Universal Dependencies za slovenščino: nadgradnja smernic, učnih podatkov in razčlenjevalnega modela

Kaja Dobrovoljc[*†‡], **Luka Terčon**[†], **Nikola Ljubešić**[‡†]

[*]Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
kaja.dobrovoljc@ff.uni-lj.si

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

[‡]Institut "Jožef Stefan"
Jamova cesta 39, 1000 Ljubljana
nikola.ljubesic@ijs.si

Povzetek

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 ter izdelavo novega strojnega modela skladenjskega razčlenjevanja v označevalnem orodju CLASSLA-Stanza. V podporo nadaljnjim aplikacijam na različnih področjih strojnega procesiranja slovenščine novi model podrobneje ovrednotimo, in sicer poleg splošne evalvacije natančnosti razčlenjevanja poročamo tudi o natančnosti na ravni posamičnih skladenjskih relacij in o najpogostejših tipih napak.

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 pragmatiko bolj primerna za jezike s prostim 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 et al. (2021).

Doslej je bilo z označevalno shemo UD ročno označenih že več kot 200 korpusov (t.i. odvisnostnih drevesnic, angl. *dependency treebanks*) v 130 svetovnih jezikih. Med njimi sta tudi univerzalni odvisnostni drevesnici pisne slovenščine SSJ (Dobrovoljc et al., 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 et al., 2018), kakor tudi raznolike primerjalnojezikoslovne raziskave (Futrell et al., 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

[1] https://universaldependencies.org/

[2] https://slovenscina.eu/

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 predstavimo v nadaljevanju prispevka, v katerem po kratki predstavitvi izhodiščne različice korpusa SSJ-UD pred začetkom projekta RSDO (2. razdelek) opišemo dokumentacijo rahlo prenovljenih označevalnih smernic UD za slovenščino (3. razdelek). Nadaljujemo s predstavitvijo označevalne kampanje (4. razdelek), v okviru katere je bilo ročno razčlenjenih več kot 5.000 novih povedi, ki skupaj z nekoliko izboljšanim prvotnih korpusom tvorijo najnovejšo različico korpusa SSJ-UD (5. razdelek). V drugem delu prispevka opišemo izdelavo na novem korpusu temelječega napovednega modela za strojno skladenjsko razčlenjevanje (6. razdelek), ki ga v sklepnem delu tudi ovrednotimo z analizo splošne natančnosti (7. razdelek) in analizo najpogostejših napak (8. razdelek).

2. Nastanek korpusa SSJ-UD

Prva različica univerzalne odvistnostne drevesnice za pisno slovenščino SSJ-UD[3] je nastala na podlagi na polavtomatske pretvorbe korpusa ssj500k (Krek et al., 2020), bogato označenega referenčnega učnega korpusa za slovenščino, ki je bil predhodno že ročno lematiziran, oblikoskladenjsko označen in skladenjsko razčlenjen po označevalnem sistemu JOS (Erjavec et al., 2010). Medtem ko so leme in oblikoskladenjske oznake JOS pripisane vsem pojavnicam korpusa ssj500k (586.248 pojavnic oz. 27.829 povedi), je skladenjsko razčlenjena zgolj slaba polovica korpusa (235.864 pojavnic oz. 11.411 povedi).

Pretvorba korpusa ssj500k iz označevalne sheme JOS v shemo UD (Dobrovoljc et al., 2016; Dobrovoljc et al., 2017) je temeljila na širokem naboru pravil za preslikavo za vse tri ravni sheme UD: besedne vrste, oblikoslovne lastnosti in odvisnostne skladenjske relacije.[4] Ker so si (z nekaj izjemami) načela označevanja obeh sistemov na ravni oblikoslovja precej podobna, je bilo mogoče s pravili za preslikavo v besedne vrste in oblikoskladenjske lastnosti UD pretvoriti celoten korpus ssj500k oz. na istem sistemu temelječi leksikon Sloleks (Dobrovoljc et al., 2019), pri čemer je bilo ročno razdvoumljanje potrebno zgolj pri besednovrstni kategorizaciji glagola *biti*.[5]

Po drugi strani pa je bil skladenjsko razčlenjeni del korpusa ssj500k v shemo UD pretvorjen le delno, saj zaradi robustnosti sistema JOS v primerjavi z UD kljub podrobnemu sistemu pravil za preslikavo vseh povedi ni bilo mogoče v celoti samodejno pretvoriti z dovolj zanesljivo natančnostjo. Med nepretvorjenimi so tako ostale zlasti povedi s strukturami, ki so bile v sistemu JOS označene kot t. i. povezave tretjega nivoja (oznaka *modra*), kot so stavčna priredja in soredja, pristavki in pojasnjevalne strukture, členki oz. nepropozicijskimi prislovi, vrivki in podobno.

Prvotna različica korpusa SSJ-UD, prvič objavljena kot del zbirke drevesnic UD v1.2 leta 2015, je tako obsegala 8.000 povedi oz. 140.670 pojavnic. Kljub kontinuiranemu izboljševanju korpusa s prilagajanjem spremembam v splošnih označevalnih smernicah in odpravljanjem posamičnih napak je njegova velikost do nedavne razširitve, ki jo predstavimo v 4. razdelku tega prispevka, ostajala ves čas nespremenjena.

3. Popis smernic UD za slovenščino

Splošne smernice UD, kakršne so dokumentirane na krovni spletni strani projekta,[6] so kot nadaljevanje predhodnih standardizacijskih iniciativ in večletnega kolaborativnega razvoja zasnovane tako, da skušajo na čim krajši način nasloviti 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 infrastuktura 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 posamične jezike prepuščena predvsem samoiniciativnosti avtorjev podatkov.

Za slovenščino so bile ob prvi objavi korpusa SSJ-UD tako dokumentirane zgolj smernice za pripisovanje besednih vrst in oblikoskladenjskih oznak, ki so odtlej ob prehodu z UD v1 na UD v2 (Nivre et al., 2020) že nekoliko zastarele, smernice za pripisovanje skladenjskih relacij UD besedilom v slovenščini pa zaradi obsežnosti niso bile podrobneje dokumentirane oz. so bile razvidne zgolj implicitno iz pretvorbenih pravil na eni strani in 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, 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 univerzalne smernice. Med njimi lahko izpostavimo predvsem spremembe v obravnavi primerjalnih struktur (lastnost kot nadredni element primerjave), poudarjalnih členkov (razlikovanje med modifikatorji samostalnikov na eni in povedkov na drugi strani), besedilnih povezovalcev (razlikovanje glede na stavčno pozicijo) in prostega morfema *se*/*si* (razlikovanje med zaimki v predmetni in ekspletivni vlogi), ki

[3] V tem prispevku namesto uradnega imena drevesnice (SSJ) zaradi podobnosti s poimenovanji sorodnih korpusov in projektov v slovenskem prostoru uporabljamo daljši akronim SSJ-UD.

[4] Pravila in skripte za pretvorbo iz sistema JOS v sistem UD so na voljo na `https://github.com/clarinsi/jos2ud`.

[5] V nasprotju s sistemom JOS, znotraj katerega so pojavitve glagola *biti* ne glede na skladenjsko vlogo ali pomen vedno označene kot glagol s podvrsto *pomožni*, sistem UD že na ravni besednih vrst ločuje med glavnimi (oznaka VERB) in pomožnimi glagoli (oznaka AUX), kamor se umeščajo glagoli v vlogi pomožnikov in veznih glagolov.

[6] `https://universaldependencies.org/guidelines.html`

so bili zaradi omejitev strojne pretvorbe iz sistema JOS prvotno označeni drugače kot predvidevajo splošne smernice UD.

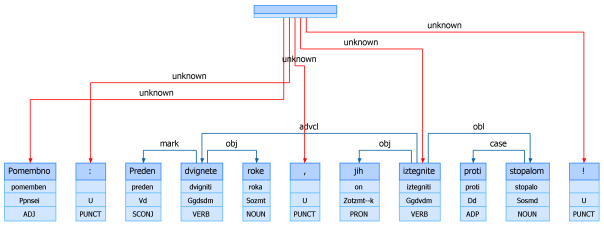
Priročnik s smernicami UD za slovenščino[7] poleg opisov posamičnih slovničnih kategorij in načel njihovega pripisovanja besedilom v slovenščini vsebuje še razdelek s podrobnejšo obravnavo težavnejših primerov, ki se je dopolnjeval tudi skozi označevalno kampanjo, opisano v 4. razdelku. V pripravi je tudi objava slovenskih smernic na uradni spletni strani UD (v angleščini) in popis odprtih vprašanj z izhodiščnimi priporočili za nadaljnje izboljšave (v sodelovanju z Univerzo v Novi Gorici) .

4. Nadgradnja korpusa SSJ-UD

V drugem koraku projekta je sledila označevalna kampanja, v okviru katere smo ročno označili več kot 5.000 novih povedi iz korpusov ssj500k oz. ELEXIS-WSD, nekoliko izboljšana pa je bila tudi označenost prvotne različice korpusa SSJ-UD. V vseh treh fazah je označevanje potekalo v označevalnem orodju Q-CAT (Brank, 2022), ki odslej podpira tudi standardni format CONLL-U, za primerjavo označenih datotek (kuriranje) pa smo uporabili lokalno inštalacijo orodja WebAnno (Eckart de Castilho et al., 2016), ki jo vzdržuje CLARIN.SI.[8] Podrobnejša analiza označevalnega procesa je opisana v prispevku Dobrovoljc in Ljubešić (2022), v nadaljevanju pa predstavimo zgolj najpomembnejše rezultate.

4.1. Razširitev s polpretvorjenimi povedmi iz ssj500k

Kot smo omenili že v 2. razdelku, nekaterih skladenjsko razčlenjenih povedi v korpusu ssj500k zaradi omejitev pretvorbenih pravil ni bilo mogoče v celoti pretvoriti v oznake UD, zato niso bile vključene v prvotno različico drevesnice SSJ-UD, predstavljale pa so logično izhodišče za nadaljnjo širitev podatkov UD za slovenščino. V prvi fazi razširitve so tako označevalci ročno pregledali teh 3.411 polpretvorjenih povedi oz. 96.194 pojavnic, med katerimi jih 22.377 (23,5 %) 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.



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

Med pojavnicami, ki v izhodišču niso imele pripisane relacije UD, je bila skoraj polovica ločil (*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*, običajno jedro povedka glavnega stavka ali drug hierarhično najpomembnejši element v povedi), 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. prireedij, torej struktur, kakršnih zgolj s pravili ni bilo mogoče pretvoriti z dovolj zanesljivo natančnostjo.

4.2. Razširitev s povedmi iz korpusa ELEXIS-WSD

V drugi fazi širitve je bil skladenjsko razčlenjen še korpus ELEXIS-WSD-SL, tj. slovenski del paralelnega korpusa ELEXIS-WSD (Martelli et al., 2021; Martelli et al., 2022), razvitega za potrebe strojnega pomenskega razdvoumljanja, ki vsebuje v več evropskih jezikov prevedena besedila iz Wikipedie (Schwenk et al., 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* pa razdvoumili ročno.

Tako označen korpus je bil izhodiščno skladenjsko razčlenjen z orodjem CLASSLA-Stanza (Ljubešić in Dobrovoljc, 2019), 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*, *advmod*, *obl*, *conj* in *punct*, kar se, kot bomo videli v nadaljevanju, sklada z najpogostejšimi tipi napak razčlenjevalnika nasploh (8. razdelek).

4.3. Izboljšanje označenosti v prvotnem korpusu

Poleg dodajanja novih razčlenjenih povedi smo glede na rahlo spremembo smernic (3. razdelek), analizo ročnih popravkov pretvorjenih relacij (razdelek 4.1.) 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,[9] nedosledno ločevanje med sorednimi in prireedno vezanimi stavki, med premimi in nepremimi predmeti, itd. Za vsako izmed kategorij smo s hevrističnimi 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 %).

[7] Priročnik je trenutno na voljo v delovni različici, uradno pa bo objavljen ob zaključku projekta RSDO.

[8] https://www.clarin.si/webanno/.

[9] 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, a vendarle redek.

5. Nova različica korpusa SSJ-UD

V zadnjem koraku smo izhodiščni korpus SSJ-UD z nekoliko izboljšano označenostjo (razdelek 4.3.) združili z novimi povedmi iz korpusov ssj500k (razdelek 4.1.) in ELEXIS-WSD (razdelek 4.2.) ter tako dobili novo različico referenčne univerzalne odvistnostne drevesnice za pisno slovenščino SSJ-UD,[10] ki je bila prvič objavljena kot del uradnega izida UD v2.10 (Zeman et al., 2022). Ob zaključku projekta RSDO bo drevesnica SSJ-UD integrirana tudi v novi referenčni korpus učne slovenščine SUK.

5.1. Sestava korpusa

Kot prikazuje tabela 5.1., 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 na 30. mesto med skupno 228 drevesnicami UD. 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.

Medtem ko besedila ssj500k kot vzorec korpusa Fida-PLUS (Arhar Holdt, 2007) vsebujejo predvsem izvorno slovenska leposlovna, neleposlovna in publicistična besedila, korpus ELEXIS-WSD vsebuje prevedena enciklopedična besedila iz Wikipedie. Po drugi strani sta si izvorni SSJ-UD in korpus ELEXIS-WSD podobna z vidika kompleksnosti (krajše in skladenjsko enostavnejše povedi), medtem ko so nove povedi iz ssj500k bistveno daljše.

Nenazadnje pa je z metodološkega vidika pomembno izpostaviti še, da se vsi trije podkorpusi razlikujejo tudi z vidika izvora pripisanih oznak UD, saj so oznake prvotnega SSJ-UD v veliki večini rezultat avtomatske pretvorbe iz sistema JOS, oznake novih povedi iz ssj500k kombinacija pretvorbe in ročnega pregleda, oznake povedi iz korpusa ELEXIS-WSD pa so bile v celoti pregledane ročno.

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

Tabela 1: Zgradba nove različice korpusa SSJ-UD (od UD v2.10 naprej).

5.2. Delitev podatkovnih množic

Del objave drevesnice v uradni zbirki UD je tudi njena delitev na učno, validacijsko in testno množico, ki se standardno uporabljajo pri razvoju in evalvaciji na teh podatkih temelječih napovednih modelov. Pri tem smo sledili načelom delitve podatkov v prvotni različici, v kateri so bile podmnožice razdeljene glede na zaporedje pojavljanja v korpusu. Glede na to, da so nove povedi iz ssj500k enakomerno razpršene po celotnem korpusu, smo te zgolj priključili k že obstoječi delitvi povedi v prvotni različici in ohranili enako razmerje (80 % učna, 10 % validacijska, 10 % testna), nato pa smo vsaki izmed treh množic v enakem razmerju dodali še povedi iz korpusa ELEXIS-WSD. Sestava podmnožic tako odslikava raznolikost nove različice korpusa SSJ-UD, kakršno opisujemo v razdelku 5.1., in z reprezentativnostjo testnih podatkov glede na učne zagotavlja ustreznejšo, besedilnozvrstno manj pristransko evalvacijo.

6. Razčlenjevalni model

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

Tako pri učenju kot evalvaciji razčlenjevalnega modela 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.

Izgradnjo napovednega modela, njegovo primerjavo z modelom, naučenim na prvotni različici SSJ-UD, in evalvacijo glede na posamične podkorpuse podrobneje opisujeta Dobrovoljc in Ljubešić (2022), ki ugotavljata, da je model, naučen na novi različici korpusa SSJ-UD, 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 (7. razdelek) na eni strani in analizo najpogostejših tipov napak (8. razdelek) na drugi.

[10] 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.

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

7. Splošna natančnost

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 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.[12]

Rezultati, predstavljeni v tabeli 7., prikazujejo, da razčlenjevalni model dosega splošno natančnost 93,21 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.[13]

Kot prikazujejo rezultati za posamične tipe relacij,[14] pa ta splošna ocena natančnosti ni reprezentativna za vse vrste skladenjskih struktur, saj je pri napovedovanju nekaterih relacij model bistveno natančnejši kot pri drugih.

Med relacijami z najvišjo natančnostjo napovedovanja so po pričakovanju funkcijske besede, kot so predlogi (*case*; 99,17), pomožni glagol *biti* (*aux*; 98,93), določilniški zaimki in prislovi (*det*; 98,79), podredni vezniki (*mark*; 98,69), ekspletivni zaimki (*expl*; 96,71) in priredni vezniki (*cc*; 96,27), 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*; 95.53) in osebki (*nsubj*; 95.28), nadpovprečno uspešen pa je tudi pri identifikaciji korena povedi (*root*; 96,26), ki je običajno jedro povedka glavnega stavka, in veznega glagola *biti* (*cop*; 95,43), ki nastopa v strukturah s povedkovimi določili.

Med relacijami, pri napovedovanju katerih model dosega najslabše rezulate, pričakovano najdemo ogovore (*vocative*; 0,0), saj se v testni množici pojavi zgolj en primer, in nedoločene strukture (*dep*; 54,55), 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*, 63,40), 'osirotelih' stavčnih členov v povedih z glagolsko elipso (*orphan*; 68,24), diskurznih členkov (*discourse*; 69,23), stavčnih soredij (*parataxis*; 70,35) in naštevalnih seznamov (*list*; 75,86) z novo različico korpusa SSJ-UD bistveno izboljšala glede na prvotni model (Dobrovoljc in 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*; 75,86), prilastkovi (*acl*; 81,73), osebkovi (*csubj*; 85,53) in predmetni odvisniki (*ccomp*; 90,67). Poleg nepremih predmetov (*iobj*; 81,66), ki jih je težavno identificirati predvsem zaradi pomanjkljivosti trenutnih označevalnih smernic,[15] modelu precejšen izziv predstavljajo tudi priredja, zlasti medstavčna (*conj*; 85,91), samostalniški prilastki (*nmod*; 87,44) in prislovna določila povedkov, samostalnikov in pridevnikov (*advmod*; 89,95).

8. Najpogostejše napake

V drugem koraku evalvacije smo analizo zanesljivosti modela pri razčlenjevanju posameznih tipov relacij dopolnili še s podrobnejšo analizo najpogostejših tipov napak. Tabela 8. 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 pet najpogostejših podtipov 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.

8.1. Napačna napoved nadrejenega elementa

Kot prikazuje tabela 8., 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. Večinoma gre za primere, kjer so napačno določena tudi

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

[13] Ta natančnost je v skladu z natančnostjo orodja Stanza za druge jezike oz. drevesnice (`https://stanfordnlp.github.io/stanza/performance.html`) oz. natančnostjo drugih sodobnih razčlenjevalnikov nasploh (`https://universaldependencies.org/conll18/results.html`), vendar neposredna primerjava zaradi specifik evalvacijske metodologije ni smiselna.

[14] V Tabeli 7. ni relacije *compound*, ki je glede na smernice v slovenščini ne uporabljamo. Pri relacijah *dislocated*, *goeswith* in *reparandum* podatkov o natančnosti ni (oznaka *n/a*), saj se v testni množici ne pojavijo. O natančnosti izpeljanih relacij oz. podoznak (npr. *flat:name*, *flat:foreign*) poročamo združeno z jedrno oznako (npr. *flat*).

[15] Zaradi kompleksnega prepletanja oblikoslovnih, skladenjskih in pomenskih razločevalnih lastnosti med premimi in nepremimi predmeti trenutne smernice UD priporočajo, da je v povedih z zgolj enim izraženim predmetom ta ne glede na sklon ali udeleženso vlogo označen kot premi predmet (*obj*). To pomeni, da se lahko tudi predmeti v dajalniku, kakršni tipično nastopajo kot nepremi predmeti, ob odsotnosti drugih predmetov označujejo kot premi.

Relacija	Izvorni opis	Slovenski prevod	LAS F1
acl	clausal modifier of noun	stavčni prilastki	81,73
advcl	adverbial clause modifier	prislovni odvisniki	75,86
advmod	adverbial modifier	prislovna določila (gl. opombo 16)	89,95
amod	adjectival modifier	pridevniški prilastki	98,9
appos	appositional modifier	pristavčna določila	63,4
aux	auxiliary verb	pomožni glagoli	98,93
case	case marking preposition	predlogi	99,17
cc	coordinating conjunction	priredni vezniki	96,27
ccomp	clausal complement	stavčna dopolnila (predmetni odvisniki)	90,67
conj	conjunct	priredno zloženi elementi	85,91
cop	copula verb	vezni glagoli	95,43
csubj	clausal subject	osebkovi odvisniki	85,53
dep	unspecified dependency	nedoločena povezava	54,55
det	determiner	določilniki	98,79
discourse	discourse element	diskurzni členki	69,23
dislocated	dislocated element	dislocirani elementi	n/a
expl	expletive	ekspletivne besede	96,71
fixed	fixed multi-word expression	funkcijske zveze	93,33
flat	flat multi word-expression	eksocentrične zveze	92,12
goeswith	disjointed token	razdruženi deli besed	n/a
iobj	indirect object	nepremi predmeti	81,66
list	list	seznami	75,86
mark	marker (subordinating conjunction)	podredni vezniki	98,69
nmod	nominal modifier	samostalniški prilastki	87,44
nsubj	nominal subject	samostalniški osebki	95,28
nummod	numeric modifier	številčna določila	94,23
obj	(direct) object	premi predmeti	95,53
obl	oblique nominal (adjunct)	odvisne samostalniške zveze	91,14
orphan	dependent of missing parent	elementi v eliptičnih strukturah	68,24
parataxis	parataxis	stavčna soredja	70,35
punct	punctuation symbol	ločila	93,08
reparandum	overriden disfluency	samopopravljanja	n/a
root	root element	koren povedi	96,26
vocative	vocative	ogovori	0
xcomp	open clausal complement	odprta stavčna dopolnila	92,87
Vse relacije			**93,21**

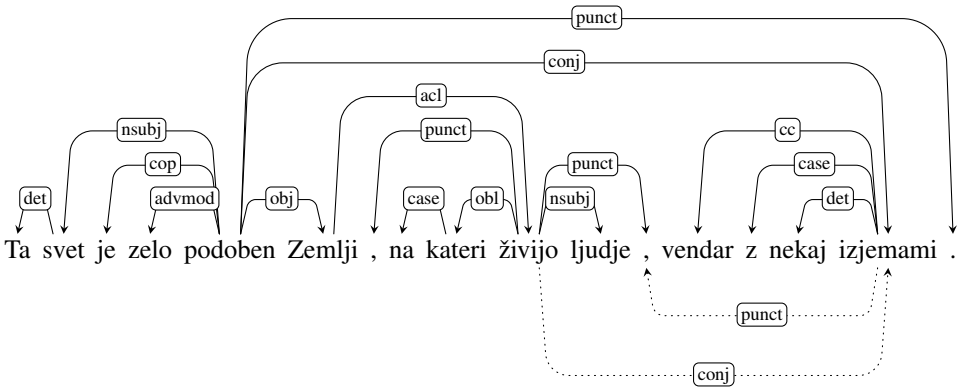
Tabela 2: Natančnost novega modela orodja CLASSLA-Stanza za skladenjsko razčlenjevanje po sistemu UD glede na metriko LAS.

jedra drugih struktur v povedi, na katera se ločila praviloma povezujejo. Napačno povezana ločila so torej predvsem 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.

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**,[16] 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 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.

Pri preostalih treh analiziranih relacijah s pogosto napačno pripisanim izvorom povezave, tj. **nmod**, **conj** in **acl**, 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).

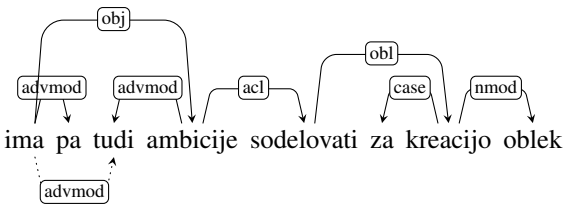
[16] Relacija *advmod* se uporablja za označevanje prislovov v vlogi modifikatorjev, kar vključuje tako prislove v vlogi okoliščinskih dopolnil povedkov (kakršna Slovenska slovnica imenuje prislovna določila, npr. *pridem takoj*) kot prislove v vlogi modifikatorjev pridevniških, prislovnih ali samostalniških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



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

Tip napake	Število napak
Napačno jedro	**914**
punct-punct	248
advmod-advmod	166
nmod-nmod	111
conj-conj	99
acl-acl	53
Napačno jedro in oznaka	**517**
obl-nmod	141
parataxis-root	37
acl-advcl	22
root-nsubj	22
nsubj-nmod	19
Napačna oznaka	**299**
conj-parataxis	23
obl-nsubj	19
appos-conj	17
obl-obj	13
iobj-obj	13
Vse napake	**1730**

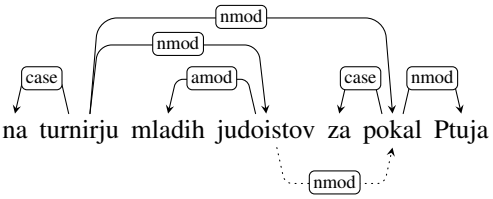
Tabela 3: Distribucija napak razčlenjevalnega modela glede na tip napake.



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

8.2. 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



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

zamenjevanje struktur z oznakama **obl**[17] in **nmod**, ki predstavlja tretji najpogostejši (pod)tip napak nasploh. Analiza primerov kaže, da gre večinoma za primere, 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.

Manj pogoste v tej kategoriji so še napake pri določanju glavnega stavka v nizu dveh ali več soredno zloženih stavkov, zlasti kadar gre za vrinjene stavke ali premi govor (**parataxis-root**), napake ločevanja med prislovno-določilnimi odvisniki in stavčnimi prilastki, pogosto v kombinaciji z veznikom *kot* (**acl-advcl**), zamenjava osebka in povedkovega določila v strukturah z veznim glagolom *biti* (**root-nsubj**) in napake določanja osebka v povedih, kjer osebek ni eksplicitno izražen (**nsubj-nmod**).

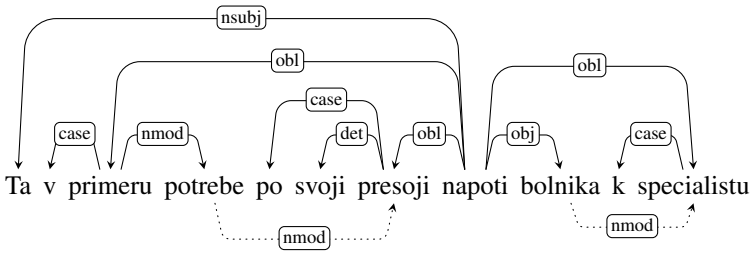
8.3. Napačna napoved relacije

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**[18] prihaja pred-

[17] 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.

[18] 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.



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

vsem pri daljših povedih, pri katerih se med dva priredno zložena stavka oz. med priredni veznik in drugi stavek v priredju vrivajo druge strukture (npr. odvisniki). Samostalniška prislovna določila (ki prejmejo relacijo **obl**) 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*).

Med drugimi tipi napačno pripisanih relacij je pogosta še dvoumnost med samostalniškimi zvezami v vlogi pristavčnih določil (**appos**) na eni in priredno povezanih elementov (**conj**) na drugi strani, zlasti kadar zadnji element v brezvezniškem priredju stoji na koncu povedi. 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**).

9. Zaključek

V prispevku smo predstavili nadgradnjo drevesnice SSJ-UD, referenčnega ročno skladenjsko razčlenjenega korpusa po medjezično usklajeni shemi Universal Dependencies, v okviru katere smo po rahli prenovi in izčrpni dokumentaciji označevalnih smernic za slovenščino korpus razširili z novimi povedmi ter nato na novi učni množici naučili tudi nov napovedni model za skladenjsko razčlenjevanje slovenskih besedil. Podrobna kvantitativna in kvalitativna analiza njegove natančnosti je pokazala, da model v splošnem dosega razmeroma dobre rezultate, pri čemer je pri členjenju nekaterih struktur mogoče pričakovati bistveno večjo zanesljivost rezultatov kot pri drugih.

Glede na mednarodno relevantnost sheme UD 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. Nguyen et al. (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 je denimo celotni korpus ssj500k oz. nastajajoči učni korpus SUK.

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, obsežen ročno označen korpus in sistematična evalvacija natančnosti na njem naučenega modela predstavljajo pomemben doprinos k nadaljnjim jezikoslovnim raziskavam ročno in strojno razčlenjenih slovenskih korpusov, pri čemer je glede na kompleksno strukturo tovrstnih korpusov nujno vzpostaviti tudi ustrezno infrastrukturo za njihovo analizo.

Seveda pa je tako z vidika jezikovnotehnološke kot jezikoslovne uporabe predstavljene rezultate smiselno kontinuirano nadgrajevati tudi v prihodnje, kar vključuje tako izboljšavo izhodiščnih smernic na eni strani kot njihovo dosledno implementacijo na drugi. Glede na v prispevku predstavljene metodološke razlike v nastanku posamičnih delov korpusa SSJ-UD in zaznane nedoslednosti med kvalitativno analizo napak je poleg nadaljnjega povečevanja korpusa vsekakor enako smiselna tudi konsolidacija obstoječega.

10. Zahvala

Predstavljeno delo sta podprla projekt Razvoj slovenščine v digitalnem okolju, ki ga financirata Ministrstvo za kulturo Republike Slovenije in Evropski sklad za regionalni razvoj, ter raziskovalni program Jezikovni viri in tehnologije za slovenski jezik (št. P6-0411), ki ga financira Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna. Zahvala gre tudi označevalcem novih podatkov (Tina Munda, Ina Poteko, Rebeka Roblek, Luka Terčon, Karolina Zgaga) ter Tomažu Erjavcu, Luku Krsniku, Cyprianu Laskowskemu in Mihaelu Šinkcu za tehnično podporo.

11. Literatura

Špela Arhar Holdt. 2007. Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. *Jezik in slovstvo*, 52(2).

Janez Brank. 2022. Q-CAT corpus annotation tool 1.3. Slovenian language resource repository CLARIN.SI.

Xinying Chen in Kim Gerdes. 2018. How do Universal Dependencies distinguish language groups. *Quantitative Analysis of Dependency Structures*, 72:277–294.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre in Daniel Zeman. 2021. Universal Dependencies. *Computational linguistics*, 47(2):255–308.

Kaja Dobrovoljc, Tomaž Erjavec in Simon Krek. 2016. Pretvorba korpusa ssj500k v univerzalno odvisnostno drevesnico za slovenščino. V: *Proceedings of the Conference on Language Technologies and Digital Humanities*.

Kaja Dobrovoljc, Tomaž Erjavec in Simon Krek. 2017. The Universal Dependencies Treebank for Slovenian. V: *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing, BSNLP@EACL 2017*, str. 33–38.

Kaja Dobrovoljc, Tomaž Erjavec in Nikola Ljubešić. 2019. Improving UD processing via satellite resources for morphology. V: *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, str. 24–34, Paris, France, August. Association for Computational Linguistics.

Kaja Dobrovoljc in Nikola Ljubešić. 2022. Extending the SSJ Universal Dependencies treebank for Slovenian: Was it worth it? V: *Proceedings of the 16th Linguistic Annotation Workshop (LAW 2022)*, June.

Kaja Dobrovoljc in Joakim Nivre. 2016. The Universal Dependencies treebank of spoken Slovenian. V: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, str. 1566–1573, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Timothy Dozat in Christopher D Manning. 2016. Deep biaffine attention for neural dependency parsing. *arXiv preprint arXiv:1611.01734*.

Richard Eckart de Castilho, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank in Chris Biemann. 2016. A web-based tool for the integrated annotation of semantic and syntactic structures. V: *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, str. 76–84, Osaka, Japan, December. The COLING 2016 Organizing Committee.

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

Richard Futrell, Kyle Mahowald in Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

Nancy Ide in James Pustejovsky. 2017. *Handbook of linguistic annotation*, zvezek 1. Springer.

Dan Jurafsky in James H. Martin. 2021. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, 3nd Edition Draft*. Prentice Hall series in artificial intelligence. Prentice Hall, Pearson Education International.

Simon Krek, Tomaž Erjavec, Kaja Dobrovoljc, Polona Gantar, Špela Arhar Holdt, Jaka Čibej in Janez Brank. 2020. The ssj500k training corpus for Slovene language processing. V: *Proceedings of the Conference on Language Technologies and Digital Humanities*, str. 24–33, Ljubljana, Slovenia, September. Institute of Contemporary History.

Nina Ledinek. 2018. Skladenjska analiza slovenščine in slovenski jezikoslovno označeni korpusi. *Jezik in slovstvo*, 63(2/3).

Nikola Ljubešić in Kaja Dobrovoljc. 2019. What does neural bring? Analysing improvements in morphosyntactic annotation and lemmatisation of Slovenian, Croatian and Serbian. V: *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, str. 29–34, Florence, Italy, August. Association for Computational Linguistics.

Nikola Ljubešić in Tomaž Erjavec. 2018. Word embeddings CLARIN.SI-embed.sl 1.0. Slovenian language resource repository CLARIN.SI.

Federico Martelli, Roberto Navigli, Simon Krek, Carole Tiberius, Jelena Kallas, Polona Gantar, Svetla Koeva, Sanni Nimb, Bolette Sandford Pedersen, Sussi Olsen, Margit Langements, Kristina Koppel, Tiiu Üksik, Kaja Dobrovolijc, Rafael-J. Ureña-Ruiz, José-Luis Sancho-Sánchez, Veronika Lipp, Tamas Varadi, András Györffy, Simon László, Valeria Quochi, Monica Monachini, Francesca Frontini, Rob Tempelaars, Rute Costa, Ana Salgado, Jaka Čibej in Tina Munda. 2021. Designing the ELEXIS parallel sense-annotated dataset in 10 European languages. V: *eLex 2021 Proceedings*, eLex Conference. Proceedings. Lexical Computing CZ.

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

Matías Guzmán Naranjo in Laura Becker. 2018. Quantitative word order typology with UD. V: *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018), December 13–14, 2018, Oslo University, Norway*, št. 155, str. 91–104. Linköping University Electronic Press.

Minh Van Nguyen, Viet Lai, Amir Pouran Ben Veyseh in Thien Huu Nguyen. 2021. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. V: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers in Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. V: *Proceedings of the 12th Language Resources and Evaluation Conference*, str.

4034–4043, Marseille, France, May. European Language Resources Association.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton in Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong in Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. V: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, str. 1351–1361, Online, April. Association for Computational Linguistics.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre in Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. V: *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, str. 1–21, Brussels, Belgium, October. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg et al. 2022. Universal dependencies 2.10. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (U´ FAL), Faculty of Mathematics and Physics, Charles University.