Universal Dependencies za slovenščino: nove smernice, ročno označeni podatki in razčlenjevalni model

Kaja DOBROVOLJC
Filozofska fakulteta, Univerza v Ljubljani; Institut Jožef Stefan

Luka TERČON
Fakulteta za računalništvo in informatiko, Univerza v Ljubljani

Nikola LJUBEŠIĆ
Institut Jožef Stefan; Fakulteta za računalništvo in informatiko, Univerza v Ljubljani

Universal Dependencies (UD) je mednarodno usklajena označevalna shema za medjezikovno primerljivo oblikoslovno in skladenjsko označevanje besedil po načelih odvisnostne slovnice, ki je bila ob več kot 130 drugih svetovnih jezikih uspešno uporabljena tudi za označevanje besedil v slovenščini. V prispevku predstavimo rezultate nedavnih aktivnosti v povezavi s shemo UD znotraj projekta Razvoj slovenščine v digitalnem okolju, v okviru katerega smo obstoječo infrastrukturo nadgradili s prenovo in podrobno dokumentacijo označevalnih smernic UD za slovenščino, razširitvijo drevesnice SSJ-UD za pisno slovenščino z novimi povedmi iz korpusov ssj500k in ELEXIS-WSD, izdelavo testne množice iz besedil korpusa SentiCoref za spletni portal SloBENCH ter polavtomatsko pretvorbo oblikoslovnih oznak referenčnih učnih korpusov SUK in Janes-Tag. Na razširjeni drevesnici SSJ-UD je bil naučen tudi novi napovedni model za skladenjsko razčlenjevanje v orodju CLASSLA-Stanza, ki ga v prispevku v

Dobrovoljc, K., Terčon, L., Ljubešić, N.: Universal Dependencies za slovenščino: nove smernice, ročno označeni podatki in razčlenjevalni model. Slovenščina 2.0, 11(1): 218–246.

1.01 Izvirni znanstveni članek / Original Scientific Article

DOI: https://doi.org/10.4312/slo2.0.2023.1.218-246

podporo nadaljnjim jezikoslovnim aplikacijam podrobneje ovrednotimo z vidika splošne natančnosti razčlenjevanja in najpogostejših tipov napak.

Ključne besede: slovnično označeni korpusi, odvisnostna slovnica, drevesnica, skladenjsko razčlenjevanje, obdelava naravnega jezika

1 Uvod

Jezikoslovno označeni korpusi, tj. digitalizirane zbirke besedil, ki poleg besed na površini vsebujejo tudi ročno pripisane podatke o njihovih slovničnih lastnostih na različnih ravneh jezikoslovnega opisa (Ide in Pustejovsky, 2017), predstavljajo enega izmed temeljnih jezikovnih virov za razvoj jezikovnotehnoloških orodij na eni strani in korpusnojezikoslovne raziskave na drugi. Slovnične lastnosti so besedilom tipično pripisane na podlagi vnaprej opredeljenih označevalnih shem oz. označevalnih sistemov, ki poleg nabora možnih oznak običajno vsebujejo tudi smernice za njihovo pripisovanje konkretnim slovničnim pojavom. Ker so v preteklosti označevalne sheme nastajale ločeno za posamezne jezike, slovnične teorije ali celo korpuse, je njihova posledična raznolikost onemogočala kakršnokoli neposredno primerjavo označenih podatkov ali na njih temelječih računalniških orodij.

Kot protiutež tovrstni razdrobljenosti je bila leta 2013 vzpostavljena označevalna shema Universal Dependencies,[1] 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. Znotraj sheme UD 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. Shema temelji na načelih odvisnostne slovnice, ki je v primerjavi s frazno skladnjo bolj primerna za jezike z bolj fleksibilnim besednim redom in za neposredno uporabo v različnih jezikovnotehnoloških aplikacijah (Jurafsky in Martin, 2021), njena teoretična izhodišča pa so podrobneje predstavljena v prispevku De Marneffe idr. (2021).

1 https://universaldependencies.org

Doslej je bilo z označevalno shemo UD ročno označenih že več kot 240 korpusov (t.i. odvisnostnih drevesnic, angl. *dependency treebanks*) v 130 svetovnih jezikih. Med njimi sta tudi univerzalni odvisnostni drevesnici pisne slovenščine SSJ (Dobrovoljc idr., 2017) in govorjene slovenščine SST (Dobrovoljc in Nivre, 2016), ki sta bili s tem neposredno vključeni v razvoj številnih najsodobnejših orodij za večjezično obdelavo naravnih jezikov (Zeman idr., 2018), kakor tudi raznolike primerjalnojezikoslovne raziskave (Futrell idr., 2015; Naranjo in Becker, 2018; Chen in Gerdes, 2018).

Glede na pomen razvoja slovenskih virov v tovrstnih mednarodnih standardizacijskih pobudah smo v okviru nacionalnega projekta Razvoj slovenščine v digitalnem okolju (RSDO),[2] ki si prizadeva za zadovoljitev potreb po računalniških izdelkih in storitvah s področja jezikovnih tehnologij za slovenski jezik, obstoječe vire in povezano infrastrukturo za označevanje slovenskih besedil po sistemu Universal Dependencies bistveno nadgradili. Potek in rezultate te aktivnosti smo delno že predstavili v prispevku na konferenci Jezikovne tehnologije in digitalna humanistika 2022 (Dobrovoljc idr., 2022), v tem članku pa ga nadgradimo s posodobljeno in bolj poglobljeno analizo prvotnih rezultatov ter predstavitvijo novih povezanih podatkovnih množic.

V nadaljevanju članka tako po predstavitvi popisa najnovejših označevalnih smernic UD za slovenščino (2. razdelek) opišemo nastanek in vsebino štirih novih ročno označenih množic po sistemu Universal Dependencies (3. razdelek) – razširjene referenčne univerzalno skladenjsko razčlenjene drevesnice SSJ-UD, nove univerzalno skladenjsko razčlenjene drevesnice SloBENCH-UD ter novih univerzalno oblikoslovno označenih učnih korpusov SUK in Janes-Tag. Na novi različici korpusa SSJ-UD je bil naučen tudi novi napovedni model razčlenjevalnika CLASSLA-Stanza, ki ga v podporo nadaljnjim jezikoslovnim in jezikovnotehnološkim aplikacijam v 4. razdelku ovrednotimo z evalvacijo splošne natančnosti in kvalitativno analizo najpogostejših napak. Prispevek v 5. razdelku sklenemo s pregledom aktualne infrastrukturne podpore za jezikoslovne analize slovenskih UD korpusov in smernicami nadaljnjih raziskav.

2 https://slovenscina.eu/

2 Popis smernic UD za slovenščino

Splošne smernice UD, kakršne so dokumentirane na krovni spletni strani projekta,[3] so kot nadaljevanje predhodnih standardizacijskih iniciativ (Zeman, 2008; Petrov idr., 2012; de Marneffe idr., 2014) in večletnega kolaborativnega razvoja zasnovane tako, da skušajo na čim krajši način nasloviti oblikoslovne in skladenjske specifike čim širšega nabora jezikov. Tako v splošnih smernicah najdemo predvsem prototipične opredelitve posameznih oznak, opis najbolj tipičnih mejnih primerov in ponazoritve na primerih izbranih jezikov, naloga avtorjev drevesnic za posamezne jezike pa je, da te splošne smernice nato prenesejo na svoje konkretne jezikovne podatke. Pri tem infrastruktura UD omogoča, da se za vsak jezik ta načela popišejo kot jezikovnospecifične smernice na uradni spletni strani, vendar to ni obvezno, zato je dokumentacija označevalnih smernic UD za posamezne jezike prepuščena predvsem samoiniciativnosti avtorjev podatkov.

Za slovenščino so bile ob prvi objavi korpusa SSJ-UD (Dobrovoljc idr., 2017, gl. razdelek 3.1) tako dokumentirane zgolj smernice za pripisovanje besednih vrst in oblikoskladenjskih oznak, ki so odtlej ob prehodu s prve na drugo različico splošnih UD smernic (Nivre idr., 2020) že nekoliko zastarele. Po drugi strani smernice za pripisovanje univerzalnih skladenjskih relacij besedilom v slovenščini zaradi obsežnosti niso bile podrobneje dokumentirane oz. so bile razvidne zgolj implicitno iz pretvorbenih pravil na eni strani in označenosti objavljenega korpusa na drugi.

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 priročnika (Dobrovoljc in Terčon, 2023), 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 sprememb na mestih, kjer je bila prvotna označenost korpusa SSJ-UD nedosledna ali neustrezna glede na splošne, jezikovno univerzalne smernice. Med njimi lahko izpostavimo predvsem spremembe v obravnavi primerjalnih struktur (jedro strukture je pridevnik

3 https://universaldependencies.org/guidelines.html

ali prislov, ki izraža primerjano lastnost), poudarjalnih členkov (razlikovanje med modifikatorji samostalnikov na eni in povedkov na drugi strani), besedilnih povezovalcev (razlikovanje glede na stavčno pozicijo) in zaimkov *se/si* (razlikovanje med zaimkom v vlogi predmeta na eni strani in prostim morfemom na drugi strani), ki so bili zaradi omejitev strojne pretvorbe iz sistema JOS (gl. razdelek 3.1) prvotno označeni drugače kot predvidevajo splošne smernice UD.

Priročnik s smernicami UD za slovenščino poleg opisov posamičnih slovničnih kategorij in načel njihovega pripisovanja slovenskim besedilom vsebuje še razdelek s podrobnejšo obravnavo težavnejših primerov (Dobrovoljc idr., 2023), ki se je dopolnjeval tudi skozi označevalne kampanje, opisane v 3. razdelku. V nekoliko strnjeni oz. tuji javnosti prilagojeni obliki so bile v angleščino prevedene slovenske smernice nato objavljene še na uradni spletni strani projekta UD,[4] kar omogoča neposredno primerjavo s smernicami za več deset drugih svetovnih jezikov, zbranimi na istem spletnem mestu.

V procesu dokumentacije slovenskih smernic so bila identificirana tudi nekatera odprta vprašanja, pri katerih bi dosledna implementacija splošnih smernic UD zahtevala znaten odmik od doslej uveljavljenih označevalnih praks v slovenskem prostoru, zlasti sistema JOS, in bi jih bilo zato smiselno 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 skladenjske analize (npr. smiselnost oz. način 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.

4 Primer slovenskih smernic za pripisovanje oznake *nsubj* (samostalniški osebek) na krovnem portalu UD: https://universaldependencies.org/sl/dep/nsubj.html.

3 Novi ročno označeni korpusi UD za slovenščino

3.1 Nadgradnja univerzalno skladenjsko razčlenjenega korpusa SSJ-UD

Korpus SSJ-UD (Dobrovoljc idr., 2016; Dobrovoljc idr., 2017) je kot prvi univerzalno skladenjski korpus za slovenščino nastal na podlagi polavtomatske pretvorbe učnega korpusa ssj500k (Krek idr., 2020), prvotno označenega po shemi JOS (Erjavec idr., 2010). Za razliko od oblikoslovne ravni, ki jo je bilo mogoče s pravili za preslikavo iz enega v drug sistem pretvoriti v celoti,[5] je bilo zaradi robustnosti sistema za skladenjsko razčlenjevanje JOS v primerjavi s sistemom UD v celoti pretvorjenih zgolj 8.000 od izvorno 11.411 skladenjsko razčlenjenih povedi korpusa ssj500k, ki so bile nato kot korpus SSJ-UD prvič objavljene v zbirki UD v1.2.

Neobjavljene, polpretvorjene skladenjsko razčlenjene povedi korpusa ssj500k (razdelek 3.1.1) so tako predstavljale logično izhodišče za napovedano povečanje učnih podatkov znotraj projekta RSDO, ki sta mu sledili še razširitev s povedmi korpusa ELEXIS-WSD (3.1.2) in izboljšanje označenosti prvotnega korpusa (3.1.3). V vseh treh fazah je označevanje potekalo v označevalnem orodju Q-CAT (Brank, 2022), ki odslej podpira tudi uvoz korpusov v formatu CONLL-U, za primerjavo označenih datotek (kuriranje) pa smo uporabili lokalno inštalacijo orodja WebAnno (Eckart de Castilho idr., 2016), ki jo vzdržuje center CLARIN.SI (Erjavec idr., 2022).[6] Označevalni proces smo podrobneje že predstavili (Dobrovoljc in Ljubešić, 2022; Dobrovoljc idr., 2022), v nadaljevanju pa povzamemo zgolj najpomembnejše rezultate.

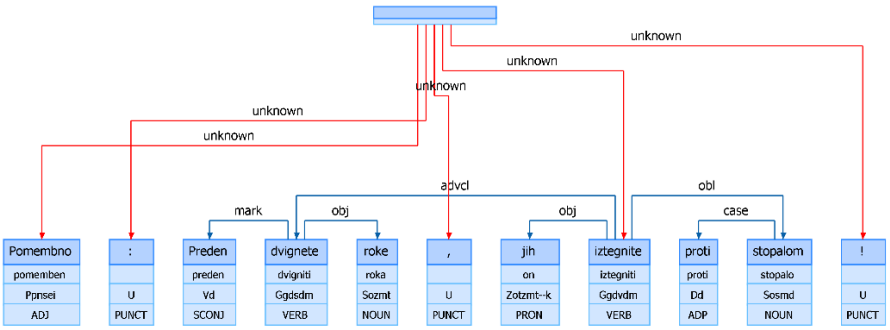
3.1.1 Razširitev s polpretvorjenimi povedmi iz korpusa ssj500k

V prvi fazi razširitve so označevalci ročno pregledali 3.411 polpretvorjenih povedi oz. 96.194 pojavnic korpusa ssj500k, med katerimi jih 22.377 (23,5 %) še ni imelo pripisane skladenjske relacije UD. Te so bile za potrebe lažje vizualizacije označene z relacijo *unknown* (Slika 1), označevalci (po dva na poved) pa so poleg ustvarjanja novih povezav preverjali tudi ustreznost že obstoječih (pretvorjenih) povezav.

5 Pravila in skripte za pretvorbo iz sistema JOS v UD so dokumentirana na povezavi https://github.com/clarinsi/jos2ud.

6 https://www.clarin.si/webanno/

Med pojavnicami, ki v izhodišču niso imele pripisane relacije UD, je bila skoraj polovica ločil (relacija *punct*), kar je bilo glede na pretvorbena pravila pričakovano, saj so bila ločila večinoma na relevantno jedro povezana šele po določitvi vseh drugih pojavnic v povedi, zlasti korena povedi (*root*), ki predstavlja tudi drugo najpogostejšo vrsto nepretvorjenih pojavnic (12 %). Tej sledita še relaciji *parataxis* (9 %) in *conj* (6 %), ki se uporabljata za povezovanje stavčnih soredij oz. priredij, torej struktur, kakršnih zgolj s pravili ni bilo mogoče pretvoriti z dovolj zanesljivo natančnostjo.



Slika 1: Primer prikaza polpretvorjene povedi iz ssj500k z manjkajočimi relacijami UD (unknown) v označevalnem orodju Q-CAT.

3.1.2 Razširitev s povedmi iz korpusa ELEXIS-WSD-SL

V drugi fazi širitve je bil skladenjsko razčlenjen še korpus ELEXIS-WSD--SL, tj. slovenski del paralelnega korpusa ELEXIS-WSD (Martelli idr., 2021; Martelli idr., 2022), razvitega za potrebe strojnega pomenskega razdvoumljanja, ki vsebuje v več evropskih jezikov prevedena besedila iz Wikipedie (Schwenk idr. 2021). Slovenski korpus ELEXIS-WSD vsebuje 2.024 povedi (31.237 pojavnic), ki so bile predhodno že ročno tokenizirane, lematizirane in oblikoskladenjsko označene po sistemu JOS, na podlagi česar smo korpus s pretvorbeno skripto samodejno pretvorili še v besedne vrste in oblikoskladenjske oznake UD, pojavitve glagola *biti*[7] pa razdvoumili ročno.

7 V primerjavi z označevalno shemo JOS, ki glagol *biti* ne glede na skladenjsko vlogo vedno označuje kot (pomožni) glagol, shema UD že na ravni določanja besednih vrst ločuje med pomožniki (AUX) in glavnimi glagoli (VERB). Podrobnejše smernice s primeri za

Tako označen korpus je bil izhodiščno skladenjsko razčlenjen z orodjem CLASSLA-Stanza (Ljubešić in Dobrovoljc 2019; Terčon in Ljubešić, 2023), pravilnost strojno pripisanih razčlemb pa so nato pregledali trije označevalci in končni kurator. Na ta način je bilo ročno popravljenih 1.534 (4,91 %) skladenjskih relacij, med katerimi so prevladovale strukture z oznakami *nmod* (samostalniški prilastki), *advmod* (prislovna določila), *obl* (odvisne samostalniške zveze), *conj* (priredja) in *punct* (ločila), kar se, kot bomo videli v nadaljevanju, sklada z najpogostejšimi tipi napak razčlenjevalnika nasploh (razdelek 4.2).

3.1.3 Izboljšanje označenosti v prvotnem korpusu SSJ-UD

Poleg dodajanja novih razčlenjenih povedi smo glede na rahlo spremembo smernic (2. razdelek), analizo ročnih popravkov pretvorjenih relacij (razdelek 3.1.1) in drugih identificiranih nedoslednosti izboljšali tudi označenost izhodiščne različice korpusa SSJ-UD.

Med približno 30 identificiranimi tipi napak oz. nedoslednosti so bile denimo pristavčne strukture, visok delež (neupravičenih) neprojektivnih povezav,[8] nedosledno ločevanje med sorednimi in priredno vezanimi stavki, med premimi in nepremimi predmeti, itd. Za vsako izmed kategorij smo s hevurističnimi poizvedbami ustvarili podkorpuse povedi s potencialno problematičnimi oznakami, ki so jih nato označevalci ročno pregledali in popravili v skladu s smernicami. Na ta način je bilo v izhodiščnem korpusu popravljenih 1.670 skladenjskih oznak, kar sicer predstavlja razmeroma majhen del celotnega korpusa (1,2 %).

3.1.4 Objava nove različice korpusa SSJ-UD

V zadnjem koraku smo izhodiščni korpus SSJ-UD z nekoliko izboljšano označenostjo (razdelek 3.1.3) združili z novimi povedmi iz korpusov ssj500k (3.1.1) in ELEXIS-WSD (3.1.2) ter tako dobili novo različico referenčne univerzalne odvisnostne drevesnice za pisno slovenščino

tovrstno razdvoumljanje glagola *biti* so na voljo tudi na GitHub repozitoriju CLARIN.SI: https://github.com/clarinsi/jos2ud/blob/master/Map/UD_biti_anno_navodila_v02.docx

8 Povezava med besedo A in besedo B je projektivna, če je beseda A posredno nadrejena tudi vsem drugim besedam med A in B – obstaja torej pot od A do vseh besed med A in B. Če si to predstavljamo grafično, se povezave v neprojektivnem drevesu med seboj križajo. To je v jezikih s prostim besednim redom, kot je slovenščina, sicer možen pojav (za primere gl. priročnik s smernicami), a vendarle redek.

SSJ-UD,[9] ki je bila s standardno delitvijo na učno, validacijsko in testno množico (več v Dobrovoljc in Ljubešić, 2022) prvič objavljena kot del uradnega izida UD v2.10 (Zeman idr., 2022).

Kot prikazuje tabela 1, nova različica 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, nenazadnje pa tudi z vidika izvora ročno pripisanih oznak (od pretvorb do popravljanja).

Tabela 1: *Zgradba nove različice korpusa SSJ-UD (od UD v 2.10 naprej)*

Podkorpus	Povedi	Pojavnice	Povp.
Prvotni SSJ-UD	8.000	140.670	17,58
Novo iz ssj500k	3.411	95.194	27,91
Novo iz ELEXIS-WSD	2.024	31.233	15,43
Skupaj novi SSJ-UD	**13.435**	**267.097**	**19,88**

Novi korpus SSJ-UD je bil obenem integriran tudi v novi referenčni učni korpus SUK (Arhar Holdt idr., 2022), v katerem univerzalno skladenjsko razčlenjene povedi po sistemu UD predstavljajo dobro četrtino celotnega korpusa. Ker korpus SUK vsebuje še številne druge ravni jezikoslovnih oznak, kot so skladenjske razčlembe po sistemu JOS, udeleženske vloge, večbesedne enote in imenske entitete, to odpira številne možnosti kompleksnejših čezravninskih analiz in raziskav.

3.2 Univerzalno skladenjsko razčlenjeni korpus SloBENCH-UD

Poleg nove, izboljšane in razširjene, drevesnice SSJ-UD smo izdelali še ločeno univerzalno odvisnostno drevesnico za spletni portal SloBENCH

9 Čeprav infrastruktura UD dopušča objavo poljubnega števila drevesnic, smo se namesto objave novih drevesnic UD za slovenščino namenoma odločili za priključitev novih povedi k že obstoječi drevesnici SSJ-UD, da bi zagotovili kar najbolj učinkovito izrabo teh podatkov v širši jezikovnotehnološki skupnosti, kjer se zaradi poenostavitve dela modeli pogosto razvijajo zgolj na izbrani, običajno največji, drevesnici nekega jezika.

(Žitnik in Dragar, 2021),[10] ki je bil razvit z namenom enotne primerjave uspešnosti orodij za različne jezikovnotehnološke naloge na slovenskih besedilih. Del vsake naloge je tudi vnaprej definirana testna množica besedil, na kateri razvijalci poženejo svoje orodje, evalvacijski sistem v ozadju pa nato strojne rezultate primerja z (javnosti skritimi) ročno pregledanimi rešitvami. V primerjavi s testno množico SSJ-UD, ki je skupaj z učno in validacijsko množico z oznakami vred javno objavljena kot del uradnega repozitorija UD, drevesnica SloBENCH-UD kot skrita testna množica zagotavlja bolj nepristransko evalvacijo, saj onemogoča namensko prilagajanje delovanja orodij specifičnim testnim podatkom.

Konkretno smo za izdelavo drevesnice SloBENCH-UD uporabili del korpusa SentiCoref (Žitnik, 2019), ki je bil za druge naloge na istem portalu že ročno pregledan na ravni segmentacije, tokenizacije, lematizacije in oblikoskladenjskih oznak JOS. Po ustaljenem postopku smo oznake JOS s pomočjo pretvorbenih pravil najprej preslikali v besedne vrste in oblikoslovne oznake UD ter nato množico strojno razčlenili z najnovejšim razčlenjevalnim modelom, naučenim na podatkih razširjenega korpusa SSJ-UD (razdelek 3.1). Strojno razčlenjene podatke je ročno pregledal in popravil ekspertni označevalec ter pri tem besednovrstno razdvoumil tudi glagol *biti* v vlogi pomožnega oz. glavnega glagola. Končna ročno pregledana množica SloBENCH-UD tako obsega 1.332 razčlenjenih povedi oz. 29.138 pojavnic in je tako po velikosti kot vsebini primerljiva s testno množico SSJ-UD. To navsezadnje potrjuje tudi zelo podobna stopnja natančnosti, ki jo na obeh testnih množicah dosega razčlenjevalni model CLASSLA-Stanza.

Množica je že bila integrirana v spletni portal SloBENCH, pri čemer je krovna naloga, tj. slovnično razčlenjevanje slovenskih besedil po shemi Universal Dependencies, zasnovana po vzoru tekmovanj CoNLL Shared Task 2017 in 2018 (Zeman idr., 2018), z evalvacijskimi metrikami vred.

3.3 Univerzalno oblikoslovno označena korpusa SUK in Janes-Tag

Kot smo omenili že v razdelku 3.1, so bila za pretvorbo korpusa ssj500k v prvo različico drevesnice SSJ-UD izdelana podrobna pravila za preslikavo

10 https://slobench.cjvt.si/

oblikoskladenjskih oznak JOS v besedne vrste in oblikoskladenjske lastnosti sistema UD. Zaradi precejšnje podobnosti med obema sistemoma je bilo s temi pravili mogoče natančno pretvoriti celotni korpus ssj500k, pozneje pa tudi druge sorodne vire z oznakami JOS, kot sta oblikoslovni leksikon Sloleks (Dobrovoljc idr., 2019) in ročno označeni učni korpus nestandardne slovenščine Janes-Tag (Erjavec idr., 2019). Čeprav tako označeni viri zaradi manka skladenjskih relacij ne morejo biti distribuirani kot del uradne zbirke drevesnic UD, kot to velja za slovenski drevesnici SSJ in SST, ti predstavljajo pomemben vir podatkov za učenje napovednih modelov na nižjih slovničnih ravneh (Dobrovoljc idr., 2019).

S tem namenom smo v univerzalne oblikoslovne oznake (besedne vrste in druge oblikoskladenjske lastnosti) pretvorili tudi novi referenčni učni korpus za standardno pisno slovenščino SUK 1.0 (Arhar Holdt idr., 2022) in razširjeni referenčni učni korpus za nestandardno pisno slovenščino Janes-Tag 3.0 (Lenardič idr., 2022), ki oba v primerjavi s prejšnjima različicama (ssj500k 2.3 in Janes-Tag 2.1) vsebujeta kar enkrat več ročno pregledanih lem in oblikoskladenjskih oznak JOS (približno 1 milijon oz. 190.000 pojavnic). 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 in opravili ustaljeni ročni pregled povedi z glagolom *biti* za razdvoumljanje med pojavitvami pomožnega in glavnega glagola (po en označevalec na primer).

4 Novi razčlenjevalni model

V drugi fazi projekta smo na novi, bistveno večji različici ročno označenega korpusa SSJ-UD (razdelek 3.1) naučili tudi nov napovedni model skladenjskega razčlenjevanja po sistemu UD v označevalnem orodju CLASSLA-Stanza (Ljubešić in Dobrovoljc, 2019; Terčon in Ljubešić 2023),[11] ki se je kot temeljno programsko orodje za označevanje besedil v slovenščini prav tako razvijalo v okviru projekta RSDO. Gre za izpeljavo odprtokodnega orodja Stanza (Qi idr., 2020), ki v primerjavi z izvornim orodjem uvaja nekatere izboljšave na ravni tokenizacije, oblikoskladenjskega označevanja in lematizacije, skladenjski razčlenjevalnik pa se od izvornega (Dozat

11 https://github.com/clarinsi/classla

in Manning, 2016), ki temelji na nadgrajeni metodi dvosmernega dolgega kratkoročnega spomina (BiLSTM), razlikuje predvsem po uporabi besednih vložitev CLARIN.SI-embed.sl (Ljubešić in Erjavec, 2018), ki so bile naučene na slovenskih besedilih v obsegu 3,5 milijard besed.

Najnovejši razčlenjevalni model (Terčon in Ljubešić, 2023) je del najnovejše različice CLASSLA-Stanza 2.0. Modeli, ki so vključeni v to različico, so bili naučeni na učnem korpusu SUK (Arhar Holdt idr., 2022). Za učenje modela za skladenjsko razčlenjevanje je bil uporabljen le tisti del korpusa, ki ustreza korpusu SSJ-UD[12].

Primerjavo novega modela s predhodnim modelom, naučenim na prvotni različici SSJ-UD, sta podrobneje opisala že Dobrovoljc in Ljubešić (2022), ki ugotavljata, da je model, naučen na novi različici korpusa SSJ-UD,[13] zaradi povečanega obsega učnih podatkov in njihove diverzifikacije bistveno izboljšan v primerjavi z modelom, naučenim na prvotni različici.

Da bi osvetlili prednosti in pomanjkljivosti uporabe novega razčlenjevalnega modela v različnih jezikovnotehnoloških in jezikoslovnih aplikacijah ter obenem identificirali prioritete za njegove nadaljnje izboljšave, v nadaljevanju prispevka te ugotovitve nadgradimo s podrobnejšo evalvacijo splošne natančnosti modela (razdelek 4.1) na eni strani in analizo najpogostejših tipov napak (razdelek 4.2) na drugi.

Pri evalvaciji smo uporabili ročno označene podatke na nižjih ravneh označevanja (tokenizacija, stavčna segmentacija, oblikoskladenjsko označevanje, lematizacija), saj nas je v tej fazi razvoja razčlenjevalnika zanimala predvsem natančnost napovednega modela v izolaciji, brez vpliva napovednih karakteristik orodja na nižjih ravneh.

4.1 Splošna natančnost modela

Za kvantitativno evalvacijo splošne natančnosti modela smo uporabili standardni protokol, po katerem smo model, naučen na učni oz. validacijski množici uporabili za razčlenjevanje testne množice, napovedane oznake pa nato primerjali z ročno pripisanimi. Za poročanje o

12 Celoten proces učenja modelov za najnovejšo različico orodja CLASSLA-Stanza je natančneje opisan na GitHub repozitoriju: https://github.com/clarinsi/classla-training

13 V prispevku sta Dobrovoljc in Ljubešić (2022) sicer evalvirala predhodno neuradno različico modela, ki se od uradne razlikuje v količini in tipu učnih podatkov na nižjih ravneh, vendar je njuna splošna natančnost povsem primerljiva, saj sta bila naučena na identičnem skladenjsko razčlenjenm korpusu SSJ-UD.

natančnosti uporabljamo uveljavljeno metriko LAS (angl. *labeled attachment score*), ki prikazuje delež pojavnic s pravilno napovedano nadrejeno pojavnico in vrsto njunega skladenjskega razmerja, pri čemer ta delež povzemamo z oceno F1, ki prikazuje harmonično sredino med preciznostjo in priklicem.[14]

Rezultati, predstavljeni v tabeli 2, prikazujejo, da razčlenjevalni model dosega splošno natančnost 93,73 LAS F1, kar nekoliko poenostavljeno pomeni, da se model v povprečju na vsakih sto označenih pojavnic zmoti pri manj kot sedmih, tj. jim pripiše napačno nadrejeno pojavnico in/ali vrsto povezave med njima.

Kot prikazujejo rezultati za posamične tipe relacij v tabeli 2,[15] pa ta splošna ocena natančnosti ni reprezentativna za vse vrste skladenjskih struktur, saj je pri napovedovanju nekaterih relacij model bistveno natančnejši kot pri drugih.

Tabela 2: *Natančnost novega modela orodja CLASSLA-Stanza za skladenjsko razčlenjevanje po sistemu UD glede na metriko LAS*

Relacija	Izvorni opis	Slovenski prevod	Učna	Testna	LAS F1
acl	clausal modifier of noun	stavčni prilastki	3377	444	**83,48**
advcl	adverbial clause modifier	prislovni odvisniki	1927	239	**76,25**
advmod	adverbial modifier	prislovna določila (gl. op. 17)	16307	1935	**90,37**
amod	adjectival modifier	pridevniški prilastki	17.628	2165	**99,12**
appos	appositional modifier	pristavčna določila	1.505	163	**66,45**
aux	auxiliary verb	pomožni glagoli	9.773	1162	**99,01**
case	case marking preposition	predlogi	19.813	2415	**99,19**
cc	coordinating conjunction	priredni vezniki	7.294	923	**96,32**

14 Izračuni temeljijo na uradni evalvacijski skripti tekmovanja CoNLL Shared Task 2018 (Zeman idr., 2018), ki smo jo dodatno prilagodili tako, da poleg splošnega izračuna natančnosti vrača tudi rezultate za posamične skladenjske relacije, besedne vrste in druge relevantne oznake.

15 V Tabeli 2 ni relacij *goeswith* (napačno razdruženi deli besed), *reparandum* (samopopravljanja) in *compound* (zloženke), saj se v korpusu SSJ-UD ne pojavljajo. Pri relaciji *dislocated* podatka o natančnosti ni (oznaka *n/a*), saj se v testni množici ne pojavi. O natančnosti izpeljanih relacij oz. podoznak (npr. *flat:name*, *flat:foreign*) poročamo združeno z jedrno oznako (npr. *flat*).

Relacija	Izvorni opis	Slovenski prevod	Učna	Testna	LAS F1
ccomp	clausal complement	stavčna dopolnila (predmetni odvisniki)	1.544	187	**92,27**
conj	conjunct	priredno zloženi elementi	9.307	1108	**86,30**
cop	copula verb	vezni glagoli	4.244	542	**96,12**
csubj	clausal subject	osebkovi odvisniki	723	78	**85,53**
dep	unspecified dependency	nedoločena povezava	169	12	**20,00**
det	determiner	določilniki	4.724	616	**98,95**
discourse	discourse element	diskurzni členki	153	15	**75**
dislocated	dislocated element	dislocirani elementi	10	0	**n/a**
expl	expletive	ekspletivne besede	2.997	361	**96,31**
fixed	fixed multi-word expression	funkcijske zveze	944	89	**92,47**
flat	flat multi word-expression	eksocentrične zveze	152	5	**95,00**
iobj	indirect object	nepremi predmeti	873	87	**83,33**
list	list	seznami	583	16	**82,35**
mark	marker (subordinating conjunction)	podredni vezniki	6.415	804	**98,38**
nmod	nominal modifier	samostalniški prilastki	16.065	1887	**87,44**
nsubj	nominal subject	samostalniški osebki	10.585	1315	**96,05**
nummod	numeric modifier	številčna določila	3.543	311	**95,13**
obj	(direct) object	premi predmeti	9.733	1140	**96,37**
obl	oblique nominal (adjunct)	odvisne samostalniške zveze	14.049	1722	**92,10**
orphan	dependent of missing parent	elementi v eliptičnih strukturah	785	83	**71,90**
parataxis	parataxis	stavčna soredja	3.273	345	**73,26**
punct	punctuation symbol	ločila	32.116	3623	**94,09**
root	root element	koren povedi	10.903	1282	**96,80**
vocative	vocative	ogovori	63	1	**0**
xcomp	open clausal complement	odprta stavčna dopolnila	1.542	198	**92,50**
Vse relacije					**93,73**

Opomba. V 4. in 5. stolpcu je navedeno število pojavnic, označenih z dano relacijo, v učni oz. testni množici.

Med relacijami z najvišjo natančnostjo napovedovanja so po pričakovanju funkcijske besede, kot so predlogi (*case*; 99,19), pridevniški

prilastki (*amod*; 99,12), pomožni glagol *biti* (*aux*; 99,01), določilniški zaimki in prislovi (*det*; 98,95), podredni vezniki (*mark*; 98,38), ekspletivni zaimki (*expl*; 96,31) in priredni vezniki (*cc*; 96,32); skratka, pojavnice, ki se pojavljajo v zelo predvidljivih oblikah in skladenjskih položajih.

Poleg navedenih relacij model razmeroma dobro natančnost dosega tudi pri napovedovanju nekaterih jedrnih skladenjskih struktur, kot so samostalniški predmeti (*obj*; 96,37) in osebki (*nsubj*; 96,05), nadpovprečno uspešen pa je tudi pri identifikaciji korena povedi (*root*; 96,8), ki je običajno jedro povedka glavnega stavka, in veznega glagola *biti* (*cop*; 96,12), ki nastopa v strukturah s povedkovimi določili.

Med relacijami, pri napovedovanju katerih model dosega najslabše rezultate, pričakovano najdemo ogovore (*vocative*; 0,0), saj se v testni množici pojavi zgolj en primer, in nedoločene strukture (*dep*; 20,0), saj se ta oznaka kot skrajna možnost uporablja predvsem za povezovanje obrobnih, iregularnih pojavov, ki jim je nemogoče pripisati katerokoli drugo povezavo (npr. ostanki oštevilčenih strani pri digitalizaciji besedil).

Čeprav se je natančnost označevanja samostalniških pristavčnih določil (*appos*, 66,45), "osirotelih" stavčnih členov v povedih z glagolsko elipso (*orphan*; 71,90), stavčnih soredij (*parataxis*; 73,26), diskurznih členkov (*discourse*; 75,0), in naštevalnih seznamov (*list*; 82,35) z novo različico korpusa SSJ-UD bistveno izboljšala glede na prvotni model (Dobrovoljc, Ljubešić 2022), te relacije ostajajo med tistimi z najnižjo natančnostjo, kar je glede na njihovo ohlapnejšo slovnično povezanost s povedkom oz. nadrejenimi stavčnimi členi tudi pričakovano.

Med drugimi relacijami s podpovprečno natančnostjo označevanja lahko izpostavimo še podredne stavke različnih tipov, kot so prislovni (*advcl*; 76,25), prilastkovi (*acl*; 83,48), osebkovi (*csubj*; 85,53) in predmetni odvisniki (*ccomp*; 92,27). Poleg nepremih predmetov (*iobj*; 83,33), ki jih je težavno identificirati predvsem zaradi pomanjkljivosti trenutnih označevalnih smernic,[16] modelu precejšen izziv predstavlja-

16 V času nastanka smernic in označenih podatkov, ki jih opisuje ta prispevek, so splošne smernice UD zaradi kompleksnega prepletanja oblikoslovnih, skladenjskih in pomenskih razločevalnih lastnosti med premimi in nepremimi predmeti priporočale, da je v povedih z zgolj enim izraženim predmetom vedno označen kot premi predmet (*obj*), ne glede na sklon ali udeleženska vlogo. To robustno pravilo, ki je bilo identificirano tudi kot eno izmed odprtih vprašanj, omenjenih v 2. razdelku, je bilo pred kratkim opuščeno. Temu bodo sledile tudi prihodnje nadgradnje korpusa SSJ in lahko pričakujemo, da se bo posledično izboljšala tudi natančnost strojnega ločevanja med premimi in nepremimi predmeti.

jo tudi priredja, zlasti medstavčna (*conj*; 86,3), samostalniški prilastki (*nmod*; 87,44) ter prislovna določila povedkov, samostalnikov in pridevnikov (*advmod*; 90,37).

4.2 Najpogostejše napake modela

V drugem koraku evalvacije smo analizo zanesljivosti modela pri razčlenjevanju posameznih tipov relacij dopolnili še s podrobnejšo analizo najpogostejših tipov napak. Tabela 3 tako povzema distribucijo napak glede na to, pri katerem izmed obeh napovedanih podatkov (identifikator nadrejene pojavnice in vrsta skladenjske relacije med njima) se je model dejansko zmotil. Za vsak tip napake navajamo tudi najpogostejše podtipe glede na relacije, pri katerih se pojavlja, pri čemer štetje prikazujemo združeno za napake v obe smeri (npr. *obl-nmod* vključuje tako napovedovanje *obl* namesto *nmod* kot napovedovanje *nmod* namesto *obl*).

Identificirane pogoste tipe napak znotraj vsake kategorije na podlagi ročne analize napačno označenih primerov opišemo v nadaljevanju, pri čemer podrobneje predstavimo predvsem najpogostejše.

Tabela 3: *Distribucija napak razčlenjevalnega modela glede na tip napake*

Tip napake	Število napak
Napačno jedro	**845**
punct-punct	214
advmod-advmod	159
nmod-nmod	121
conj-conj	94
acl-acl	47
parataxis-parataxis	30
obl-obl	27
advcl-advcl	25
cc	23
cop-cop	20
Napačno jedro in oznaka	**493**
obl-nmod	140
parataxis-root	39
acl-advcl	19
root-nsubj	16
parataxis-appos	15

Tip napake	Število napak
Napačna oznaka	**257**
conj-parataxis	19
obl-nsubj	16
appos-conj	16
obj-iobj	14
obl-obj	11
Vse napake	**1595**

4.2.1 Napačna napoved nadrejenega elementa

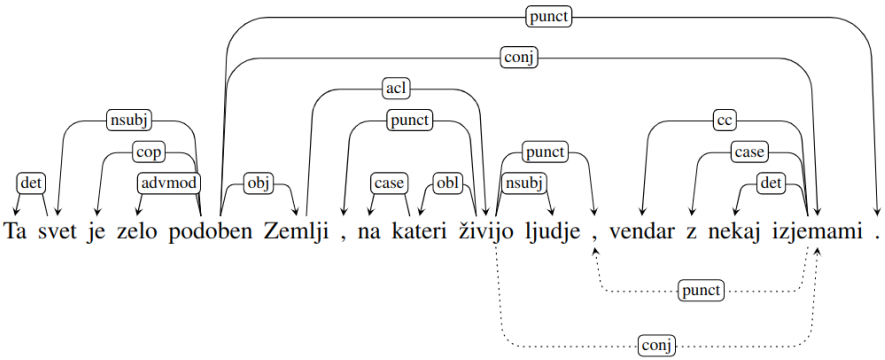
Kot prikazuje tabela 3, dobro polovico (52,8 %) predstavljajo napake, pri katerih je model pravilno napovedal skladenjsko vlogo pojavnice (pravilno relacijo oz. oznako), zmotil pa se je pri napovedi njenega nadrejenega elementa (jedra oz. izvora relacije).

Najpogostejša napaka pri določanju nadrejenega elementa je povezana z relacijo **punct**, ki označuje ločila. Po večini gre za primere, kjer so napačno določena tudi jedra drugih struktur v povedi, na katera se ločila praviloma povezujejo. Napačno povezana ločila so torej večinoma posledica napak razčlenjevanja njihovih nadrejenih struktur, kot prikazuje primer na Sliki 2, pri katerem razčlenjevalnik zadnji stavek zmotno interpretira kot priredje pred njim stoječega odvisnika, čemur ustreza tudi (napačno) označena vejica. Pri dolgih povedih, v katerih nastopa veliko podrednih elementov, obdanih z vejicami, se pojavljajo tudi napake, kjer razčlenjevalnik povzroči neprojektivnost z vezanjem vejic na napačno jedro.

Druga pogosta skupina je povezana s t.i. poudarjalnimi členki oz. prislovi, kot so besedice *tudi*, *še*, *le*, *že* idr., ki jim pripisujemo relacijo **advmod**,[17] njihova stava pa je v slovenščini razmeroma prosta – modificirajo lahko tako povedek kot posamezne stavčne člene, kar je pogosto mogoče razbrati šele iz konteksta ali prozodičnih poudarkov pri branju. Kot prikazuje primer na Sliki 3, razčlenjevalnik te besede namesto na poudarjeni samostalnik pogosto veže na povedek stavka. To ni

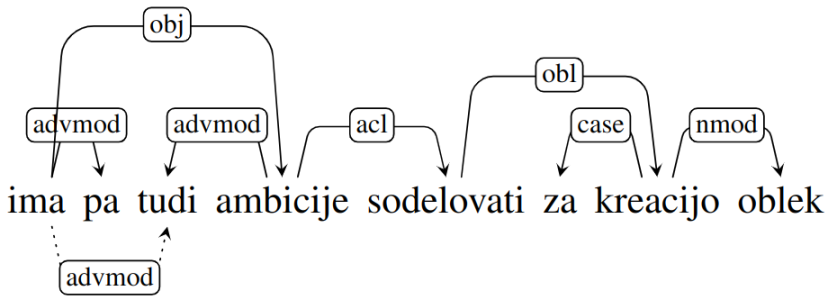
17 Relacija *advmod* se uporablja za označevanje prislovov v vlogi modifikatorjev, kar vključuje tako prislove v vlogi okoliščinskih dopolnil povedkov (kakršna Slovenska slovnica (Toporišič 2000) imenuje prislovna določila, npr. pridem *takoj*) kot prislove v vlogi modifikatorjev pridevniških, prislovnih ali samostalniških besednih zvez (prislovni prilastki, npr. *izjemno* prilagodljiv).

presenetljivo, glede na to, da gre za eno izmed kategorij, pri kateri so se označevalci najpogosteje razhajali, prav tako pa je bila nedosledno označena v prvotnem korpusu, v katerem so bile ob pretvorbi te pojavnice ne glede na vlogo vedno povezane na povedek.



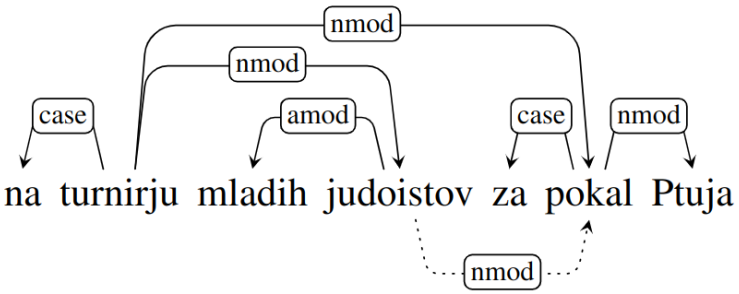
Slika 2: Primer razhajanja med ročno (zgoraj) in strojno (spodaj) pripisanim jedrom relacije punct.

Pri relaciji **advmod** se napake pojavljajo, tudi ko nekemu prislovu sledi pridevnik. Razčlenjevalnik za nadrejeni element pogosto določi ta pridevnik, čeprav je pravi nadrejeni element povedek stavka. Zgodi pa se lahko tudi obratno: kot nadrejeni element je označen povedek, v resnici pa bi to moral biti sledeči pridevnik.



Slika 3: Primer napačne razčlembe poudarjalnih členkov (advmod zgoraj) kot prislovnih določil povedka (advmod spodaj).

Pri štirih relacijah s pogosto napačno pripisanim izvorom povezave, tj. **nmod**, **conj**, **acl** in **obl**, prihaja do podobne napake: razčlenjevalnik zanesljivo prepozna vrsto nadrejene strukture (npr. samostalniške zveze, pridevniške zveze ali povedki), vendar namesto prave strukture kot jedro izbere najbližjo ustrezno zvezo na levi, kar ni vedno prav, saj se včasih pravi izvor relacije v povedi pojavi že prej (Slika 4).



Slika 4: Primer razhajanja med ročno (zgoraj) in strojno (spodaj) identificirano odnosnico predložne zveze v vlogi desnega prilastka (*nmod*).

Napake relacije **conj** se pogosto pojavijo, ko gre za zaporedje treh ali več priredno zloženih stavkov, kjer model ne ujame pravilnega nanašanja. Pojavijo se tako napake, kjer je pravilen izvorni stavek relacije pred ciljnim, kot tudi napake, kjer se pravilen izvoren stavek pojavi za ciljnim. Enaka napaka se pojavlja tudi pri stavčnih soredjih, ki so označena z relacijo **parataxis**.

Ko gre pri prislovnih določilih, označenih z **obl**, za izbiro med več kot enim možnim nadrejenim povedkom, označevalnik kot izvor relacije pogosto določi napačen povedek. Ti povedki se pojavijo tako pred ciljem relacije kot tudi za njim in sestavljajo najrazličnejše strukture.

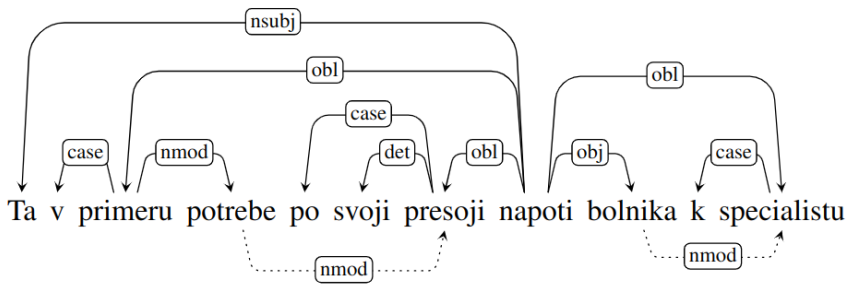
Pri relaciji **advcl** pogosto pride do napak v povezavi s stavčnimi primerjavami. Razčlenjevalnik za izvor primerjave ne določi primerjane lastnosti, pač pa povedek glavnega stavka, kar ni v skladu s trenutnimi smernicami. Do napak prihaja tudi pri strukturah z modalnimi in faznimi glagoli, kjer se odvisnik včasih pomotoma veže na nedoločnik in ne na modalni oz. fazni glagol, kot predpisujejo smernice. To nakazuje, da so med učnimi podatki pri teh strukturah morda nedoslednosti.

Včasih se zgodi, da razčlenjevalnik ne prepozna vseh enot priredja. Ker se priredni veznik vedno veže na drugi element priredja, ta napaka hkrati povzroči napako pri določanju jedra relacije **cc** (primer: *bilo je namenjeno predvsem in samo moškim* - tu sta v priredju besedi *predvsem* in *samo*, model pa je za drugi element napačno določil besedo *moškim*, ki je posledično dobila vlogo jedra relacije **cc**).

Napake relacije **cop** se pogosto pojavijo v obliki zamenjave osebka in povedkovega določila v povedih, v katerih je težko določiti, kateri člen izpolnjuje katero od teh dveh vlog. Te napake se pojavijo le ob veznem glagolu *biti*.

4.2.2 Napačna napoved nadrejenega elementa in relacije

Po pogostosti sledijo napake, pri katerih se je model zmotil tako pri napovedi nadrejene pojavnice kot njune skladenjske relacije (29,9 %). Med njimi najbolj izstopa zamenjevanje struktur z oznakama **obl**[18] in **nmod**, ki predstavlja tretji najpogostejši (pod)tip napak nasploh. Analiza primerov kaže, da gre večinoma za povedi, v katerih predložna zveza v vlogi prislovnega določila povedka (**obl**) stoji tik za neko samostalniško zvezo, model pa prislovno določilo napačno tolmači kot njen desni prilastek, za katere se uporablja relacija **nmod**, kot prikazuje primer na Sliki 5.



Slika 5: Primer napačne razčlembe predložnih prislovnih določil (*obl* zgoraj) kot desnih prilastkov (*nmod* spodaj).

18 Relacija *obl* se uporablja za odvisne samostalniške in predložne zveze, ki nastopajo v vlogi nejedrnih argumentov povedka. Poleg teh se s to relacijo označujejo tudi neglagolske strukture s primerjalnimi vezniki.

Manj pogoste v tej kategoriji so napake drugih kombinacij relacij. Pri **parataxis-root** gre za napake pri določanju glavnega stavka v nizu dveh ali več soredno zloženih stavkov, zlasti kadar gre za vrinjene stavke ali premi govor. Pri kombinaciji **acl-advcl** gre za napake ločevanja med prislovnodoločilnimi odvisniki in stavčnimi prilastki, pogosto v kombinaciji z veznikom *kot,* model pa včasih tudi ne prepozna določenih stavkov kot stavčnih primerjav in jih namesto tega veže na najbližjo samostalniško zvezo z relacijo **acl**. Kombinacija **root-nsubj** pogosto pomeni zamenjavo osebka in povedkovega določila v strukturah z veznim glagolom *biti,* včasih pa se te napake pojavijo tudi pri eliptičnih stavkih z nejasno strukturo. To so pogosto krajši stavki z več izpuščenimi členi.

4.2.3 Napačna napoved relacije

Med vsemi tremi kategorijami napak pa je najmanj takih, pri katerih je razčlenjevalnik pojavnico povezal s pravim nadrejenim elementom, a tej relaciji pripisal napačno oznako (17,3 %). V primerjavi s prvima dvema kategorijama so tukaj tipi glede na relacije razpršeni bolj enakomerno.

Do zamenjav oznak **conj** in **parataxis**[19] prihaja predvsem pri daljših povedih, pri katerih se med dva prirediono zložena stavka oz. med priredni veznik in drugi stavek v priredju vrivajo druge strukture (npr. odvisniki). Samostalniška prislovna določila (ki prejmejo relacijo **obl**)[20] so napačno označena kot osebki (**nsubj**) predvsem v zvezah z glagoli, kot so *imenovati*, *praviti*, idr., v katerih se pojavljajo v imenovalniku (npr. *pravimo jim mikroznaki*). Do zamenjave **obl-nsubj** velikokrat pride tudi pri strukturah z glagolom *biti* in samostalnikom v rodilniku (npr. v stavku *on je mnenja, da…* in *So bolj vesele narave…*). V takih primerih model pogosto določi samostalnik v rodilniku za osebek, kar lahko delno pojasnimo z dejstvom, da primerov tovrstnih struktur v učnih podatkih ni prav veliko.

Med drugimi tipi napačno pripisanih relacij je pogosta še dvoumnost med samostalniškimi zvezami v vlogi pristavčnih določil (**appos**)

19 Relacija *parataxis* se uporablja za označevanje stavčnih soredij različnih vrst. To so razmerja med besedo (običajno jedrom glavnega stavka) in drugimi elementi, ki z njo niso v priredju, podredju ali kateremkoli drugem jedrnem slovničnem razmerju.

20 Eno izmed odprtih vprašanj za prihodnje nadgradnje smernic je tudi jasnejša opredelitev kategorije *obl* in njene razmejitve glede na druge vrste samostalniških dopolnil glagola. Trenutno smernice namreč sledijo načelom opredelitve samostalniških ‘prislovnih določil’ znotraj sheme JOS-SSJ.

na eni in priredno povezanih elementov (**conj**) na drugi strani, zlasti kadar gre za naštevanje in zadnji element v brezvezniškem priredju stoji na koncu povedi (primer: *to je popoln poraz nekega koncepta, popoln poraz vseh nas*).

Pojavljajo se tudi napake ločevanja med prislovnimi določili in predmeti, predvsem pri samostalniških zvezah, ki izražajo časovni oz. prostorski okvir dogodka (**obl-obj**) in pa napačno določanje premega (**obj**) in nepremega predmeta (**iobj**).

5 Sklep

V prispevku smo predstavili nadgradnjo slovenskih slovnično razčlenjenih korpusov, ročno označenih po medjezikovno primerljivi shemi Universal Dependencies, v okviru katere smo po rahli prenovi in izčrpni dokumentaciji označevalnih smernic za slovenščino referenčno drevesnico pisne slovenščine SSJ-UD razširili z več kot 5.000 novimi povedmi, izdelali povsem novo testno množico za uporabo na evalvacijskem portalu SloBENCH in referenčna ročno oblikoskladenjsko označena korpusa SUK in Janes-Tag pretvorili v oblikoslovne oznake UD. Na novi različici drevesnice SSJ-UD smo naučili tudi nov napovedni model za skladenjsko razčlenjevanje slovenskih besedil, ki v splošnem dosega razmeroma visoko stopnjo natančnosti, pri čemer naša analiza kaže, da je pri členjenju nekaterih struktur mogoče pričakovati bistveno večjo zanesljivost rezultatov kot pri drugih.

Glede na mednarodno relevantnost sheme UD ti rezultati predstavljajo pomemben doprinos k nadaljnjemu razvoju jezikovnih tehnologij za slovenščino tako v slovenskem kot mednarodnem prostoru, saj je glede na odprti dostop in standardizirano distribucijo drevesnic UD mogoče pričakovati, da bodo novi podatki za slovenščino kmalu integrirani tudi v številna druga razčlenjevalna orodja oz. na njih temelječe aplikacije (npr. Honnibal in Montani, 2017; Nguyen idr., 2021). Poleg modelov za skladenjsko razčlenjevanje, kakršnega smo predstavili v tem prispevku, je skoraj enkrat večja količina učnih podatkov za slovenščino neprecenljiva tudi za nadaljnji razvoj modelov za lematizacijo in oblikoslovno označevanje po sistemu UD, ki v mednarodnem prostoru večinoma temeljijo zgolj na uradno izdanih drevesnicah UD, kot je SSJ-UD,

ne pa virih, ki so bili razviti oz. distribuirani v lokalnem kontekstu, kot sta denimo univerzalno oblikoslovno označena učna korpusa (ne)standardne slovenščine, SUK in Janes-Tag .

Čeprav je bila shema UD prvotno vzpostavljena predvsem za potrebe jezikovnotehnoloških raziskav, pa številne odmevne primerjalnojezikoslovne študije dokazujejo tudi njeno relevantnost na področju jezikoslovja, vključno s slovenistiko, kjer metodološki potencial skladenjsko razčlenjenih korpusov doslej še ni bil polno izkoriščen (Ledinek, 2018). Verjamemo, da izčrpno dokumentirane smernice, obsežni ročno označeni korpusi in sistematična evalvacija natančnosti na njih naučenih modelov predstavljajo pomemben doprinos k nadaljnjim jezikoslovnim raziskavam ročno in strojno razčlenjenih slovenskih korpusov

Pri tem je glede na kompleksno strukturo tovrstnih korpusov za doseganje tega cilja nujno vzpostaviti tudi ustrezno infrastrukturo za njihovo analizo. Poleg vključevanja korpusov SSJ-UD, SUK in Janes-Tag v konkordančnike CLARIN.SI (Erjavec, 2013), ki niso prilagojeni specifikam iskanja po odvisnostnih drevesih in njihovi vizualizaciji, je bilo v zadnjem času razvitih več namenskih orodij za analizo univerzalno skladenjsko razčlenjenih korpusov v slovenščini, kot sta orodje za statistično analizo skladenjsko razčlenjenih korpusov STARK (Krsnik idr., 2019) in spletni portal Drevesnik za napredno brskanje po slovenskih UD drevesnicah (Štravs in Dobrovoljc, 2022). Temu tipu oznak je bilo prilagojeno tudi označevalno orodje Q-CAT (Brank, 2022), za strojno slovnično označevanje novih besedil pa je bil pred kratkim vzpostavljen tudi spletni portal CJVT Označevalnik,[21] ki temelji na orodju CLASSLA-Stanza, a omogoča tehnično manj podkovanemu uporabniku prijaznejšo izbiro nastavitev in prikaz rezultatov.

Ne glede na v prispevku predstavljeno nadgradnjo jezikovne infrastrukture za medjezikovno primerljivo slovnično analizo slovenskih besedil pa je tako z vidika jezikovnotehnološke kot jezikoslovne uporabe te rezultate smiselno kontinuirano nadgrajevati tudi v prihodnje, kar vključuje tako izboljšavo izhodiščnih smernic na eni strani kot nadgradnjo in razvoj novih virov in tehnologij na drugi. Med drugim lahko pomembne rezultate pričakujemo tudi v okviru več tekočih nacionalnih projektov, ki se ukvarjajo s slovničnim označevanjem govorjenega jezika, ter v okviru

21 https://orodja.cjvt.si/oznacevalnik

evropske mreže COST UniDive,[22] ki se ukvarja z jezikovno raznolikostjo in univerzalnostjo v kontekstu jezikovnih tehnologij.

Zahvala

Predstavljeno delo so podprli projekt *Razvoj slovenščine v digitalnem okolju*, ki sta ga financirala Ministrstvo za kulturo Republike Slovenije in Evropski sklad za regionalni razvoj, ter raziskovalni program *Jezikovni viri in tehnologije za slovenski jezik* (št. P6-0411) in raziskovalni projekt *Na drevesnici temelječ pristop k raziskavam govorjene slovenščine* (št. Z6-4617), ki ju financira Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna. Zahvala gre tudi označevalcem novih podatkov ter Tomažu Erjavcu, Luku Krsniku, Cyprianu Laskowskemu in Mihaelu Šinkcu za tehnično podporo.

Literatura

Arhar Holdt, S., & Gorjanc, V. (2007). Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. *Jezik in Slovstvo*, *52*, 95–110.

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

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

Chen, X., & Gerdes, K. (2018). How Do Universal Dependencies Distinguish Language Groups? In J. Jiang & H. Liu (Eds.), *Quantitative Analysis of Dependency Structures* (pp. 277–294). De Gruyter Mouton. doi: 10.1515/9783110573565-014

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

de Castilho, R., Mújdricza-Maydt, É., Yimam, S. M., Hartmann, S., Gurevych, I., Frank, A., & Biemann, C. (2016). A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)* (pp. 76–84). Pridobljeno s https://aclanthology.org/W16-4011

22 https://www.cost.eu/actions/CA21167/

de Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., & Manning, C. D. (2014). Universal Stanford dependencies: A cross-linguistic typology. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (pp. 4585–4592). Pridobljeno s http://www.lrec-conf.org/proceedings/lrec2014/pdf/1062_Paper.pdf

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

Dobrovoljc, K., Erjavec, T., & Krek, S. (2016). Pretvorba korpusa ssj500k v Univerzalno odvisnostno drevesnico za slovenščino. *Zbornik Konference Jezikovne Tehnologije in Digitalna Humanistika, 29. September - 1. Oktober 2016, Filozofska Fakulteta, Univerza v Ljubljani, Ljubljana, Slovenija* (str. 190–192). Pridobljeno s http://www.sdjt.si/wp/wp-content/uploads/2016/09/JTDH-2016_Dobrovoljc-et-al_Pretvorba-korpusa-ssj500k.pdf

Dobrovoljc, K., Erjavec, T., & Krek, S. (2017). The Universal Dependencies Treebank for Slovenian. *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing* (pp. 33–38). doi: 10.18653/v1/W17-1406

Dobrovoljc, K., Erjavec, T., & Ljubešić, N. (2019). Improving UD processing via satellite resources for morphology. *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, (pp. 24–34). Pridobljeno s https://doi.org/10.18653/v1/W19-8004

Dobrovoljc, K., Krek, S., Holozan, P., Erjavec, T., Romih, M., Arhar Holdt, Š., Čibej, J., Krsnik, L., & Robnik-Šikonja, M. (2019). *Morphological lexicon Sloleks 2.0*. Pridobljeno s http://hdl.handle.net/11356/1230

Dobrovoljc, K., & Ljubešić, N. (2022). Extending the SSJ Universal Dependencies Treebank for Slovenian: Was It Worth It? *Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) within LREC2022*, (pp. 15–22). Pridobljeno s https://aclanthology.org/2022.law-1.3

Dobrovoljc, K., & Nivre, J. (2016). The Universal Dependencies Treebank of Spoken Slovenian. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 1566–1573). Pridobljeno s https://aclanthology.org/L16-1248

Dobrovoljc, K., & Terčon, L. (2023). Universal Dependencies: Smernice za označevanje besedil v slovenščini. Pridobljeno s https://wiki.cjvt.si/attachments/23

Dobrovoljc, K., Marušič, F., Mišmaš, P. & Žaucer, R. (2023). Odprta vprašanja pri prenosu označevalne sheme Universal Dependencies na slovenska besedila: Priloga k smernicam. Pridobljeno s https://wiki.cjvt.si/attachments/25

Dozat, T., & Manning, C. D. (2016). Deep Biaffine Attention for Neural Dependency Parsing. *5th International Conference on Learning Representations, ICLR 2017 – Conference Track Proceedings*. doi: 10.48550/arxiv.1611.01734

Erjavec, T. (2013). Korpusi in konkordančniki na strežniku nl.ijs.si. *Slovenščina 2.0: Empirične, Aplikativne in Interdisciplinarne Raziskave*, *1*(1), 24–49. doi: 10.4312/slo2.0.2013.1.24-49

Erjavec, T., Dobrovoljc, K., Fišer, D., Javoršek, J. J., Krek, S., Kuzman, T., Laskowski, C. A., Ljubešić, N., & Meden, K. (2022). Raziskovalna infrastruktura CLARIN.SI. In D. Fišer & T. Erjavec (Eds.), *Jezikovne tehnologije in digitalna humanistika: zbornik konference* (pp. 47–54). Inštitut za novejšo zgodovino. Pridobljeno s https://nl.ijs.si/jtdh22/pdf/JTDH2022_Erjavec-et-al_Raziskovalna-infrastruktura-CLARIN.SI.pdf

Erjavec, T., Fišer, D., Čibej, J., Arhar Holdt, Š., Ljubešić, N., Zupan, K., & Dobrovoljc, K. (2019). *CMC training corpus Janes-Tag 2.1*. Pridobljeno s http://hdl.handle.net/11356/1238

Erjavec, T., Fišer, D., Krek, S., & Ledinek, N. (2010, May). The JOS Linguistically Tagged Corpus of Slovene. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Pridobljeno s http://www.lrec-conf.org/proceedings/lrec2010/pdf/139_Paper.pdf

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(33), 10336–10341. doi: 10.1073/PNAS.1502134112/SUPPL_FILE/PNAS.1502134112.ST01.PDF

Guzmán Naranjo, M., & Becker, L. (2018). Quantitative Word Order Typology with UD. *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018), December 13–14, 2018, Oslo University, Norway* (pp. 91–104).

Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*.

Ide, N., & Pustejovsky, J. (2017). Handbook of linguistic annotation / Nancy Ide, James Pustejovsky, editors. In *Handbook of linguistic annotation*. Springer.

Jurafsky, D., & Martin, J. H. (2021). *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, 3rd Edition Draft*. Prentice Hall, Pearson Education International.

Krek, S., Erjavec, T., Dobrovoljc, K., Gantar, P., Arhar Holdt, Š., Čibej, J., & Brank, J. (2020). The ssj500k training corpus for Slovene language processing. *Jezikovne Tehnologije in Digitalna Humanistika*, 24–33. Pridobljeno s http://nl.ijs.si/jtdh20/pdf/JT-DH_2020_Krek-et-al_The-ssj500k-Training-Corpus-for-Slovene-Language-Processing.pdf

Krsnik, L., Dobrovoljc, K., & Robnik-Šikonja, M. (2019). *Dependency tree extraction tool STARK 1.0*. Pridobljeno s http://hdl.handle.net/11356/1284

Ledinek, N. (2018). Skladenjska analiza slovenščine in slovenski jezikoslovno označeni korpusi. *Jezik in Slovstvo*, *63*(2/3), 103–116. Pridobljeno s http://www.dlib.si/details/URN:NBN:SI:doc-N94NNL3K

Lenardič, J., Čibej, J., Arhar Holdt, Š., Erjavec, T., & Fišer, D. (2022). *CMC training corpus Janes-Norm 3.0*. Pridobljeno s http://hdl.handle.net/11356/1733

Ljubešić, N., & Dobrovoljc, K. (2019). What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian. *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing* (pp. 29–34). doi: 10.18653/v1/W19-3704

Ljubešić, N., & Erjavec, T. (2018). *Word embeddings CLARIN.SI-embed.sl 1.0*. Pridobljeno s http://hdl.handle.net/11356/1204

Martelli, F., Navigli, R., Krek, S., Kallas, J., Gantar, P., Koeva, S., Nimb, S., Sandford Pedersen, B., Olsen, S., Langemets, M., Koppel, K., Üksik, T., Dobrovoljc, K., Ureña-Ruiz, R., Sancho-Sánchez, J.-L., Lipp, V., Váradi, T., Győrffy, A., László, S., … Munda, T. (2022). *Parallel sense-annotated corpus ELEXIS-WSD 1.0*. Pridobljeno s http://hdl.handle.net/11356/1674

Martelli, F., Navigli, R., Krek, S., Tiberius, C., Kallas, J., Gantar, P., Koeva, S., Nimb, S., Pedersen, B. S., Olsen, S., Langements, M., Koppel, K., Üksik, T., Dobrovolijc, K., Ureña-Ruiz, R.-J., Sancho-Sánchez, J.-L., Lipp, V., Varadi, T., Györffy, A., … Munda, T. (2021). Designing the ELEXIS Parallel Sense--Annotated Dataset in 10 European Languages. *ELex 2021 Proceedings*. Pridobljeno s https://elex.link/elex2021/

Nguyen, M. van, Lai, V. D., Pouran Ben Veyseh, A., & Nguyen, T. H. (2021). Trankit: A Light-Weight Transformer-based Toolkit for Multilingual Natural Language Processing. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations* (pp. 80–90). doi: 10.18653/v1/2021.eacl-demos.10

Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., & Zeman, D. (2020). Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 4034–4043). Pridobljeno s https://aclanthology.org/2020.lrec-1.497

Petrov, S., Das, D., & McDonald, R. (2012). A Universal Part-of-Speech Tagset. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 2089–2096). Pridobljeno s http://www.lrec-conf.org/proceedings/lrec2012/pdf/274_Paper.pdf

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). *Stanza: A Python Natural Language Processing Toolkit for Many Human Languages*. arXiv. doi: 10.48550/ARXIV.2003.07082

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., & Guzmán, F. (2021). WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 1351–1361). doi: 10.18653/v1/2021.eacl-main.115

Štravs, M., & Dobrovoljc, K. (2022). *Service for querying dependency treebanks Drevesnik 1.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1715

Terčon, L., & Ljubešić, N. (2023). The CLASSLA-Stanza model for UD dependency parsing of standard Slovenian 2.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1769

Terčon, L. & Ljubešić, N. (2023). CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages. arXiv. doi: 10.48550/arXiv.2308.04255

Toporišič, J. (2000). *Slovenska slovnica*. Založba Obzorja Maribor.

Zeman, D. (2008). Reusable Tagset Conversion Using Tagset Drivers. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Pridobljeno s http://www.lrec-conf.org/proceedings/lrec2008/pdf/66_paper.pdf

Zeman, D., Hajič, J., Popel, M., Potthast, M., Straka, M., Ginter, F., Nivre, J., & Petrov, S. (2018). CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 1–21. doi: 10.18653/v1/K18-2001

Zeman, D., Nivre, J., Abrams, M., Ackermann, E., Aepli, N., Aghaei, H., Agić, Ž., Ahmadi, A., Ahrenberg, L., Ajede, C. K., Aleksandravičiūtė, G., Alfina, I., Algom, A., Andersen, E., Antonsen, L., Aplonova, K., Aquino, A., Aragon, C., Aranes, G., … Ziane, R. (2022). *Universal Dependencies 2.10*, http://hdl.handle.net/11234/1-4758

Žitnik, S. (2019). *Slovene corpus for aspect-based sentiment analysis - SentiCoref 1.0*, http://hdl.handle.net/11356/1285

Žitnik, S., & Dragar, F. (2021). *SloBENCH evaluation framework*, http://hdl.handle.net/11356/1469

Universal Dependencies for Slovenian: An Upgrade to the Guidelines, Annotated Data and Parsing Model

Universal Dependencies (UD) is an internationally coordinated annotation scheme for cross-linguistically comparable morphosyntactic annotation of corpora, which has been applied to more than 130 other languages worldwide, including Slovenian. In this paper, we present the results of recent activities related to Slovenian UD annotation within the Development of Slovene in a Digital Environment project. During the project, we upgraded the existing infrastructure with reviewed and detailed documentation of the Slovenian UD annotation guidelines and produced four new datasets, manually annotated in accordance with the scheme. Specifically, we expanded the SSJ-UD treebank for written Slovenian with new sentences from the ssj500k and ELEXIS-WSD corpora, and created a new hidden UD treebank based on the SentiCoref corpus to be used on the SloBENCH evaluation platform. In addition, the SUK and Janes-tag reference training corpora, originally annotated using the language-specific JOS annotation scheme, have been semi-automatically converted to UD part-of-speech categories and morphological features. The new version of the reference SSJ-UD treebank with more than 5,000 new sentences and double the original number of tokens was used to train a new dependency parsing model in the CLASSLA-Stanza annotation tool. This paper gives an in-depth evaluation of its performance with respect to the overall parsing performance, the relation-specific parsing performance and the most common types of errors produced.

Keywords: linguistic annotation, dependency grammar, treebanks, dependency parsing, natural language processing