Pretvorba korpusa ssj500k v Univerzalno odvisnostno drevesnico za slovenščino

Kaja Dobrovoljc,* Tomaž Erjavec,† Simon Krek‡

* Zavod za uporabno slovenistiko Trojina
Trg republike 3, 1000 Ljubljana
kaja.dobrovoljc@trojina.si
† Odsek za tehnologije znanja, Institut »Jožef Stefan«
Jamova cesta 39, 1000 Ljubljana
tomaz.erjavec@ijs.si
‡ Laboratorij za umetno inteligenco, Institut »Jožef Stefan«
Jamova cesta 39, 1000 Ljubljana
simon.krek@ijs.si

1 Uvod

Tako kot na drugih področjih procesiranja naravnih jezikov se tudi na področju skladenjskega razčlenjevanja pojavlja vse večja potreba po poenotenju označevalnih sistemov, ki so bili razviti za označevanje posameznih jezikov ali besedilnih zbirk, saj njihova raznolikost onemogoča neposredno primerjavo podatkov in na njih temelječih orodij. Kot protiutež tovrstni razdrobljenosti je bila nedavno vzpostavljena mednarodna pobuda Universal Dependencies (UD),[1] ki si prizadeva za medjezično usklajeno skladenjsko razčlenjevanje besedil z namenom primerjalnih evalvacij, razvoja večjezičnih razčlenjevalnikov, medjezičnega učenja jezikovnih modelov in kontrastivnih jezikoslovnih analiz (Nivre et al., 2016), njeni glavni principi (Nivre 2015) pa v veliki meri temeljijo na drugih predhodnih standardizacijskih projektih (de Marneffe et al., 2014; McDonald et al., 2013; Petrov et al., 2012; Zeman, 2008). Doslej je bilo z označevalno shemo UD označenih že več kot 50 korpusov različnih svetovnih jezikov, med njimi tudi *Univerzalna odvisnostna drevesnica za slovenščino*, skladenjsko razčlenjeni korpus pisne slovenščine.

Univerzalna odvisnostna drevesnica za slovenščino predstavlja tretji samostojni označevalni sistem in četrto prosto dostopno zbirko s formaliziranimi podatki o skladenjskih razmerjih v ročno označenih besedilih v slovenskem jeziku. Ker se je sistem prvega skladenjsko razčlenjenega korpusa, Slovenske odvisnostne drevesnice (SDT, pribl. 30.000 pojavnic) (Džeroski et al., 2006; Erjavec in Ledinek, 2006), ki je izhajal iz modela Praške odvisnostne drevesnice (Hajič et al., 2001), glede na kadrovske in finančne omejitve izkazal za preveč kompleksnega, je bil v okviru projekta Jezikoslovno označevanje slovenščine (JOS) v letih 2007–2009 načrtno razvit robustnejši nabor skladenjskih kategorij (Ledinek in Erjavec 2009). Po tej shemi je bil najprej razčlenjen korpus jos100k (pribl. 100.000 pojavnic), ki je bil nato v okviru projekta Sporazumevanje v slovenskem jeziku (SSJ) v letih 2009–2011 razširjen v korpus ssj500k (Krek et al., 2015), v katerem skladenjsko razčlenjeni del predstavlja nekaj manj kot polovico celotnega korpusa (pribl. 235.000 pojavnic). Ta najobsežnejša zbirka ročno razčlenjenih besedil v slovenščini je bila tako izbrana kot osnova za izdelavo Univerzalne odvisnostne drevesnice za slovenščino, ki jo na kratko predstavljamo v nadaljevanju.

2 Pretvorba iz sistema JOS v sistem UD

Pretvorba korpusa ssj500k v Univerzalno odvisnostno drevesnico je bila zasnovana kot povsem avtomatiziran proces, pri čemer pa je ta glede na številne razlike med obema označevalnima sistemoma, zlasti na ravni skladenjskega opisa, zahteval izdelavo kompleksnega sistema pretvorbenih pravil. Po pretvorbi formata XML TEI, v katerem je izvorno zapisan korpus ssj500k, v tabelarični format CONLL-U, kakršnega predvideva sistem UD, je sledila vsebinska pretvorba na dveh ločenih ravneh: oblikoslovni, ki prinaša informacije o besedni vrsti pojavnic in njihovih oblikoslovnih lastnostih, ter skladenjski ravni, ki vključuje pripis podatka o skladenjski vlogi te pojavnice v stavku. Segmentacijska, tokenizacijska in lematizacijska načela korpusa ssj500k so ostala nespremenjena.

2.1 Oblikoskladenjska raven

Sistem UD razlikuje med 17 univerzalnimi besednimi vrstami, pri čemer se načela besednovrstnega razvrščanja posameznih tradicionalno problematičnih skupin (npr. glagolnikov in deležnikov) večinoma ujemajo z načeli sistema JOS. Med temeljnimi spremembami glede na nabor besednih vrst JOS lahko izpostavimo delitev samostalnikov na občno- in lastnoimenske, delitev veznikov na priredne in podredne,

[1] Spletna stran: http://universaldependencies.org/.

delitev glagolov na pomožne in vse ostale, delitev ločil na pravopisna ločila in simbole, premik kategorije okrajšav v kategorijo 'drugo' in premik nekaterih skupin številnikov med pridevnike. Največja novost glede na sistem JOS in dosedanje slovnične opise slovenskega jezika nasploh pa je kategorija določilnikov (*DET*), v katero se umeščajo besede, ki modificirajo samostalniške zveze in izpostavljajo njihovo referenco v kontekstu. Čeprav se ta kategorija še naprej kaže kot eden izmed preizkusnih kamnov predlaganega standarda, se v slovenski drevesnici vanjo po vzoru drugih slovanskih jezikov umeščajo nekatere skupine zaimkov (npr. *moj, ta, enak*) in prislovov (npr. *nekaj, toliko*), kadar so rabljeni v vlogi določil samostalniških besednih zvez.

Tudi po shemi UD se lahko posameznim besednim oblikam poleg besedne vrste pripisujejo podrobnejše leksikalne in slovnične lastnosti v obliki parov oblikoslovnih lastnosti in njihovih vrednosti. Za razliko od besednovrstnih kategorij njihov nabor ni končen, saj je odvisen od nabora lastnosti, med katerimi razlikujejo posamezni jeziki ali izvorne označevalne sheme, mednarodno usklajena pa so njihova poimenovanja. Seznam 22 oblikoslovnih lastnosti, ki jih vsebuje Univerzalna odvisnostna drevesnica za slovenščino, je skupaj s podrobnejšim opisom in razmerji glede na sistem JOS predstavljen v slovenski dokumentaciji na projektni spletni strani, med večjimi spremembami v primerjavi z naborom lastnosti JOS pa lahko izpostavimo predvsem uvedbo glagolske lastnosti načina (povedni, pogojni, velelni), podrobnejšo členitev številskih tipov (glavni, vrstni, množica, splošno), ki se lahko pripisujejo tudi pridevnikom, ter ukinitev kategorije dvovidskosti.

Na podlagi primerjave podobnosti in razlik v obeh sistemih je bil nato izdelan sistem pravil v obliki medtabeličnih preslikav besednih vrst in lastnosti, pri čemer se ena lastnost sistema JOS lahko prevede v različne lastnosti UD, izbira ustreznega pravila pa je v teh primerih lahko odvisna od leme besede, od drugih oblikoskladenjskih lastnosti ali njene skladenjske vloge.[2] V primeru, da neki vhodni pojavnici ustreza več pravil, imajo specifična pravila prednost pred splošnimi.

2.1 Skladenjska raven

Čeprav oba sistema skladenjskega razčlenjevanja v izhodišču temeljita na teoriji odvisnostne slovnice (Tesniere, 1959; Kübler et al., 2009), se shema UD v nekaterih vidikih od sistema JOS bistveno razlikuje. Najpomembnejša razlika izhaja iz samega obsega skladenjske analize, saj je bil sistem JOS zasnovan predvsem za razčlenjevanje vezljivostnih dopolnil povedka (stavčnih členov) in strukture besednih zvez, medtem ko sistem UD v skladenjsko analizo vključuje tudi vse druge tipe stavčnih struktur, kot so členki (*advmod*), pristavki (*appos*), nagovori (*vocative*), medmeti, pastavki in drugi elementi interakcije (*discourse*), tujejezični elementi (*foreign*), ločila (*punct*), soredja (*parataxis*) itd. Ker je taksonomija UD s 40 univerzalnimi skladenjskimi oznakami[3] bistveno obsežnejša kot nabor 10 oznak sistema JOS, tudi pri razčlenjevanju jedrnih skladenjskih struktur predvideva natančnejše opredelitve skladenjskih razmerij in izdelavo kompleksnejših skladenjskih dreves kot robustni JOS. Tipični primer so denimo besedne zveze, pri katerih sistem UD razlikuje med več različnimi tipi prilastkov (npr. *amod, nmod, nummod, advmod; det; acl*), funkcijskih modifikatorjev (npr. *case, neg, expl, cc*) in razmerij znotraj stalnih besednih zvez (npr. *mwe, name, compound, goeswith*).

Druga temeljna razlika med obema sistemoma izhaja iz deleža vključevanja semantičnih interpretacij v samo skladenjsko analizo, saj sistem JOS pri razčlenjevanju stavčnih členov na določenih mestih upošteva tudi njihovo pomensko vlogo (npr. ločevanje med načinovnimi in drugimi prislovnimi določili), medtem ko sistem UD razlikuje zgolj med t. i. jedrnimi argumenti (osebek in direktni/indirektni predmet) na eni strani ter vsemi ostalimi argumenti povedka na drugi, ne glede na stopnjo njihove vezljivostne ali pomenske obveznosti. Pri tem sistem UD argumente podrobneje razvršča tudi glede na njihovo skladenjsko strukturo, torej ločuje med besednozveznimi in stavčnimi ubeseditvami osebkov (*nsubj* proti *csubj*), predmetov (*dobj/iobj* proti *ccomp*) in drugih dopolnil (*nmod/advmod* proti *advcl*).

Skripto za samodejno pretvorbo skladenjske ravni korpusa ssj500k tako sestavlja niz številnih podrobnih pravil, ki vsaki pojavnici korpusa določijo tip skladenjske povezave in njen nadrejeni element po sistemu UD.[4] Ker zaradi robustnosti sistema JOS vseh neopredeljenih struktur (tj. struktur, vezanih na korenski element) v korpusu ssj500k ni bilo mogoče z dovolj zanesljivo natančnostjo samodejno preslikati v sistem UD, ki obenem dopušča le eno korensko povezavo v povedi, trenutna različica Univerzalne odvisnostne drevesnice za

[2] Primer pravila za pretvorbo na oblikoslovni ravni je denimo ukaz, da se po sistemu UD besedna vrsta pomožnik (*AUX*) pripiše vsem pojavnicam, ki imajo po sistemu JOS na oblikoslovni ravni pripisano kategorijo 'glagol' in vrsto 'pomožni', na skladenjski ravni pa so označeni s povezavo 'del' povezani na drugo glagolsko pojavnico.

[3] V slovenski drevesnici se trenutno pojavlja 31 različnih univerzalnih ali jezikovnospecifičnih skladenjskih oznak, saj ta poleg oznak za strukture, ki se v korpusu niso pojavljale, ne vsebuje tudi povezav, ki jih ni bilo mogoče razdvoumiti ali prepoznati samodejno.

[4] Primer pravila za pretvorbo na skladenjski ravni je denimo ukaz, da se po sistemu UD skladenjska povezava prirednega veznika (*cc*) pripiše vsem pojavnicam, ki so na oblikoslovni ravni UD kategorizirane kot priredni veznik (*CONJ*), podredni veznik (*SCONJ*), členek (*PART*) ali drugo (*SCONJ*), na skladenjski ravni JOS pa so s povezavo 'vez' povezane na pojavnico, ki je sama cilj povezave 'prir'. Za razliko od sistema JOS, kjer je v priredjih veznik podrejen zadnjemu elementu priredja, se po sistemu UD tej pojavnici kot nadrejena pojavnica pripiše prvi element priredja.

slovenščino vsebuje manj povedi kot izhodiščni korpus ssj500k (7.996 proti 11.411), a je tako po obsegu (140.418 pojavnic) kot povprečni dolžini povedi (17,6 pojavnic na poved) primerljiva z univerzalnimi drevesnicami za druge jezike.

3 Dostopnost in nadaljnje raziskave

Najnovejša, druga različica Univerzalne odvisnostne drevesnice za slovenščino je bila skupaj s 54 drevesnicami za 40 drugih svetovnih jezikov, vključno s komplementarno Univerzalno odvisnostno drevesnico govorjene slovenščine (Dobrovoljc in Nivre, 2016), objavljena kot del zbirke Universal Dependencies 1.3 (Nivre et al., 2016), pod licenco CC BY-NC-SA 4.0. Po njej in drugih drevesnicah je mogoče brskati preko dveh spletnih konkordančnikov,[5] poleg številnih večjezičnih razčlenjevalnih sistemov, ki temeljijo na tej ali predhodnih različicah te korpusne zbirke in potrjujejo pomen vpetosti slovenskih jezikovnih virov v širši jezikovnotehnološki prostor, pa je bil nedavno razvit tudi namenski spletni servis[6] za večnivojsko označevanje neoznačenih besedil z jezikovnimi modeli UD, ki tudi za slovenščino dosegajo zelo dobro natančnost (Straka et al., 2016). V prihodnosti nameravamo obstoječo različico slovenske drevesnice nadgraditi v skladu s posodobljenimi smernicami za označevanje, dopolniti njeno spletno dokumentacijo in razširiti nabor pravil za vključitev manjkajočih delov korpusa ssj500k. Poleg podrobnejših jezikoslovnih analiz skladenjskih specifik slovenskega jezika pa bi bilo z jezikovnotehnološkega vidika prioritetno raziskati vpliv spremembe označevalne sheme na natančnost referenčnih označevalnih orodij za slovenščino (Grčar et al., 2012; Dobrovoljc et al., 2012) in na podlagi rezultatov ovrednotiti potrebo po nadaljnjem vzdrževanju samostojnega sistema JOS.

4 Literatura

Kaja Dobrovoljc, Simon Krek in Jan Rupnik. 2012. Skladenjski razčlenjevalnik za slovenščino. V: *Zbornik Osme konference Jezikovne tehnologije*, str. 42–47.

Kaja Dobrovoljc in Joakim Nivre. 2016. The Universal Dependencies Treebank of Spoken Slovenian. V: *Proceedings of LREC'16*, str. 1566–1573.

Sašo Džeroski et al. 2006. Towards a Slovene Dependency Treebank. V: *Proceedings of LREC'06*, str. 1388–1391.

Tomaž Erjavec in Nina Ledinek. 2006. Slovenska odvisnostna drevesnica: prvi rezultati. V: *Jezikovne tehnologije: zbornik 9. mednarodne multikonference Informacijska družba IS 2006*, str. 162–167.

Miha Grčar, Simon Krek in Kaja Dobrovoljc. 2012. Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. V: *Zbornik Osme konference Jezikovne tehnologije*, str. 89–94.

Jan Hajič et al. 2001. The Prague Dependency Treebank: Annotation Structure and Support. V: *Proceedings of the IRCS Workshop on Linguistic Databases*, str. 105–114.

Simon Krek, Kaja Dobrovoljc, Tomaž Erjavec, Sara Može, Nina Ledinek in Nanika Holz. 2015. Training corpus ssj500k 1.4, *Slovenian language resource repository CLARIN.SI*, http://hdl.handle.net/11356/1052.

Sandra Kübler, Ryan McDonald in Joakim Nivre. 2009. *Dependency Parsing*. Morgan and Claypool.

Nina Ledinek in Tomaž Erjavec. 2009. Odvisnostno površinskoskladenjsko označevanje slovenščine: specifikacije in označeni korpusi. V: *Infrastruktura slovenščine in slovenistike (Obdobja 28)*, str. 219–224.

Marie-Catherine de Marneffe et al., 2014. Universal Stanford Dependencies: A cross-linguistic typology. V: *Proceedings of LREC'14*, str. 4585–4592.

Ryan McDonald et al., 2013. Universal Dependency Annotation for Multilingual Parsing. V: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, str. 92–97.

Joakim Nivre. 2015. Towards a Universal Grammar for Natural Language Processing. *Computational Linguistics and Intelligent Text Processing*, (9041):3–16.

Joakim Nivre et al. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. V: *Proceedings of LREC'16*, str. 1659–1666.

Joakim Nivre et al. 2016. Universal Dependencies 1.3, *LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague*, http://hdl.handle.net/11234/1-1699.

Slav Petrov, Dipanjan Das in Ryan McDonald. 2012. A universal part-of-speech tagset. V: *Proceedings of LREC'12*, str. 2089–2096.

Milan Straka, Jan Hajic in Jana Straková. 2016. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. V: *Proceedings of LREC'16*, str. 4290–4297.

Lucien Tesnière. 1959. *Eléments de Syntaxe Structural*, Paris: Librairie C. Klincksieck.

Daniel Zeman. 2008. Reusable Tagset Conversion Using Tagset Drivers. V: *Proceedings of Sixth International Conference on Language Resources and Evaluation (LREC'08)*, str. 213–218.

[5] SETS: http://bionlp-www.utu.fi/dep_search; PML Tree Query: http://lindat.mff.cuni.cz/services/pmltq/.
[6] UDPipe: http://lindat.mff.cuni.cz/services/udpipe/.