevanje[10] JOS (Holozan et al. 2008), ki vsebujejo načela za ročni pregled lematizacije in oblikoskladenjskih oznak ter so bile v projektu RSDO prvič po njihovi pilotni izvedbi ponovno uporabljene. Obstoječe smernice smo postavili v okolje *Google Dokumenti* (angl. *Google Docs*), kjer smo jih glede na sprotne težave in rešitve nanje dopolnjevali. Njihov konceptualni okvir ostaja isti – morebitni korenitejši poseg v označevalni sistem bi se, glede na izsledke v pričujočem poročilu, dotikal označevanja lastnih imen (gl. razdelek 4.2) in tujega besedišča.[11]

Postavi se konceptualno vprašanje, ali je na ravni oblikoskladnje smiselno označevanje kategorije občno- in lastnoimenskosti, ki inherentno vsebuje tudi dileme, vezane na lematizacijio (iskanje prekrivnosti (delov) lastnoimenskih entitet z občnoimenskimi; gl. razdelek 4.2). Prvi pomislek se pojavi ob dejstvu, da trenutno ta kategorija v MSD-oznaki obstaja samo za samostalnike, ne pa tudi za pridevnike, čeprav je pripisljiva tudi njim (izlastnoimenski pridevniki, recimo *Aškerčeva [cesta]*, *Delova [dopisnica]*, *Magov [novinar]*). Obenem v stvarnih lastnih imenih prihaja do razhajanja pri iskanju prekrivnosti z občnoimenskimi entitetami; pri samostalnikih (vendar ne tistih, ki so del osebnega/zemljepisnega lastnega imena) jo upoštevamo, pri pridevnikih tudi, le po nekoliko drugačni logiki, medtem ko prekrivnosti pri ostalih besednih vrstah ne iščemo. Različna obravnava osebnih in zemljepisnih lastnih imen na eni strani in stvarnih na drugi, pri čemer pri prvih ne iščemo prekrivnosti, pri drugih pa jo, in tudi izlastnoimenskih svojilnih pridevnikov kaže na nedobrodošel prodor semantičnega kriterija na raven oblikoskladnje.

[10] https://wiki.cjvt.si/books/oblikoskladnja-multext-east/page/oznacevalne-smernice

[11] V dosedanjih smernicah sta bili ti dve kategoriji slabše zastopani, predvsem zato, ker besedilne vrste, ki sestavljajo korpus ssj500k, ne vsebujejo toliko lastnih in tujih imen kot besedila, zastopana v množicah za nadgradnjo v SUK; sploh SentiCoref, največja podmnožica nadgradnje, vsebuje odstopajoče veliko število raznovrstnih lastnih imen, saj je bil s tem namenom tudi zgrajen.

Operacijo Razvoj slovenščine v digitalnem okolju 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.

Ločena težava je zapis leme v primerih, ko avtor besedila ne upošteva pravopisnih pravil. Sistem namreč takih situacij ne predvideva, kar pomeni, da se na raven oblikoskladenjskega označevanja prenašajo vprašanja pravopisa – ali kar pravopisov, če štejemo zraven tujejezične elemente, na katere se težave neizogibno transponirajo. Ker je bilo največ popravkov opravljenih ravno pri kategoriji občno-/lastnoimenskosti, hkrati pa v zadnjih letih lastna imena uspešno odkrivamo s prepoznavanjem imenskih entitet (NER), bi bilo smiselno znova premisliti vrednost te kategorije na ravni oblikoskladnje. Če se vendarle izkaže za koristno, bi nekatere dileme lahko bile odpravljene recimo z opuščanjem iskanja prekrivnosti lastnih imen z občnoimenskimi, kar bi predstavljalo korenitejši poseg v označevalne smernice.

Slednje bi se lahko uveljavilo tudi pri tujem občnoimenskem besedišču (npr. *fair play*, *club*, *missing trader*), ki ga trenutno – ob vrzelih v jezikovnih virih z jezikovnim opisom in predpisom – med slovensko besedišče uvrščamo precej neenotno. Drugi potencialno smiseln scenarij za tujejezične občnoimenske lekseme, ki v slovenščino pronicajo po nepredvidljivih vzorcih, bi bila opredelitev smiselnih kategorij, če se izkaže za smotrno, ločeno po jezikih.

Še ena težavna skupina so bile enakopisne oblike, ki zavzemajo drugačna skladenjska mesta in s tem drugačne MSD-oznake. Kljub temu je pri odločanju za ustrezno oznako oblikoskladenjski vidik mestoma izkazal manko in se je tudi tu med kriterije za presojanje vrinila semantična komponenta (npr. pri *torej* – priredni veznik proti prislov). V skupino težav, vezanih na enakopisnice, spada tudi ločevanje med enakopisnimi pridevniki in prislovi v določenih skladenjskih strukturah. Izkazalo se je, da tovrstni primeri v korpusu ssj500k niso povsem dosledno označeni in bi jih bilo v prihodnosti smiselno popraviti.

Za nadaljnje delo je treba upoštevati, da bi bilo obravnavo in reševanje naštetih težav treba zasnovati celovito z ozirom na vse označevalne ravni, ne le na lematizacijsko in oblikoskladenjsko.

Poglavitna ugotovitev te označevalne kampanje je, da sta strojna lematizacija in strojno pripisovanje oblikoskladenjskih oznak že tako natančna, da bi v prihodnje smiselno celostne preglede omejiti na delne (več o podatkih o izboljšani natančnosti strojnega označevanja v razdelku 9). Delni ročni pregledi bi se osredotočali le na težavne kategorije, popisane v pričujočem poročilu, za to pa bi bilo treba razviti postopke za (pol)avtomatsko identifikacijo težavnih kategorij.

Operacijo Razvoj slovenščine v digitalnem okolju 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.

5 Kaja Dobrovoljc: UD oblikoslovje in skladnja

Kot protiutež razdrobljenosti označevalnih shem, ki so dotlej nastajale ločeno za posamezne jezike, slovnične teorije ali celo korpuse, je bila leta 2013 na podlagi predhodnih standardizacijskih iniciativ vzpostavljena označevalna shema Universal Dependencies (UD), ki si prizadeva za mednarodno oz. medjezično usklajeno slovnično označevanje besedil na oblikoslovni in skladenjski ravni, da bi pospešila razvoj večjezičnih jezikovnih tehnologij, medjezičnega strojnega učenja in kontrastivnih jezikoslovnih analiz (de Marneffe et al. 2021). Znotraj sheme UD, ki temelji na načelih odvisnostne slovnice, je bil tako vzpostavljen univerzalni nabor kategorij in smernic (17 besednih vrst, 24 oblikoskladenjskih lastnosti, 37 odvisnostnih skladenjskih relacij), ki odslej omogoča enotno označevanje podobnih slovničnih pojavov v različnih svetovnih jezikih, obenem pa dovoljuje tudi jezikovnospecifične izpeljave, če je to potrebno.

Pobuda je v raziskovalni skupnosti naletela na dober odziv, saj je bilo doslej z označevalno shemo UD ročno označenih že več kot 240 korpusov (t. i. odvisnostnih drevesnic, angl. dependency treebanks) v več kot 130 jezikih s celega sveta, med katerimi sta tudi univerzalni odvisnostni drevesnici za pisno oz. govorjeno slovenščino (Dobrovoljc et al. 2017; Dobrovoljc in Nivre 2016). Glede na pomen razvoja slovenskih virov v tovrstnih mednarodnih standardizacijskih pobudah smo v okviru nacionalnega projekta Razvoj slovenščine v digitalnem okolju (RSDO) obstoječe vire in povezano infrastrukturo za označevanje slovenskih besedil po sistemu Universal Dependencies bistveno nadgradili z:

- prenovo in izčrpno dokumentacijo označevalnih smernic sheme UD za slovenščino
- povečanjem količine ročno označenih pisnih besedil po shemi UD za slovenščino, tj.
    - več kot 430.000 novimi ročno oblikoslovno označenimi pojavnicami
    - več kot 5.000 novimi ročno skladenjsko razčlenjenimi povedmi
- razvojem temeljne infrastrukture za izdelavo in analizo drevesnic UD za slovenščino

V nadaljevanju potek in rezultate teh aktivnosti strnjeno povzamemo, podrobneje pa so bili predstavljeni tudi v več znanstvenih prispevkih (Dobrovoljc in Ljubešić 2022; Dobrovoljc et al. 2022).

Operacijo Razvoj slovenščine v digitalnem okolju 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.

5.1 Prenova in dokumentacija označevalnih smernic

Prvi korak znotraj projekta RSDO je bil tako namenjen izčrpnemu popisu smernic UD za slovenščino na vseh treh ravneh označevanja (besedne vrste, oblikoskladenjske lastnosti in skladenjske relacije) v obliki samostojnega priročnika s smernicami za označevanje po sistemu UD, ki na slovenskih primerih razlaga in ponazarja uporabo posameznih oznak UD za označevanje besedil v slovenščini. Pri tem smo poleg opisa prvotnih smernic uvedli tudi nekaj manjših izboljšav na mestih, kjer je bila prvotna označenost korpusa SSJ-UD nedosledna ali neustrezna glede na splošne, jezikovno univerzalne smernice.

Priročnik je razdeljen na štiri krovna poglavja, v katerih so opisane smernice za pripisovanje oznak za besedne vrste (npr. NOUN, ADJ, VERB …), smernice za pripisovanje oznak za oblikoskladenjske lastnosti (npr. Gender, Number, Case …) in smernice za pripisovanje odvisnostnih relacij (npr. aux, nsubj, obj, advmod …), dodatno pa še smernice za označevanje kompleksnejših skladenjskih struktur, ki podrobneje pojasnjujejo posebne ali mejne primere znotraj posamičnih relacij ter strukture, ki so se kot težavnejše pokazale med samim označevanjem (npr. elipse, primerjave, poudarjalni členki …).

V nekoliko strnjeni oz. tuji javnosti prilagojeni obliki so bile nato smernice prevedene še v angleščino in objavljene na uradni spletni strani projekta UD[12], kar omogoča njihovo neposredno primerjavo s smernicami za več deset drugih svetovnih jezikov, zbranimi na tem spletnem mestu. Ker je spletna stran zasnovana kot kolaborativno spletišče, lahko k izboljševanju smernic prispeva kdorkoli, bodisi z neposrednim urejanjem zalednih besedil na platformi GitHub bodisi z odpiranjem zahtevkov s komentarji in predlogi za izboljšave (GitHub issues).

V procesu dokumentacije smernic za slovenščino so bila identificirana tudi nekatera odprta vprašanja, pri katerih splošne smernice UD odstopajo od doslej uveljavljenih označevalnih praks v slovenskem prostoru, zlasti označevanja po sistemu JOS (gl. 4. in 6. poglavje), in bi jih bilo zato pred implementacijo smiselno najprej nasloviti s širšo strokovno diskusijo. Nekaj več kot trideset tovrstnih vprašanj, ki segajo na vse ravni slovničnega opisa, od tokenizacije (npr. smiselnost razvezovanja naveznih zaimkov tipa *nanj* → *na njega*) do besednovrstne kategorizacije (npr. smiselnost premika členkov med prislove) in

[12] https://universaldependencies.org/

Operacijo Razvoj slovenščine v digitalnem okolju 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.

skladenjske analize (npr. smiselnost ločevanje trpniških struktur od povedkovodoločilnih), smo popisali v ločeni prilogi h krovnim smernicam, pri čemer so bila v sodelovanju z Univerzo v Novi Gorici za približno tretjino izbranih vprašanj že oblikovana nekatera izhodiščna priporočila za nadaljnje izboljšave.

5.2 Povečanje ročno označenih podatkov

Na podlagi izčrpno dokumentiranih smernic smo v drugi fazi projekta povečali obseg ročno označenih podatkov po shemi UD tako na ravni oblikoslovnih oznak kot na ravni skladenjskih razčlemb.

5.2.1 Oblikoslovno označevanje

Ker sta si označevalna sistema JOS in UD na ravni pripisovanja besednih vrst in oblikoskladenjskih 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 lastnostisistema UD,[13] s katerimi je bil v celoti pretvorjen tudi učni korpus ssj500k v2.2. 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 (gl. 4. poglavje). 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).

5.2.2 Skladenjsko razčlenjevanje

Poleg zgoraj opisanega označevanja celotnega korpusa SUK na oblikoslovni ravni smo v okviru projekta RSDO prvotni univerzalno skladenjsko razčlenjeni podkorpus ssj500k v obsegu 8.000 povedi dodatno povečali še za 5.435 novih ročno razčlenjenih povedi.

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. Označevalci so poleg pripisovanja novih oz. manjkajočih povezav (22.377 oz. 23,5 %

[13] https://github.com/clarinsi/jos2ud

Operacijo Razvoj slovenščine v digitalnem okolju 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.

vseh pojavnic) preverjali tudi ustreznost že obstoječih, pri čemer je bilo popravljenih 4.623 oz. 4,8 % vseh pojavnic. Vsako poved sta pregledala po dva neodvisna označevalca, morebitna nestrinjanja pa je razrešil tretji, ekspertni označevalec.

V drugi fazi širitve je bil skladenjsko razčlenjen še podkorpus ELEXIS-WSD-SL, ki vsebuje 2.024 povedi oz. 31.237 pojavnic. Korpus je bil izhodiščno skladenjsko razčlenjen z orodjem CLASSLA-Stanza, pravilnost strojno pripisanih razčlemb pa so nato pregledali trije označevalci in končni kurator, s čimer je bilo ročno popravljenih 1.534 (4,91 %) skladenjskih relacij.

V sklepni fazi smo glede na nekoliko spremenjene izhodiščne smernice in druge identificirane nedoslednosti izboljšali tudi označenost izhodiščne različice korpusa SSJ-UD, tj. univerzalno skladenjsko razčlenjenega podkorpusa ssj500k v2.3. Za vsako izmed približno 30 identificiranih tipov napak oz. nedoslednosti smo s hevrističnimi poizvedbami ustvarili podkorpuse povedi s potencialno problematičnimi oznakami, ki so jih nato označevalci ročno pregledali in popravili v skladu z najnovejšimi smernicami. Na ta način je bilo v izhodiščnem korpusu popravljenih 1.670 skladenjskih oznak (1,2 % celotnega korpusa).

5.2.3 Objava podatkov

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 (Zeman et al. 2022), kot nova, razširjena in izboljšana, različica drevesnice SSJ-UD. Nova različica SSJ-UD v primerjavi s prvotno vsebuje 5.435 novih razčlenjenih povedi (+67,9 %) oz. skoraj enkrat večje število pojavnic (126.427, +89,9 %), s čimer se korpus SSJ-UD po številu pojavnic danes umešča v zgornjo osmino vseh UD drevesnic po svetu. Z razširitvijo je korpus SSJ-UD postal tudi bolj raznolik, saj se vsi trije podkorpusi (izvorne povedi iz ssj500k, nove povedi iz ssj500k, povedi iz ELEXIS-WSD) med seboj razlikujejo tako z vidika vrste vsebovanih besedil kot njihove skladenjske kompleksnosti.

Operacijo Razvoj slovenščine v digitalnem okolju 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.

5.3 Razvoj povezane infrastrukture

Tekom projekta RSDO je v sodelovanju z raziskovalnim programom Jezikovni viri in tehnologije za slovenski jezik, Centrom za jezikovne vire in tehnologije UL in raziskovalno infrastrukturo CLARIN.SI potekalo tudi več drugih aktivnosti, povezanih z izdelavo, analizo in izrabo univerzalno slovnično označenih korpusov za slovenščino, kot so:

- razvoj komplementarnih jezikovnih virov, označenih po shemi UD:
  - polavtomatska pretvorba korpusa Janes-Tag 3.0[14] v oblikoslovne oznake UD
  - avtomatska pretvorba oblikoslovnega leksikona Sloleks 3.0[15] v oblikoslovne oznake UD
  - izdelava skrite drevesnice UD za evalvacijski portal SloBENCH[16] v obsegu 1.332 povedi
- razvoj orodij za ročno označevanje slovenskih besedil po shemi UD:
  - prilagoditev orodja Q-CAT[17] za oznake UD in format CoNLL-U
  - prilagoditev instalacije orodja WebAnno[18] CLARIN.SI za kuriranje skladenjskih oznak
- razvoj orodij za strojno razčlenjevanje slovenskih besedil po shemi UD:
  - izdelava novih modelov označevalnika CLASSLA-Stanza[19]
  - izdelava spletnega vmesnika CJVT Označevalnik[20]
- razvoj orodij za analizo in vizualizacijo razčlenjenih besedil po shemi UD:
  - izdelava spletnega vmesnika za brskanje po skladenjskih drevesnicah Drevesnik[21]
  - izdelava orodja za statistično analizo skladenjskih drevesnic STARK[22].

5.4 Smernice in priporočila za nadaljnje delo

Predstavljeni rezultati predstavljajo pomemben doprinos k nadaljnjemu razvoju jezikovnih tehnologij za slovenščino tako v slovenskem kot mednarodnem prostoru, saj je glede na odprti dostop in

[14] http://hdl.handle.net/11356/1732
[15] http://hdl.handle.net/11356/1745
[16] https://slobench.cjvt.si/
[17] http://hdl.handle.net/11356/1684
[18] https://www.clarin.si/webanno/login.html
[19] https://pypi.org/project/classla/
[20] https://orodja.cjvt.si/oznacevalnik/slv/
[21] https://orodja.cjvt.si/drevesnik/
[22] https://gitea.cjvt.si/lkrsnik/STARK

Operacijo Razvoj slovenščine v digitalnem okolju 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.

standardizirano distribucijo drevesnic UD mogoče pričakovati, da bodo novi učni podatki za slovenščino v prihodnosti integrirani v številna razčlenjevalna orodja oz. na njih temelječe aplikacije. Vendarle pa je glede na relevantnost sheme UD v mednarodnem prostoru in potrebe po kvalitetno slovnično označenih korpusih nasploh nujno zagotoviti kontinuiran razvoj tega področja tudi v prihodnje. Prioritetne smernice za nadaljnji razvoj med drugim vključujejo:

- izboljševanje kakovosti učnih korpusov UD za slovenščino: kontinuirano odpravljanja nedoslednosti v označenosti korpusa SUK in drugih sorodnih ročno označenih virov (SST, Janes-Tag, Sloleks);
- izboljševanje smernic UD za slovenščino: kontinuirane izboljšave smernic UD za slovenščino na podlagi širšega strokovnega konsenza; usklajevanje z nadaljnjim mednarodno usklajenim razvojem splošnih smernic UD, tudi v kontekstu evropske raziskovalne mreže COST UniDive;
- povečanje obsega in diverzifikacija učnih korpusov UD za slovenščino: kontinuirano povečevanje deleža skladenjsko razčlenjenih povedi v učnem korpusu standardne pisne slovenščine SUK; razvoj komplementarnih učnih korpusov za druge jezikovne zvrsti (npr. spletna besedila, govorjeni jezik, starejša besedila);
- razvoj orodij za razčlenjevanje besedil po sistemu UD: kontinuirana izdelava računalniških orodij za univerzalno slovnično razčlenjevanje slovenskih besedil in na njih temelječih aplikacij;
- razvoj infrastrukture za delo s korpusi UD: kontinuiran razvoj orodij in spletnih servisov za označevanje, analizo in vizualizacijo univerzalno skladenjsko razčlenjenih korpusov.

6 Špela Arhar Holdt, Luka Terčon: JOS-SYN skladnja

6.1 Obseg in namen označevanja

6.1.1 Označevalni sistem JOS-SYN

Sistem JOS-SYN, ki je bil zasnovan v projektu Jezikoslovno označevanje slovenščine (Erjavec et al. 2010), je namenjen ročnemu označevanju skladenjskih odnosov v slovenskih povedih. Sistem sledi spoznanjem uveljavljenega slovenskega jezikoslovja (zlasti slovnici Toporišič 2004), obenem pa temeljnim idejam, ki jih zarisujejo obstoječi sistemi odvisnostnega označevanja. Ključna lastnost

Operacijo Razvoj slovenščine v digitalnem okolju 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.

sistema je, da upošteva informacije, ki jih v 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 in pregleden označevalni sistem.

V sistemu JOS-SYN imajo imena za povezave med deli stavka oz. povedi nevtralne oznake zaradi izogibanja terminološki zmedi pri poimenovanju delov stavka, ki bi jih tradicionalno imenovali osebek, predmet, prislovno določilo ipd. Sistem oznak s kratkimi opisi strnjeno predstavlja Tabela 2.

Tabela 2: Sistem oznak v označevalnem sistemu JOS-SYN

Oznaka	Opis povezave
dol	S povezavo dol povezujemo določujoči in določani del podredne besedne zveze, tj. jedro in različne vrste prilastkov. Izhodišče je jedro zveze, cilj pa beseda, ki jedro določa. Tipično jo uporabljamo pri samostalniških, pridevniških oz. prislovnih besednih zvezah ter za povezovanje delov povedka pri modalnih glagolih, povedkovem določilu in povedkovem prilastku. Na medstavčni ravni s to povezavo povezujemo prilastkove odvisnike z ustreznim glavnim stavkom.
del	S povezavo del povezujemo dele besednih zvez, pri katerih ne govorimo o prototipnem jedru ter določilu zveze, pač pa zgolj o delih zveze. Tipično jo uporabljamo za dele zloženega povedka, pri čemer je izvor osebna glagolska oblika oz. deležnik na -l, cilj so morfemi »ne«, »se«, »si«, »bi« oz. oblike glagola »biti« v sestavljenih nesedanjiških časih, tj. »bo«, »je« itd.
prir	S povezavo prir povezujemo dele priredij na besednozvezni ravni. Pri besednih zvezah povezujemo jedro prvega dela priredja z jedrom drugega dela priredja, pri čemer je izhodišče v levem delu, cilj pa v desnem delu priredja.
vez	Povezavo vez uporabljamo v kombinaciji s povezavo prir tako, da trije elementi, povezani s prir in vez, tvorijo trikotnik. S povezavo vez povezujemo jedro drugega dela besednozveznega priredja in priredni veznik ali priredno ločilo, če veznika ni.
skup	S povezavo skup povezujemo besede, ki imajo zelo močno tendenco po sopojavljanju in tvorijo neke vrste večbesedne enote, ne gre pa za klasične podredne ali priredne besedne zveze. S to povezavo tipično označujemo besede, ki imajo variantni zapis skupaj ali narazen, nekatere večbesedne veznike in podobne večbesedne enote.

Operacijo Razvoj slovenščine v digitalnem okolju 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.

ena	S povezavo ena povezujemo dele stavkov oz. povedi, ki bi jih tradicionalno lahko opredelili kot osebek, vendar tako označena razmerja ne ustrezajo povsem klasični definiciji tega pojava. Na znotrajstavčni ravni povezuje povedkovo vozlišče z osebkom stavka, pri čemer je izhodišče razmerja jedro povedka, cilj razmerja pa jedro osebka. Na medstavčni ravni s to povezavo povezujemo osebkove odvisnike z ustreznim glavnim stavkom. Izhodišče je jedro povedka nadrednega stavka, cilj je jedro povedka v odvisniku.
dve	S povezavo dve povezujemo dele stavkov oz. povedi, ki bi jih tradicionalno lahko opredelili kot predmet stavka ali predmetni odvisnik, vendar tako označena razmerja ne ustrezajo povsem klasični definiciji teh dveh pojavov. Na znotrajstavčni ravni povezuje vozlišče povedka in predmet stavka. Izhodišče razmerja je jedro povedka, cilj je jedro predmeta. Na medstavčni ravni s to povezavo povezujemo predmetne odvisnike z ustreznim glavnim stavkom. Izhodišče je jedro povedka glavnega stavka, cilj je jedro povedka v odvisniku.
tri	S povezavo tri povezujemo dele stavkov oz. povedi, ki bi jih tradicionalno lahko opredelili kot prislovno določilo lastnosti ali lastnostni odvisnik, vendar tako označena razmerja ne ustrezajo povsem klasični definiciji teh dveh pojavov. Na znotrajstavčni ravni povezuje vozlišče povedka in prislovno določilo lastnosti. Izhodišče razmerja je jedro povedka, cilj je jedro prislovnega določila. Na medstavčnem nivoju s to povezavo povezujemo lastnostne odvisnike oz. polstavke z ustreznim glavnim stavkom. Izhodišče je jedro povedka glavnega stavka, cilj je jedro povedka v odvisniku.
štiri	S povezavo štiri povezujemo dele stavkov oz. povedi, ki bi jih tradicionalno lahko opredelili kot prislovno določilo ali prislovni odvisnik (razen prislovnega določila lastnosti ali lastnostnega odvisnika), vendar tako označena razmerja ne ustrezajo povsem klasični definiciji teh pojavov. Na znotrajstavčni ravni povezuje vozlišče povedka in prislovno določilo. Izhodišče razmerja jedro povedka, cilj je jedro prislovnega določila. Na medstavčni ravni s to povezavo povezujemo odvisnike ustrezajočih tipov z ustreznim glavnim stavkom. Izhodišče je jedro povedka glavnega stavka, cilj je jedro povedka v odvisniku.
modra	S povezavo modra povezujemo abstraktno vozlišče stavka oz. povedi z elementi, iz katerih potekajo nadaljnje drevesnične povezave. S to povezavo so tipično povezani stavčni povedki, eliptični deli povedi brez povedka ali členkovni elementi znotraj stavka. S to povezavo so na vozlišče povezane tudi vse pojavnice (besede ali ločila), ki nimajo eksplicitne skladenjske vloge v povedi.

6.1.2 Označevalna kampanja JOS-SYN

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

Operacijo Razvoj slovenščine v digitalnem okolju 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.

(Krek et al. 2020: 25–26). Na teh podatkih je že bil naučen skladenjski razčlenjevalnik za slovenščino, ki je dosegal 90,43 % za pravilno določeno mesto povezave oziroma 87,52 % za pravilno določena mesto in tip povezave (Dobrovoljc et al. 2012). Zato smo na projektu RSDO za označevanje JOS-SYN izbrali le podatkovno množico ELEXIS-WSD, pri kateri želimo zagotoviti karseda visoko možno pokritost z raznolikimi označevalnimi sistemi za treniranje strojnega pomenskega razdvoumljanja. Cilj označevalne kampanje JOS-SYN na RSDO je bil tako skladenjsko označiti 2.024 novih povedi ELEXIS-WSD (s tem povečati obseg učnega gradiva za 18 %), pri tem pa natančneje oceniti ter nadgraditi označevalne smernice (Holozan et al. 2008).

Pri skladenjskem označevanju so sodelovali trije jezikoslovci. Povedi množice ELEXIS-WSD, v katerih je predhodno 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 (neodvisno eden od drugega) s pomočjo orodja Q-CAT (Brank 2022) vsako od povedi ročno pregledala in popravila strojno pripisane skladenjske oznake. 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 natančno popisali in jezikoslovno analizirali. Za analize smo pripravili izvoze označenih podatkov iz ssj500k, ki smo jih s specializirano programsko skripto še uredili za lažje branje.

Celotna kampanja je trajala približno štiri mesece, od tega dva meseca intenzivnega označevanja in dva meseca za pripravo analiz in nadgradnjo smernic. Poleg vrzeli v smernicah, ki so popisane v nadaljevanju poročila, smo med delom identificirali tudi številna problematična mesta, kjer so obstoječi podatki v ssj500k označeni neskladno. Nekatera od teh neskladij izvirajo iz nejasnosti smernic, ki so sedaj odpravljene, nekaj pa jih je širših, vezanih na težave označevanja nižjih ravni, čemur se bo treba posvetiti v nadaljnjih projektih.

6.2 Nadgradnja označevalnih smernic

6.2.1 Oblikovna in strukturna poenostavitev

Z željo, da postanejo označevalne smernice preproste za nadaljnje nadgrajevanje ter prenos v digitalno obliko, smo jih oblikovno in strukturno poenostavili. Odstranili smo napredno formatiranje, npr. opozorila v barvnih okencih, in poenotili navajanje zgledov, naslovov ter členjenje alinej. Vsebino smo

Operacijo Razvoj slovenščine v digitalnem okolju 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.

členili na podpoglavja četrtega nivoja, da je vsak tematski sklop oz. označevali problem na voljo v svoji zaključeni enoti. Oštevilčevanje zgledov, ki je bilo prej bralcu skrito, smo prikazali, da se je na zglede mogoče neposredno sklicevati. Nekaterim poglavjem in podpoglavjem smo spremenili zaporedni vrstni red, da si informacije vedno sledijo zaporedno in postopoma.

Ker v projektu RSDO prehajamo na nov način hranjenja in dostopnosti označevalnih smernic, smo iz njih odstranili projektno specifične informacije in informacije o delotokih označevalnega procesa (kako naj se označevalci organizirajo in komunicirajo; kakšno gradivo in orodja imajo na voljo in kako naj do njega dostopajo). Popis predhodnih projektov in kampanj je smiselno pripraviti ločeno od smernic, ki se osredotočajo na označevalni sistem in dileme.

6.2.2 Vsebinska poenostavitev

Na več mestih smernic smo poenostavili ubeseditve in zagotovili dodatne zglede označevanja. Poenostavili smo npr. poglavji *Povedkovo določilo in povedkov prilastek* ter *Vezljivost,* ki sta vsebovali natančne primerjave označevalnega sistema s tradicionalnim (strukturalističnim) slovničnim opisom, kar je oteževalo preglednost dejanskih označevalnih navodil. Izboljšali smo tudi berljivost poglavij *Presojanje sestavljenosti zveze, Besedi 'ne' in 'ni'* ter predhodno težko razumljivo poglavje *Strukture s 'kot'.* Zamenjali smo napačno označeni (torej napačen) zgled v poglavju *Pasivne konstrukcije* in dodali zglede v poglavjih *Zveze tipa 'od'-'do', Večdelni vezniki* ter *Uvajanje odvisnikov s kazalnimi zaimki.*

6.2.3 Nove vsebine

V smernice smo dodali tudi nove vsebine, ki natančneje pojasnjujejo obravnavo pojavov, ki so bili predhodno slabše opisani ali drugače obravnavani. Večja sprememba je poglavje *Povezovanje simbolov in ločil,* v katerem natančno popišemo povezovanje, kadar v povedi naletimo na znake, 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*). Po starem sistemu smo vsa ta znamenja povezovali z metaelementa, zaradi česar je razpadla drevesnica vseh povedi, ki so takšne zapise vsebovala. Nove smernice, ki ločujejo povezljive znake od nepovezljivih, so skladne s primerljivimi skladenjskimi sistemi, npr. UD za slovenščino (Dobrovoljc et al. 2022). Na projektu RSDO smo pregledali in posodobili obravnavo povezljivih znakov tudi v ssj500k.

Operacijo Razvoj slovenščine v digitalnem okolju 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.

Obširnejša posodobitev je bila pripravljena tudi za poglavje *Lastna imena in tujejezični elementi.* Navodila za označevanje lastnih imen so bila predhodno precej skopa, posledično pa je v ssj500k opaziti velike neskladnosti označevanja, tako pri določanju, ali lastnoimensko zvezo sploh označevati (ali vse njene dele obravnavati kot fragment v tujem jeziku in posledično vezati z metaelementa), kot tudi odločanje, kaj je jedro pri zvezah, ki jih povezujemo. Precej nedoslednosti je tudi pri povezovanju tujejezičnih členov, ki naj bi jih načeloma vezali na metaelement, 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 optimalno zasnovane (Pori et al. 2022). Poglavje o lastnih imenih smo notranje strukturirali na *Imena in priimki; Večdelna lastna imena; Tujejezični predlogi oz. členki tipa 'von', 'de', 'la'; Tujejezične občne zveze* in *Tujejezična lastna imena*, za vsako temo pripravili opis, temelječ na analizah predhodnega označevanja, pa tudi posebna opozorila, kje je v preteklosti prihajalo do zmede. Na RSDO neskladno označenih primerov v ssj500k nismo popravljali, saj bi bilo pred popravki treba analizirati in izboljšati sisteme označevanja na različnih nivojih. Slednje predlagamo za bodoče delo.

V poglavje, ki se posveča strukturam tipa *nujno je* smo dodali obravnavo struktur *treba je,* saj je za označevalce koristno na enem mestu videti, da so pridevniki v teh primerih iz 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 *Prilogo C – informativni seznam zvez, ki jih povezujemo z SKUP* ter seznam členkov, ki jih namesto z metaelementa povezujemo s povezavo DOL v poglavju *Členkovna modifikacija.* 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.

Operacijo Razvoj slovenščine v digitalnem okolju 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.

6.2.1 Nedoslednosti v označenih podatkih

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 števnike (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 ravni (SRL) ali so posledica nerešenih vprašanj na nižjih ravneh (oblikoskladnja).

6.3 Rezultat in nadaljnje delo

6.3.1 Nova učna množica za JOS-SYN

Rezultat označevalne kampanje za JOS-SYN je razširjena učna množica, ki je sestavljena iz dveh delov: tistega dela že obstoječega učnega korpusa ssj500k, ki vsebuje oznake JOS-SYN, in slovenske različice paralelnega korpusa ELEXIS-WSD. Skupaj obsega 13.435 povedi, 267.097 pojavnic in 25.677 lem. Poleg skladenjskih oznak množica vsebuje tudi ročno pregledane oznake na ravneh lematizacije, oblikoskladnje, imenskih entitet, glagolskih večbesednih enot in udeleženjskih vlog. Povečani obseg nove množice je prikazan v Tabeli 3.

Tabela 3: Pregled obsega stare (ssj500k) in nove (ssj500k + ELEXIS-WSD) množice za JOS-SYN

	ssj500k	**ssj500k + ELEXIS-WSD**
Št. pojavnic	235,864	267,097
Št. povedi	11.411	13.435

Glede na vrsto povezave je v novi množici največ povezav tipov *dol* (92.044) in *modra* (72.957). Nato po številčnosti sledijo *vez* (21.666), *del* (18.200), *dve* (17.465), *štiri* (16.141), *ena* (13.689), *prir* (7.578) in *tri* (6.358). Pričakovano najmanj je povezav tipa *skup* (999), saj se ta tip uporablja zgolj za omejen nabor večbesednih enot.

Operacijo Razvoj slovenščine v digitalnem okolju 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.

6.3.2 Uspešnost označevanja JOS-SYN

Razširjeno učno množico za označevalni sistem JOS-SYN smo uporabili za učenje označevalnega orodja CLASSLA-Stanza (Ljubešić in Dobrovoljc 2019), ki podpira tudi skladenjsko označevanje po sistemu JOS-SYN.

Pri učenju razčlenjevalnega modela smo uporabili delitev na učno, validacijsko in testno podmnožico, ki izhaja iz delitve celotnega učnega korpusa SUK.[23] Novi skladenjski model pri evalvaciji na testni množici dosega splošno natančnost 0,93, izraženo z metriko LAS (angl. *Labeled attachment score*), ki se pogosto uporablja za ocenjevanje natančnosti pri sistemih odvisnostnega označevanja (Nivre in Fang 2017).

Tabela 4 spodaj prikazuje še podrobneje, kako uspešen je bil razčlenjevalni model pri vsakem tipu povezave posebej. Vrednosti so izražene z oceno F1 za vsak tip povezave.

Tabela 4: Ocena F1 za vsak tip povezave

modra	0,97
dol	0,97
prir	0,92
vez	0,97
skup	0,95
ena	0,91
del	0,99
dve	0,90
štiri	0,88
tri	0,82

Najvišjo natančnost model dosega pri povezavi *del,* ki povezuje dele zloženih povedkov. Tu gre predvsem za pomožne glagole v sestavljenih časih, ki ponavadi sledijo zelo predvidljivim strukturnim

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

Operacijo Razvoj slovenščine v digitalnem okolju 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.

vzorcem. Zelo visoko natančnost (>0,95) model dosega tudi pri povezavah *modra, dol* in *vez.* Sledijo povezave *skup, prir, ena* in *dve*, vse z vrednostmi nad 0,90. Najmanjšo natančnost (0,80 – 0,90) model dosega pri povezavah *štiri* in *tri,* ki se uporabljata za lastnostna prislovna določila (*tri*) in ostale tipe prislovnih določil (*štiri*). Tekom označevalne kampanje se je izkazalo, da se v praksi pogosto pojavijo mejni primeri, ki nosijo lastnosti obeh kategorij, zato sta ti dve povezavi še najbolj odvisni od pomenske interpretacije povedi.

6.3.3 Strategija nadaljnjega razvoja

Po projektu RSDO učna množica z ročno pregledanimi oznakami JOS-SYN obsega 13.435 povedi, označevalnik pa dosega splošno natančnost 0,93 (izraženo z metriko LAS). Nadaljnje povečevanje podatkovnega obsega je dobrodošlo, vendar so prioritete zlasti izboljšava kvalitete podatkov, urejanje njihovega boljšega dostopa za lažjo nadaljnjo obdelavo in analize ter obravnava specializiranega jezika.

a. Skladenjske podatke je trenutno mogoče vizualizirati v programu Q-CAT, ki omogoča tudi napredno iskanje podatkov; prav tako je mogoče po skladenjskih oznakah iskati v konkordančnikih, ki jih ponuja infrastruktura CLARIN.SI. Vendar pa sta vizualizacija in izvoz skladenjskih podatkov v obeh orodjih omejena, zato niso enostavno, pregledno dostopni – čeprav so skladenjske oznake že vključene v številne slovenske besedilne korpuse, se redno uporabljajo. V nadaljevanju je treba razviti možnosti za enostavno pregledovanje ter izvažanje skladenjsko označenih podatkov v berljivem formatu za jezikoslovne analize in druge nadaljnje rabe.

b. Na osnovi identificiranih šibkih mest in nedoslednosti je treba izboljšati kvaliteto označenih podatkov, pri čemer je ključno usklajevanje smernic med označevalnimi ravninami.

c. Zagotoviti je treba analize mest, kjer se strojni označevalnik moti, in v primeru, da so določeni jezikovni pojavi v učni množici preveč redki ali razpršeni, da bi bilo možno strojno učenje, dodati (ročno izbrane) povedi, ki pojav reprezentirajo.

d. V primeru nadaljnjega označevanja bi bilo smiselno dodati vzorce specializiranega jezika, npr. iz korpusa KAS, korpusov jezikovnega usvajanja, govorjene slovenščine, nestandardne slovenščine, historičnih besedil itd. Pilotski preizkus, označevanje 168 tvitov v nestandardni

Operacijo Razvoj slovenščine v digitalnem okolju 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.

slovenščini (Arhar Holdt et al. 2016), je pokazal številne jezikovne značilnosti, ki v označevalnih smernicah JOS-SYN niso bile naslovljene, zato je pri tovrstnem označevanju treba zagotoviti tudi sredstva za (žanrsko specifično) nadgrajevanje smernic.

7 Polona Gantar: SRL udeleženske vloge

Poročilo o semantičnem označevanju učnega korpusa SUK, ki je nastal v okviru projekta RSDO, vsebuje opis izgradnje podkorpusa SRL ter njegovih sestavnih delov z vidika predhodnega avtomatskega procesiranja in z vidika poteka ročnega označevanja. V poročilu so podrobno predstavljena izhodišča semantičnega označevanja, nabor oznak za slovenščino ter semantični sistem, ki temelji na Praški odvisnostni drevesnici in je realiziran v češkem vezljivostnem leksikonu Vallex. Na sistemetičnem prikazu udeleženskih vlog so opisana odstopanja od češkega sistema in posebnosti pri slovenskem naboru semantičnih oznak. V samostojnem poglavju so prikazani rezultati označevanja, in sicer zastopanost posamezne udeleženske vloge v korpusu SUK ter preliminarna evalvacija uspešnosti avtomatskega pripisovanja udeleženskih vlog v podlkorpusu SRL-3 glede na ročne oznake. Poročilo vsebuje tudi podroben opis vsebinske nadgradnje korpusa, ki se nanaša tako na pomenski kot na pomenskoskladenjski vidik ter zagotavlja konsistentnost pri označevalnih odločitvah. S tem smo želeli zagotoviti večjo natačnost učne množice in njeno uporabnost v jezikovnotehnoloških nalogah. Poročilo zaključujemo s smernicami za izboljšavo korpusa na več ravneh: z aktualizacijo semantičnih oznak glede na označevalni sistem praške odvisnostne drevesnice; 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, ELEXIS).

7.1 Obseg in namen označevanja

7.1.1 Izgradnja podkorpusa SRL

Učni korpus SUK (Arhar Holdt et al. 2022) je sestavljen iz več delov: SentiCoref, ELEXIS-WSD, Tviti, Ambiga in ssj500k 2.3. V tem delu dokumenta opisujemo izgradnjo podkorpusa SRL (SRL-1, SRL-2 in

Operacijo Razvoj slovenščine v digitalnem okolju 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.

SRL-3), ki je sestavljen iz korpusa ssj500k 2.3 (Krek et al. 2021) in korpusa ELEXIS-WSD (Martelli et al. 2022) ter je ročno označen s semantičnimi kategorijami na ravni udeleženskih vlog.

Korpus ssj500k 2.3 smo za namene semantičnega označevanja dodatno pripravili, in sicer smo najprej izdelali korpus SRL-1, ki vsebuje stavke, ki so bili že ročno označeni pri pripravi predhodne različice učnega korpusa (ssj500k 2.3). Ta korpus vsebuje 5.491 stavkov z ročno pripisanimi semantičnimi oznakami na ravni udeleženskih vlog in je bil uporabljen kot učna množica za učenje SRL razčlenjevalnika (Björkelund 2009), s katerim smo predhodno avtomatsko semantično označili slovenski in hrvaški učni korpus (Gantar et al. 2018). Ločeno smo pripravili korpus SRL-2, ki vsebuje stavke, ki so bili v ssj500k 2.3 že označeni na morfološki in skladenjski ravni (JOS-SYN, UD-SYN), niso pa bili označeni na semantični ravni. Ta korpus vsebuje 11.411 stavkov.

Tretji del semantično označenega novega učnega korpusa SUK (SRL-3) predstavlja korpus ELEXIS_WSD_SL (Martelli et al. 2021), ki je eden od 10 paralelno morfološko, skladenjsko in semantično označenih korpusov, ki so bili v okviru projekta ELEXIS izdelani za 10 evropskih jezikov. Korpus ELEXIS-WSD-SL vsebuje 2.024 stavkov in je bil v projektu RSDO vključen v učni korpus SUK z ročnimi oznakami na vseh omenjenih nivojih. Korpusa SRL-2 in SRL-3 sta bila nato avtomatsko označena še na semantični ravni s pomočjo SRL parserja, korpus SRL-3 pa tudi na morfološki in skladenjski ravni z orodjem CLASSLA po sistemu JOS in UD.

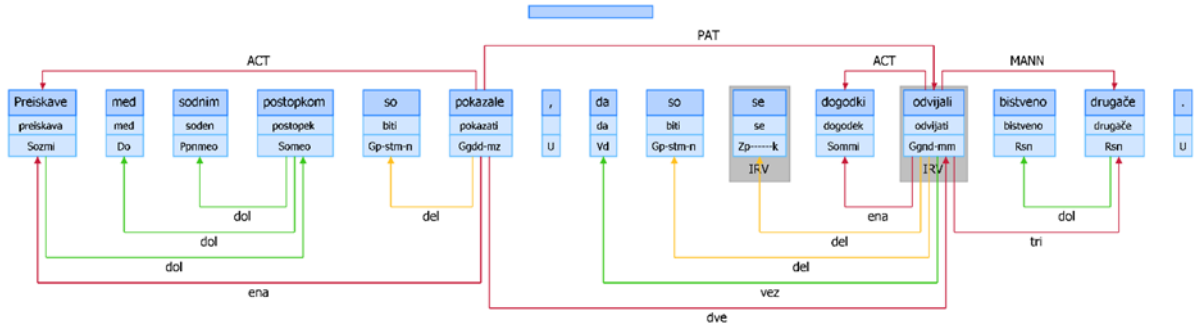
7.1.2 Postopek ročnega označevanja

Glede na projektne cilje je bilo za ročno označevanje predvidenih približno 5.000 na novo označenih stavkov na vseh predvidenih nivojih, hkrati pa naj bi se – glede na nadgradnjo jezikoslovnih izhodišč – ustrezno sistematično popravili tudi predhodno že ročno označeni stavki. V ta namen je bilo za označevalno kampanjo na semantični ravni pripravljenih 18.926 stavkov, od tega je bilo 5.491 stavkov ponovno pregledanih, 13.435 stavkov pa je bilo najprej povsem na novo avtomatsko označenih nato pa ročno pregledanih. Odločitve so bile na koncu usklajene v celotnem podkorpusu SRL učnega korpusa SUK.

Označevanje semantične ravni za razliko od skladenjske in morfološke ni potekalo kampanjsko. Celoten delež na novo označenega in popravljenega predhodno že označenega korpusa je izvedla podiplomska

Operacijo Razvoj slovenščine v digitalnem okolju 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.

študentka slovenistike na podlagi prve različice smernic, pri katerih je že sodelovala, in na podlagi sprotnih konzultacij in navodil. Celoten na novo označen korpus je bil nato pregledan s strani jezikoslovca, in sicer na podlagi sistematičnih in ciljnih preverjanj. Za označevanje je bilo uporabljeno orodje Q-Cat (Brank 2022), kor prikazuje Slika 1.



Slika 1: Prikaz označevalnih ravni v orodju Q-CAT

Za izhodišče semantičnega označevanja smo vzeli posamezni glagol ter določili semantične kategorije njegovih (skladenjskih) udeležencev v okviru stavka. To nam je omogočilo prepoznavanje tipičnih glagolskih vzorcev in večjo konsistentnost pri označevanju.

Z označevanjem smo začeli pri frekventnejših 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 stavku. Takih glagolov je bilo v korpusu pribl. 1200. Na ta način smo v največji možni meri zajeli tisti del korpusnih stavkov, za katere je bilo mogoče izpeljati čim bolj sistematične in usklajene jezikovne rešitve. Odločitev je pomembna tudi z vidika nadaljnje uporabe učnega korpusa za učenje jezikovno-analitičnih orodij.

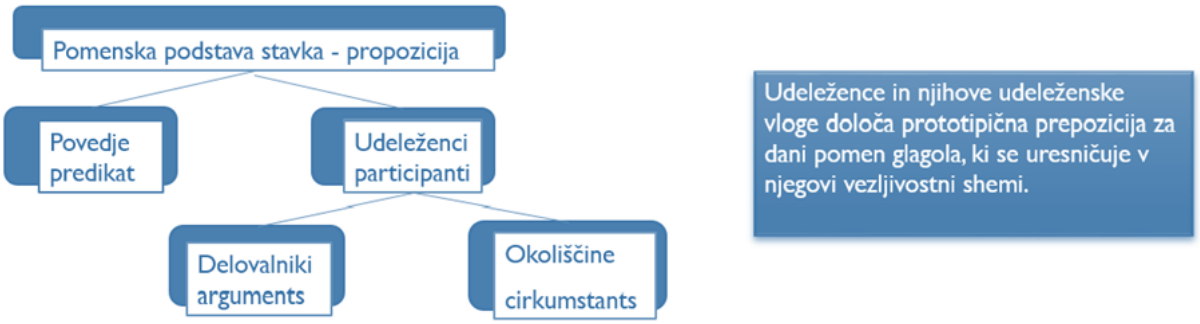
7.2 Opis označevalne ravni

7.2.1 Semantično označevanje udeleženskih vlog

Označevanje udeleženskih vlog s semantičnimi kategorijami (angl. *semantic role labeling*, SRL) je postopek, ki je z jezikoslovnega vidika namenjen prepoznavanju udeleženskih vlog, z jezikovnotehnološkega pa razvoju sistemov za luščenje informacij, sistemov za odgovarjanje na vprašanja (angl. *question answering system*), izboljšavi delovanja skladenjskih razčlenjevalnikov ter

Operacijo Razvoj slovenščine v digitalnem okolju 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.

strojnih prevajalnikov ipd. (Shen in Lapata 2007; Christensen et al. 2011). Izhodišče za določanje udeleženskih vlog (in njihovih semantičnih oznak) predstavlja t. i. propozicija ali pomenska podstava stavka (Slika 2), ki ima dve temeljni sestavini: predikat (ali povedje) in udeležence oz. participante. Udeležence lahko nadalje delimo na delovalnike (aktante) in okoliščine (cirkumstante). Povedje je na skladenjski ravni povedek (oz. zloženi povedek: povedkovo določilo in povedkov prilastek), delovalnika sta osebek in predmet (argumenti – določila), okoliščine pa prislovna določila (tj. adjuncts – dopolnila).



Slika 2: Elementi pomenske podstave povedi

Semantične oznake je torej glede na elemente propozicije mogoče pripisovati sestavnim elementom zloženega povedka, delovalnikom in okoliščinam. Konkretno bi lahko rekli, da predvideva glagol *narediti* v enem od svojih pomenov tako delovalniške kot okoliščinske udeležence, ki se na oblikoskladenjski ravni realizirajo v obliki argumentov, ki jih je mogoče zapisati kot: *kdo naredi komu kaj (kdaj, kje, kako, zakaj)*, kar predstavlja vezljivostno shemo konkretnega glagolskega pomena.

Pri snovanju slovenskega modela semantičnega označevanja smo sledili naslednjim izhodiščem (Krek et al. 2016): (a) zagotoviti nabor kategorij, ki je kar najbolj optimalen, tj. pokriva vse (v našem primeru za slovenščino) ključne udeleženske vloge in hkrati (b) ne vsebuje kategorij, ki so prepodrobne ali medsebojno prekrivne, (c) temeljiti primarno na semantičnih lastnostih glagola, (d) omogočiti formalni opis oz. uporabnost v jezikovnotehnoloških aplikacijah ter (e) biti čim bolj kompatibilen s kategorijami in merili, ki veljajo za druge jezike (prim. Petukhova in Bunt, 2008: 39). V ta namen smo pri izbiri metode semantičnega označevanja in določanju semantičnih kategorij za slovenščino najprej analizirali posamezne pristope, ki so bili razviti in uporabljeni za druge jezike, npr. PropBank (Palmer et al. 2005), Verbnet (Kipper et al. 2006) in FrameNet (Backer et al. 1998) za angleščino, AnCora (Taulé et al. 2011)

Operacijo Razvoj slovenščine v digitalnem okolju 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.

za španščino, SoNaR (Schuurman et al. 2010) za nizozemščino. Poleg tega pa še nabor oznak za hrvaščino (Filko et al. 2012) in češki valenčni leksikon Vallex[24]. Glede na analizo označevalnih sistemov in zgoraj opredeljena izhodišča smo se odločili, da bomo pri naboru udeleženskih vlog in njihovih formalnih opisov izhajali iz funkcijskega generativnega pristopa Praške odvisnostne drevesnice (angl. *Prague Dependency Treebank*, PDT; Mikulová et al. 2006), na katerem temelji že omenjeni Vezljivostni leksikon čeških glagolov – VALLEX.

7.2.2 Specifikacija semantičnih oznak

Osnovo za nabor udeleženskih vlog in njihovih oznak nam je, kot rečeno, predstavljal nabor oznak praške odvisnostne drevesnice PDT 2.0[25]. Z vidika optimizacije pomenske razdrobljenosti, upoštevanja slovenskih specifik in hkrati prekrivnosti oznak med posameznimi sistemi, smo nabor ustrezno zreducirali, kot je podrobneje opisano v nadaljevanju. Podroben opis semantičnih oznak in pravila za njihovo uporabo je izdelan v Smernicah za semantično označevanje učnega korpusa[26], ki so bile v okviru projekta RSDO deloma nadgrajene in posodobljene glede na vsebinske analize.

7.2.2.1 Delovalniki

Med delovalniškimi udeleženci, ki se v sistemu PDT obravnavajo kot obligatorni,smo prevzeli vseh pet kategorij (Tabela 5): vršilec (ACT), prizadeto (PAT), prejemnik (REC), izvor (ORIG) in učinek oz. rezultat dejanja (RESLT). Udeleženska vlogo naslovnika (addressee) in koristnika (benefactor) smo glede na PDT združili v enotno oznako prejemnik (REC), ki tako združuje prejemnike (podčrtano) kot posredne udeležence dejanja (*Prijatelju sem poslal darilo*) in nedelovalniške udeležence, ki jim je dejanje v škodo ali v prid (*Miha je Maji ujel pobeglega kanarčka; To sliko je narisal otrokom/za otroke*; *Glava mu počasi pada na prsa*; *Ura mi dobro dela*; *To ti je žurka).* Z udeleženska vlogo izvor (ORIG) smo v slovenskem sistemu združili kategoriji izvora (*Kaj se je zgodilo, sem izvedel iz časopisja*) in podedovanosti (*Plašč je dobil po očetu*).

[24] https://ufal.mff.cuni.cz/vallex
[25] https://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/html/index.html
[26] https://wiki.cjvt.si/books/udelezenske-vloge-srl/page/oznacevalne-smernice

Operacijo Razvoj slovenščine v digitalnem okolju 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.

Tabela 5: Nabor oznak za delovalniške udeleženske vloge za slovenščino glede na PDT

Oznaka	**PDT**		**SLO**	
ACT	ACT	actor		**vršilec, aktant**
PAT	PAT	patient		**prizadeto**
REC	ADDR	addressee	prejemnik	**prejemnik**
	BEN	benefactor	koristnik	
ORIG	ORIG	origin	izvor	**izvor**
	HER	inheritence	podedovanost	
RESLT	EFF	effect		**učinek**

7.2.2.2 Okoliščine

Praška odvisnostna drevesnica predvideva med okoliščinskimi udeleženci naslednje kategorije: čas, prostor, vzročnost in način, ki smo jih upoštevali tudi v slovenskem naboru (Tabela x).

Pri notranji razčlenjenosti smo težili k združevanju pomenskih kategorij pod eno oznako. Tako smo za različne časovne kategorije (skupaj 9), ki imajo v PDT ločene oznake, v slovenskem naboru združili v tri: TIME (čas), DUR (trajanje) in FREQ (pogostnost). Hkrati smo določili, da oznaka TIME zajema semantične povezave, ki ustrezajo opredelitvam kot: kdaj (*Odrinemo ob zori*), sočasnost (*Med počitnicami ni niti enkrat posijalo sonce*), od kdaj (*Sestanek je prestavil s četrtka na petek*), do kdaj (*Sestanek je prestavil s četrtka na petek*). Oznaka DUR določa povezave, ki opredeljujejo trajanje stanja ali dejanja: kako dolgo (*Prišel je za en mesec*), koliko časa (*To mu je uspelo narediti v enem tednu*), konkretni trenutek začetka (*Od jutri dalje sem na dopustu)* in konca dejanja (*Do četrtka uredim še papirje*). Oznaka FREQ pa pogostnost (kako pogosto, kolikokrat: *Vsak dan se mučimo s tem*).

Med prostorskimi oznakami smo glede na PDT pod oznako LOC združili kraj oz. mesto dejanja (kje: *V bližini vasi stoji kozolec*) in smer v prostoru (kam, kje, v katero smer: *Sprehajajo se po gozdu*) ter ohranili oznaki SOURCE za označevanje začetne lokacije (od kod: *S stropa odpada omet*) in GOAL za označevanje končne točke v prostoru (kam: *Prišel je domov*).

Operacijo Razvoj slovenščine v digitalnem okolju 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.

Med vzročnostnimi oznakami smo pod skupno oznako AIM združili namen dejanja (s kakšnim namenom: *Telovadi, da bi shujšala*) in namen gibanja (čemu: *Poslali so ga po malico*). Pod oznako CONTR smo združili dopustnost in protivnost (kljub čemu: *Študijskim uspehom navkljub službe ni dobil; Medtem ko se plače nižajo, cene rastejo*). Nespremenjene pa smo ohranili oznake za vzrok CAUSE (zakaj: *Umrl je zaradi srčne kapi*) in pogoj COND (pod katerim pogojem: *Če neha deževati, gremo na kavo*).

Za razliko od PDT, ki predvideva tri semantične oznake za označevanje ozira (REG, CRIT in CPR), smo ohranili le prvo, pod katero združujemo ozir (glede na kaj: *Glede na vreme ni mogoče ugotoviti …*), merilo (v skladu s čim: *Preiskava je potekala v skladu z zakonom*) in primerjavo (v primerjavi s čim: *Za razliko od mene ima posluh*).

Med t. i. načinovnimi oznakami ohranjamo glede na PDT nespremenjeno oznake ACMP (spremstvo: *Mama je s sinom odšla v cirkus*), RESTR (omejitev: *Vsi so bili tam razen tebe*), MEANS (sredstvo/orodje za izvedbo dejanja: *Piše s peresom*) in MANN, ki združuje oznaki za način in rezultat ob koncu dejanja (kako; na kakšen način: *Dela prepočasi; Govori tako potiho, da ničesar ne razumem*).

Med količinskimi oznakami, ki jih predvideva PDT (DIFF in EXT) v slovenskem sistemu ohranjamo le eno (QUANT), s katero označujemo kakovostno razliko med dogodki, stanji in predmeti (*Cena goriva se je podražila za 3 cente*) ter mero, razpon ali intenziteto dejanja (koliko; v kolikšni meri: *Bil sem star kakšna tri, štiri leta*; *To je priložnost za vse, ki veliko kupujete*). Med načinovnimi oznakami PDT v slovenskem sistemu ne upoštevamo oznake COMPL, ki je namenjena označevanju povedkovih prilastkov, saj smo jo nadomestili z oznako RESLT, ki označuje rezultat glagolskega dejanja (*Zdela se mi je vsakdanja*).

Samostojna v slovenskem sistemu je oznaka EVENT (dogodek), ki označuje časovno-prostorsko določen dogodek (*Sodelovati v skupnih akcijah*), ki ustreza Framenetovski pomenski shemi za Event.

Operacijo Razvoj slovenščine v digitalnem okolju 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.

Tabela 6: Nabor oznak za okoliščinske udeleženske vloge za slovenščino glede na PDT

	Oznaka	PDT		SLO	
ČAS	**TIME**	TWHEN	when		**čas**
		TPAR	parallel	sočasnost	
		TFRWH	from when	začetek	
		TOWH	to when	konec	
	DUR	TFHL	for how long		**trajanje**
		THL	how long		
		TSIN	since when		
		TTILL	till		
	FREQ	THO	how often		**pogostnost**
PROSTOR	**LOC**	LOC	locative		**kraj**
		DIR2	which way		**smer**
	SOURCE	DIR1	from	začetna lokacija	
	GOAL	DIR3	where to	končna lokacija	
VZROČNOST	**AIM**	AIM	aim	namen	**namen**
		INTT	intent	namera	
	CAUSE	CAUS	cause		**vzrok**
	CONTR	CNCS	concession	dopustnost	**protivnost**
		CONTRD	contradiction	protivnost	
	COND	COND	condition		**pogojnost**
OZIR	**REG**	REG	regard	ozir	**ozir**
		CRIT	criterion	merilo	
		CPR	comparison	primerjava	
	ACMP	ACMP	accompaniment		**spremstvo**
NAČIN	**RESTR**	RESTR	restriction	omejitev	**omejitev**
	MANN	MANN	manner	način	**način**
		RESL	result	rezultat	
	MEANS	MEANS	means		**sredstvo**
KOLIČINA	**QUANT**	DIFF	difference	razlika	**količina**
		EXT	extent	količina	

7.2.2.3 Glagolske zveze

Največ odstopanj glede na praški sistem predstavljajo oznake za semantične vloge znotraj glagolskih zvez (Tabela 7). Za označevanje zveze glagola s pomensko neprozorno zvezo smo ohranili oznako PHRAS (v PDT: DPHR=dependent part of phraseme; *Iti na živce; mi je beseda ostala v grlu*). Na novo pa smo uvedli oznaki MWPRED (multi-word predicate) za zveze z nedoločniki (*Dati vedeti*) ter za fazne in nemodalne glagole (*Klinov ni videti nikjer*) ter oznako MODAL za zveze modalnega glagola in

Operacijo Razvoj slovenščine v digitalnem okolju 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.

nedoločnika (*Za neposredni prenos jih niso želeli prikrajšati*) ter za zveze glagola *biti* in modalnega prislova (*je treba poudariti*).

Tabela 7: Nabor oznak za udeleženske vloge v glagolskih zvezah za slovenščino glede na PDT

Oznaka	PDT		SLO	
PAT	CPHR	compound phraseme		**prizadeto**
RESLT				**učinek**
PHRAS	DPHR	dependant part of phraseme		**frazem**
MWPRED			večbesedni predikat	**glagolska zveza**
MODAL			modalna zveza	

V procesu ročnega označevanja so se pojavila vprašanja o možnostih izboljšave označevanja npr. z dodajanjem novih kategorij (gl. smernice za naprej) ter z iskanjem univerzalnih kategorij na večjezikovni ravni.

7.2.3 Rezultati označevanja

Korpus SRL vsebuje skupno 47.650 semantičnih oznak. Kot prikazuje Graf x, po pričakovanju prevladujejo glagolski vzorci z delovalniškima oznakama ACT in PAT, sledi RESLT, ki vsebuje produktivne strukture z glagolom *biti*, ki je tudi sicer najpogostejši glagol v korpusu. Sledijo okoliščinske oznake za čas (TIME), način (MANN) in kraj (LOC). Med pogostejšimi kategorijami izstopajo še delovalniška oznaka za prejemnike (REC), krajevna okoliščinska oznaka za smer (GOAL) in oznaka MODAL za sicer zelo pogost vzorec pomožnega glagola in modalnega prislova (*biti + treba*).

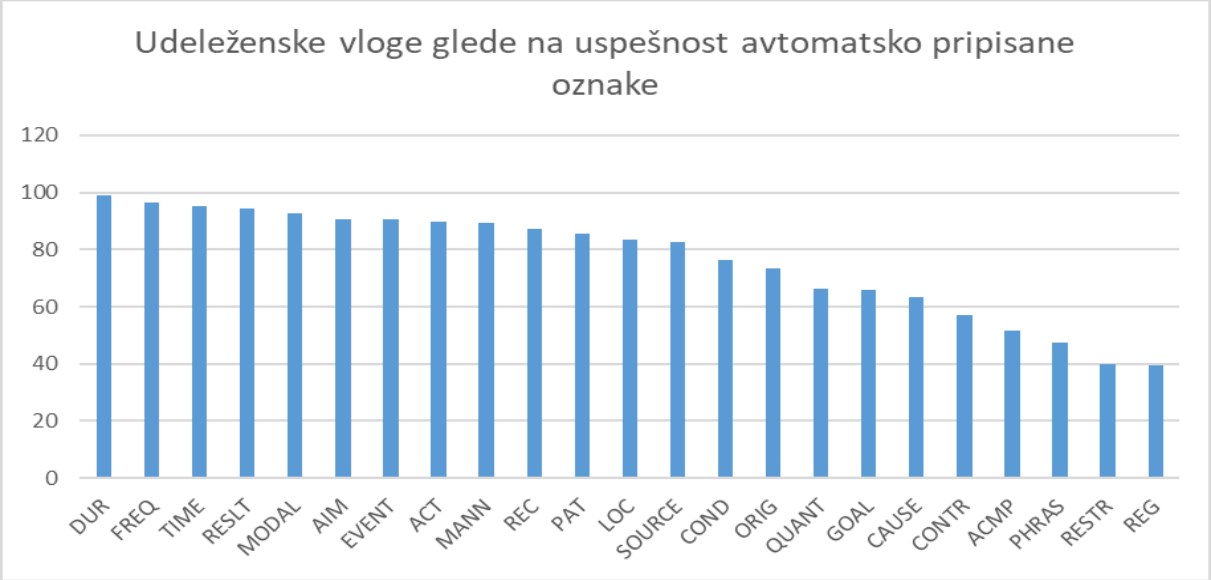
Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Graf 4: Število pojavitev posamezne udeleženske vloge v korpusu SUK.

Kot smo omenili, smo del podkorpusa SRL avtomatsko označili s pomočjo SRL razčlenjevalnika (Björkelund 2009), in sicer skupno 13.435 stavkov oz. celotni korpus SRL-3 (2024 stavkov) in SRL-2 (11.411 stavkov), ki vsebuje stavke, ki so bili v ssj500k 2.3 že označeni na morfološki in skladenjski ravni (JOS-SYN, UD-SYN), niso pa bili označeni na semantični ravni.

Poskusno smo za primerjavo med avtomatsko pripisanimi oznakami in odločitvami jezikoslovca kvantitativno ovrednotili rezultate na korpusu SRL-3, kot je prikazano v grafu X. Avtomatski označevalnik v povprečju deluje z več kot 76-% natančnostjo, pri čemer obstajajo precejšnje razlike pri posameznih udeleženskih vlogah. Pri časovnih udeleženskih vlogah, kot so DUR, FREQ in TIME, je avtomatsko označevanje prekrivno z ročnim v več kot 95 %. Več kot 90-% natančnost dosega avtomatski označevalnik tudi pri oznakah RESLT in MODAL, ki tvorita pomenskoskladenjsko predvidljive strukture, ter pri oznaki EVENT. Med delovalniškimi oznakami dosegajo ACT, REC in PAT več kot 85-% natančnost, načinovna oznaka MANN ter prostorski LOC in SOURCE pa več kot 80-% zanesljivost. Manj kot 50-% zanesljivost avtomatski označevalnik dosega pri oznakah za udeleženske vloge, ki so v korpusu najredkeje zastopane (CONTR, ACMP, PHRAS, RESTR) oz. je v procesu označevanja prišlo do večjih sprememb v označevalnih smernicah (npr. pri REG).

Operacijo Razvoj slovenščine v digitalnem okolju 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.



Graf 5: Odstotek uspešnosti avtomatskega označevalnika pri posamezni udeleženski vlogi v korpusu SRL-3.

Na podlagi ročnih oznak, ki so v novi obsežnejši učni množici pri najpogostejših udeleženskih vlogah bolj sistematične in konsistentnejše, pričakujemo v prihodnje izboljšanje avtomatskega označevanja tudi na tej ravni.

7.2.4 Vsebinska nadgradnja korpusa

V procesu označevanja je bil korpus nadgrajen tudi z vsebinskega vidika, pri čemer dodana vrednost temelji predvsem na jezikoslovnem premisleku že obstoječih odločitev v skladu z novimi spoznanji na področju izdelave semantičnih virov (npr. nove različice leksikona Vallex), analize vezljivostnih vzorcev pri izdelavi Vezljivostnega leksikona (Gantar 2021, Gantar v tisku) in na upoštevanju potreb jezikovnotehnološke skupnosti. Vsebinska nadgradnja temelji na semantičnem izhodišču in je usmerjena v jezikovno problematiko, ki omogoča čim bolj sistematične in konsistentne odločitve tudi na skladenjskem in morfološkem nivoju.

7.2.4.1 Semantično izhodišče: glagoli

Izhodišče semantičnega označevanja so predstavljale skupine glagolov glede na svoje pomensko polje, npr. glagoli govorjenja, premikanja, kognitivnih procesov ipd., kar nam je omogočilo prepoznati tipične

Operacijo Razvoj slovenščine v digitalnem okolju 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.

udeleženske vloge, ki se povezujejo s posameznimi pomeni glagolov znotraj skupnega pomenskega polja.

Glagoli rekanja

V približno ¾ 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. Pri odločitvah smo upoštevali obravnavo podobnih in zlasti večvalentnih glagolov v češkem vezljivostnem leksikonu Vallex. To je mogoče ponazoriti z vzorci glagola *prepričati* (Tabela 6), za katerega je v enem od svojih osnovnih pomenov mogoče reči, da tipično potrebuje vršilca (ACT) in naslovnika (REC) ter rezultat (RESLT) ali vsebino (PAT) izjavljanja, opcijsko pa še način, sredstvo čas in kraj, kot prikazuje z vsemi evidentiranimi različnimi vezljivostnimi vzorci za glagol *prepričati*.

Tabela 6: Vezljivostni vzorci za glagol *prepričati* v korpusu SUK
*Vzorci, ki so poudarjeni, se v korpusu pojavijo več kot 1x.

prepričati							
ACT*	**V**	**REC**	**RESLT**				
	V	**REC**	**RESLT**				
	V	**REC**					
	V		**RESLT**				
ACT	V	REC					
	V	REC		MANN	MEANS		
ACT	V	REC		MANN			
	V	REC	RESLT		MEANS		
ACT	V		RESLT	MANN			
ACT	V		RESLT			TIME	
	V		RESLT	MANN			
	V		RESLT			TIME	
	V		RESLT				LOC

Pri pogostejših glagolih, npr. z več kot 50 pojavitvami v korpusu, je mogoče vzorce uporabiti za analizo obligatornosti udeleženskih vlog, kot jih denimo predvideva češki vezljivostni leksikon Vallex, ki prav tako temelji na praški odvisnostni drevesnici. Po tem modelu bi denimo prototipični vzorec za glagol *prepričati* s podatki o obveznosti (obligatory), tipičnosti (typical) in opcijskosti (optional) posamezne

Operacijo Razvoj slovenščine v digitalnem okolju 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.

udeleženske vloge, v vzorcu za pomen 'doseči, da ima kdo povedano, mišljeno … za resnično, pravilno', izgledal takole:

prepričati [ACT_1^{OBL} REC_4^{OBL} $RESLT_{da;v}^{OBL}$ (PAT o^{TYP}, $MANN^{TYP}$, $MEANS_s^{OPT}$, TIME, LOC …)]

Prototipični vezljivostni vzorec glagolskega pomena po vzoru sistema Vallex predvideva tudi podatek o sklonu (ACT – imenovalnik, REC - tožilnik), ter tipična predložna in vezniška sredstva, s pomočjo katerih se tipično realizira udeleženska vloga (*prepričati, da …; o čem; v kaj*), kar potrjujejo tudi korpusni zgledi:

Morilcu ACT je uspelo prepričati pravosodne oblasti REC

Grofica ACT ga REC je prepričala, da bo Henrik umrl RESLT
Njegova naloga je, da človeštvo REC prepriča o blagoslovu gensko manipuliranih organizmov PAT

7.2.4.2 Pomensko-skladenjsko izhodišče: skladenjske strukture

Druge pomembne vsebinske izboljšave korpusa temeljijo na analizi nekaterih problematičnih skladenjskih struktur in poenotenju odločitev v povezavi z označevanjem skladenjskega nivoja.

Skladenjsko enakovredni stavki

Sem najprej sodi poenotenje in uskladitev opredeljevanja udeležencev oz. razmerja med njimi v skladenjsko enakovrednih stavkih tipa: *kdo ali kaj je kdo ali kaj*. Na podlagi smernic predhodnega semantičnega označevanja učnega korpusa ssj500k smo z udeleženso 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 PAT (»prizadeto«): *območje medenice*(ACT) *je središče telesa*(PAT); *problem beguncev*(ACT) *je stvar države*(PAT); *naprava BICOM*(ACT) *je zbirka impulzov*(PAT). Glede na omenjena izhodišča smo že na ravni prvotnega označevanja učnega korpusa 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. neponopomenski vlogi glagola *biti*, ki odloča med osebkovo in povedkovodoločilno vlogo na skladenjski ravni. V zvezi s tem smo pri nadgradnji korpusa v okviru

Operacijo Razvoj slovenščine v digitalnem okolju 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.

projekta RSDO sprejeli odločitev, da v skladenjsko enakovrednih stavkih, 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 sistemu označevanja na skladenjski ravni, kjer se temu, kar je na pomenski ravni aktant, pripisuje odvisni del povedka: 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 semantični ravni na pribl. 90 % stavkov 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 konsens za v prihodnje.

Udeleženske vloge glagola *biti*

S to 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 stavkih, s čimer smo želeli doseči enotnost označevanja pri nakaterih najfrekventnejših semantičnih vzorcih.

Agentne in deagentne rabe (aktiv-pasiv)

Prav tako so bile deloma poenotene odločitve - aplicirane na korpusne stavke v približno 80 %, tudi na tistem delu korpusa, ki je bil predhodno ročno že označen - pri razumevanju posebnosti, ki se nanašajo na agentne (tvorne) in deagentne (trpne) rabe. 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 (pričakovane) 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

Operacijo Razvoj slovenščine v digitalnem okolju 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.

smiselno upoštevati tudi nekatere neenotnosti v pomenski interpretaciji, ki niso bile sistematično odpravljene, npr. *med njimi so se širile govorice (ACT) : potem je začela širiti govorice (PAT)*.

7.3 Smernice in priporočila za nadaljnje delo

Poročilo zaključujemo s smernicami in priporočili za izboljšavo korpusa na več ravneh: z aktualizacijo semantičnih oznak glede na označevalni sistem praške odvisnostne drevesnice; 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, ELEXIS).

Znotraj Praške odvisnostne drevesnice je bilo od različice PDT 2.0 izvedenih več nadgradenj,[27] kar se kaže tudi v dodajanju novih semantičnih vlog, npr. OBST (ovira): *Z roko je zadela ob mizo*; RCMP (zamenjava): *Otrokom je kupil za krono bonbonov*.; SUBS (Substitution): *Plačal mu je (znesek) za pripravljenost*, ki jih v slovenskem sistemu nimamo. Glede na analizo ročnega označevanja se je npr. pokazala potreba po ponovni uvedbi oznake CHPR, ki označuje zveze pomensko izpraznjenega glagola in samostalnika, npr. *Imeti načrt*. V slovenskem sistemu je obveljala odločitev, da tovrstne udeležence označujemo s prizadeto (PAT), s tem pa smo izgubili možnost prepoznavanja posebnega tipa večbesednih enot, t. i. zvez z glagoli z oslabljenim pomenom, ki so bile prepoznane kot samostojen podtip glagolskih večbesednih enot tudi v okviru cost akcije Parseme (Ramisch et al. 2018) in analizirane v slovenskem jeziku (Gantar et al. 2019).

Analiza pomenskih razmerij v glagolskem vzorcu ja pokazala tudi potrebo po razlikovanju med prejemniki (REC) in koristniki (BEN), npr. pri glagolih *ohraniti, braniti, varovati, ščititi*, npr. *ženo (BEN) je branil pred njegovimi izbruhi (PAT)*. Smiselno pa bi bilo v celoti izvesti tudi spremembo oznake MWPRED v ACT v zvezah glagola *biti* in nedoločnika, kjer je nedoločnik v vlogi aktanta: *zagotoviti (ACT) je potrebno; paziti (ACT) je treba*. S tem bi bila semantična raven usklajena tudi s skladenjskim nivojem.

Pri nadaljnji nadgradnji učnega korpusa bi bilo smiselno upoštevati tudi nekatere neenotnosti, ki niso bile sistematično odpravljene, smo jih pa zaznali pri ročnem označevanju. Sem sodi neenotna interpretacija oznak ACT in REC pri aktivno pasivnih razmerjih, npr. *pogosto se ji (REC) je sanjalo : kdo*

[27] Prvotna različica slovenskega sistema je izhajala iz različice Vallex 3.0, trenutno je v izgradnji Vallex 4.5.

Operacijo Razvoj slovenščine v digitalnem okolju 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.

(ACT) sanja; Nadalje neenotno označevanje t. i. lokacijskih metaforičnih pomenov, npr. *v osnovni šoli (LOC) bodo pripravili srečanje : v bolnišnici (ACT) bodo uvedli šolo za starše.* Neenotna je tudi interpretacija kazalnih in nedoločnih zaimkov z vidika pripisovanja udel. vloge ali ne. Načeloma naj bi odločitve sledile ločevanju med referenčnostjo – brez oznake in argumentom, tj. udeležencem – z oznako, kar pa ni sistematično pregledano.

Temeljna vrednost korpusa SUK z vidika označenosti na semantični ravni je predvsem v zagotavljanju možnosti vključevanja vira v večjezikovne projekte, ki se ukvarjajo s povezovanjem semantičnih konceptov in iskanjem semantičnih univerzalij. Posebej uporaben v tem smislu je korpus ELEXIS-WSD (na semantični ravni podkorpus SRL-3), ki je nastal v okviru projekta ELEXIS in je v korpus SUK vključen zaradi večnivojske označenosti v desetih jezikih). Prav tako bo ta segment učnega korpusa uporabljen znotraj evropske Cost akcije UniDive, in sicer v delovnih sklopih, ki se ukvarjajo z označevanjem korpusov in oblikovanjem leksikonskih in korpusnih vmesnikov.

8 David Bordon, Nejc Robida, Slavko Žitnik, Tina Munda, Špela Arhar Holdt: Imenske entitete in koreference

8.1 Imenske entitete

8.1.1 Obseg in namen označevanja

V okviru projekta so bile v 20.166 povedih oz. v 96,31 % novega gradiva ročno pregledane imenske entitete. Pri delu smo sledili označevalnim smernicam Zupan et al.[28] (2017). Korpus SentiCoref 1.0 je že vseboval strojno pripisane oznake za imenske entitete, vendar te niso bile dosledno pripisane, zato je bil pred vključitvijo v SUK nujen celovit ponoven pregled. Pri množici ELEXIS-WSD imenske entitete še niso bile označene.

Na obeh množicah so bile imenske entitete predoznačene z orodjem CLASSLA-Stanza[29], pri čemer smo v korpusu SentiCoref za pohitritev označevalnih odločitev posebej nakazali tudi predhodno obstoječe oznake. Sledil je ročni pregled obeh podmnožic. Kampanja pregledovanja je potekala v spletnem okolju

[28] https://wiki.cjvt.si/books/imenske-entitete/page/oznacevalne-smernice
[29] https://github.com/clarinsi/classla

Operacijo Razvoj slovenščine v digitalnem okolju 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.

INCEpTION, ki je enostavno za uporabo, hkrati pa nudi dober pregled nad že opravljenim delom. Pregled podmnožice ELEXIS-WSD je potekal nekoliko počasneje, ker je bila vsaka od 2024 povedi samostojna datoteka, ki jo je strežnik moral posamično naložiti.

Gradivo so pod vodstvom koordinatorja pregledovale tri študentke jezikoslovnih smeri. Pred začetkom pregledovanja so jim bile predstavljene smernice in ponazorjeno delo z orodjem, slednje v obliki videoposnetka. Vsako datoteko/poved so pregledale vse tri študentke, neujemanja med pripisanimi oznakami pa so bile naslovljene v fazi kuracije, kjer jih je koordinator kampanje znova obravnaval in jim pripisal končno oznako.

8.1.2 Označevalne dileme in rešitve

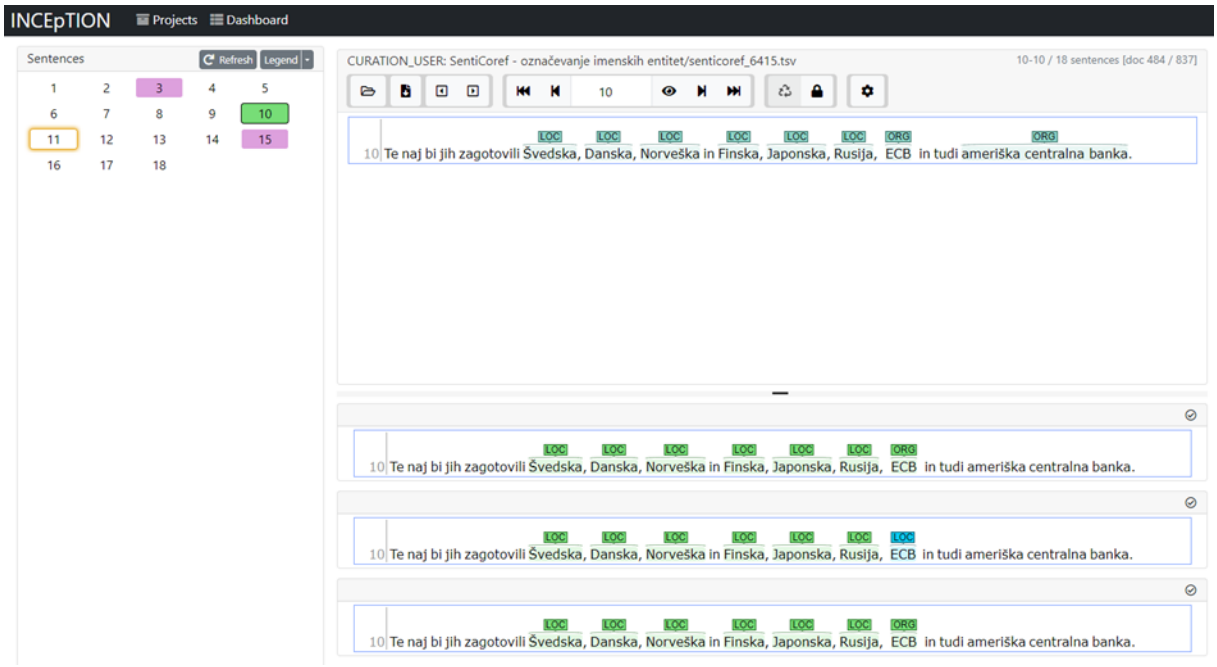
Poleg v smernicah predvidenih oznak PER (osebna lastna imena), DERIV-PER (svojilni pridevniki iz osebnih lastnih imen), ORG (organizacije), LOC (zemljepisna lastna imena) in MISC (stvarna lastna imena) se je pri obravnavi dilem pokazal manko za oznake pridevnikov iz stvarnih lastnih imen (*Mercatorjev*), za katere bi po vzoru DERIV-PER lahko uvedli oznako DERIV-ORG. Enako velja za svojilne pridevnike iz entitet z oznako LOC (*Lunin*), ki bi jim lahko pripisali oznako DERIV-LOC. To bi predstavljalo radikalnejši poseg v obstoječe smernice, ki ga bi ga bilo smotrno v prihodnje premisliti.

Še en pomislek, izražen med pregledovanjem, je zadeval ženske oblike priimkov, ki so morfološko svojilni pridevniki, tvorjeni iz moških priimkov (*Kresalova*). Morfološko so resda pridevniki in bi jim po tem kriteriju prisodili oznako DERIV-PER, ker pa pomensko ustrezajo priimku, je bila tovrstnim primerom pripisana oznaka PER. Kot problematično se je izkazalo tudi

Kot problematično se je izkazalo tudi določanje začetka imenske entitete v primerih, kjer je prvi del (uradnega) imena organizacije zapisan z malo začetnico, ker ga pisec besedila smatra za vrstno poimenovanje (npr. *občina Gornja Radgona*). Obveljalo je pravilo, da je glavni kazalnik, da celo enoto označimo kot imensko entiteto, velika začetnica ([*Občina Gornja Radgona*]ORG). Vendar pa kot imensko entiteto obravnavamo tudi primere, ki so zapisani z malo začetnico, a vsebujejo vse elemente uradnega imena te institucije ([*ameriška centralne banka*]ORG; uradno slovensko poimenovanje: 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 (*Karavanški predor*; uradno

Operacijo Razvoj slovenščine v digitalnem okolju 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.

ime: predor Karavanke). Vezano na slednje je problematika označevalnega sistema, ki predpostavlja, da pisci besedil vedno upoštevajo pravopisna pravila, a v praksi to ne drži. Na to smo nakazali že v poglavju o označevanju lem in MSD-oznak (gl. razdelek 4.3).



Slika 3: Prikaz faze kuracije v orodju INCEpTION (spodnje vrstice: odločitve treh pregledovalk, zgornja vrstica: končna odločitev kuratorja)

8.2 Koreference

8.2.1 Obseg in namen označevanja

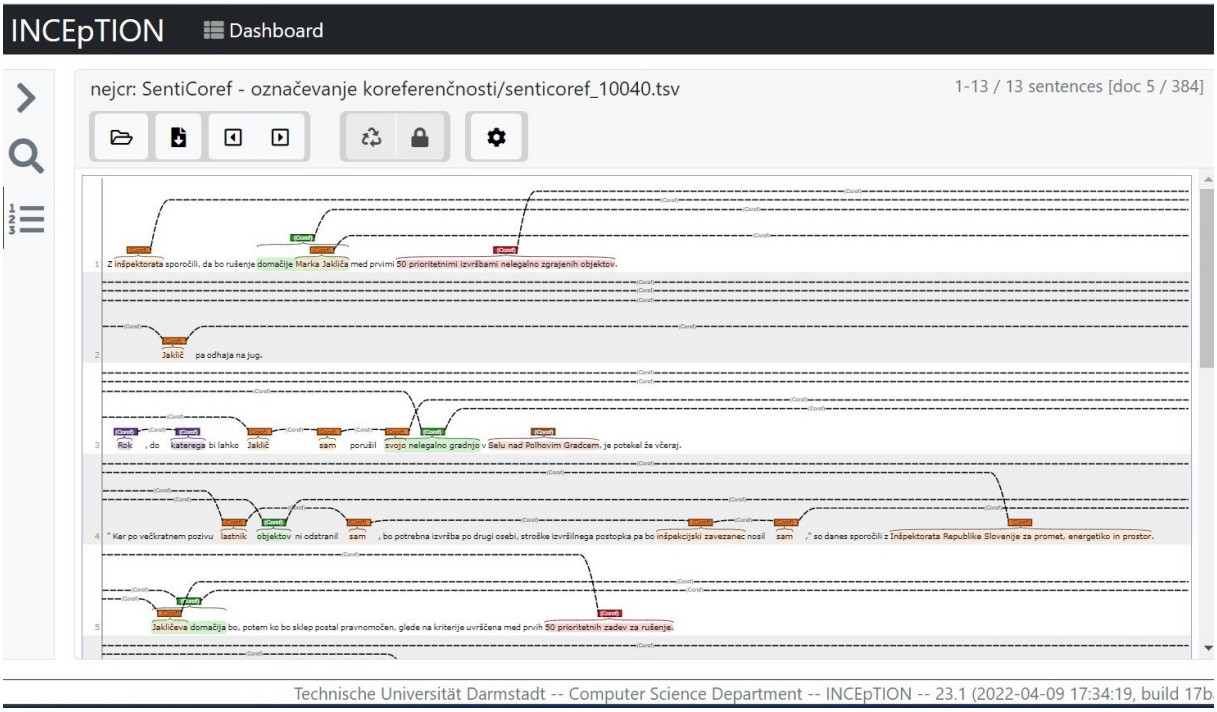
Koreference so bile v okviru projekta označene v 837 besedilih korpusa SentiCoref, v katerih je 18.142 povedi oz. 391.962 pojavnic, kar predstavlja 89,21 % novega gradiva. Množica SentiCoref je trenutno edina primerna za označevanje koreferenčnosti, saj vsebuje celotna besedila, za razliko od drugega vključenega gradiva, ki je razdeljeno na odstavke ali krajše enote.

Gradivo je bilo s koreferencami označeno že predhodno (Žitnik in Bajec 2018), vendar se je izkazalo, da bi bilo označevalni sistem treba nadgraditi, da bo bolje prilagojen značilnostim slovanskih jezikov, ki referenčnost pogosteje izražajo tudi morfemsko. Zato smo se odločili za označevanje uporabiti

Operacijo Razvoj slovenščine v digitalnem okolju 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.

označevalne smernice RELDI: Uputstvo za anotiranje koreferenci, ki so nastale za srbščino v sklopu iniciative RELDI 2008. Smernice smo prevedli v slovenščino, jih uredili in prilagodili, pri čemer je bila najpomembnejša odločitev, da za razliko od srbske kampanje na ravni koreferenc ne označujemo skladenjskih značilnosti - za slovenščino so te informacije namreč že (dosledneje in celoviteje) pokrite z oblikoskladenjskimi in skladenjskimi oznakami. Smernice so bile pripravljene v *Google Dokumenti*, kjer smo lahko sproti beležili opažanja in odpravljali nejasnosti ter pomanjkljivosti. Končna različica smernic je na voljo na portalu Wiki CJVT[30].

Tako kot kampanja označevanja imenskih entitet je tudi to označevanje potekalo na platformi INCEpTION. Gradivo sta pregledovala dva raziskovalca, izmed katerih je eden kampanjo tudi koordiniral. Osnovne dileme so bile večinoma razrešene v uvajalni fazi, nekaj pa tudi pozneje.



Slika 4: Označevanje koreferenc v orodju INCEpTION

[30] https://wiki.cjvt.si/books/odkrivanje-koreferencnosti/page/oznacevalne-smernice

Operacijo Razvoj slovenščine v digitalnem okolju 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.

8.2.2 Označevalne dileme in rešitve

Glavne konceptualne dileme so bile pri označevanju koreferenčnosti, kadar je bila potencialna povezava zanikana ali je bil o njej izražen dvom. Predvsem se je to kot problem pokazalo v npr. novinarskih besedilih, ko je bila celotna novica spisana s poudarjenim dvomom, ko je šlo za govorice, namige in podobno. Pravilo v smernicah, da se koreferenčnosti ob dvomu o povezavi znotraj samega besedila ali njenem zanikanju ne označuje, se je izkazalo za neizvedljivo in nam je posledično povzročalo precej težav. Pri nadaljnjem razvoju označevalnega sistema bo ta izziv treba upoštevati.

Dogovoriti smo se morali tudi o vrstnem redu označevanja, saj smo lahko označili samo eno omenitev istega referenta znotraj enega stavka. Zato smo se odločili, da imajo prednost samostalniki, nato zaimki itd. Nekaj težav je bilo tudi pri veriženju vseh izpeljanih pridevnikov iz lastnoimenskih samostalnikov, saj je prihajalo do kopičenja koreferenčnih povezav.

Predvsem pa smo morali na začetku v smernicah uskladiti terminologijo, odločili smo se, da bomo kot osnovne termine uporabljali: koreferenčnost, koreferenca in omenitev. Prav tako smo iz izvornih smernic odstranili velik del gradiva, ki je primerjalo označevalni sistem z alternativnimi pogledi na koreferenčnost ter poskrbeli za natančno členjenost vseh poglavij ter označenost zgledov. To je precej olajšalo našo komunikacijo in poenostavilo smernice, pri katerih je ključno jasno in nedvoumno predstaviti izbrani označevalni sistem in omogočiti, da se da na navodila jasno referirati.

Nekatere dele izvornih smernic, ki za slovenščino niso bili relevantni, smo umestili na konec smernic, opozorili, da v kampanji niso bili upoštevani, vendar trenutno ostajajo, saj prinašajo dragocene zglede označevanja. V nadaljnjih kampanjah je smiselno evalvirati uspešnost in ustreznost posameznih označevalnih odločitev in smernice še enkrat posodobiti. Pri tej posodobitvi bi veljalo trenutne zglede nadomestiti z realnimi zgledi iz korpusnih besedil, saj označevalna praksa razkije številne izzive, na katere teoretične smernice ne dajejo natančnih odogovorov.

Po končnem pregledu in kontroli kvalitete je bila označena množica izvožena v formati TSV in vključena v SUK.

Operacijo Razvoj slovenščine v digitalnem okolju 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.

9 Luka Terčon, Špela Arhar Holdt: Ocena uspešnosti in strategija nadaljnjega razvoja

9.1 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,[31] ki je bilo v pretekli različici že naučeno na učnem korpusu ssj500k (Ljubešić in Dobrovoljc 2019).

Kot pripravo na učenje modelov smo najprej izvedli delitev korpusa SUK na učno, validacijsko in testno podmnožico v razmerju 8:1:1.[32] Na učni in validacijski množici 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 udeleženjskih vlog. Naučene modele smo nato ovrednotili na testni množici. Rezultati evalvacije modelov so prikazani v Tabeli 7.[33]

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

Tabela 7: Uspešnost modelov za vsako označevalno raven

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

[31] https://pypi.org/project/classla/

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

[33] Tabela prikazuje rezultate modelov, ki so pri učenju in evalvaciji za napovedovanje oznak uporabljali tudi novo različico slovenskega oblikoslovnega leksikona Sloleks 3.0 (Čibej et al. 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.

Operacijo Razvoj slovenščine v digitalnem okolju 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.

Nekoliko nižje rezultate od ostalih dosega model za označevanje udeleženjskih vlog, za katerega je bilo v učni množici na voljo najmanj učnih podatkov od vseh zgoraj omenjenih označevalnih ravni. Vsi ostali modeli dosegajo F1 vrednosti nad 90. Vsi jezikovni modeli so že vključeni v najnovejšo različico orodja CLASSLA-Stanza kot privzeti jezikovni modeli za označevanje standardne slovenščine.

9.2 Strategija nadaljnjega razvoja

Modele strojnega učenja, ki jih vsebuje cevovod za jezikoslovno označevanje, je potrebno za njihovo delovanje naučiti na človeško označenih učnih množicah. V okviru projekta RSDO sta bila nadgrajena učni korpus ssj500k (po novem Slovenski učni korpus *SUK 1.0*) za učenje označevanja standardnega jezika ter učni korpus Janes-Tag (po novem *Janes-Tag 2.0*) za učenje označevanja nestandardnega jezika. Za njen nadaljnji razvoj in nadaljnji razvoj jezikoslovne obdelave je potrebno naslednje.

- Kontinuirane izboljšave učnih korpusov SUK in Janes-Tag

  Učna korpusa SUK in Janes-Tag sta temeljna učna korpusa za označevalne modele strojnega učenja, zato ju je potrebno kontinuirano izboljševati in dopolnjevati z novimi primeri ter novimi označevalnimi ravninami. Posodabljati je potrebno tudi smernice za njuno označevanje ter tako smernice kot učna korpusa usklajevati z razvojem mednarodnih standardov.
- Učne množice za sistematično popravljanje napak

  Pri nekaterih nalogah jezikovne obdelave prihaja do napak, ki so posledica premajhne vsebovanosti nekaterih jezikovnih fenomenov v učnih množicah (redkejše dvoumne besedne oblike, redkeje zastopane kategorije, kot sta srednji spol, dvojina in podobno). Na podlagi kvalitativne analize modelov je k odpravljanju zaznanih pomanjkljivosti potrebno pristopiti sistematično in dopolniti učne množice s primeri, ki naslavljajo te pomanjkljivosti. Dopolnjene učne množice se po evalvaciji vključijo v osnovne učne korpuse, npr. SUK ali Janes-Tag.
- Učne množice za specifične jezikovne potrebe

  Trenutne učne množice (SUK 1.0), pokrivajo predvsem sodobni standardni pisni jezik. Treba je pokriti tudi druge oblike jezikovne rabe in zanje pripraviti ločene manjše učne množice, ki bodo omogočale podporo jezikoslovnemu cevovodu za:
  - starejši jezik od 16. pa do konca 19. stoletja, kot tudi slabo pokriti del 20. stoletja do leta 1900 do 1990, npr. specifično jugoslovansko besedišče;

Operacijo Razvoj slovenščine v digitalnem okolju 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.

  - narečni jezik, kjer je potrebno določiti vrstni red, po katerem se bodo obravnavala posamezna narečna besedišča;
  - zamejski jezik, npr. jeziki slovenskih skupnosti na avstrijskem Koroškem, Italiji, Argentini in Severni Ameriki;
  - govorjeni jezik, ki je za nekatere naloge, npr. segmentacijo stavkov in skladenjsko razčlenjevanje, bistveno težji od standardnega jezika;
  - določene domene pisnega jezika, ki se glede na raziskave jezikovnih značilnosti pomembneje razlikujejo od splošne rabe, npr. pravna in uradovalna besedila, znanstveni jezik, pisanje učečih se.

- Učne množice za označevanje semantičnih ravnin jezika

  Dosedanje učne obstoječe množice v zelo omejeni obliki pokrivajo nekatere naprednejše semantične koncepte, kot so relacije, koreference, razdvoumljanje, metafore itd. Tudi te nivoje je potrebno pokriti s kakovostnimi in dovolj velikimi učnimi množicami.

- Učne množice za označevanje diskurznih ravnin jezika

  Dosedanje učne obstoječe množice ne pokrivajo diskurznih ravnin jezika, kot je npr. ugotavljanja mnenj v diskurzu, povzemanje diskurza in tudi posebni lastnostih spletnega diskurza, kot so oblike žaljivega in sovražnega govora. Zanje je potrebno po zgledu mednarodnih iniciativ pripraviti nove učne množice.

- Podpora učnih množic kakovostni in javno dostopni evalvaciji

  Za evalvacijo orodij za jezikoslovno označevanje smo zgradili portal SloBENCH, ki podatke za vsako evalvacijsko nalogo razbije na javni in skriti del. Javni del učne množice je na voljo javnosti za učenje modelov, pri skritem delu pa so na voljo le učni primeri brez oznak. Manjkajoče oznake napovedo naučeni modeli, sistem SloBENCH pa jih primerja s skritimi pravilnimi oznakami in na tej podlagi ovrednoti in rangira različne pristope k danemu problemu. Vključitev v SloBENCH in podobno delitev na javni in skriti del za potrebe evalvacije je potrebno zagotoviti tudi za vse nove javno financirane učne množice. V SloBENCH je potrebno vključiti tudi tiste ravnine jezikovnega označevanja, ki niso bile pokrite v projektu RSDO, npr. JOS-SYN skladnja in udeleženske vloge.

- Harmonizacija smernic za jezikoslovno označevanje

  Priprava učnih korpusov na RSDO je razkrila določene nekonsistentnosti med različnimi nivoji

Operacijo Razvoj slovenščine v digitalnem okolju 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.

jezikoslovnega označevanja, ki jih je v nadaljevanju treba jezikoslovno raziskati in odpraviti. Obstoječe smernice je potrebno poenotiti, z njimi uskladiti obstoječe učne množice in jih vključiti v enotne povečane učne množice.

- Mednarodna usklajenost in sodelovanje

  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 evalvacijskih množic.

10 Literatura

Arhar Holdt et al. 2022 = Arhar Holdt, Špela et al., 2022, Training corpus SUK 1.0, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1747.

Arhar Holdt et al. 2016 = Arhar Holdt, Špela; Fišer, Darja; Erjavec, Tomaž in Krek, Simon. Syntactic annotation of Slovene CMC: first steps. **V:** *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities*, 27.–28. september 2016, Ljubljana, Slovenia, 2016, str. 3–6. http://nl.ijs.si/janes/cmc-corpora2016/proceedings/.

Backer et al. 1998 = Backer, Collin F., Charles J. Fillmore in John B. Lowe. 1998. The Berkeley FrameNet project. Proceedings of the COLING-ACL. Montreal, Canada. 86–90.

Bjorkelund et al. 2009 = Bjorkelund, Anders; Hafdell, Love in Nugues, Pierre. 2009. Multilingual semantic role labeling. V: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning* (CoNLL 2009): Shared Task. 43–48, Boulder, Colorado, June. Association for Computational Linguistics.

Brank 2022 = Brank, Janez, 2022, *Q-CAT Corpus Annotation Tool 1.4*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1684.

Operacijo Razvoj slovenščine v digitalnem okolju 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.

Christensen et al. 2011 = Christensen, Janara, Mausam, Stephen Soderland, and Oren Etzioni. 2011. *An Analysis of Open Information Extraction based on Semantic Role Labeling*. International Conference on Knowledge Capture (KCAP), str. 113–120. Banff, Alberta, Canada.

Čibej et al. 2022 = Čibej, Jaka et al. *Morphological lexicon Sloleks 3.0*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1745.

Čibej et al. 2018 = Čibej, Jaka; Arhar Holdt, Špela; Erjavec, Tomaž in Fišer, Darja. 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.

de Marneffe et al. 2021 = de Marneffe, Marie-Catherine; Manning, Christopher D.; Nivre, Joakim in Zeman, Daniel (2021). Universal Dependencies. Computational Linguistics, 47(2): 255–308. https://doi.org/10.1162/coli_a_00402.

Dobrovoljc in Ljubešić 2022 = Dobrovoljc, Kaja in Ljubešić, Nikola (2022). Extending the SSJ Universal Dependencies Treebank for Slovenian: Was It Worth It? **V:** *Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) within LREC2022*, str. 15–22. https://aclanthology.org/2022.law-1.3.

Dobrovoljc in Nivre 2016 = Dobrovoljc, Kaja in Nivre, Joakim (2016). The Universal Dependencies Treebank of Spoken Slovenian. **V:** P*roceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC'16), str. 1566–1573. https://aclanthology.org/L16-1248.

Dobrovoljc et al. 2022 = Dobrovoljc, Kaja; Terčon, Luka in Ljubešić, Nikola. (2022). Universal Dependencies za slovenščino: nadgradnja smernic, učnih podatkov in razčlenjevalnega modela. V: D. Fišer in T. Erjavec (ur.), *Jezikovne tehnologije in digitalna humanistika: zbornik konference*, str. 30–39. Inštitut za novejšo zgodovino. https://nl.ijs.si/jtdh22/pdf/JTDH2022_Dobrovoljc-et-al_Universal-Dependencies-za-slovenscino.pdf.

Operacijo Razvoj slovenščine v digitalnem okolju 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.

Dobrovoljc et al. 2017 = Dobrovoljc, Kaja; Erjavec, Tomaž in Krek, Simon. (2017). The Universal Dependencies Treebank for Slovenian. **V:** *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, str. 33–38. https://doi.org/10.18653/v1/W17-1406.

Dobrovoljc et al. 2012 = Dobrovoljc, Kaja; Krek, Simon in Rupnik, Jan. 2012. Skladenjski razčlenjevalnik za slovenščino. V: *Proceedings of the 8th Language Technologies Conference, volume C*, str. 42–47, Ljubljana, Slovenia, Oktober. IJS.

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

Erjavec et al. 2019 = Erjavec, Tomaž et al., 2019, *CMC training corpus Janes-Tag 2.1*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1238.

Erjavec et al. 2010 = Erjavec, Tomaž; Fišer, Darja; Krek, Simon in Ledinek, Nina. 2010. The JOS Linguistically Tagged Corpus of Slovene. V: *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valeta, Malta, Maj. European Language Resources Association (ELRA).

Filko et al. 2012 = Filko, Matea, Farkaš, Daša in Merkler, Danijela. 2012. SRL Tagset for Croatian. Institute of Linguistics, Faculty of Humanities and Social Sciences, Zagreb. http://hobs.ffzg.hr/static/docs/SRL_tagset.pdf.

Gantar v tisku = Gantar, Polona. Analiza udeleženskih vlog s skladenjskega, pomenskega in leksikalnega vidika (v tisku). Zbornik Simpozij o skladnji.

Gantar 2021 = Gantar, Polona. Strojno berljiv Vezljivostni leksikon slovenskih glagolov. V: Arhar Holdt, Špela (ur.). Nova slovnica sodobne standardne slovenščine : viri in metode. 1. izd. Ljubljana: Znanstvena založba Filozofske fakultete, 2021. Str. 259-297, ilustr. Zbirka Sporazumevanje. ISBN 978-961-06-0547-8. ISSN 2738-4527. https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/325/477/7313-1.

Gantar et al. 2019 = Gantar, Polona, Čibej, Jaka in Bon, Mija. Slovene multi-word units: identification, categorization, and representation. V: Corpas Pastor, Gloria (ur.), Mitkov, Ruslan (ur.). Computational

Operacijo Razvoj slovenščine v digitalnem okolju 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.

and corpus-based phraseology: Third International Conference, Europhras 2019, Malaga, Spain, September 25-27, 2019: proceedings. Cham: Springer, cop. 2019. Str. 99-112, ilustr. Lecture notes in computer science (Internet), 11755, Lecture notes in artificial intelligence.

Gantar et al. 2018 = Gantar, Polona; Strkalj Despot, Kristina; Krek, Simon in Ljubešić, Nikola. 2018. Towards semantic role labeling in Slovene and Croatian. V: FIŠER, Darja (ur.), PANČUR, Andrej (ur.). Zbornik konference Jezikovne tehnologije in digitalna humanistika, 20. september - 21. september 2018, Ljubljana, Slovenija. 1. izdaja. Ljubljana: Znanstvena založba Filozofske fakultete, 2018. Str. 93-98, ilustr. ISBN 978-961-06-0111-1. http://nl.ijs.si/jtdh18/JTDH-2018-Proceedings.pdf.

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

Kipper et al. 2006 = Kipper, Karin, Anna Korhonen, Neville Ryant in Martha Palmer. 2006. Extensive Classifications of English verbs. Proceedings of the 12th EURALEX International Congress. Turin, Italy. September. 1–15.

Klemen in Žitnik 2022 = Klemen, Matej in Žitnik, Slavko. *Neural coreference resolution for Slovene language*. Computer Science and Information Systems 2022 Volume 19, Issue 2, 495-521. https://doi.org/10.2298/CSIS201120060K.

Krek et al. 2021 = Krek, Simon et al., 2021, Training corpus ssj500k 2.3, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1434.

Krek et al. 2020 = Krek, Simon; Erjavec, Tomaž; Dobrovoljc, Kaja; Gantar, Polona; Arhar Holdt, Špela; Čibej, Jaka in Brank, Janez. The ssj500k training corpus for Slovene language processing. V: Fišer, D. in Erjavec, T. *Jezikovne tehnologije in digitalna humanistika: zbornik konference: 24.-25. september 2020, Ljubljana, Slovenija.* Ljubljana: Inštitut za novejšo zgodovino, 2020. Str. 24–33. http://nl.ijs.si/jtdh20/pdf/JT-DH_2020_Krek-et-al_The-ssj500k-Training-Corpus-for-Slovene-Language-Processing.pdf.

Operacijo Razvoj slovenščine v digitalnem okolju 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.

Krek et al. 2016 = Krek, Simon, Gantar, Polona, Dobrovoljc, Kaja, Škrjanec, Iza. Označevanje udeleženskih vlog v učnem korpusu za slovenščino. V: Erjavec, Tomaž (ur.), Fišer, Darja (ur.). *Zbornik konference Jezikovne tehnologije in digitalna humanistika*, 29. september-1. oktober 2016, Filozofska fakulteta, Univerza v Ljubljani, Ljubljana, Slovenija. 1. izd. V Ljubljani: Znanstvena založba Filozofske fakultete, 2016. Str. 106-110, ilustr. http://www.sdjt.si/wp/wp-content/uploads/2016/09/JTDH-2016_Krek-et-al_Oznacevanje-udelezenskih-vlog-v-ucnem-korpusu-za-slovenscino.pdf.

Lenardič et al. 2022 = Lenardič, Jakob et al., 2022, CMC training corpus Janes-Tag 3.0, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1732.

Ljubešić in Dobrovoljc 2019 = Ljubešić, Nikola in Dobrovoljc, Kaja. 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. Association for Computational Linguistics. https://aclanthology.org/W19-3704.

Martelli et al. 2021 = Martelli, Federico; Navigli, Roberto; Krek, Simon; Tiberius, Carole; Kallas, Jelena; Gantar, Polona; Koeva, Svetla; Nimb, Sanni; Pedersen, Bolette Sandford; Olsen, Sussi; Langements, Margit; Koppel, Kristina; Üksik, Tiiu; Dobrovolijc, Kaja; Ureña-Ruiz, Rafael-J.; Sancho-Sánchez, José-Luis; Lipp, Veronika; Varadi, Tamas; Györffy, András … Munda, Tina. (2021). Designing the ELEXIS Parallel Sense-Annotated Dataset in 10 European Languages. V: Kosem, I. in Cukr, M. *eLex 2021 Proceedings: Proceedings of the eLex 2021 conference.* Lexical Computing CZ. Str. 377-395. https://elex.link/elex2021/wp-content/uploads/2021/08/eLex_2021_33_pp514-528.pdf.

Mikulová et al. 2006 = Mikulová, Marie et al. 2006. Annotation on the tectogrammatical level in the Prague Dependency Treebank. Aannotation manual. Technical Report 30. 5–11.

Nivre in Fang 2017 = Nivre, Joakim in Fang, Chiao-Ting. 2017. Universal Dependency Evaluation. V: *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, str. 86–95, Gothenburg, Sweden. Association for Computational Linguistics. https://aclanthology.org/W17-0411/.

Operacijo Razvoj slovenščine v digitalnem okolju 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.

Palmer et al. 2005 = Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. Computational linguistics, 31(1): 71–106.

Pori et al. 2022 = Pori, Eva; Čibej, Jaka; Munda, Tina; Terčon, Luka in Arhar Holdt, Špela. Lematizacija in oblikoskladenjsko označevanje korpusa SentiCoref. V: Fišer, Darja (ur.), Erjavec, Tomaž (ur.). *Jezikovne tehnologije in digitalna humanistika: zbornik konference.* 15.–16. september 2022, Ljubljana, Slovenija. Ljubljana: Inštitut za novejšo zgodovino, 2022. Str. 162–168. https://nl.ijs.si/jtdh22/pdf/JTDH2022_Proceedings.pdf.

Ramisch et al. 2018 = Ramisch, Carlos, Gantar, Polona et al. Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. V: Proceedings : LAW-MWE-CxG 2018. The 12th Linguistic Annotation Workshop (LAW XII) and the 14th Workshop on Multiword Expressions (MWE 2018). Santa Fe: [s. n.], 2018. Str. 222-240

Schuurman et al. 2010 = Schuurman, Ineke, Véronique Hoste in Paola Monachesi. 2010. Interacting semantic layers of annotation in sonar, a reference corpus of contemporary written dutch. Proceedings of LREC'10, Valletta, Malta. ELRA. 2471–2477.

Shen in Lapata 2007 = Shen, Dan in Mirella Lapata. 2007. Using Semantic Roles to Improve Question Answering. *Proceedings of the Conference on Empirical Methods in Natural Language Processing and on Computational Natural Language Learning*. Prague. 12–21.

Taulé et al. 2011 = Taulé, Mariona, Antònia M. Martí in Oriol Borrega. 2011. AnCora 2.0: Argument Structure Guidelines for Catalan and Spanish. Working paper 4: TEXT-MESS 2.0 (Text-Knowledge 2.0). Universitat de Barcelona. Barcelona.

Toporišič 2004 = Toporišič, Jože. (2004): *Slovenska slovnica*. Maribor: Obzorja.

Zeman et al. 2022 = Zeman, Daniel et al., 2022, *Universal Dependencies 2.10*, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-4758.

Operacijo Razvoj slovenščine v digitalnem okolju 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.

Zupan et al. 2017 = Zupan, Katja; Ljubešić, Nikola in Erjavec, Tomaž. elektronski vir. 2017. *Smernice Janes-NER za označevanje imenskih entitet v slovenskem jeziku v1.1*. Dostopno na: https://wiki.cjvt.si/books/imenske-entitete/page/oznacevalne-smernice.

Žitnik in Bajec 2018 = Žitnik, Slavko. in Bajec, Marko (2018). Odkrivanje koreferenčnosti v slovenskem jeziku na označenih besedilih iz coref149. Slovenščina 2.0: Empirične, Aplikativne in Interdisciplinarne Raziskave, 6(1), 37–67. https://journals.uni-lj.si/slovenscina2/article/view/7967/8253.

Operacijo Razvoj slovenščine v digitalnem okolju 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.