Slovenska odvisnostna drevesnica: prvi rezultati

Tomaž Erjavec*, Nina Ledinek†

*Odsek za tehnologije znanja, Institut "Jožef Stefan"
Jamova 39, 1000 Ljubljana
tomaz.erjavec@ijs.si
†Šoštanj, Slovenija
nina.ledinek@siol.net

Povzetek

Članek obravnava rezultate prve faze gradnje korpusa Slovenske odvisnostne drevesnice (Slovene Dependency Treebank, SDT), ki trenutno obsega 2.000 povedi oz. približno 30.000 besed. Zaradi skladenjskih sorodnosti med češčino in slovenščino, dostopnosti izčrpnega priročnika za površinskoskladenjsko označevanje češčine in inteligentnega urejevalnika dreves je označevalni sistem SDT oblikovan po modelu korpusa Prague Dependency Treebank (PDT). Korpus SDT zaenkrat sestavlja del oblikoskladenjsko označenega vzporednega korpusa MULTEXT-East, tj. prvi del prevoda romana *1984* Georgea Orwella. Korpus je bil najprej označen avtomatsko, nato pa so bile jezikovnoanalitične skladenjske oznake s pomočjo urejevalnika TrEd popravljene še ročno. Vzporedno je potekalo tudi prilagajanje češkega priročnika za označevanje za slovenščino. Trenutna verzija korpusa je dostopna v formatih XML/TEI in izpeljanih formatih in je že bila vključena v več raziskav, predvsem tisto v okviru CoNLL-X, ki je evalvirala natančnost dvajsetih dependenčnih razčlenjevalnikov na drevesnicah trinajstih jezikov. Prispevek predstavlja tudi načrte za nadaljnje delo, zlasti v zvezi s korpusom, ki bo Slovenski odvisnostni drevesnici dodan v prihodnje.

Slovene Dependency Treebank: first results

The paper presents the first release of the Slovene Dependency Treebank, currently containing 2,000 sentences or 30,000 words. Our approach to annotation is based on the Prague Dependency Treebank, which serves as an excellent model due to the similarity of the languages, the existence of a detailed annotation guide and an annotation editor. The initial treebank contains a portion of the MULTEXT-East parallel word-level annotated corpus, namely the first part of the Slovene translation of Orwell's "1984". This corpus was first parsed automatically, to arrive at the initial analytic level dependency trees. These were then hand corrected using the tree editor TrEd. The current version is available in XML/TEI, as well as derived formats, and has been used in several comparative evaluations, e.g. as part of the dataset for the CoNLL-X shared task on dependency parsing. Further work, in the first instance the composition of the corpus to be annotated next is also discussed.

1. Uvod

Skladenjsko označeni korpusi[1] postajajo pomembni jezikovni viri, saj omogočajo statističen pregled distribucije skladenjskih kategorij na velikem vzorcu besedil dejanske jezikovne rabe – pri čemer skladenjsko analizo (relativno) velikega vzorca realnih besedil navadno predpostavljajo in jo, predvsem, napovedujejo – in tako olajšujejo raziskave teoretičnega jezikoslovja ter skladnje posameznih jezikov. Poleg tega potrebujemo podatke, ki jih skladenjsko označeni korpusi nudijo, tudi za razvoj jezikovnih tehnologij, saj je na njih mogoče testirati in predvsem šolati avtomatske skladenjske označevalnike, pri čemer dajejo v zadnjem času zelo obetavne rezultate zlasti statistični označevalniki.

Zaenkrat obstaja kar nekaj problemsko zamejenih opisov skladnje slovenščine, ki sledijo različnim jezikoslovnim usmeritvam, in (nesočasnih) slovnic slovenskega jezika.[2] Najbolj celovita je Slovenska slovnica (Toporišič, 1984), najizčrpnejša opisa različnih vidikov slovenske skladnje pa sta Nova slovenska skladnja (Toporišič, 1982) ter Vezljivost v slovenskem jeziku (s poudarkom na glagolu) (Žele, 2001), poleg tega pa so bili številni skladenjski fenomeni raziskani zlasti v okviru generativne paradigme. Vendar pa za slovenščino še vedno ne obstaja nobena strogo formalna, računalniška (tj. primerna za računalniško obravnavo jezika) in obenem izčrpna slovnica. Do nedavnega tudi skladenjsko označenega korpusa slovenskega jezika še nismo imeli, saj so bili dostopni samo oblikoskladenjsko označeni in lematizirani korpusi slovenščine (Jakopin in Bizjak, 1996; Lönneker, 2005; Erjavec et al., 1998; Erjavec, 2006).

Z gradnjo korpusa Slovene Dependency Treebank[3] smo začeli leta 2003. V prvi fazi smo izbrali teoretični model označevanja, usposobili programsko platformo in pripravili korpus, tako da smo ga avtomatsko skladenjsko označili. V naslednji fazi smo se posvetili ročnemu površinskoskladenjskemu označevanju 2.000 povedi oz. 30.000 besed korpusa in pripravi priročnika za površinskoskladenjsko označevanje slovenskega jezika. Rezultati dela, ki je bilo zaključeno pred kratkim (Džeroski et al., 2006), so že vidni, saj je bil korpus že uporabljen v dveh raziskavah.

Čeprav je zaenkrat označen relativno majhen nabor povedi, smo v času od začetka projekta uspeli ustvariti za skladenjsko označevanje potrebno infrastrukturo. V razdelku 2 bomo zato prikazali, kakšen teoretični model smo za gradnjo korpusa izbrali, razdelek 3 pojasnjuje, kako smo ga operacionalizirali in prilagodili za slovenščino, v razdelku 4 pa bomo predstavili rezultate tekmovanja CoNLL-X v zvezi s korpusom SDT. Razdelek 5 prinaša nekaj zaključkov.

[1] Za dobronamerne pripombe in komentarje, ki so pripomogli k izboljšanju prvotne verzije besedila, se avtorja zahvaljujeta anonimnemu recenzentu. Za morebitne napake, ki se v članku še vedno pojavljajo, sta odgovorna sama.

[2] Npr. Breznikova (1916–1934), čitankarska (Bajec et al.; 1940–1956, 1964), Toporišičeva (1976–2004) ipd. Pogosto so nastajale (tudi) kot nekakšni nadomestki za srednješolske učbenike ali pa kot njihova nadgradnja.

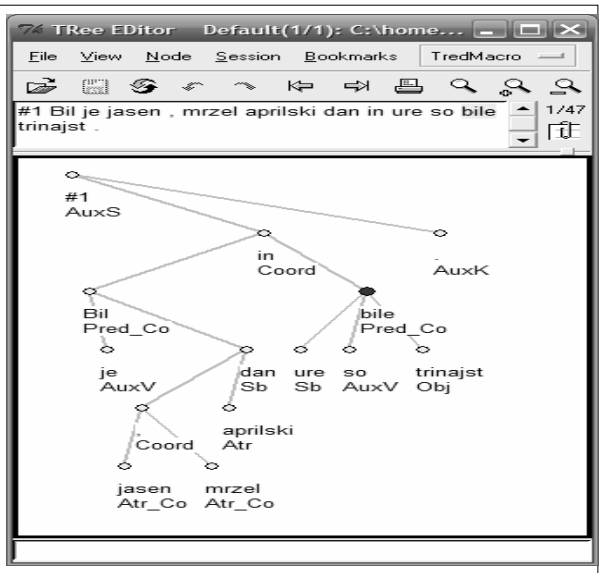
[3] http://nl.ijs.si/sdt/

2. Ozadje

Pri gradnji skladenjsko označenega korpusa, zlasti za jezik, ki takšnega korpusa še nima, je odločilnega pomena, kakšen teoretičen (in praktičen) model označevanja razvijemo oz. prevzamemo. Projekt Slovenska odvisnostna drevesnica namenskega vira financiranja zaenkrat nima, zato razvijanje lastnega modela označevanja ni bilo mogoče, hkrati pa bi bila priprava takšnega sistema otežena zaradi kadrovskih in časovnih omejitev. Glede na tipološke sorodnosti med jeziki in teoretične modele njihove obravnave smo torej med javno dostopnimi označevalnimi modeli izbrali najbolj ustreznega in prevzete jezikoslovne rešitve prilagodili za slovenščino.

Čeprav za veliko germanskih ter romanskih jezikov in za nekatere slovanske jezike, npr. bolgarščino (Simov et al., 2002) in ruščino (Boguslavsky et al. 2000), skladenjsko označeni korpusi in programska orodja za njihovo analizo že obstajajo, smo za slovenščino našli vzornika v projektu Prague Dependency Treebank[4] (LDC, 2001). Gre za enega najbolj ambicioznih in najbolje dokumentiranih projektov skladenjskega označevanja morfološko bogatih jezikov s prostim besednim redom, poleg tega pa je za komparativno analizo že na voljo zelo obsežen, na dveh ravneh označen korpus. Projekt PDT je za slovenščino posebej relevanten še zlasti zato, ker smo lahko zaradi pomensko-, funkcijsko- in strukturnoskladenjske podobnosti med slovenščino in češčino poleg teoretičnega modela (dependenčna skladnja, funkcijska generativna slovnica) v prvi fazi dela neposredno prevzeli tudi sistem ročnega površinskoskladenjskega označevanja korpusa, definiranega v priročniku *Annotations at Analytical Level: Instructions for Annotators* (Bémová et al., 1999) (v nadaljevanju: *AAL*), poleg tega pa smo imeli na voljo še urejevalnik dreves, ki zelo olajšuje ročno označevanje korpusa in omogoča njegovo vizualizacijo.

Na analitični, tj. površinskoskladenjski ravni, pri čemer so upoštevana zlasti funkcijskoskladenjska razmerja, je struktura vsake povedi v korpusu PDT predstavljena s skladenjskim drevesom, v katerem je razločevalno opredeljen tip površinskoskladenjske odvisnosti vsake pojavnice v povedi v razmerju do njenega neposredno nadrejenega elementa. Tektogramatična raven oz. pomenskoskladenjska raven že vključuje globlja, semantična razmerja med besedami, upoštevana pa so tudi koreferenčna razmerja, rematsko-tematska struktura povedi oz. členitev po aktualnosti. Korpus SDT je zaenkrat označen le na površinskoskladenjski ravni.



Slika 1: Skladenjsko drevo iz korpusa SDT v TrEdu.

2.1. Priročnik za površinskoskladenjsko označevanje

Ena od prednostnih nalog pri gradnji in označevanju skladenjsko označenega korpusa, zlasti če je označevalcev korpusa več, je pripraviti priročnik za označevanje, ki bo predpisoval način označevanja za čim več skladenjskih struktur. Opis skladenjskih razmerij mora biti zaradi zahtev avtomatske obdelave jezika skrajno formaliziran in izčrpen, poleg tega pa mora ilustrirati označevalne konvencije tudi z zgledi (delov) skladenjskih dreves. Priprava takšnega priročnika je zelo zahtevna in zamudna, potekati pa mora hkrati z ročnim označevanjem, kot nekakšna sinteza analize, opisa in, posledično, "izčiščenja" označevalnega sistema na podlagi pridobljenih izkušenj.

Za korpus PDT so javno dostopen priročnik za površinskoskladenjsko označevanje (v češčini) pripravili leta 1997 (Hajič et al., 1997). Zaradi skladenjske sorodnosti med češčino in slovenščino smo začeli korpus SDT označevati po pravilih priročnika *AAL*, istočasno pa smo ga začeli primerjati z obstoječimi opisi slovenske skladnje in ga nato za potrebe slovenščine prilagajati glede na izkušnje pri ročnem označevanju ter glede na razumevanje pomensko-, funkcijsko- ter strukturnoskladenjske vloge struktur v novejšem slovenskem jezikoslovju. Proces prilagajanja sistema za površinskoskladenjsko označevanje razumemo kot permanenten proces, ki bo, glede na to, da korpusi kažejo, da je jezik v skladenjskem smislu veliko bolj raznolik, pravila pa mnogo manj določljiva, kot kažejo mnogi sedanji jezikoslovni opisi, trajal še kar nekaj časa. Definirati bo treba način označevanja struktur, specifičnih za slovenščino,[5] češke zglede bo treba nadomestiti s

[4] The Prague Dependency Treebank, 1.0 in 2.0β, http://ufal.mff.cuni.cz/pdt/

[5] Te strukture je večinoma mogoče odkriti le naključno, pri primerjavi opisov in ročnem označevanju povedi (ko naletimo na strukturo, za katero menimo, da v slovenskem jezikoslovju v

slovenskimi, vse druge spremembe priročnika *AAL*, zlasti tiste, ki so posledica razlik (v interpretaciji) jezikovnih sistemov, pa bomo morali zelo natančno dokumentirati.

V trenutni fazi dela prilagajamo samo označevanje površinskoskladenjske ravni, načrtujemo pa, da bomo korpus kasneje označili tudi pomenskoskladenjsko. Glede na spremembe na površinskoskladenjski ravni bo treba prilagoditi tudi nekatere pomenskoskladenjske oznake, vendar pričakujemo, da bo zaradi večje stopnje "univerzalnosti" te ravni prilagoditev manj kot sprememb na površinskoskladenjski ravni.[6]

Vpeljani označevalni sistem je primeren predvsem za skladenjsko označevanje pisnih tekstov, za analizo govornih korpusov pa ga bo treba dodatno spremeniti (ali morda vpeljati novega). Skladnja govornega teksta je namreč veliko bolj kompleksna oz. manj "urejena" kot skladnja pisnih besedil (Halliday, 1989), njenih pojavnih oblik zato ne smemo obravnavati kot "odstopov" od pisne norme. Ker za slovenščino pregleden in izčrpen opis skladnje govornih tekstov še ne obstaja in ker niti dileme v zvezi z označevanjem govornih korpusov na "nižjih" ravneh, ki so predpogoj za skladenjsko označevanje oz. njegova predstopnja, večinoma še niso razrešene, priprave skladenjsko označenega govornega korpusa pri projektu Slovenska odvisnostna drevesnica zaenkrat ne načrtujemo,[7] kljub temu da se glede na tendenco razvoja korpusnega jezikoslovja pomembnosti področja zavedamo.

2.2. Urejevalnik dreves TrEd in skladenjski razčlenjevalnik za slovenščino

Pomembno orodje pri gradnji skladenjsko označenega korpusa je urejevalnik dreves, ki omogoča vizualizacijo in ročno korekcijo (avtomatsko že označenih) skladenjskih dreves. Dober urejevalnik označevalcu korpusa označevanje zelo olajša, hitrost dela se poveča, število napak pri označevanju pa je znatno manjše.

Za delo s korpusom PDT je bil razvit urejevalnik TrEd (Hajič et al., 2001), ki je javno dostopen, zato ga uporabljamo tudi pri projektu Slovenska odvisnostna drevesnica. Napisan je v programskem jeziku Perl/Tk in deluje tako na operacijskem sistemu Linux kot na sistemu Windows. Omogoča navigacijo med datotekami in povedmi, označevanje struktur z operacijo "primi in spusti" ter hitro izbiro analitičnih oznak s seznamov. Program je zelo konfigurabilen ter podpira precejšnje število vhodnih in izhodnih formatov (npr. XML/TEI, omogoča pa tudi prikaz skladenjskih dreves v formatu GIF). Slika skladenjskega drevesa, kot jo vidimo na računalniškem ekranu v programu TrEd, je prikazana na Sliki 1.

Da bi se označevalcem korpusa delo olajšalo, je bil razvit skladenjski razčlenjevalnik za slovenščino (Džeroski et al., 2006), ki deluje na osnovi majhnega števila ročno napisanih pravil. Program izkorišča oznake, ki so bile prvotnim besednim oblikam pripisane na oblikoskladenjski ravni in ki dajejo sorazmerno dobro informacijo o (potencialni) skladenjski strukturi povedi, in s pomočjo te informacije prvotnim besednim oblikam pripiše analitične oznake in odvisnostna razmerja.

3. Korpus SDT

V prvi fazi gradnje korpusa smo izvorne datoteke v formatu XML (glej razdelek 3.1) pretvorili v format programa TrEd. Razdelili smo jih na manjše datoteke, ki vsebujejo približno 50 povedi.[8] Te so bile nato najprej označene avtomatsko, nato pa ročno pregledane in popravljene. Kasneje smo analitične oznake TrEdovih datotek pridružili izvornim podatkom, nastali korpus pa je bil nato zopet pretvorjen v dokument XML (glej razdelek 3.2) in trenutno inačico korpusa SDT. Nadaljnji načrti za razširitev korpusa so predstavljeni v razdelku 3.3.

3.1. Korpus MULTEXT-East "1984"

V prvi fazi dela se nam je zdelo najpomembneje, da je korpus, ki naj bi služil kot osnova za skladenjsko označevanje, čim bolje dokumentiran in da je oblikoskladenjsko čim natančneje označen. Kot izvorni korpus za površinskoskladenjsko označevanje smo zato izbrali slovenski del oblikoskladenjsko označenega vzporednega korpusa MULTEXT-East (Erjavec, 2004), ki vsebuje oblikoskladenjsko označen prevod romana *1984* Georgea Orwella.

Korpus MULTEXT-East je zapisan v formatu XML, upošteva priporočila iniciative TEI P4 (Sperberg-McQueen in Burnard, 2002) ter je stavčno poravnan z angleškim originalom in prevodi romana v nekatere druge jezike. Razdvoumljanje oblikoskladenjskih oznak in lem glede na kontekst je potekalo v dveh fazah, najprej avtomatsko, nato pa so bile oznake pregledane še ročno. Njihova definicija je bila prevzeta po načelih projekta MULTEXT in oblikovana v sodelovanju z iniciativo EAGLES, kar omogoča večjo izmenljivost podatkov, poleg tega pa zagotavlja tudi možnost njihove avtomatske analize.

V SDT je trenutno zajet prvi del romana, ki vsebuje tretjino besedila, tj. okoli 30.000 besed oz. 2.000 povedi. Korpus glede na kvaliteto in obseg že vsebovanih oznak sicer nudi dobro osnovo, ima pa takšen izbor tudi nekaj očitnih pomanjkljivosti: korpus sestavlja eno samo prevodno umetnostno besedilo, ki vsebuje tudi izmišljen jezik (novorek), poleg tega pa so za tekst značilni dolgi stavki in premi govor, roman pa je na nekaterih mestih tudi nekoliko slabše preveden in zlektoriran, kar njegovo označevanje zelo otežuje.

3.2. Korpus SDT 0.4

Verzija 0.4 korpusa SDT[9] obsega 1998 povedi (29.991 besed in 6.563 ločil), ki so bile ročno površinskoskladenjsko označene, že izvorni MULTEXT-

---

funkcijskoskladenjskem smislu še ni bila opisana), zato jih bomo v priročnik za označevanje dodajali vedno znova.

[6] Upoštevati pa moramo tudi dejstvo, da se označevalna sistema PDT in SDT razlikujeta že na oblikoskladenjski ravni, saj prvi predvideva približno 4700 oznak, drugi pa le okrog 2100 oznak.

[7] Kolikor je avtorjema članka znano, za noben jezik skladenjsko označeni govorni korpus še ne obstaja, obstajajo le načrti zanj (npr. pri projektu PDT).

[8] Takšna razdelitev teksta je pomembna zlasti iz psiholoških razlogov, saj lahko označevalec na dan označi le približno 50 povedi. Število označenih datotek je za označevalca pomembno merilo napredka dela.

[9] Kolofon SDT 0.4 je dosegljiv na naslovu http://nl.ijs.si/sdt/sdtHeader-2006-05-17.html

East pa ročno lematiziran in oblikoskladenjsko označen. Dostopen je v nekaj različnih formatih, kanoničen format je format korpusa MULTEXT-East TEI P4 z dodanimi atributi pojavnic, v katerih je kodirana kazalka na neposredno nadrejeno pojavnico (parent node, atribut dep) in tip skladenjske odvisnosti med starševsko in hčerinsko pojavnico (vloga pojavnice na površinskoskladenjski ravni, atribut afun in, za koordinacije, parallel). Primer z začetka korpusa je podan na Sliki 2.

```
<text id="Osl." lang="sl">
 <body>
  <div type="part" id="Osl.1">
   <div type="chapter" id="Osl.1.2">
    <p id="Osl.1.2.2">
     <s id="Osl.1.2.2.1">
      <w id="s1t1" dep="s1t8"
          afun="Pred" parallel="Co"
          ana="Vcps-sma"
          lemma="biti">Bil</w>
      <w id="s1t2" dep="s1t1"
          afun="AuxV"
          ana="Vcip3s--n"
          lemma="biti">je</w>
      <w id="s1t3" dep="s1t4"
           afun="Atr" parallel="Co"
           ana="Afpmsnn"
           lemma="jasen">jasen</w>
      <c id="s1t4" dep="s1t7"
           afun="Coord">,</c>
```

Slika 2: SDT v kanoničnem formatu TEI; začetek korpusa "Bil je jasen, mrzel aprilski dan …".

Korpus SDT je sestavljen iz kolofona TEI, ki korpus dokumentira, in treh dokumentov TEI. Prvi vsebuje formalne oblikoskladenjske specifikacije korpusa MULTEXT-East, ki definirajo nabor oznak za oblikoskladenjske kategorije, uporabljene pri označevanju korpusa (torej vrednosti atributa ana). Drugi dokument prinaša seznam možnih analitičnih oznak (afun in parallel). Tretji dokument obsega zaenkrat edino (dokončno urejeno) komponento skladenjsko označenega korpusa, kot rečeno, 1. tretjino slovenskega prevoda romana *1984.*

3.3. Razširitev korpusa

V nadaljnjih fazah projekta se bomo osredotočili predvsem na dva segmenta dela. Prioriteta bo še naprej prilagajanje priročnika za površinskoskladenjsko označevanje (v prvi fazi smo pozornost posvečali predvsem prilagajanju sistema označevanja struktur, ki jim v slovenskem jezikoslovju navadno pripisujemo vlogo povedka), k delu bo zato treba pritegniti tudi več novih označevalcev, poleg tega pa bomo pripravili nov tekst za označevanje, s katerim bomo korpus SDT razširili. Ob tem se seveda pojavlja dilema, kateri tekst naj kot novo komponento korpusa izberemo.

Kot pomemben dejavnik je treba upoštevati Penn Treebank (Marcus et al., 1993), enega najrelevantnejših skladenjsko označenih korpusov. Oblikovanje korpusa, ki bi bil glede dokumentiranosti in označevanja primerljiv s korpusom Penn Treebank, bi omogočilo komparativne analize in poenostavilo druge raziskave. Poleg tega se je z objavo korpusa Prague Czech-English Dependency Treebank, ki vsebuje prevod dela korpusa Penn Treebank v češčino, oba dela vzporednega korpusa pa sta označena z analitičnimi oznakami, pojavila nova priložnost za raziskave. Prevod istega dela korpusa Penn Treebank v slovenščino in njegova označitev bi pomenila nastanek trijezičnega vzporednega korpusa, tak jezikovni vir pa bi bil odličen za medjezikovne raziskave in raziskave strojnega prevajanja. Zagotovljena bi bila tudi možnost za raziskave učenja avtomatskega skladenjskega označevanja s pomočjo prenosa pravil med jeziki (Kuhn, 2004), pri čemer bi se skušali naučiti označevanja slovenščine prek skladenjskega razčlenjevanja češčine.

Slovensko odvisnostno drevesnico bi radi uporabili čim prej – za to je pomembno, da podatke za razvoj računalniških aplikacij za specifično rabo pridobimo iz tekstov, povezanih z istim specifičnim področjem. Zato bomo v naslednji fazi označili vzorec povedi iz dveh korpusov, SVEZ-IJS (Erjavec, 2006) in korpusa časopisnih člankov. Vzorec iz vzporednega angleško-slovenskega korpusa SVEZ-IJS, ki obsega približno 800 povedi in 15.000 besed, je površinskoskladenjsko že označen, vendar pa je zaenkrat dostopen le v formatu fs. Za njegovo označevanje smo se odločili, ker so bile oblikoskladenjske oznake vzorca ročno popravljene in ker je bil korpus dostopen, hkrati pa je aplikativno zanimiv – z nastankom velikega vzporednega skladenjsko označenega korpusa bi bile mogoče raziskave strojnega prevajanja in druge medjezikovne raziskave. Z vidika reprezentativnosti je izbira korpusa seveda manj ustrezna, saj ga sestavljajo prevodni teksti, zato je tipologija dobljenih stavčnih vzorcev za slovenščino (lahko) vprašljiva, poleg tega pa so upravno-pravni teksti v skladenjskem smislu specifični (obsežne naštevalne enote ter "tabelarnost" in siceršnja skrajna formaliziranost določenih delov teksta npr. povzročajo, da je precejšen del povedi interpretiran kot niz elips, poleg tega pa velik del vzorca sestavljajo zelo obsežne in kompleksne samostalniške zveze). Status normativne reference, rečeno pogojno, bi lažje pripisali korpusu slovenskih časopisnih besedil, ki ga bomo označevali v naslednji fazi dela, vzorec reprezentativnejših besedilnih tipov pa bo (verjetno) vzet iz korpusov Fida ali FidaPlus.

4. Šolanje in evalvacija razčlenjevalnikov na korpusu SDT

Drevesnice so po eni strani uporabne za jezikoslovne raziskave jezikov, po drugi pa za razvoj jezikovnih tehnologij, saj avtomatsko skladenjsko razčlenjevanje besedil omogoča bistveno boljšo osnovo za nadaljnje obdelave, npr. strojno prevajanje, iskanje informacij, avtomatsko sumarizacijo itd.

Tradicionalni skladenjski razčlenjevalniki iz 70. in 80. let so temeljili na ročno napisanih pravilih in leksikonu, vendar pa je bilo zanje tipično majhno pokritje pravil, poleg tega niso bili odporni na napake v besedilih, na neznane besede in konstrukcije, niso pa tudi semantično oz. kontekstno razdvoumili besedila, kar pomeni, da je dobila ena poved veliko število različnih analiz. Samo za nekaj največjih jezikov so bili razčlenjevalniki (pravila,

leksikon) izdelani do te mere, da so postali uporabni za analize odprtega besedila, saj zahteva izdelava potrebne infrastrukture ogromno dela in sredstev.

V zadnjih letih se je izredno okrepilo zanimanje za metode obravnave jezika, ki temeljijo na pristopih strojnega oz. statističnega učenja. Skupno jim je to, da se orodja, ki te metode uporabljajo, določenega modela jezika induktivno naučijo iz vnaprej pripravljenih podatkov, v našem primeru skladenjsko označenega korpusa. Ti pristopi so robustni in (seveda ob izdelanem označenem korpusu) ceneni, vendar pa dostikrat delajo "neumne" napake, naučeni modeli pa so netransparentni.

Aktualnost večjezičnega induktivnega skladenjskega razčlenjevanja se je pokazala v precejšnjem zanimanju za uporabo korpusa SDT, kljub njegovemu majhnemu obsegu in dejstvu, da je sredi razvoja. Že prototipna inačica SDT 0.1 je bila uporabljena v raziskavi o stopnji natančnosti skladenjskega razčlenjevanja (Chanev, 2005), kasneje pa tudi v drugih raziskavah z razčlenjevalnikom MALT (Nivre in Hall, 2005).

SDT 0.3 je bil nato vključen v dosti širšo evalvacijo, ki se je dogajala v sklopu konference CoNLL-X "10th Conference on Computational Natural Language Learning"[10] (CoNLL, 2006). CoNLL vsako leto organizira, po vzoru sedaj že številnih drugih konferenc, odprto tekmovanje (shared task) iz nekega področja strojnega učenja jezikovnih podatkov, pri čemer je bilo tekmovanje v letu 2006 posvečeno učenju odvisnostnih slovnic.[11] Naloga je zajemala testiranje na korpusih več jezikov, saj je moral vsak tekmovalec svoj razčlenjevalnik preizkusiti na vseh ročno označenih drevesnicah jezikov, ki so bile v tekmovanje vključene. Kot baze podatkov so bili testirani korpusi[12] 13 jezikov, od daljno- in bližnjevzhodnih do obilice evropskih, tudi češčina in slovenščina. Češki korpus je bil korpus PDT, z milijon besedami je bil eden večjih, slovenski SDT pa je bil najmanjši korpus, ki je na tekmovanju sodeloval.

Na tekmovanje je bilo prijavljenih 20 sistemov, potekalo pa je tako, da so tekmovalci dobili večji del korpusa za učenje svojega razčlenjevalnika, delovanje naučenega sistema pa je bilo potem preizkušeno na skritem delu korpusa. Tako učni kot skriti korpus sta vsebovala tudi oblikoskladenjske oznake in leme, kar je sistemom razčlenjevanje lahko olajšalo. Ocenjevanje je potekalo z enotnim programom, ki je meril rezultat (v odstotokih) za:

1. označeno povezanost (OP): pravilne so tako odvisnostne povezave kot tudi oznake (labeled attachment score);
2. neoznačeno povezanost (NP): pravilne so odvisnostne povezave, pravilnost oznak ni relevantna (unlabeled attachment score);
3. oznake (OZ): pravilne so oznake, pravilnost odvisnostnih povezav ni relevantna (label accuracy).

[10] New York, 8–9 junij 2006, http://www.cnts.ua.ac.be/conll/

[11] Opis vseh sistemov (zbornik) in rezultati so dostopni na http://nextens.uvt.nl/~conll/

[12] Tekmuje se v natančnosti, ki jo razčlenjevalniki pri analizi veliko različnih drevesnic dosežejo, vendar pa tovrstno tekmovanje daje posredno tudi informacije o natančnosti označevanja korpusov samih in o tipu razčlenjevalnikov, ki pri določeni predstavitvi jeziko(slo)vnih podatkov dosegajo najboljše rezultate.

Če je npr. zveza "bela hiša" v neosebkovi vlogi označena tako, da obstaja povezava med podrejeno pojavnico "bela" in nadrejeno pojavnico "hiša" in je ta povezava označena kot »Subj«, je neoznačena povezanost pravilna, označena povezanost in oznaka pa ne.

Rezultati posameznih sistemov se zelo razlikujejo, tako med seboj kot glede na jezik. Najboljše rezultate, tako v povprečju za vse jezike kot tudi za češčino in slovenščino, sta imela dva sistema:

1. dvostopenjski razlikovalni razčlenjevalnik (two-stage discriminative parser) (McDonald et al., 2006);
2. označeno psevdoprojektivno odvisnostno razčlenjevanje z uporabo SVM (labeled pseudo-projective dependency parsing with support vector machines) (Nivre et al., 2006).

Natančnost obeh sistemov se zelo razlikuje glede na obravnavane jezike. McDonald et al. (2006) imajo npr. najboljši rezultat za japonščino in bolgarščino, najslabšega pa za češčino, danščino, slovenščino, turščino in, končno, arabščino, pri čemer so razlike odraz ne samo različnosti jezikov, temveč tudi velikosti in raznovrstnosti korpusov, teoretičnih izhodišč in doslednosti označevanja.

V Tabeli 1 vidimo dosežene stopnje natančnosti za SDT in PDT, in sicer posebej za oba razčlenjevalnika in v povprečju za vseh 20 prijavljenih sistemov. Tabela kaže, da rezultati za slovenski jezik zaostajajo za rezultati za češčino, vendar je to delno pričakovano zaradi ogromne razlike v velikosti korpusov. Češki je za sodobne sisteme včasih celo prevelik, saj nekateri sistemi zaradi časovno zahtevnega učenja niso mogli uporabiti celega korpusa. Vendar pa je češki korpus tudi bolj raznolik, saj vsebuje besedila iz mnogih virov, zato je njegovo označevanje v primerjavi s SDT, ki vsebuje samo en roman, v splošnem verjetno težje.

	SDT	PDT
OP McDonald et al.	73.44	80.18
OP Nivre et al.	70.30	78.42
OP povprečno	65.16	67.17
NP McDonald et al.	83.17	87.30
NP Nivre et al.	78.72	84.80
NP povprečno	76.53	77.01
OZ McDonald et al.	82.51	86.72
OZ Nivre et al.	80.54	85.40
OZ povprečno	76.31	76.59

Tabela 1: Rezultati CoNLL-X za slovenski in češki jezik. OP = označena povezanost, NP = neoznačena povezanost, O = oznake

Vseeno so rezultati za slovenski jezik spodbudni, saj je bilo z avtomatskimi orodji, ki so se učila na zelo majhnem vzorcu jezika, mogoče pravilno označiti skoraj tri četrtine povezav skupaj z njihovimi oznakami. Seveda pa se je v zvezi s temi rezultati treba zavedati, da so oblikoskladenjske oznake v SDT ročno pregledane. V realnih sistemih se te oznake določajo strojno, pri čemer pride tudi do napak. Natančnost razčlenjevalnika, ki bi potem take oznake uporabljal, bi bila brez dvoma bistveno manjša.

5. Zaključek

Prispevek prikazuje rezultate prve faze gradnje korpusa Slovene Dependency Treebank, ki je oblikovan po korpusu Prague Dependency Treebank. Čeprav korpus ni obsežen, je že bil uporabljen v nekaj raziskavah. Da bi bil maksimalno uporaben, smo ga pretvorili v tri formate, format TEI P4, format urejevalnika dreves TrEd, tj. format fs, in v tabularno datoteko (tabular file), format, uporabljen v raziskavah v okviru CoNLL-X. SDT je opisan na domači strani projekta http://nl.ijs.si/sdt/ in je prosto dostopen za raziskovalne namene.

Predstavili smo tudi načrte za nadaljnje delo, ki pa so odvisni tudi od možnosti financiranja projekta. Še naprej se bomo ukvarjali s prilagajanjem priročnika za označevanje, razširitvijo korpusa z novimi teksti, poleg tega pa se bomo začeli posvečati tudi raziskavam indukcije pravil z avtomatskim skladenjskim razčlenjevalnikom.

Literatura

Bémová, A., Buráňová, E., Hajič, J., Panevová, J., Urešová Z. (1999). Annotations at Analytical Level: Instructions for Annotators. Praga, UK MFF UFAL.

Boguslavsky, I., Grigorieva, S., Grigoriev, N., Kreidlin, L., & Frid, N. (2000). Treebank for Russian: Concept, tools, types of information. COLING-2000.

Chanev, A. (2005). Portability of Dependency Parsing Algorithms - an Application for Italian. V: Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT'05). Barcelona.

CoNLL (2006). CoNNL-X "10th Conference on Computational Natural Language Learning", New York, 8–9 junij 2006.

Čmejrek, M., Curín, J., Havelka, J., Hajič, J., Kuboň, V. (2004). Prague Czech-English Dependency Treebank: Syntactically Annotated Resources for Machine Translation. LREC'04.

Džeroski, S., Erjavec, T., Ledinek, N., Pajas, P., Žabokrtsky, Z., Žele, A. (2006). Towards a Slovene Dependency Treebank. V: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'2006). Pariz, ELRA.

Erjavec, T., Gorjanc, V., Stabej, M. (1998). Korpus FIDA. Konferenca Jezikovne tehnologije za slovenski jezik. Ljubljana, Institut Jožef Stefan.

Erjavec, T. (2006). The English-Slovene ACQUIS corpus. V: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'2006). Pariz, ELRA.

Erjavec, T. (2004). MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. V: Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'2004). Pariz, ELRA.

Hajič, J., Pajas, P. in Vidová Hladká, B. (2001). The Prague Dependency Treebank: Annotation Structure and Support. IRCS Workshop on Linguistic databases, 2001 (pp. 105--114).

Hajič, J., Panevová, J., Buráňová, E., Urešová, Z., Bémová, A. (1997). A Manual for Analytic Layer Tagging of the Prague Dependency Treebank. UFAL Technical Report TR-1997-03, Karlova univerza, Češka republika.

Halliday, M. A. K. (1989). Spoken and written language. Oxford, University Press.

Jakopin, P., Bizjak, A. (1997). O strojno podprtem oblikoslovnem označevanju slovenskega besedila. Slavistična revija, 45(3–4), 513--532.

Kuhn, J. (2004). Experiments in Parallel-Text Based Grammar Induction. V: Proceeding of the ACL'04.

Ledinek, N. (2005) Površinskoskladenjsko označevanje korpusa Slovene Dependency Treebank (s poudarkom na predikatu). Diplomska naloga. Univerza v Ljubljani.

Ledinek, N., Žele, A. (2005). Building of the Slovene Dependency Treebank According to the Prague Dependency Treebank. V: Zbornik konference Gramatika & korpus. Praga, Ústav pro jazyk český. [V tisku].

Linguistic Data Consortium, (2001). Prague Dependency Treebank 1. LDC2001T10.

Linguistic Data Consortium, (2004). Prague Czech-English Dependency Treebank Version 1.0, LDC2004T25.

Lönneker, B. (2005). Strojno oblikoslovno označevanje slovenskih besedil: Kako daleč smo, Slavistična revija, 53(2), 193--210.

Marcus, M., Beatrice, P. S. & Markiewicz, M. A. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. Computational Linguistics, 19/2.

McDonald, R., Lerman, K., Pereira, F. (2006). Multilingual Dependency Analysis with a Two-Stage Discriminative Parser. CoNLL-X Shared Task: Multi-lingual Dependency Parsing. New York, 8–9 junij 2006. http://nextens.uvt.nl/~conll/slides/McDonald.pdf

Nivre, J., Hall, J. (2005). MaltParser: A Language-Independent System for Data-Driven Dependency Parsing. V: Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT'05). Barcelona.

Nivre, J., Hall, J., Nilsson, J., Eryiğit, G., Marinov, S. (2006). Labeled Pseudo-Projective Dependency Parsing with Support Vector Machines. CoNLL-X Shared Task: Multi-lingual Dependency Parsing. New York, 8–9 junij 2006. http://nextens.uvt.nl/~conll/slides/Nivre.pdf

Simov, K., Osenova, P., Slavcheva, M., Kolkovska, S., Balabanova, E., Doikoff, D., Ivanova, K., Simov, A., & Kouylekov, M. (2002). Building a linguistically interpreted corpus of Bulgarian: the BulTreeBank. LREC'02.

Sperberg-McQueen, C. M. in Burnard, L. (ur.) (2002). Guidelines for Electronic Text Encoding and Interchange, the XML Version of the TEI Guidelines. The TEI Consortium.

Toporišič, J. (1982). Nova slovenska skladnja. Ljubljana, DZS.

Toporišič, J. (1984). Slovenska slovnica. Maribor, Obzorja.

Žele, A. (2001). Vezljivost v slovenskem jeziku (s poudarkom na glagolu). Ljubljana, Založba ZRC, ZRC SAZU.

Žele, A. (2003). Glagolska vezljivost: iz teorije v slovar. Ljubljana, Založba ZRC, ZRC SAZU.