Nina Ledinek
Šoštanj
nina.ledinek@siol.net

UDK 811.163.6'367.4:81'322.2

SLOVENSKA ODVISNOSTNA DREVESNICA V RAZISKAVAH O INDUKTIVNEM ODVISNOSTNEM RAZČLENJEVANJU

Prispevek prinaša nekaj temeljnih pojasnil o področju skladenjskega označevanja korpusov ter o metodologiji gradnje drevesnic. Opozori na jezikovnoteoretične modele, na podlagi katerih so označevalni sistemi večine drevesnic oblikovani, ter na uporabnost skladenjskega označevanja korpusov za raziskave opisnega in teoretičnega jezikoslovja in za področje procesiranja naravnih jezikov. Predstavlja tudi rezultate treh raziskav v zvezi z večjezičnim induktivnim odvisnostnim razčlenjevanjem na prvi drevesnici za slovenski jezik, Slovenski odvisnostni drevesnici.

1 Skladenjsko označevanje korpusov in njegova uporabnost

O skladenjskem označevanju korpusov govorimo, kadar jezikovnim elementom v korpusih dodajamo glede na izbrani jezikovnoteoretični model predvidene interpretativne skladenjske analitične oznake (Gorjanc 2005: 64). Skladenjsko označeni korpusi oz. drevesnice tako nudijo kvantitativni pregled distribucije predpostavljenih skladenjskih kategorij na velikem vzorcu realnih besedil, s čimer olajšujejo raziskave opisnega in teoretičnega jezikoslovja, v določenih primerih pa jih uporabljamo tudi za empirično testiranje in potrjevanje specifičnih jezikoslovnih paradigem, npr. modela HPSG (*head-driven phrase structure grammar*), ki ga uporabljajo pri označevanju bolgarskega korpusa BulTreeBank[1] (Simov et al. 2002: 1729–1736) in testnega vzorca povedi iz poljskih pisnih besedil (Marciniak et al. 2003: 129–146); delno tudi funkcijskega generativnega opisa, ki je teoretična podstava za gradnjo korpusa Prague Dependency Treebank (LDC 2006). Hkrati skladenjsko označeni korpusi omogočajo razvoj različnih jezikovnih tehnologij (npr. avtomatskih prevajalnikov, sintetizatorjev in prepoznavalnikov govora, črkovalnikov, iskalnikov različnih podatkov ipd.), saj lahko z njimi testiramo skladenjske razčlenjevalnike, rezultati njihove analize so namreč odlična osnova za nadaljnjo strojno obdelavo jeziko(slo)vnih podatkov.

[1] BulTreeBank: <http://www.bultreebank.org/>.

Ker je potencialnih rab drevesnic veliko, področja njihove uporabe pa so zelo raznorodna, si bomo v nadaljevanju na nekaj primerih ogledali, v katerih raziskavah so bile drevesnice že uporabljene oz. glede na kakšne raziskovalne cilje so bile oblikovane. Pravzaprav so vse drevesnice vključene v testiranje različnih skladenjskih razčlenjevalnikov, lematizatorjev, oblikoskladenjskih označevalnikov ipd., nemška drevesnica NEGRA/TIGER pa je bila poleg tega uporabljena npr. še v raziskavi o premiku oziralniških stavkov na konec povedi in o besednem redu v nemščini, na njeni podlagi pa so bili pridobljeni tudi podatki za raziskavo o kolokacijah (Brants et al. 2003: 83). Francoska drevesnica je bila vključena v raziskavo o tem, katerim skladenjskim strukturam dajejo govorci prednost (pri raziskavi so poleg korpusnih jezikoslovcev sodelovali tudi psiholingvisti) (Abeillé et al. 2003: 165), poljski testni korpus pa je bil zgrajen z namenom testirati že obstoječe formalne računalniške ter odvisnostne slovnice poljščine glede na empirične podatke, hkrati pa bo uporabljen tudi v raziskavah o avtomatski indukciji slovnic (Marciniak et al. 2003: 130). Podatki Italijanske skladenjsko-pomenske drevesnice (ISST) so imeli ključno vlogo pri izboljšavi prevajalnega sistema za jezikovni par italijanščina – angleščina, ki je sestavljen iz več programskih orodij (Montemagni et al. 2003: 204–205), nemška drevesnica TüBa-D/Z[2] je bila npr. vključena v raziskavo o avtomatski interpretaciji anafor (kot jih opredeljuje tvorbeno jezikoslovje v enem od modulov vezalno-navezovalne teorije) v nemščini (Hinrichs et al. 2005). Drevesnica Twente Niuews Corpus je bila uporabljena za raziskavo, ali je premik predložne besedne zveze na začetek stavka v nizozemščini premik iz samostalniške ali glagolske besedne zveze, na podlagi te analize pa naj bi bila omogočena tudi prilagoditev oz. poenotenje označevalnih sistemov dveh drugih nizozemskih drevesnic (Bouma 2004: 15–26). Za Slovensko odvisnostno drevesnico načrtujemo, da bo v prihodnosti omogočala tako oblikovanje novih jezikoslovnih opisov kot raziskave na področju obdelave naravnih jezikov. Zaenkrat je zlasti besedilna sestava drevesnice takšna, da tovrstnih analiz še ne omogoča, izhodiščne raziskave, ki jih v začetni fazi projekta oblikovanja drevesnice izvajamo, in njihovi rezultati pa so namenjeni analizi metodologije avtomatskega skladenjskega označevanja korpusov slovenščine in so ključni za razmislek o tem, kako kompleksen naj bo nabor skladenjskih oznak in posledično pravil, ki jih je pri označevanju smiselno upoštevati. Zaenkrat je bila drevesnica vključena v tri raziskave o večjezičnem induktivnem odvisnostnem razčlenjevanju (Chanev 2005; Gjorgjioski 2006; CoNLL-X 2006).

1.1 Ravni označevanja in kompleksnost označevalnih modelov

Korpusi se skladenjsko navadno označujejo na dveh ravneh, na površinskoskladenjski in na pomenskoskladenjski oz. pomenski ravni. Na površinskoskladenjski ravni so večinoma predstavljena strukturna in/ali funkcijska razmerja med pojavnicami v povedi, pri pomenskoskladenjski analizi pa se pogosto označujejo pomenska razmerja med glagolom oz. povedkom in njegovimi določili ter dopolnili, tematsko-rematska struktura povedi oz. členitev po aktualnosti, koreferenčna razmerja ipd. Osnovna analitična enota je, vsaj pri korpusih, ki jih sestavljajo pisna besedila, (stavčna) poved.

[2] TüBa-D/Z: <http://www.sfs.uni-tuebingen.de/en_tuebadz.shtml>.

Modeli označevanja in nabori skladenjskih oznak za različne korpuse se zelo razlikujejo, vendar pa lahko glede na kompleksnost procesa izpostavimo dva osnovna tipa označevanja:[3] popolno (*full parsing*) in skeletno (*skeleton/skeletal parsing*) oz. plitko (*shallow parsing*) skladenjsko označevanje oz. razčlenjevanje.[4] Popolno skladenjsko označevanje omogoča zelo natančno in podrobno analizo skladenjskih struktur, saj so pri takem tipu označevanja predstavljeni strukturni ali odvisnostni (ter tudi pomenski) odnosi med vsemi pojavnicami v povedi. S skeletnim oz. plitkim označevanjem pa shematično predstavimo samo najpomembnejša skladenjska razmerja v povedi, podatkov o natančni skladenjski vlogi vseh pojavnic v povedi pa takšno označevanje ne nudi, saj več pojavnic združujemo v večje enote, analitične oznake (njihov nabor je manjši kot pri popolnem skladenjskem označevanju) pa pripisujemo le njim.

1.2 Načini skladenjskega označevanja korpusov

Korpus lahko skladenjsko označimo ročno, polavtomatsko (z interaktivnimi tehnikami) in avtomatsko. Pri polavtomatskem označevanju označevalcem pomagajo inteligentni urejevalniki, pri avtomatskem označevanju pa je rezultate pogosto treba popraviti ročno, saj skladenjski razčlenjevalniki še ne dosegajo zadovoljive stopnje natančnosti (Abeillé 2003: XVIII). Glede na način označevanja korpusov včasih ločujemo med drevesnicami (*treebanks*), o katerih govorimo, kadar korpuse skladenjsko označujemo ročno ali kadar pravilnost oznak v avtomatsko označenih korpusih ročno preverimo, in skladenjsko označenimi oz. razčlenjenimi korpusi (*parsed corpora*), pri katerih ročni pregled analitičnih oznak ni nujen (Schulte im Walde in Zinsmeister 2006). Mi bomo termin drevesnica uporabljali v širšem smislu, tj. za korpus, ki je skladenjsko označen na kateri koli od zgoraj navedenih načinov.

Poznamo dva osnovna tipa avtomatskega (površinsko)skladenjskega označevanja korpusov: označevanje na podlagi vnaprej pripravljene slovnice (*rule-based parsing*) in statistično označevanje (*stochastic/probabilistic/data-driven parsing*). Pri skladenjskem označevanju po pravilih vnaprej pripravljene slovnice (tem sistemom je na voljo tudi leksikon besednih oblik) navadno izberemo eno od uveljavljenih skladenjskih teorij in na njeni podlagi pripravimo sistem zelo formaliziranih jezikovnospecifičnih pravil o funkciji ali zgradbi skladenjskih struktur v določenem skladenjskem kontekstu. Nato jih vnesemo v računalnik, razčlenjevalnik pa pri analizi jezika v korpusu sistem pravil pregleduje in ugotavlja, s katerim(i) bi določeno strukturo lahko opisal. Če ne najde nobenega ustreznega pravila, strukturo opredeli kot neslovnično – v tem primeru nastopijo težave z robustnostjo razčlenjevalnika.

---

[3] Razločevanje je pomembno zlasti pri korpusih, pri katerih na površinskoskladenjski ravni označujemo sestavniško strukturo povedi.

[4] Skladenjsko označevanje in razčlenjevanje korpusov sta sicer dva različna postopka, pri gradnji skladenjsko označenih korpusov pa vključujemo oba. Postopka v različnih fazah dela drug drugega predpostavljata in dopolnjujeta. Termina označevanje in razčlenjevanje zato v tem prispevku večinoma uporabljamo kot nekakšna sinonima, ki pravzaprav pomenita sintezo obeh procesov. Odločitev za enega od njiju je odvisna od tega, kateri postopek je pri delu na nekem področju gradnje korpusov odločilnejši. Kljub temu razumemo termin označevanje kot bolj splošen, zato ga bomo uporabljali tudi takrat, kadar bi se lahko nanašali na katerega koli od obeh postopkov.

Slovnice, ki jih razčlenjevalniki pri takšnem označevanju uporabljajo, so zelo podobne opisnim slovnicam, saj je vsem vnaprej pripravljenim sistemom slovničnih pravil skupno, da skušajo ljudje pri njihovi pripravi upoštevati podatke o človeškem znanju jezika in na tej podlagi izdelati opise, ki jih bo lahko uporabil tudi računalnik. Vendar se tovrstni razčlenjevalniki niso izkazali za zelo učinkovite. Težava je predvsem v premajhnem pokritju pravil, zelo težko je namreč pripraviti primerno veliko število pravil, ki opišejo, kako naj razčlenjevalniki opravijo analizo natančno in konsistentno (McEnery in Wilson 1996: 131–132; Kennedy 1998: 232–233). Hkrati ti razčlenjevalniki niso uspešni pri analizi neznanih struktur in besed, težave pa jim povzročajo tudi napake v korpusu, saj delujejo po principu strogega ločevanja med slovničnostjo in neslovničnostjo. Njihova še večja pomanjkljivost je, da besedila niso sposobni razdvoumiti pomensko, glede na kontekst, zato lahko dobi ena poved, če jo je mogoče opisati z več pravili v slovnici, večje število možnih analiz. Relativno robustni skladenjski razčlenjevalniki, ki delujejo po principu pregledovanja vnaprej pripravljenih pravil, so bili zaradi finančnih in časovnih omejitev, priprava pravil in leksikona je namreč zelo kompleksen projekt, razviti samo za nekaj ekonomsko močnih jezikov oz. jezikov z velikim številom govorcev (Erjavec in Ledinek 2006: 165–166).

Obetavnejše rezultate v zadnjem času dosegajo razčlenjevalniki, ki delujejo po principu statističnega učenja. Pri statističnem skladenjskem označevanju korpusov razčlenjevalnikom ne posredujemo nikakršnega metajezikovnega znanja. Uporabljajo samo abstraktne statistične modele, s katerimi razbirajo in opisujejo slovnico jezika, kot se kaže v ročno označenih učnih korpusih. Na podlagi statistične verjetnosti sopojavljanja pojavnic določenega oblikoskladenjskega tipa – oblikoskladenjske analitične oznake dajejo sorazmerno dobro informacijo o potencialni skladenjski strukturi povedi – v takem korpusu ocenjujejo, katere strukture so glede na tipično besedilno umeščenost skladenjskih enot bolj verjetne in katere manj, in na osnovi tovrstnih izračunov skladenjskim strukturam določajo zgradbo ali funkcijo. Avtomatski označevalniki pri takem označevanju inducirajo slovnico, ki ni podobna opisnim slovnicam jezika. Pristopi k statističnemu skladenjskemu označevanju so raznovrstni, vendar pa razčlenjevalniki za uspešno analizo neoznačenega besedila vedno potrebujejo natančno označen učni korpus, katerega izgradnja pa je izredno zahtevna, dolgotrajna in draga (McEnery in Wilson 1996: 132–133; Kennedy 1998: 234). Učinkoviti statistični razčlenjevalniki[5] že dosegajo razmeroma dobre rezultate,[6]

[5] Eden takšnih je npr. Collinsov razčlenjevalnik za češčino, ki dosega pri razčlenjevanju korpusa Prague Dependency Treebank več kot 80-odstotno natančnost (Collins et al. 1999: 505).

[6] Podatek o stopnji natančnosti razčlenjevalnika sam po sebi sicer ni zadosten, saj je dosežena natančnost odvisna zlasti od razmerja med naborom skladenjskih in oblikoskladenjskih analitičnih oznak in velikostjo učnega korpusa. Dobri rezultati so načeloma zagotovljeni, kadar se razčlenjevalnik uri (in je kasneje testiran) na korpusu, ki obsega vsaj milijon besed, pri čemer na vsaki od označevalnih ravni razločujemo manj kot 50 analitičnih oznak (Abeillé 2003: XXII). Kaj od razčlenjevalnika pri analizi neoznačenega teksta pričakujemo, je navadno odvisno od namena, za katerega drevesnico gradimo. Za razčlenjevalnik, katerega analiza bo pomembna za razvoj jezikovnih tehnologij, je pogosto dovolj, če uspe jezikovnim elementom natančno in konsistentno pripisovati majhen nabor skladenjskih oznak. Razčlenjevalnik, ki bo pripomogel k jezikoslovnim analizam, pa bo primeren, če bo strukture natančno označeval z velikim številom zelo podrobnih skladenjskih oznak.

tj. 60- do 80-odstotno natančnost pri analizi drevesnic, označenih po kompleksnih označevalnih modelih, vendar pa niso odporni na »neumne« napake.

1.3 Teoretični modeli skladenjskega označevanja korpusov

Kot podstavo za skladenjsko označevanje korpusov navadno uporabljamo dva jezikovnoteoretična modela: odvisnostno slovnico (glavno razmerje v tem tipu je asimetrično binarno razmerje podrednost – nadrednost, navadno se povezuje s funkcijsko analizo) in frazno gramatiko (v sestavniški strukturi so elementi razvrščeni glede na razmerje del – celota). O tretjem, hibridnem modelu označevanja govorimo takrat, kadar analitične skladenjske oznake pojavnic združujejo tako informacije o odvisnostnih kot sestavniških razmerjih, pri čemer je eno od razmerij pri predstavitvi dominantno (Schulte im Walde in Zinsmeister 2006; Abeillé 2003: XVII). Glede na *Priporočila za skladenjsko označevanje korpusov* iniciative EAGLES[7] je uporaba označevalnih modelov, ki so razviti na podlagi uveljavljenih jezikoslovnih teorij, koristna zato, ker imajo te teorije že dolgo tradicijo v jezikoslovju nasploh, ker so bili ti modeli pri ročnem in avtomatskem označevanju korpusov tipološko različnih jezikov empirično že testirani in ker za analizo korpusov, označenih na njihovi podlagi, že obstajajo robustni razčlenjevalniki. Obstajajo tudi bolj specifični teoretični modeli označevanja[8] (npr. HPSG), vendar se iz praktičnih razlogov uporabljajo redkeje (Verdonik 2004: 208–211; Schulte im Walde in Zinsmeister 2006; Priporočila za skladenjsko označevanje korpusov iniciative EAGLES). Drevesnica, označena na podlagi frazne gramatike, je npr. Penn Treebank[9] (Taylor et al. 2003: 5–22), primerka odvisnostnih drevesnic sta Prague Dependency Treebank (PDT)[10] (LDC 2006) ter Slovenska odvisnostna drevesnica[11] (Erjavec in Ledinek 2006: 162–167; Džeroski et al. 2006: 1388–1391), po hibridnih principih so npr. označene drevesnice NEGRA/TIGER,[12] SUSANNE[13] idr.

---

[7] Glej <http://www.ilc.cnr.it/EAGLES96/segsasg1/segsasg1.html>.

[8] Že Leech se je v svojih priporočilih za označevanje korpusov, ki naj bi pripomogla k temu, da pri interpretaciji analitičnih oznak ne bi prihajalo do napačnega razumevanja, opredelil proti specifičnim označevalnim modelom. Predlagal je, naj se nivo označevanja v največji možni meri približa tistemu razumevanju jezikovnih pojavov, ki je v določenem okolju najbolj uveljavljeno in ki je glede navezanosti na teorije najbolj »nevtralno«, hkrati pa naj osnova za označevalni sistem ne bo prekompleksna. Tako npr. predlaga, da naj kot podstavo za označevalni model izberemo najosnovnejše predpostavke frazne gramatike in ne morda kakšne od njenih bolj specifičnih paradigem, npr. teorije načel in parametrov (povzeto po McEnery in Wilson 1996: 25–26). Temeljne predpostavke pri označevanju odvisnostnih drevesnic (tudi Slovenske odvisnostne drevesnice) pa so navadno, da so skladenjske strukture sestavljene iz elementov, ki jih povezujejo asimetrična binarna razmerja. Osnovno razmerje tega modela je opozicija nadrejeni – podrejeni element, med katerima nastopi odvisnost, kriteriji (skladenjski in pomenski) za določanje nadrejenega elementa pa so pri različnih modelih lahko različni. Eden od teoretičnih modulov odvisnostne slovnice je tudi teorija vezljivosti. Odvisnostni modeli, po katerih drevesnice označujemo, navadno upoštevajo tudi tri bistvena načela: besedna oblika je podrejena natanko eni drugi besedni obliki, vse besedne oblike v odvisnostna razmerja morajo vstopati, vendar ne smejo biti povezane ciklično (tj. odvisnosti ne smejo biti predstavljene tako, da tvorijo krog) (Nivre 2005: 1–11; Samuelsson 2000: 684).

[9] Penn Treebank: <http://www.cis.upenn.edu/~treebank/>.

[10] Prague Dependency Treebank: <http://ufal.mff.cuni.cz/pdt/>.

[11] Slovenska odvisnostna drevesnica: <http://nl.ijs.si/sdt/>.

[12] Projekt TIGER: <http://www.ims.uni-stuttgart.de/projekte/TIGER/>.

[13] Korpus SUSANNE: <http://www.grsampson.net/RSue.html>.

Kakšen teoretičen model pri skladenjskem označevanju korpusa izberemo, je vsaj do določene mere odvisno tudi od jezika, katerega korpus označujemo. Označevanje po načelih frazne gramatike je bolj primerno za jezike s stalnim besednim redom in jasno sestavniško strukturo, odvisnostne drevesnice pa navadno gradimo za morfološko bogate jezike s prostim besednim redom. Navadno velja, da so drevesnice, zgrajene na podlagi frazne gramatike, bolj berljive od odvisnostnih, saj besedni red pri predstavitvi odvisnostnih razmerij ni vedno ohranjen (Verdonik 2004: 208–211; Priporočila za skladenjsko označevanje korpusov iniciative EAGLES; Abeillé 2003: XVII).

2 Slovenska odvisnostna drevesnica

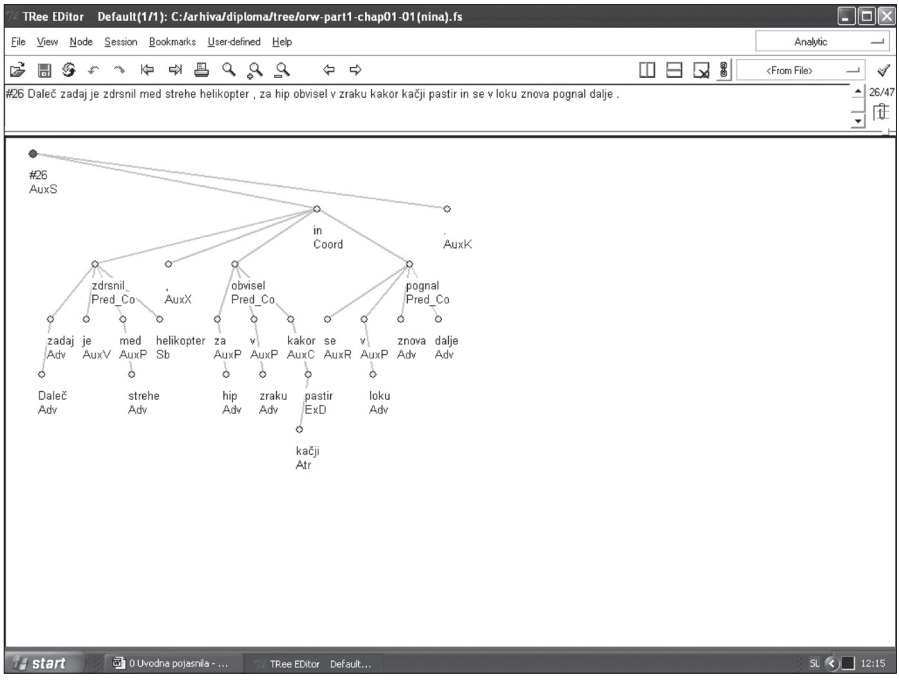
Ker je slovenščina morfološko bogat jezik s prostim besednim redom in ker ima v slovenskem jezikoslovju odvisnostna slovnica daljšo tradicijo kot frazna gramatika, smo se odločili, da za slovenščino pripravimo drevesnico odvisnostnega tipa. Z gradnjo Slovenske odvisnostne drevesnice (projekt *Slovene Dependency Treebank*, SDT), drevesnice pisnih tekstov, smo začeli leta 2003. Zaradi finančnih omejitev, projekt namreč ni imel namenskih sredstev financiranja, izdelava lastnega označevalnega modela ni bila mogoča, poleg tega pa bi bila priprava vse za skladenjsko označevanje potrebne infrastrukture otežena tudi zaradi kadrovskih in časovnih omejitev. Glede na tipološke sorodnosti med jeziki in modele njihove obravnave smo zato med javno dostopnimi označevalnimi modeli izbrali najustreznejšega, predpostavljene jezikoslovne rešitve pa prilagodili za slovenščino.

Odvisnostne drevesnice skupaj z javno dostopno dokumentacijo o jezikovno-teoretičnih modelih ter programsko opremo za njihovo izgradnjo ter analizo danes obstajajo za številne tipološko zelo različne jezike. Kot model korpusa, po katerem gradimo Slovensko odvisnostno drevesnico, pa smo izbrali PDT, saj je češčina po skladenjskih lastnostih slovenščini zelo podobna, hkrati pa je vsaj del slovenistične jezikoslovne teorije tesno povezan s češkim jezikoslovjem. PDT je posebej relevanten tudi zato, ker gre za eno največjih in najbolje dokumentiranih drevesnic sploh, ustreznost označevalnega modela, ki temelji na teoriji funkcijskega generativnega opisa, pa je bila tudi empirično že preverjena. Drevesnica je označena na dveh ravneh, na površinsko- in pomenskoskladenjski ravni. Na površinskoskladenjski ravni je funkcijskoskladenjska vloga vsake pojavnice v povedi, tj. njeno odvisnostno razmerje do neposredno nadrejene pojavnice, razločevalno prikazana v acikličnem grafu – skladenjskem drevesu. Na pomenskoskladenjski ravni pa so prikazana pomenska skladenjska razmerja, koreferenčna razmerja, členitev po aktualnosti oz. rematsko-tematska struktura povedi, fonetično neizražene pojavnice povedi ipd., v skladenjsko drevo pa so na tej ravni umeščene le polnopomenske besede.

Slovenska odvisnostna drevesnica je trenutno označena le na površinskoskladenjski ravni, v nadaljnjih fazah projekta pa načrtujemo tudi označevanje pomenskoskladenjske ravni. Sistem ročnega površinskoskladenjskega označevanja smo prevzeli neposredno po priročniku za označevanje korpusa PDT (Bémová et al. 1999: 1–312),

vendar smo ga glede na razlike v jezikovnih sistemih slovenščine in češčine in v njihovi interpretaciji takoj začeli prilagajati. Označevanje je potekalo v prvi fazi avtomatsko, nato pa smo rezultate avtomatske analize ročno pregledali in popravili. Pri delu smo si pomagali z urejevalnikom dreves TrEd (Hajič et al. 2001: 105–114) in razčlenjevalnikom za slovenščino (Džeroski et al. 2006: 1388–1391). Z vidika projekta gradnje Slovenske odvisnostne drevesnice ima izbrani označevalni model zlasti to pomanjkljivost, da je zelo kompleksen, zato smo se pri njegovem prilagajanju za slovenščino odločili za nekatere poenostavitve (zaenkrat smo se večinoma ukvarjali s pripravo označevalnega sistema za strukture, ki jim v slovenskem jezikoslovju navadno pripisujemo funkcijskoskladenjsko vlogo povedka) (Ledinek 2005).

Pri površinskoskladenjskem označevanju Slovenske odvisnostne drevesnice vsakemu izrazno ločenemu ali drugačnemu elementu povedi (besedi, delu besede, ločilu, simbolu, številki ipd.), ki je v skladenjskem drevesu označen z vozlom, pripišemo posebno analitično oznako glede na njegovo vlogo na površinskoskladenjski ravni. Poseben vozel z oznako AuxS dobi še baza drevesa. Analitične oznake se pogosto ujemajo s funkcijskoskladenjskimi oznakami, kot bi jih besede, ki funkcijskoskladenjsko vlogo imajo, dobile v večini opisov novejšega slovenskega jezikoslovja (Sb = osebek, Obj = predmet, Pred = povedek ipd.). Kot lahko vidimo na **Sliki 1**, pa se označevalni sistem od funkcijskoskladenjskega opisa razlikuje zlasti pri opredeljevanju vloge funkcijskih besed in načinu predstavljanja neodvisnostnih razmerij (priredja, vrinjeni stavki, pristavki ipd.) (prim. Ledinek 2005: 31–67).



Slika 1: *Primer skladenjskega drevesa iz Slovenske odvisnostne drevesnice, kot ga vidimo v programu* TrEd.

Ker so označevalni model, potek priprave za gradnjo drevesnice potrebne infrastrukture ter številni drugi relevantni podatki natančneje predstavljeni v dostopni literaturi (Erjavec in Ledinek 2006: 162–167; Džeroski et al. 2006: 1388–1391; Ledinek 2005: 21–67, 132–167) in na domači spletni strani projekta <http://nl.ijs.si/sdt/>, se bomo v nadaljevanju opredelili zlasti glede besedilne sestave drevesnice, saj ima ta velik vpliv na rezultate raziskav, o katerih bomo poročali v nadaljevanju.

Slovenska odvisnostna drevesnica trenutno obsega približno 2.800 povedi oz. 45.000 besed. Sestavljena je iz vzorcev povedi, vzetih iz dveh korpusov, korpusa MULTEXT-East »1984« in korpusa SVEZ-IJS. Vzorec iz korpusa »1984« obsega prvo tretjino romana *1984* Georgea Orwella, ki jo sestavlja 2.000 povedi in približno 30.000 besed. Različica 0,4 tega korpusa je prosto dostopna za raziskovalne namene. Drugi korpus drevesnice sestavljajo povedi iz vzporednega angleško-slovenskega korpusa SVEZ-IJS, ki vsebuje pravni red EU (Erjavec 2006: 2138–2141; Erjavec in Sárossy 2006: 169). V vzorec za površinskoskladenjsko označevanje smo vključili tri zaporedne segmente na 1000 segmentov oz. približno 3 ‰ slovenskega dela korpusa, kar je približno 800 povedi in okrog 15.000 besed.

Čeprav je izbira korpusov z vidika reprezentativnosti manj ustrezna, smo ju izbrali, ker so bile njune oblikoskladenjske oznake ročno pregledane, kar prispeva k večji stopnji natančnosti pri avtomatskem površinskoskladenjskem označevanju, hkrati pa je bilo na ta način mogoče izpustiti fazo oblikoskladenjskega označevanja, kar je močno zmanjšalo stroške začetne faze projekta. Korpusa imata sicer nekaj očitnih pomanjkljivosti. »1984« sestavlja eno samo umetnostno besedilo, ki vključuje tudi izmišljen jezik, za upravno-pravna besedila v korpusu SVEZ-IJS pa je tipična skrajna formaliziranost določenih odsekov, »tabelarnost« ter naštevalnost, kar povzroči, da so posamezni deli besedil interpretirani kot niz elips, za besedila pa je značilen tudi visok delež zelo obsežnih in kompleksnih samostalniških zvez. Poleg tega analitična enota, ki jo v korpusu SVEZ-IJS označujemo, pogosto ni poved, ampak manjši segment, katerega funkcijskoskladenjska vloga brez dodatnega jezikovnega konteksta ni vedno enoznačno določljiva. Oba korpusa drevesnice torej sestavljajo v skladenjskem smislu relativno specifična prevodna besedila, zato je pri njuni analizi ključno vprašanje, ali in v kolikšni meri je tipologija vzorcev pojavljanja skladenjskih struktur v njih za slovenščino tipična. Da bi dobili okviren odgovor na to vprašanje, se bomo ob razširitvi drevesnice posvetili označevanju vzorca slovenskih časopisnih besedil, ki jih bomo (verjetno) pridobili iz referenčnih korpusov *Fida* oz. *FidaPLUS*, vključiti pa bomo skušali čim bolj reprezentativne besedilne tipe z različnih tematskih področij. Šele bolj uravnotežen in obsežnejši korpus bo namreč omogočal relevantnejše jezikoslovne raziskave, ki bodo dejansko lahko opozorile na še neodkrite jezikovne pojavnosti in na regularnosti v vzorcih pojavljanja skladenjskih struktur in njihovi skladenjski vlogi v realnih besedilih, hkrati pa bodo ročno označevanje korpusa in njegove analize omogočali natančnejši razmislek o tem, kako je smiselno označevalni sistem prilagajati, da ga bo mogoče uspešno uporabiti pri označevanju in analizi katerega koli (pisnega) besedila v slovenščini.

2.1 Slovenska odvisnostna drevesnica v raziskavah o induktivnem odvisnostnem razčlenjevanju

Za jezike z manjšo ekonomsko močjo in z manjšim številom govorcev, ki so posledično z vidika korpusnega jezikoslovja oz. celotnega področja procesiranja naravnih jezikov manj razis(kov)ani, se v zadnjem času pojavljajo nove priložnosti za analize. Spoznanja na področju prenosljivosti programskih orodij za avtomatsko obdelavo jeziko(slo)vnih podatkov, zlasti statističnih razčlenjevalnikov, namreč omogočajo, da z relativno majhnimi sredstvi tudi za te jezike zgradimo kvalitetne jezikovne vire in jezikovne tehnologije, ki jih sicer morda ne bi bilo mogoče pripraviti.

V tri raziskave o večjezičnem induktivnem odvisnostnem razčlenjevanju je bila vključena tudi Slovenska odvisnostna drevesnica. Takšne raziskave so pomembne, ker omogočajo razvoj učinkovitih orodij za analizo slovenske skladnje, hkrati pa dobimo z njimi tudi povratno informacijo predvsem o konsistentnosti označevanja korpusa, primernosti izbranega teoretičnega modela označevanja ter nenazadnje podatke, relevantne za jezikoslovje.

V raziskavi o natančnosti pri skladenjski analizi zlasti italijanščine (Chanev 2005) z razčlenjevalnikom Malt (Nivre et al. 2005) smo za slovenski korpus kljub urjenju in testiranju razčlenjevalnika na zelo majhnem vzorcu besedila iz prototipne inačice 0,1 korpusa »1984« (testni korpus je obsegal 335 povedi oz. približno 7000 pojavnic) in dejstvu, da učni in testni vzorec še nista bila označena povsem konsistentno, dobili relativno vzpodbudne rezultate. Dosežena stopnja natančnosti za označeno povezanost (razčlenjevalnik je moral pravilno določiti tako hierarhična razmerja med besednimi oblikami kot tip odvisnostnih razmerij, tj. analitične oznake) je bila približno 58 %, stopnja natančnosti pri določanju neoznačene povezanosti (pravilno so morala biti določena le hierarhična razmerja, pravilnost analitičnih oznak pa ni bila relevantna) je bila približno 69 %.

Inačica 0,3 korpusa »1984«, ki je obsegala približno 1.500 povedi, je bila nato vključena v mnogo obširnejšno raziskavo, v tekmovanje o natančnosti statističnih odvisnostnih razčlenjevalnikov, ki je potekalo v okviru *10. Konference o računalniškem učenju naravnih jezikov* (*10th Conference on Computational Natural Language Learning*, *CoNLL-X*). Dvajset različnih razčlenjevalnikov je bilo testiranih na 13 drevesnicah tipološko različnih jezikov (danščina, portugalščina, češčina, japonščina, slovenščina, turščina itd.). Učni korpusi, na katerih so tekmovalci svoje razčlenjevalnike urili, so bili oblikoskladenjsko označeni in lematizirani, vendar so bili različno veliki, češki korpus PDT je bil največji in je obsegal 1.250.000 besed, Slovenska odvisnostna drevesnica pa je bila daleč najmanjši korpus, njegovo učno množico je sestavljalo le 29.000 besed. S posebnim programom so merili stopnjo natančnosti razčlenjevalnikov pri določanju označene povezanosti, neoznačene povezanosti in oznak. Najboljši rezultati so bili doseženi z dvostopenjskim razlikovalnim razčlenjevalnikom[14] (*two-stage discriminative parser*) (McDonald

[14] Opis sistema in natančnejši rezultati so dostopni na spletni strani <http://www.cnts.ua.ac.be/conll/pdf/21520.pdf>.

et al. 2006). Za slovensko drevesnico je zaradi sorodnosti označevalnih sistemov najzanimivejša primerjava rezultatov s češkim korpusom PDT. Rezultati raziskave so prikazani v **Tabeli 1** (OP = označena povezanost, NP = neoznačena povezanost, O = oznake).

	SDT	**PDT**
OP McDonald et al.	73.44 %	80.18 %
OP povprečno	65.16 %	67.17 %
NP McDonald et al.	83.17 %	87.30 %
NP povprečno	76.53 %	77.01 %
OZ McDonald et al.	82.51 %	86.72 %
OZ povprečno	76.31 %	76.59 %

Tabela 1: *Rezultati raziskave* CoNLL-X *za* SDT *in* PDT.

Rezultati pri razčlenjevanju Slovenske odvisnostne drevesnice so le za nekaj odstotkov slabši od rezultatov za drevesnico PDT, kar bi najbrž lahko pripisali zelo veliki razliki v velikosti učnih množic. Vendar pa je treba upoštevati, da so v češko drevesnico zajeti zelo različni besedilni tipi, razčlenjevalniki morajo biti zato za takšno analizo bolj robustni kot za označevanje enega besedila. Kljub temu je raziskava dokazala, da lahko drevesnico sorodnih slovenskih besedil, ki je oblikoskladenjsko označena zelo natančno, s primernimi razčlenjevalniki površinskoskladenjsko avtomatsko označimo s približno 75-odstotno natančnostjo. Hkrati je treba pričakovati, da bo stopnja natančnosti pri strojnem razčlenjevanju povsem avtomatsko označenih in »odprtih« korpusov precej nižja.

Tretja raziskava[15] (Gjorgjioski 2006), raziskava o natančnosti skladenjske analize z razčlenjevalnikom Malt, v katero sta bila vključena oba korpusa Slovenske odvisnostne drevesnice, je opozorila na za jezikoslovje zelo zanimivo vprašanje, kako pomembno vpliva na frekvenco pojavljanja in vzorce pojavljanja posameznih skladenjskih struktur besedilni tip.[16] V prvi fazi raziskave[17] je bilo na obeh korpusih drevesnice opravljeno desetkratno križno preverjanje. Druga faza raziskave je bila zastavljena tako, da je bil razčlenjevalnik izurjen na korpusu »1984«, testiran pa na korpusu SVEZ-IJS in obratno. V tretji fazi raziskave pa je bil razčlenjevalnik izurjen na združenih podatkih iz obeh korpusov, na celotnem korpusu »1984« in devetih desetinah korpusa SVEZ-IJS in obratno, ena desetina posameznega korpusa pa je služila kot testni korpus. Postopek je bil ponovljen 20-krat, tako da je bil kot testni korpus uporabljen vsak od na dvajset delov razdeljene drevesnice. Nato je bila izračunana povprečna natančnost, ki jo je razčlenjevalnik dosegel pri analizi vsakega od obeh korpusov. Rezultati vseh analiz so prikazani v spodnji tabeli. Raziskovali

[15] Še neobjavljeno raziskavo (objavljena naj bi bila kot delovno poročilo na Institutu Jožef Stefan) je kot del izpitnih obveznosti na podiplomskem študijskem programu opravil Valentin Gjorgjioski.

[16] Pojem besedilni tip razumemo kot krovni pojem za varianto oz. register jezika, ki ga opredeljujejo podobne besedilne vrste in ki je definiran s sorodnimi tematskimi področji ter, do določene mere, s primerljivimi situacijskimi karakteristikami.

[17] Učni in razčlenjevalni algoritem razčlenjevalnika ter parametri so bili v vseh fazah enaki, in sicer takšni, kot so bili uporabljeni tudi na tekmovanju v sklopu konference CoNLL-X.

smo stopnjo natančnosti razčlenjevalnika pri določanju označene povezanosti (OP) in neoznačene povezanosti (NP).

	»1984«		**SVEZ-IJS**	
	OP	NP	OP	NP
»1984«, SVEZ-IJS (1. faza)	62,03 %	73,25 %	64,01 %	74,70 %
»1984« : SVEZ-IJS (2. faza)	15,22 %	23,24 %	46,57 %	64,00 %
»1984« + SVEZ-IJS (3. faza)	61,62 %	73,12 %	63,88 %	75,30 %

Tabela 2: *Rezultati analiz z razčlenjevalnikom* Malt.

Glede na to, da so v korpusa drevesnice vključena zelo različna besedila, smo pričakovali, da bo stopnja natančnosti pri navzkrižnem avtomatskem razčlenjevanju testnih korpusov (2. faza) nizka, vendar pa so nas rezultati – ugotovljen je bil izjemen razkorak med rezultati druge in prve ter tretje faze raziskave – nekoliko presenetili. Podatki kažejo, da je obema korpusoma skupen samo minimalen nabor skladenjskih struktur, saj se v korpusu SVEZ-IJS samo približno 15 % struktur, ki jih najdemo tudi v korpusu »1984«, pojavlja dovolj sistematično in pogosto, da so statistično signifikantne pri indukciji razčlenjevalnikove slovnice, s katero je mogoče razčlenjevati tudi drugi korpus drevesnice. Boljše rezultate pri testiranju na korpusu SVEZ-IJS je mogoče pripisati dejstvu, da je nabor vzorcev pojavljanja različnih struktur v korpusu »1984« bistveno večji[18] kot v korpusu SVEZ-IJS, zato je verjetnost, da bo razčlenjevalnik s pomočjo tega učnega korpusa uspel izdelati slovnico, s katero bo mogoče označiti večji nabor struktur v korpusu SVEZ-IJS kot obratno, večja. Rezultati 1. faze raziskave hkrati kažejo, da sta oba korpusa ročno označena s približno enako stopnjo natančnosti, vendar je treba pri interpretaciji rezultatov upoštevati še dejstvo, da je korpus »1984« približno dvakrat obsežnejši od korpusa SVEZ-IJS. Če razčlenjevalnik urimo na učnem korpusu, v katerega so vključeni podatki obeh korpusov, se natančnost pri avtomatski analizi posameznih korpusov sicer bistveno izboljša, vendar primerjava rezultatov prve in tretje faze raziskave hkrati kaže, da je specifičnost besedil v obeh korpusih tolikšna, da večja in besedilno raznovrstnejša učna množica glede na množico, ki jo sestavlja eno besedilo oz. enoten besedilni tip, rezultate pri razčlenjevanju v povprečju celo minimalno poslabša.

Dosežena stopnja natančnosti (zlasti prva in tretja faza) je posledica dejstva, da je označevalni sistem Slovenske odvisnostne drevesnice tako na oblikoskladenjski kot površinskoskladenjski ravni zelo kompleksen. Rezultati bi bili nekoliko boljši tudi, če uporabljeni razčlenjevalni algoritem ne bi bil omejen na analizo projektivnih struktur. Označevalni sistem Slovenske odvisnostne drevesnice kot drevesnice jezika s prostim besednim redom neprojektivne strukture dovoljuje, saj bi bilo načelo projektivnosti[19] pri predstavitvi slovenske skladnje preveč omejevalno. Analiza je pokazala še, da je treba, ko je govora o slovenski skladnji, posplošitve v opisu

[18] Podatke o tem so dale izkušnje pri ročnem površinskoskladenjskem označevanju, posredno pa na to opozarja že nabor različnih oblikoskladenjskih oznak, uporabljenih v korpusih (»1984«: 774, SVEZ-IJS: 452).
[19] Načelo projektivnosti zahteva, da je v vsakem odvisnostnem razmerju $a \rightarrow b$, če se med strukturama *a* in *b* glede na besedni red pojavi struktura *c,* ta odvisna od strukture *a*. Strukturo *c*, ki je odvisna od strukture *a*, mora sestavljati glede na besedni red neprekinjen niz besednih oblik.

zrelativizirati vsaj v tem smislu, da opozorimo, da se od vseh možnih sintagmatskih razmerij, ki jih jezik dovoljuje, v posameznih besedil(n)ih (tipih) (lahko) uresničuje njihov sorazmerno omejen in specifičen nabor.

3 Sklep

Kljub temu da so se z gradnjo skladenjsko označenih korpusov začeli ukvarjati že pred več kot tridesetimi leti, je kompleksnost tega raziskovalnega področja botrovala k temu, da ostajajo vprašanja o načinih gradnje skladenjsko označenih korpusov in njihovih označevalnih modelih še vedno aktualna. Tudi pri oblikovanju Slovenske odvisnostne drevesnice je vprašanje, kakšen naj bo označevalni model drevesnice, da ga bo mogoče uporabiti pri označevanju katerega koli slovenskega (pisnega) besedila, bistvenega pomena. Izhodiščne raziskave na v skladenjskem smislu relativno specifičnih besedilih so pokazale, da je nujno, da v drevesnico vključimo večje število reprezentativnejših besedil. Cilj projekta gradnje drevesnice, uresničitev katerega pa je v veliki meri odvisna od možnosti financiranja projekta, namreč je, da oblikujemo takšen korpus, ki bo omogočal relevantne jezikoslovne analize na realnih besedilih in razvoj jezikovnih tehnologij za slovenščino, za kar pa bo potreben bolj obsežen in uravnotežen korpus.

Literatura

Abeillé, Anne, 2003: Introduction. Abeillé, Anne (ur.): *Treebanks*: *Building and Using Syntactically Annotated Corpora*. Dordrecht: Kluwer Academic Publishers. XIII–XXVI.

Abeillé, Anne et al., 2003: Building a Treebank for French. Abeillé, Anne (ur.): *Treebanks*: *Building and Using Syntactically Annotated Corpora*. Dordrecht: Kluwer Academic Publishers. 165–187.

Bémová, Alla et al., 1999: *Annotations at Analytical Level*: *Instructions for Annotators*. Praga: UK MFF UFAL.

Bouma, Gosse, 2004: Treebank Evidence for the Analysis of PP-Fronting. Kübler, Sandra et al. (ur.): *Proceedings of the Third Workshop on Treebanks and Linguistic Theories*, *TLT'04*. 25–26.

Brants, Thorsten et al., 2003: Syntactic Annotation of a German Newspaper Corpus. Abeillé, Anne (ur.): *Treebanks*: *Building and Using Syntactically Annotated Corpora*. Dordrecht: Kluwer Academic Publishers. 73–87.

Chanev, Atanas, 2005: Portability of Dependency Parsing Algorithms: An Application for Italian. *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories*, *TLT'05*.

Collins, Michael et al., 1999: A Statistical Parser for Czech. *Proceedings of the 37th ACL'99*. College Park: University of Maryland. 505–512.

Džeroski, Sašo et al., 2006: Towards a Slovene Dependency Treebank. *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, LREC'06. Pariz: ELRA. 1388–1391.

Erjavec, Tomaž, 1996/1997: Računalniške zbirke besedil. *Jezik in slovstvo* 42/2–3. 81–95.

Erjavec, Tomaž, 2006: The English-Slovene ACQUIS Corpus. *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, LREC'06. Pariz: ELRA. 2138–2141.

Erjavec, Tomaž in Ledinek, Nina, 2006: Slovenska odvisnostna drevesnica: prvi rezultati. Erjavec, Tomaž in Žganec Gros, Jerneja (ur.): *Zbornik 9. mednarodne multikonference Informacijska družba IS 2006*, *Zbornik 5. slovenske in 1. mednarodne konference Jezikovne tehnologije IS-LTC*. Ljubljana: Institut Jožef Stefan. 162–167.

Erjavec, Tomaž in Sárossy, Bence, 2006: Oblikoslovno označevanje slovenskega jezika: primer korpusa SVEZ-IJS. Erjavec, Tomaž in Žganec Gros, Jerneja (ur.): *Zbornik 9. mednarodne multikonference Informacijska družba IS 2006*, *Zbornik 5. slovenske in 1. mednarodne konference Jezikovne tehnologije IS-LTC*. Ljubljana: Institut Jožef Stefan. 168–173.

Gjorgjioski, Valentin, 2006: *Evaluating MaltParser performance on Slovene Dependency Treebank*. Ljubljana: Institut Jožef Stefan [neobjavljeno].

Gorjanc, Vojko, 2005: *Uvod v korpusno jezikoslovje*. Domžale: Izolit.

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

Hinrichs, Erhard W. et al., 2005: What Treebanks Can Do for You: Rule-Based and Machine-Learning Approaches to Anaphora Resolution in German. *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories*, *TLT'05*.

Kennedy, Graeme, 1998: *An Introduction to Corpus Linguistics*. London: Longman.

Ledinek, Nina, 2005: *Površinskoskladenjsko označevanje korpusa Slovene Dependency Treebank* (*s poudarkom na predikatu*). Diplomska naloga. Ljubljana: Filozofska fakulteta Univerze v Ljubljani.

Marciniak, Małgorzata et al., 2003: An HSPG-Annotated Test Suite for Polish. Abeillé, Anne (ur.): *Treebanks*: *Building and Using Syntactically Annotated Corpora*. Dordrecht: Kluwer Academic Publishers. 129–146.

McDonald, Ryan et al., 2006: *Multilingual Dependency Analysis with a Two-Stage Discriminative Parser. Proceedings of the 10th Conference on Computational Natural Language Learning*, *CoNNL-X*.

McEnery, Tony in Wilson, Andrew, 1996: *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Montemagni, Simonetta et al., 2003: Building the Italian Syntactic-Semantic Treebank. Abeillé, Anne (ur.): *Treebanks*: *Building and Using Syntactically Annotated Corpora*. Dordrecht: Kluwer Academic Publishers. 189–210.

Nivre, Joakim et al., 2005: MaltParser: A Language Independent System for Data-Driven Dependency Parsing. *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories*, *TLT'05*.

Nivre, Joakim, 2005: Dependency Grammar and Dependency Parsing: *MSI Report 05133*. Växjö: Växjö University, School of Mathematics and Systems Engineering.

Samuelsson, Christer, 2000: A Statistical Theory of Dependency Syntax. *Proceedings of the 18th International Conference on Computational Linguistics*, *COLING 2000.* Morgan Kaufmann. 684–690.

Simov, Kiril et al., 2002: Building a Linguistically Interpreted Corpus of Bulgarian: The BulTreeBank. Rodríguez, Manuel González, Suarez Araujo, Carmen Paz (ur.): *Proceedings of the Third International Conference on Language Resources and Evaluation*, *LREC'O2*. Grand Canaria: ELRA. 1729–1736.

Taylor, Ann et al., 2003: The Penn Treebank: An Overview. Abeillé, Anne (ur.): *Treebanks*: *Building and Using Syntactically Annotated Corpora*. Dordrecht: Kluwer Academic Publishers. 5–22.

Verdonik, Darinka, 2004: Parsed Corpora: An Overview and Posibility of Their Adaptation to Slovenian. Horvat, Bogomir in Kačič, Zdravko (ur.): *Advances in Speech Technology*. Maribor: Fakulteta za elektrotehniko, računalništvo in informatiko. 208–211.

Spletne strani

BulTreeBank: <http://www.bultreebank.org/>. (Dostopno 19. 2. 2007.)

10th Conference on Computational Natural Language Learning (CoNNL-X), 2006: <http://www.cnts.ua.ac.be/conll2006/proceedings.html>. (Dostopno 19. 2. 2007.)

Korpus SUSANNE: <http://www.grsampson.net/RSue.html>. (Dostopno 19. 2. 2007.)

Linguistic Data Consortium (LDC), 2006: *Prague Dependency Treebank 2.0. LDC2006T01*. <http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T01>. (Dostopno 19. 2. 2007.)

Schulte im Walde, Sabine in Zinsmeister, Heike, 2006: Introduction to Corpus Resources, Annotation and Access. *Foundational Course at ESSLLI 2006*, 18th European Summer School in Logic, Language and Information. Málaga, Španija, 31. 7.–11. 8. 2006. <http://www.coli.uni-saarland.de/~schulte/Teaching/ESSLLI-06/Slides/syntax.pdf>. (Dostopno 19. 2. 2007.)

Priporočila za skladenjsko označevanje korpusov iniciative EAGLES: <http://www.ilc.cnr.it/EAGLES96/segsasg1/segsasg1.html>. (Dostopno 19. 2. 2007.)

Penn Treebank: <http://www.cis.upenn.edu/~treebank/>. (Dostopno 19. 2. 2007.)

Prague Dependency Treebank: <http://ufal.mff.cuni.cz/pdt/>. (Dostopno 19. 2. 2007.)

Projekt TIGER: <http://www.ims.uni-stuttgart.de/projekte/TIGER/>. (Dostopno 19. 2. 2007.)

Slovenska odvisnostna drevesnica: <http://nl.ijs.si/sdt/>. (Dostopno 19. 2. 2007.)

TüBa-D/Z: <http://www.sfs.uni-tuebingen.de/en_tuebadz.shtml>. (Dostopno 19. 2. 2007.)