

ODVISNOSTNO POVRŠINSKOSKLADENJSKO OZNAČEVANJE SLOVENŠČINE: SPECIFIKACIJE IN OZNAČENI KORPUSI

Nina Ledinek

Inštitut za slovenski jezik Frana Ramovša, ZRC SAZU, Ljubljana

Tomaž Erjavec

Institut "Jožef Stefan", Ljubljana

UDK 811.163.6'367:811.163.6'322

Prispevek predstavi prve rezultate projektov JOS in SSJ s področja skladnje, in sicer nabor oznak za odvisnostno površinskoskladenjsko označevanje ter dva skladenjsko označena korpusa. Korpusa sta bila vzorčena iz referenčnega korpusa FidaPLUS ter imata ročno označene oz. pregledane leme, oblikoskladenjske ter površinskoskladenjske oznake. Viri bodo kot podatkovna zbirka na voljo za raziskovalne namene po licenci Creative Commons, namenjeni pa so zlasti razvoju jezikovnih tehnologij za slovenščino.

skladijsko označevanje, korpusi slovenskega jezika, Creative Commons

The paper introduces the first results of the JOS and SSJ projects from the area of syntax, comprising the framework for surface dependency annotation of Slovene texts and two annotated corpora. The corpora have been sampled from the Slovene reference corpus FidaPLUS and contain hand validated lemmas, morphosyntactic and surface-syntactic annotations. These resources will be made available as downloadable datasets under a Creative Commons licence, targeted primarily at language technology research for Slovene.

Syntactic annotation, Slovene corpora, Creative Commons

1 Uvod

Skladenjsko označeni korpusi predpostavljena skladijska razmerja eksplicirajo na velikem vzorcu besedil dejanske rabe in omogočajo statističen pregled vzorcev distribucije skladijskih struktur, zato so lahko izhodišče za nadaljnjo, bolj poglobljeno jezikoslovno obravnavo skladijskih fenomenov, hkrati pa so pomembni za razvoj jezikovnih tehnologij, kar je temeljnega pomena za ohranjanje konkurenčnosti ter polnofunkcionalnosti slovenščine med ostalimi jeziki.

Za slovenščino je bil do zdaj na voljo en površinskoskladijsko označen korpus: Slovenska odvisnostna drevesnica (SDT), ki obsega približno 2800 povedi oz. 45.000 besed (Erjavec, Ledinek 2006, Džeroski idr. 2006, Ledinek 2007). Zaradi skladijskih sorodnosti med češčino in slovenščino ter dostopnosti

izčrpnega priročnika za površinskoskladijsko označevanje češčine (Bémová idr. 1999) in inteligentnega označevalnika povedi (Hajič idr. 2001) je bil označevalni sistem slovenskega korpusa oblikovan po modelu korpusa Prague Dependency Treebank. Korpus SDT sestavlja ta dva besedilno homogena ter ročno oblikoskladijsko ter površinskoskladijsko označena podkorpusa: del vzorednega korpusa MULTEXT-East (Erjavec 2004), tj. del prevođa romana *1984* Georgea Orwella, ter del vzorednega korpusa SVEZ-IJS (Erjavec 2006). Tako s tehnološkega kot jezikoslovnega vidika je slabost korpusa njegova majhnost ter skromna besedilnotipska sestava, s čimer je korpus lahko le pomožni vir za razvoj in šolanje skladijskih razčlenjevalnikov ter jezikoslovne analize. Kot največjo slabost korpusa pa bi, glede na finančne, kadrovske in časovne

okolščine nastanka skladiščno označenih korpusov slovenščine, lahko opredelili izjemno kompleksnost označevalnega sistema korpusa.

Zaradi navedenih pomanjkljivosti smo se odločili oblikovati nove površinskoskladiščno označene korpusne slovenščine. Njihova izgradnja kot rezultat oblikovanja in implementacije premišljenega jezikovnospecifičnega označevalnega sistema se odvija pri projektu Jezikoslovno označevanje slovenščine (JOS), v okviru katerega sta pred tem že potekali revizija in nadgradnja nabora oznak za avtomatsko oblikoskladiščno označevanje slovenščine (Erjavec, Krek 2008, Arhar, Ledinek 2008), isti označevalni sistem pa bo uporabljen tudi pri izgradnji skladiščno označenih korpusov projekta Sporazumevanje v slovenskem jeziku (SSJ).

2 Korpus jos100k

Za površinskoskladiščno¹ označevanje smo vzeli korpus jos100k (Erjavec, Krek 2008), ki je enojezični vzorčeni in uravnoteženi korpus slovenskega jezika s 100.000 besedami in ročno označenimi oz. pregledanimi lemmami ter oblikoskladišnimi oznakami. Korpus sestavljajo vzorčeni odstavki iz 620-milijonskega korpusa FidaPLUS (Arhar, Gorjanc 2007), pri čemer so bili kriteriji za vzorčenje uravnoteženost, reprezentativnost in kvaliteta besedil. Oblikoskladišne oznake, ki se uporabljajo v korpusu, so nadgradnja oznak MULTEXT-East za slovenski jezik (te se uporabljajo v korpusu FidaPLUS). Podrobneje so opisane v Arhar, Ledinek (2008), v celoti pa dostopne na <http://nl.ijs.si/jos/josMSD-sl.html>. Izvirne oznake in leme v korpusu so bile prepisane iz korpusa FidaPLUS, nato pa revidirane in v več korakih ročno pregledane.

Korpus jos100k je zapisan v formatu XML, v skladu s smernicami konzorcija za označe-

vanje besedil TEI (2007), ter vsebuje bibliografske podatke o posameznih besedilih in njihovi umestitvi v taksonomijo besedil, oznake za odstavke, povedi in besede oz. ločila, za vsako besedo pa njeno ročno preverjeno oblikoskladiščno oznako in lemo. Korpus je na voljo preko spletnega konkordančnika, v celoti pa je za raziskovalne namene dostopen tudi kot zbirka podatkov po licenci Creative Commons.

3 Oblikovanje specifikacij za površinskoskladiščno označevanje slovenščine JOS

Oblikovanje označevalnega sistema je vodilo načelo, da je treba izhajati iz spoznanj o (tipoloških) značilnostih slovenščine in pripraviti jezikovnospecifičen označevalni sistem ter da je nujno oblikovati konsistenten ter dovolj robusten nabor oznak, pri čemer smo skušali upoštevati potrebe raziskovalne skupnosti (izraba virov tako za razvoj metod obdelave naravnega jezika in jezikovnih tehnologij kot tudi za jezikoslovne raziskave), predpostavljeno stopnjo avtomatske pripisljivosti oznak in stopnjo določljivosti mej potencialnih jezikoslovnih kategorij znotraj potencialnih širših kategorij.

V prvi fazi oblikovanja označevalnega sistema in nabora oznak JOS je potekala analiza prednosti in slabosti različnih površinskoskladiščnih označevalnih sistemov (Prague Dependency Treebank, Minipar, Link Grammar itn.), pri čemer so bile upoštevane zlasti namenskost gradnje korpusov in tipološke lastnosti jezikov, katerih korpusi so bili označeni z omenjenimi sistemi. Na podlagi te analize in interpretacije skladiščnih razmerij v nekaj sto vzorčnih povedih iz korpusa FidaPLUS je bila razvita izhodiščna različica specifikacij za odvisnostno površinskoskladiščno označevanje slovenščine.

¹ Z izrazom *površinskoskladiščno označevanje* poimenujemo funkcijskoskladiščno ter, deloma, strukturnoskladiščno analizo povedi v korpusu (nasproti globinsko- oz. pomenskoskladišni analizi). Izraz *površinskoskladiščno označevanje* se torej ne nanaša na kompleksnost označevalnega modela in ni ustreznik terminov *plitko oz. skeletno skladiščno označevanje* (*shallow parsing, skeletal parsing, chunking*).

Nato smo iz korpusa jos100k izbrali 500 povedi (jos500s), ki služijo kot testni korpus za označevanje. Ustreznost izhodiščnega označevalnega sistema je bila nato testirana pri ročnem označevanju korpusa jos500s s strani razvijalcev označevalnega sistema, pri čemer so bile specifikacije revidirane, vsaka poved korpusa pa večkrat pregledana. Rezultat opisanega dela je nabor površinskoskladenjskih oznak JOS, različica 1.0 priročnika za označevanje in majhen, a natančno označen učni korpus zlati standard jos500s.

Nastali robustni trinivojski sistem odvisnostnega površinskoskladenjskega označevanja slovenščine predvideva 10 analitičnih oznak in je namenjen raziskovanju zlasti jedrnih skladenjskih pojavov. Površinskoskladenjska oznaka je pripisana vsaki (besedni) pojavnici v povedi, poseben element v označevalni strukturi pa je t. i. metaelement oz. korenski element povedi, na katerega so povezani z vidika skladenjske zgradbe hierarhično najvišji elementi stavkov povedi (navadno jedrni glagol povedka glavnega stavka) ter skladenjsko manj predvidljive strukture, ki bi sicer ostale nepovezane (npr. ob elipsi). Pri vzpostavitvi predpostavljenih kategorij smo se opirali na spoznanja (slovenskega) jezikoslovja, vendar smo pri analizi možnosti prenosa v jezikoslovju že uveljavljenih kategorij v označevalni sistem upoštevali zglede dejanske jezikovne rabe. Hkrati je že sama avtomatska analiza jezika (težnja po robustnosti označevalnega sistema)

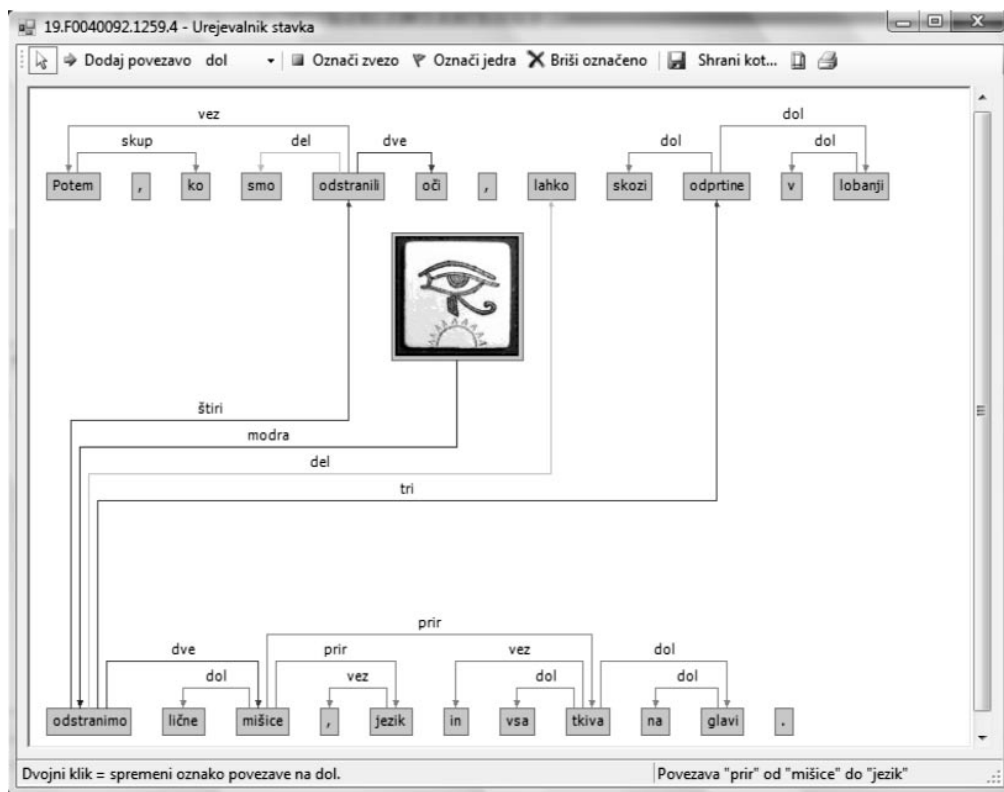
zahtevala, da se nabor oznak in sistem njihovega pripisovanja skladenjskim strukturam od v jezikoslovju uveljavljenih kategorij in interpretacij nekoliko odmakneta, na kar kažejo tudi imena analitičnih oznak, ki se od uveljavljenih jezikoslovnih terminov zavestno razlikujejo (Tabela 1).

Oznake združujemo v tri skupine glede na strukturoskladenjski nivo struktur, ki jih z oznakami določamo, in predvidljivost skladenjskih razmerij, ki jih oznake opredeljujejo. Oznake prvega nivoja so večinoma namenjene označevanju znotrajbesednozveznih skladenjskih razmerij, oznake drugega nivoja so navadno pripisane jedrnim elementom struktur, ki jim v jezikoslovju v glavnem pripisujemo stavčnočlensko vlogo, oznaka tretjega nivoja pa se uporablja za označevanje zlasti nadstavčnih, skladenjsko manj predvidljivih (priredno povezani stavki, pastavki, vrivki, členki, eliptične enote) ter medsebojno zelo oddaljenih struktur. Strukture, katerih jedrni elementi so označeni z oznako tretjega nivoja, so povezane na metaelement.

Vtis o rabi nabora površinskoskladenjskih oznak je mogoče dobiti ob ogledu ene od označenih povedi iz zlatega standarda jos500s (Slika 1), natančnejše podatke o načinu pripisovanja oznak specifičnim skladenjskim strukturam pa nudi priročnik za označevanje, ki ga v smislu specifikacij za površinskoskladenjsko označevanje slovenščine JOS smiselno dopolnjuje zlati standard jos500s.

Nivo označevanja	Analitične oznake
Prvi nivo	DOL; DEL; VEZ; PRIR; SKUP
Drugi nivo	ENA; DVE; TRI; ŠTIRI
Tretji nivo	MODRA

Tabela 1: Nabor površinskoskladenjskih oznak JOS.



Slika 1: Primer označene povedi iz korpusa jos500s, kot jo vidimo v programu za ročno označevanje povedi.

Označevalni sistem je rezultat preudarnega jezikoslovnega razmisleka ob označevanju avtentičnih primerov jezikovne rabe in upoštevanju namenskosti označenih korpusov ter predvidene frekvence oz. razpršenosti analitičnih oznak v korpusu. Upoštevaty skuša naslednje smernice:

1. Oblikoskladenjske oznake dajejo razmeroma dobro informacijo o potencialni skladenjski strukturi povedi, zato je ob interpretaciji jezikovnih pojavov predvideno njihovo kombiniranje s površinskoskladenjskimi oznakami.

2. Robusten označevalni sistem omogoča identifikacijo zlasti jedrnih znotrajstavčnih struktur (t. i. chunkov).

3. Podrobnejša analiza zelo oddaljenih ter skladenjsko manj predvidljivih struktur glede na trenutne zmožnosti avtomatske analize jezika (še) ni smiselna.

4 Postopek označevanja

Za potrebe ročnega označevanja je bil razvit program za označevanje in vizualizacijo povedi,² ki ponuja grafični vmesnik, s katerim lahko pregledujemo površinskoskladenjske oznake povedi ter jih ročno popravljamo oz. dopolnjujemo. S pomočjo tega urejevalnika, ob hkratnem pisanju priročnika za označevanje, je bil najprej označen korpus jos500s, ki je v nadaljevanju služil kot učni korpus za avtomatsko

2 Avtor programa je Janez Brank, primer slike zaslona pa je podan na Sliki 1.

skladenjsko označevanje, na podlagi katerega je bil skladdenjsko predoznačen korpus jos100k, kar je bistveno pohitrilo postopek ročnega označevanja. S povečevanjem števila ročno označenih povedi se večja tudi učni korpus, ki ga je mogoče uporabiti pri avtomatskem označevanju korpusa jos100k, zato bo postopek predoznačitve tega korpusa na določen časovni interval ponovljen.

Trenutno poteka ročno pregledovanje avtomatsko pripisanih oznak v korpusu jos100k. Vsak segment besedila pregledujeta najmanj dva označevalca, kjer pri njunem označevanju prihaja do razlik, oznake preveri še tretji označevalec. Z obsežnejšim korpusom se povečuje tudi nabor skladdenjskih struktur, ki jih je treba sistematično označiti. Vzporedno z ročnim označevanjem zato poteka tudi nadgradnja specifikacij za površinskoskladdenjsko označevanje, ki je deloma tudi rezultat evalvacije, ki sledi analizi skladnosti ročnega označevanja pri označevalcih ter analizi razlik med ročnim in avtomatskim označevanjem. Pri analizi razlik nam je v pomoč program Whatswrong, ki prikaže razlike med dvema označitvama povedi istega korpusa in tako grafično izpostavi razlike v označevanju med označevalcema.

5 Zaključek

Korpusi JOS s specifikacijami so standardizirani in po licenci Creative Commons za raziskovalne namene prosto dostopni jezikovni viri za slovenščino; oblikoskladdenjsko označena korpusa sta že dostopna na spletni strani projekta JOS, v nadaljevanju pa bodo na enak način dostopni tudi skladdenjsko in pomensko označeni korpusi projekta. Površinskoskladdenjsko označena korpusa JOS bosta v nadaljnjih fazah dela vključena tudi v učni korpus projekta SSJ ssj400k. Ta korpus bo dovolj velik, da bo omogočil razvoj robustnega skladdenjskega razčlenjevalnika za slovenščino, ki bo nadalje tudi služil za izgradnjo površinskoskladdenjsko avtomatsko označena korpusa ssj100M. Na osnovi skladdenjsko

označenih virov SSJ je predvideno tudi oblikovanje novih jezikovnih priročnikov za slovenščino.

Literatura

- ARHAR, Špela, GORJANC, Vojko, 2007: Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. *Jezik in slovstvo* 52/2. 95–110.
- ARHAR, Špela, LEDINEK, Nina, 2008: Oblikoskladdenjske oznake JOS: revizija in nadgradnja nabora oznak za avtomatsko oblikoskladdenjsko označevanje slovenščine. Erjavec, Tomaž, Žganec Gros, Jerneja (ur.): *Zbornik Šeste konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. 49–53.
- BĚMOVÁ, Alla idr., 1999: *Annotations at Analytical Level: Instructions for Annotators*. Praga: UK MFF UFAL.
- DŽEROSKI, Sašo idr., 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, 2004: MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC'04*. Pariz: ELRA. 1535–1538.
- 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, KREK, Simon, 2008: Oblikoskladdenjske specifikacije in označeni korpusi JOS. Erjavec, Tomaž, Žganec Gros, Jerneja (ur.): *Zbornik Šeste konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. 49–53.
- ERJAVEC, Tomaž, LEDINEK, Nina, 2006: Slovenska odvisnostna drevesnica: prvi rezultati. Erjavec, Tomaž, Žganec Gros, Jerneja (ur.): *Zbornik 5. slovenske in 1. mednarodne konference Jezikovne tehnologije IS-LTC*. Ljubljana: Institut Jožef Stefan. 162–167.
- FidaPLUS: <<http://www.fidaplus.net/>>. (Dostop 7. 7. 2009.)
- HAJIČ, Jan idr., 2001: The Prague Dependency Treebank: Annotation Structure and Support. *Proceedings of the IRCS Workshop on Linguistic Databases*. 105–114.
- JOS, *Jezikoslovno označevanje slovenščine*: <<http://nl.ijs.si/jos/>>. (Dostop 7. 7. 2009.)
- LEDINEK, Nina, 2007: Slovenska odvisnostna drevesnica v raziskavah o induktivnem odvisnostnem označevanju. *Jezik in slovstvo* 52/1. 3–16.
- Link Grammar: <<http://www.link.cs.cmu.edu/link/>>. (Dostop 7. 7. 2009.)

- Minipar*: <<http://www.cs.ualberta.ca/~lindek/minipar.htm>>. (Dostop 7. 7. 2009.)
- Oblikoskladenjske oznake JOS*: <<http://nl.ijs.si/jos/josMSD-sl.html>>. (Dostop 7. 7. 2009.)
- PDT, Prague Dependency Treebank*: <<http://ufal.mff.cuni.cz/pdt2.0/>>. (Dostop 7. 7. 2009.)
- Priročnik za površinskoskladenjsko označevanje korpusov JOS*: <http://www.slovenscina.eu/Media/Kazalniki/Kazalnik2/SSJ_Kazalnik_2_Specifikacije-ucni-korpus_v1.pdf>. (Dostop 27. 9. 2009.)
- SDT, Slovenska odvisnostna drevesnica*: <<http://nl.ijs.si/sdt/>>. (Dostop 7. 7. 2009.)
- SSJ, Sporazumevanje v slovenskem jeziku*: <<http://www.slovenscina.eu/>>. (Dostop 7. 7. 2009.)
- TEI Consortium, 2007: *TEI P5: Guidelines for Electronic Text Encoding and Interchange*.
- TEI Consortium*: <<http://www.tei-c.org/>>. (Dostop 7. 7. 2009.)
- Whatswrong*: <<http://code.google.com/p/whatswrong/>>. (Dostop 7. 7. 2009.)