UDK 811.163.6:004
Birte Lönneker
Institut für Germanistik II, Hamburg

STROJNO OBLIKOSLOVNO OZNAČEVANJE SLOVENSKIH BESEDIL: KAKO DALEČ SMO

Članek obravnava oblikoslovno označevanje in lematizacijo slovenskih besedil. Prvo poglavje razlaga izvedbo teh postopkov. Drugo poglavje predstavi rezultate poskusov strojnega označevanja slovenskih besedil z uporabo milijonskega že označenega učnega korpusa. Za slovenščino prilagojen strojni označevalnik TreeTagger je dosegel točnost okoli 85 % in označil ter lematiziral 100 milijonov besed slovenskega korpusa Nova Beseda.

The article deals with part-of-speech tagging and lemmatization of Slovene texts. The first section explains how these procedures are performed. The second section presents results of experiments in automated tagging of Slovene texts, using a pre-tagged training corpus of one million words. The TreeTagger, a statistical tagger, was trained for Slovene and achieved a precision of about 85%. It tagged and lemmatized 100 million running words of the Slovene corpus *Nova beseda*.

Ključne besede: korpus, oblikoslovno označevanje, ocenjevanje označevanja, TreeTagger, Nova beseda, lematizacija
Key words: corpus, part-of-speech tagging, evaluation of tagging, TreeTagger, Nova beseda, lemmatization

1 Uvod

Izvedba številnih jezikoslovnih in slovaropisnih nalog danes ni več predstavljiva brez digitaliziranih besedilnih zbirk oziroma korpusov. Namen tega članka je pregled trenutnih možnosti jezikoslovnega obdelovanja slovenskih korpusov in dejanskega stanja na tem področju; še posebej članek pri tem obravnava oblikoslovno označevanje in lematizacijo. Razprava je ločena v dva glavna dela: 2. poglavje razlaga osnove in postopke izvedbe oblikoslovnega označevanja in lematizacije, 3. poglavje pa predstavi rezultate strojnega označevanja slovenskih besedil z uporabo milijonskega že označenega korpusa. V 4. poglavju so podani zaključki.

Za slovenski jezik obstajajo korpusi različnih vrst in obsegov. Vrsta korpusa se lahko opisuje po več kriterijih, npr. po besedilni vrsti ali po nivoju jezikoslovne obdelave. Različni nivoji obsegajo npr. ročno popravljanje napak po zajemanju, označevanje mej poglavij in povedi ter oblikoslovno označevanje in lematizacijo.

Najobsežnejša slovenska korpusa sta *Nova beseda* Inštituta za slovenski jezik Frana Ramovša ZRC SAZU s trenutno 148 milijoni besed[1] in FIDA, izdelek skupnega projekta Filozofske fakultute, Instituta Jožef Stefan, založbe DZS in podjetja Amebis, v obsegu 100 milijonov besed.[2] Novo besedo lahko po Brew/Moens (1999: 72) opišemo kot »monitorski korpus«, saj je vsako leto posodobljena z novimi besedili. FIDA je po

[1] http://bos.zrc-sazu.si/s_beseda.html [6. januar 2005]
[2] http://www.fida.net/slo/ [6. januar 2005]

Brew/Moens »referenčni korpus«, saj sta nastanek in sestava tega korpusa natančno opisana in se v določeni, dosegljivi različici korpusa ne spreminjata več. Med drugimi zanimivimi korpusi naj tukaj omenimo le še vzporedna slovensko-angleška korpusa IJS-ELAN[3] (1 milijon besed; Erjavec (2002a, 2002b)) in EVROKORPUS[4] (16 milijonov besed, december 2004; Željko (2002)).

2 Oblikoslovno označevanje in lematizacija

Postopek oblikoslovnega označevanja pripiše oblikoslovne oznake vsem besednim oblikam v besedilni zbirki. V izrazu »seminar slovenskega jezika« bi lahko npr. pripisali besedni obliki »slovenskega« oznako P za pridevnik in besedni obliki »jezika« oznako S za samostalnik. V naslednjih podpoglavjih sta natančneje opisana izbira oznak (2.1) in format pripisovanja oznak besedilu (2.2). V podpoglavju 2.3 so predstavljene najpomembnejše informacije o lematizaciji. Načini izvajanja obeh postopkov so opisani v podpoglavju 2.4, medtem ko podpoglavje 2.5 prikaže načine ocenjevanja rezultatov označevanja.

2.1 Izbira oblikoslovnih oznak

Oblikoslovne oznake lahko vsebujejo poleg kategorije besed (pridevnik, samostalnik, glagol itd.) še natančnejše podatke o obliki, kot so podkategorije (npr. za zaimek: kazalni, osebni, oziralni itd.) in skladenjske atribute (npr. za samostalnik: sklon, spol in število). Zato so pri jezikoslovno bogatejših jezikih oznake pogosto sestavljene iz dveh delov, in sicer iz kategorije/podkategorije besed – prikazane ponavadi z velikimi črkami – in skladenjskih atributov, ponavadi prikazanih z malimi črkami ali s številkami. Primer je oznaka `Sme2` za »**s**amostalnik **m**oškega spola **e**dnine v drugem (**2.**) sklonu«.

Pri izbiri seznama uporabljenih oznak je koristno upoštevati mednarodno prenosljivost z uporabo standardiziranega nabora oznak za več jezikov, upoštevati pa je treba tudi jezikoslovnoopisno tradicijo obravnavanega jezika. Kateri vidik najbolj vpliva na izbiro uporabljenega nabora oznak, je seveda odvisno predvsem od namena korpusa (Jakopin/Bizjak 1997:514–517; Erjavec 2003: 74).

Pri slovenskih virih so danes v uporabi različni sistemi oznak. Prvi nabor oblikoslovnih oznak za slovenščino sta verjetno opisala Jakopin/Bizjak (1997); njune oznake slonijo na slovenski slovnici Jožeta Toporišiča in uporabljajo kratice slovenskih slovničnih izrazov, kot je prikazano v zgornjem primeru. Jakopin/Bizjak (1997: 523) sta izračunala, da obsega njun sistem 4.797 različnih oznak. Posodobljena različica tega nabora oznak je še vedno v uporabi na Inštitutu za slovenski jezik Frana Ramovša. Oznake korpusov FIDA (Erjavec 1998) in IJS-ELAN so v formatu MULTEXT-East, ki se uporablja za več evropskih jezikov. Kratice za besedne vrste in oblikoslovne atribute v oznakah MULTEXT-East slonijo na angleško-latinskih slovničnih izrazih; npr. oznaka Ncnsg (**N**oun, **c**ommon, **n**euter, **s**ingular, **g**enitive) je ena izmed možnih

[3] http://nl.ijs.si/elan/ [6. januar 2005]
[4] http://www.gov.si/evrokor/ [6. januar 2005]

oznak za besedno obliko *ministrstva*. Za slovenščino nudi MULTEXT-East 2.083 različnih oblikoslovnih oznak (Džeroski idr. 2000: 1101). Tudi oznake slovenskih jezikovnih virov evropskega projekta LC-STAR (Verdonik/Rojc 2004) so angleške. V LC-STAR-u je na primer ena izmed možnih oznak za besedno obliko *agenciji* naslednja: `NOM class = »common« number = »dual« gender = »feminine« case = »nominative«.`

Oznake lahko tudi prevajamo oz. lokaliziramo (Džeroski idr. 2000: 1100). Natančno ločevanje in dokumentacija opisanih oblikoslovnih kategorij in atributov sta zato pomembnejši kot jezik, v katerem z okrajšavami opišemo oblikoslovne podatke. Snovalci naborov oznak se morajo vprašati, do katere jezikoslovne natančnosti želijo ločevati jezikoslovne pojave. Tako bi se morali pri slovenščini odločiti, katere vrste in koliko zaimkov bi ločili, katere atribute bi ločili (npr. živost pri samostalnikih, poudarjenost pri osebnih zaimkih), ali bi ločili lastna imena od drugih samostalnikov, ali bi zanje določili podkategorije itd. Razvrščanje lastnih imen v podkategorije (npr. osebna, veroslovna, živalska, stvarna imena itd.) pravzaprav ne spada k oblikoslovnemu označevanju, temveč h kasnejšemu, semantičnemu postopku obdelovanja, ki se imenuje identificiranje informacij (angl. *Information Extraction*). Čim več imenskih podkategorij imamo, tem težje je označevanje: sta npr. imeni *Sodoma* in *Gomora* zemljepisni imeni (`IZ`), veroslovni imeni (`IV`) ali mitološki imeni (`IM`)?

2.2 Format pripisa oblikoslovne oznake besedilu

Druga značilnost označenega korpusa je uporabljen format za pripis oznak besedni obliki. Na področju računalniškega jezikoslovja sta v ta namen najbolj razširjena vertikalni format in jezik XML. V vertikalnem formatu je vsaka besedna oblika v svoji vrstici, skupaj s svojo oznako in morebitnimi drugimi podatki, od katerih je ločena s tabulatorjem (gl. sliko 1). V formatu XML pripada vsaki besedni obliki ustrezna oznaka kot npr. `<w>` (za angl. *word* 'beseda') z atributom kot npr. ana (pri projektu MULTEXT-East), katerega vrednost je oblikoslovna oznaka (gl. sliko 2). Vertikalni format je primeren za uporabo v programih za procesiranje korpusov in poizvedovanje v njih, kot sta npr. *Manatee/Bonito* (Masarykova Univerza, Brno) in klasični CWB (*Stuttgart IMS Corpus Workbench*). Tudi za procesiranje korpusov v formatu XML obstaja več programov; eden izmed njih je sistem CLaRK (Univerza v Tübingenu/ Bolgarska akademija znanosti).

```
seminar      Sme1
slovenskega  Pme2
jezika       Sme2
```

Slika 1: Del oblikoslovno označenega besedila v vertikalnem formatu

```
<w ana="Sme1">seminar</w> <w ana="Pme2">slovenskega</w> <w ana="Sme2">jezika</w>
```

Slika 2: Del oblikoslovno označenega besedila v formatu XML

2.3 Lematizacija

S postopkom lematizacije pripisujemo besednim oblikam še osnovno obliko oz. lemo; npr. v izrazu *seminar slovenskega jezika* bi pripisali besedni obliki *slovenskega* lemo *slovenski* in besedni obliki *jezika* lemo *jezik*. Korpus je lematiziran, če vsebuje za vsako besedno obliko svojo lemo. Sliki 3 in 4 prikazujeta dela besedil oblikoslovno označenih in lematiziranih korpusov v dveh formatih.

```
seminar      Sme1 seminar
slovenskega  Pme2 slovenski
jezika       Sme2 jezik
```

Slika 3: Del oblikoslovno označenega in lematiziranega besedila v vertikalnem formatu

```
<w lemma="seminar" ana="Sme1">seminar</w> <w lemma="slovenski"
ana="Pme2">slovenskega</w> <w lemma="jezik" ana="Sme2">jezika</w>
```

Slika 4: Del oblikoslovno označenega in lematiziranega besedila v formatu XML

V lematiziranem korpusu lahko uporabnik poišče vse besedne oblike za iskano lemo in lahko zanjo izračuna tudi frekvence in kolokacije, namesto da bi te podatke izračunal za vsako besedno obliko posebej. Lematiziran korpus je zato še bolj uporaben pri leksikološkem, jezikoslovnem in jezikovnotehnološkem delu.

2.4 Načini izvajanja označevanja in lematizacije

Denimo, da bi imeli v digitalni obliki popoln slovar vseh slovenskih besednih oblik skupaj z ustreznimi oblikoslovnimi oznakami, torej da bi bile vse besedne oblike »znane«. Tudi v tem hipotetičnem primeru ne bi bilo mogoče nedvoumno označiti besedila zaradi večpomenskosti besednih oblik. Večpomenska besedna oblika v slovenščini je npr. beseda *prej*. Ta je lahko prislov (z oznako A) ali pa oblika ženskega samostalnika *preja* v drugem sklonu dvojine ali množine (z oznako Sžd2 ali Sžm2). Katera je prava izmed teh treh možnosti označevanja, je razvidno le iz sobesedila, kot kažeta primera v stavkih 1 a), kjer je *prej* prislov, in 1 b), kjer je oblika samostalnika.

1 a) To je že prej vedel.
1 b) To je stroj za izdelavo finih prej.

Večpomenskost oz. dvoumnost besedne oblike *prej* je pri označevanju treba razločevati oz. razdvoumiti. Ročno izvajanje je ponavadi najnatančnejši način označevanja, a hkrati dolgotrajna naloga, na dolgi rok tako dolgočasna kot tudi naporna, in za naročnika zelo draga. Prvi uspešni poskusi strojnega oblikoslovnega označevanja so se začeli izvajati po letu 1980, najprej za angleščino. Danes ločimo glede na način izvajanja tri vrste odtlej nastalih programov za strojno označevanje, ti. označevalnikov:

1. uporaba ročno sestavljenih jezikoslovnih pravil (angl. *rule based approach*);
2. strojno učenje pravil iz že označenih besedil (angl. *machine learning approach*);
3. združitev obeh zgoraj omenjenih načinov.

Popolnoma brez ročnega dela torej še ni mogoče priti do oblikoslovno označenega besedila, saj zahteva tudi strojno učenje čimbolj pravilno označeni učni korpus. Pri učenju verjetnosti pojavljanja določenega oblikoslovnega vzorca (oz. modela jezika) strojni označevalniki upoštevajo različne značilnosti učnega korpusa. Ponavadi ne izračunajo le pogostosti kombinacije oblik in njihovih oblikoslovnih oznak (npr. *prej* z oznako `A`, `Sžd2` ali `Sžm2`), ampak izračunajo tudi pogostosti zaporedij oznak v stavkih. Denimo, da je označevalnik že označil stavek 1 a) do vključno besede *že* in poskuša označiti besedo *prej*. Med drugim bi lahko za vsako morebitno oznako besedne oblike *prej* (`A`, `Sžd2`, `Sžm2`) iz učnih podatkov izračunal statistično verjetnost, da taka oznaka neposredno sledi oznakama `GPce` in `A`, ki označujeta besedi *je* in *že*. V stavku 1 b) bi na podoben način izračunal verjetnost, da taka oznaka sledi zaporedju oznak `Sže4 Pžm2`, ki označujeta besedi *izdelavo* in *finih*.

Katere značilnosti učnega korpusa se upoštevajo in kako so utežene, je odvisno od vsakega posameznega označevalnika. Pomembno pa je razumeti, da s številom oznak v uporabljenem naboru raste tudi število možnih in dejanskih zaporedij oznak v besedilu. Zlato pravilo za učna gradiva pravi, naj ta vsebujejo vsak učni pojav vsaj desetkrat (Weischedel idr. 1993: 363–364). Če označevalnik upošteva trojčke oznak, to pomeni, naj vsebuje učni korpus vsako različno zaporedje iz treh oznak najmanj desetkrat.

Iz tega je jasno, da mora biti učni korpus za oblikoslovno bogato slovenščino veliko večji kot za angleščino, pri kateri so Weischedel idr. (1993) zaradi manjšega pregibanja uporabili samo 47 oznak. Odkrili so, da se večina trojčkov teh oznak dovolj pogosto pojavi že v korpusu z le 64.000 besedami. Ker so pa nabori oznak za slovenščino nekajkrat večji od angleških (gl. 2.1), se močno poveča tudi število možnih kombinacij treh oznak. Slovenščina ima v primerjavi z angleščino tudi manj restriktiven besedni red, kar spet omogoča več dejanskih kombinacij zaporedij oznak. Največja ovira za strojno oblikoslovno označevanje slovenskih besedil je torej prav pomanjkanje velikega, dostopnega učnega korpusa (Erjavec 2003: 71). Doslej je bil v ta namen uporabljen označeni del korpusa MULTEXT-East,[5] ki je javno dostopen v raziskovalne namene, toda majhen in po besedilni vrsti omejen: Vsebuje samo prevod romana *1984*, tj. 100.000 besed (Erjavec 2003: 71).

Pri strojnih označevalnikih ločimo še robustne in nerobustne. Robustni program se ne ustavi pri nepričakovanih vhodnih podatkih (Menzel 1995); robustni označevalnik torej pripiše oznako tudi neznanim besedam in celo »slovnično napačnim« izrazom (Cutting idr. 1992: 133). Nerobustni označevalnik pa nasprotno v takih primerih konča z napako ali počaka, da uporabnik ročno vnese manjkajoče oznake. Vsi štirje označevalniki, ki so jih Džeroski idr. (2000) poskusili in ocenili za slovenščino, so bili robustni. Jakopin (2002: 35–45) opisuje dva nerobustna označevalnika za slovenščino, integrirana v urejevalniku Eva. V interaktivni obliki po potrebi povabita uporabnika, naj ročno dopolni oznake; uporabnik pa ima tudi možnost popravljanja napačno določenih oznak (Jakopin 2002: 38–44; Srdanović-Erjavec 2004).

Z oblikoslovnim označevanjem je pogosto vezana še lematizacija. Digitalni slovar, t. i. lematizacijski slovar, bi pri tem za vsako besedno obliko razen ustreznih oblikoslo-

[5] http://nl.ijs.si/ME/ [6. januar 2005]

vnih oznak vseboval še lemo. Pri označevanju je s tem mogoče pripisati vsaki slovarju znani obliki še osnovno obliko, tako da dobimo kot rezultat ne le označen, ampak tudi lematiziran korpus.

2.5. Preverjanje in ocenjevanje

Zadnje področje tega poglavja so postopki preverjanja in ocenjevanja rezultatov ročnega in strojnega označevanja.

Merilo ročnega označevanja se ponavadi imenuje skladnost (angl. *agreement*), ki se lahko določi za delo bodisi enega označevalca ali več označevalcev. Če ista oseba označuje ali preverja ob različnih trenutkih isti pojav (jezikovni izraz ali oznako), lahko za ta pojav določimo notranjo skladnost (angl. *intra-annotator agreement*); če pa več označevalcev označuje ali preverja isti pojav, lahko zanj določimo medoznačevalsko skladnost (angl. *inter-annotator agreement*). Raziskave o skladnosti pri oblikoslovnem označenju (npr. Voutilainen 1999, Brants 2000a) kažejo, da se medoznačevalska skladnost približa 100 % le v primeru, da je nabor oznak jasno in natančno dokumentiran, da so označevalci izkušeni in se o označevanju posebnih primerov sporazumejo. Brants (2000a) poroča, da diskusije med označevalci in nadaljnji popravki izboljšajo medoznačevalsko skladnost za nemščino z uporabo nabora STTS (54 oznak) z 98,57 % na 98,80 %.

O kakovosti ročnega označevanja za slovenščino obstaja malo podatkov. Kot je razvidno iz Lönneker/Jakopin (2004), je pri ročnem preverjanju in popravljanju rezultatov interaktivnega označevalnika najbolj moteča odsotnost dokumentacije nabora oznak, ki bi natančno predpisala dovoljene oznake ter opisala in ločila problematične primere. V izvirnem opisu nabora (Jakopin/Bizjak 1997) ni bilo natančno opredeljeno, v katerih primerih naj bo ročno označevanje natančnejše ali manj natančno, kot je razvidno iz značilnosti izoliranih besed. Ko označevalec doda neko informacijo v oblikoslovno oznako, ki je označevana besedna oblika ne kaže, pride do preoznačevanja (angl. *overspecification*). To se je zgodilo pri oznaki besede *tri* in je prikazano v primerih 2 a) in 2 b), prevzetih iz Lönneker/Jakopin (2004: 54), kjer ŠG označuje besedno vrsto »glavni številnik« (Jakopin/Bizjak 1997: 520).

2 a) Ura odbije tri<pos>ŠGžp4</pos>.

2 b) [...] kjer je delala za tri<pos>ŠGmp4</pos>.

Izolirana beseda *tri* ne kaže informacije o spolu; vendar je ta značilnost v primerih označena v atributih kot ž(enski) ali m(oški). Tudi če pogosto lahko ugotovimo ustrezen spol besede *tri* iz sobesedila, je to razpoznavanje v nekaterih primerih nemogoče (ali pa le s težavami), kot v primerih 2 a) in 2 b). Obratno pride lahko prav tako tudi do podoznačevanja (angl. *underspecification*), kjer označevalec izpusti neko informacijo iz oblikoslovne oznake, ki pa jo označevana besedna oblika kaže. Brez podanih navodil se označevalec pri označevanju besede *tri* torej lahko odloči za oznako ŠGmp4, ŠGp4 (brez spola), ŠG4 (brez spola in števila) ali celo le za oznako ŠG. Take različne odločitve imajo dve posledici:

1. seznam uporabljenih oznak raste, ker hkrati vsebuje oznake različne natančnosti;
2. skladnost se zmanjša.

Ocenjevanje rezultatov strojnega označevanja se ponavadi izvede tako, da ročno označene ali preverjene podatke pred poskusom razdelimo v učni in testni korpus. Točnost označevalnika je stopnja ujemanja oznak, ki jih strojni označevalnik določi za testni korpus, s tistimi, ki so mu jih pripisovali označevalci. Testni korpus je lahko bistveno manjši kot učni korpus, paziti pa moramo na to, da se v njem besedila iz učnega korpusa ne ponavljajo, saj bi s tem dobili zavajajoče visoko stopnjo ujemanja oznak. Tudi če se celotna besedila ne ponovijo, imata učni in testni korpus lahko več ali manj ujemajočih se lastnosti, tj. znane oz. neznane elemente kot npr. lastna imena ali zaporedja oznak. Če imamo npr. pet označenih člankov in jih razdelimo v učni in testni korpus tako, da vsebuje testni korpus petino vsakega članka, bodo rezultati verjetno boljši kot v primeru, da bi kot testni korpus uporabili le enega od člankov v celoti. Zato se Brants (2000b: 227) opredeli za uporabo celotnih, pri učenju še ne »videnih« besedil (npr. člankov) kot bolj realističen testni korpus.

Ocenjevanje strojnega označevanja je bolj razvito in pogostejše uporabljeno kot preverjanje ročnega označevanja. Džeroski idr. (2000) so opisali učenje štirih jezikovno neodvisnih označevalnikov za slovenščino. Pri tem so najprej izmerili obseg korpusa v številu samostojnih besedilnih enot, kot so besede, številke, večbesedne enote, kratice in stavčna ločila. Te enote se imenujejo tokeni (angl. *tokens)*. Učni korpus vsebuje 81.805 tokenov (izmed njih 12.980 stavčnih ločil) in testni korpus 10.594 tokenov (izmed njih 1.647 stavčnih ločil). Oba izhajata iz označenega dela korpusa MULTEXT-East (gl. 2.4). Točnost označevalnikov, ki so jih naučili s celotnim naborom oznak MULTEXT-East, ocenijo za vse tokene testnega korpusa, torej vključno s stavčnimi ločili, za katere je točnost vedno 100 %. Za točnost pri vseh tokenih dobijo rezultate od 85,95 % do 89,22 %. Brez upoštevanja 1.647 ločil bi točnost najuspešnejšega označevalnika, TnT (Brants 2000b), bila 87,18 %. Za angleški jezik je leta 1993 z omenjenim manjšim učnim korpusom (gl. 2.4) in naborom 47 oznak bila točnost označevanja slovarju znanih besednih oblik 96,13 % (Weischedel idr. 1993).

3 Strojno označevanje slovenskih besedil z uporabo *POSBesede*

Od leta 1996 (Jakopin/Bizjak 1997: 514) je bilo na Inštitutu slovenskega jezika z Jakopinovimi označevalniki (gl. 2.4) označenih ter ročno dopolnjenih in popravljenih več kot milijon in pol besed. Leta 2004 je bila prvič ocenjena primernost takrat sodobne različice označenega gradiva z imenom POSBeseda/2003 kot učni korpus za robustni označevalnik (Lönneker/Jakopin 2004). V tem poglavju[6] bo predstavljeno, kako so bila POSBeseda in druga gradiva Inštituta slovenskega jezika pretvorjena, dopolnjena in uporabljena za učenje strojnega označevalnika. Učni in testni korpus obravnavamo v podpoglavju 3.1, nato označitev besed in stavkov (3.2), tujejezično gradivo v korpusih (3.3) ter označevalni slovar (3.4). Temu sledijo v podpoglavju 3.5 predstavljeni rezultati različnih poskusov učenja robustnega označevalnika.

[6] Vsebina tega poglavja je razširjen povzetek treh poročil, ki jih je avtorica napisala o svojem delu v Laboratoriju za korpus slovenskega jezika Inštituta za slovenski jezik Frana Ramovša ZRC SAZU med oktobrom in decembrom 2004.

3.1 Učni in testni korpus

Učni korpus POSBeseda/2003 vsebuje naslednje gradivo: slovenske prevode romanov *1984* (George Orwell) in *Bouvard in Pécuchet* (Gustave Flaubert), Platonove *Države* in *Nove zaveze*; kot izvirno slovensko leposlovje zbrana dela Cirila Kosmača[7] in roman Toma Križnarja *O iskanju ljubezni*; ter še nekaj več kot 50.000 besed iz slovenskega dnevnika *Delo* iz leta 1997. Prispevek vsakega romana, tudi Države, je nekaj nad 100.000 besed, le Križnar jih prispeva okoli 130.000. Nova zaveza prispeva nekaj manj kot 150.000 označenih besed, zbrana dela Kosmača pa so z več kot 400.000 besedami največja sestavina. Testni korpus je delo Marka Uršiča *Štirje časi – Pomlad*, *Filozofski pogovori in samogovori* in vsebuje malo več kot 170.000 besed. Vsi podatki so brez upoštevanja stavčnih ločil.

Oba korpusa sta bila označena z enim izmed Jakopinovih označevalnikov. Interaktivno popravljanje so opravili različni označevalci, predvsem Aleksandra Bizjak (samo učni korpus, in sicer v prvi, doslej najobsežnejši fazi) in Lučka Uršič (manjši del učnega korpusa ter testni korpus). Vsak del korpusov je samostojno popravil samo en označevalec.

Interaktivno ročno popravljanje oznak je potekalo v urejevalniku besedil Eva v dveh vrsticah (Jakopin 2002), kot prikazuje slika 5. Pred pretvorbo v druge formate sta bili obe vrstici v istem urejevalniku združeni v eno vrstico tako, da je vsaki besedi sledila ustrezna oznaka. Pri posebnih znakih takšen prepis žal ni vedno uspešen, saj razdeli oznake in iz njih izreže črko, kot kaže primer 3), ki predstavlja nepravilno preoblikovano besedilo s slike 5.

3) [...] v d<pos>P</pos>'Orbignyjevem<pos>Ome5</pos> * priročniku.

Nepopolno delovanje urejevalnika je bilo opaženo šele po združevanju vrstic. Žal do sedaj tega še ni bilo mogoče spremeniti. Zato so bile napačne razdelitve v učnem in v testnem gradivu po možnosti popravljene ročno ali s pomočjo avtomatične nadomestitve. Tako popravljen učni korpus vsebuje 1.268.973 oznak z upoštevanjem ločil oziroma 1.031.658 oznak brez ločil; testni korpus pa 212.950 (171.598 brez ločil). Učni in testni podatki so bili potem prepisani v format XML in v vertikalni format (gl. 2.2).

```
Poiskala sta_ jo____ v_ d'Orbignyjevem * priročniku.
 GLmd     GPcd ZOcže4 E5 PIOme5              Sme5
```

Slika 5: Izvirnik označenega primera 3) v urejevalniku Eva.

Med pripravo vhodnih datotek za učenje in ocenjevanje so se pokazale različne neskladnosti. Neskladnosti v ročno označenih ali popravljenih jezikovnih podatkih tudi drugim raziskovalcem jezikovnih tehnologij niso neznane (gl. 2.5). Kot je opisano v članku (Lönneker/Jakopin 2004), so v učnem in testnem korpusu oznake, ki jih Jakopin/Bizjak (1997) še nista opisala. Takrat je bil nabor oznak še odprt (Jakopin/Bizjak 1997: 521) in se je razširil vsakič, ko je bila odkrita nova besedna vrsta: danes npr.

[7] http://www.ff.uni-lj.si/hp/pj/disertacija/prilogab.html [6. januar 2005]

obstajajo oznake za več podkategorij imen kot leta 1997. Razen že omenjenih neskladnih oznak (gl. tudi 2.5), najdemo še redke tipkarske napake. Do tega prihaja, ker zaenkrat urejevalnik Eva še ne preverja ročno vpisanih oznak z izčrpnim seznamom dovoljenih oznak. Lönneker/Jakopin (2004) opisujeta izčrpen seznam 3.218 možnih oznak v POSBesedi, tako kot tudi opažene pojave preoznačevanja in podoznačevanja (gl. 2.5) v gradivu.

Zaradi teh pomanjkljivosti so bili razviti nadomestni postopki, s katerimi lahko popravimo neskladnosti, da bi standardizirali označevanje korpusa. V učnem korpusu so bile nadomeščene oznake 644 besed v besednih nizih (med njimi je 59 predložnih zvez), 13.886 primerov podoznačevanja, 10.530 primerov preoznačevanja ter oznake 2.734 veznikov, 30 številk in 87 kratic. Z istim postopkom so bile v testnem korpusu popravljene oznake 553 besed v nizih besed (med njimi so 3 predložne zveze), 2.436 primerov podoznačevanja, 1.106 primerov preoznačevanja ter oznake 843 veznikov, 27 številk in 7 kratic. Tabela 1 povzame te podatke.

Nadomestitve	**Učni korpus**	**Testni korpus**
Oznake v besednih nizih	585	550
Oznake v predložnih zvezah	59	3
Primeri podoznačevanja	13.886	2.436
Primeri preoznačevanja	10.530	1.106
Oznake veznikov	2.734	843
Oznake številk	30	27
Oznake kratic	87	7
Skupaj	27.911	4.972

Tabela 1: Standardizacija oznak korpusov

3.2 Označitev besed in stavkov

Deljenje besedil v besede oz. tokene imenujemo tokenizacija (angl. *tokenization*), deljenje v stavke in druge stavkom podobne enote pa imenujemo segmentacija (angl. *segmentation*). Označevalnik pričakuje, da sta tokenizacija in segmentacija pri neoznačenem besedilu izvedeni in označeni na enak način kot pri učnem korpusu. Zato bi za učne korpuse bila potrebna dokumentacija, ki bi obravnavala pogostejše nejasne primere na teh področjih. Tukaj bodo dodani primeri iz POSBesede.

Primer 1 V učnem korpusu so besedne zveze s *koli* kot npr. *kadar koli*, *kdor koli* večbesedne enote z notranjim presledkom. To pomeni, da dobita tidve besedi skupaj eno oblikoslovno oznako.

Primer 2 Deljenje večbesednih številnikov, npr. s *tisoč* in *milijon*, je manj jasno. Ponavadi je bilo učno gradivo tokenizirano tako, da so iz več besed sestavljeni številniki bili obravnavani kot večbesedne enote, kot v primeru 4 a), ki ga slika 6 prikazuje v vertikalnem formatu. Deli številnikov pa so lahko tudi samostojne enote: v primeru 4 b) ima številnik *sto* isti status kot številnik *deset*. Oba sta bila tokenizirana tako, da je vsak

izmed njiju ločena enota. Na tak način je bil tokeniziran tudi primer 4 c) (gl. sliko 7), kar pa je v nasprotju s podobnim primerom 4 a).

4a) [...] znižal številko na sedeminpetdeset milijonov [...]
4b) [...] sto ali deset milijonov
4c) [...]omejeno na šest milijonov

znižal	[...]
številko	[...]
na	E4
sedeminpetdeset milijonov	ŠG2

Slika 6: Večbesedni številnik kot večbesedna enota (primer 4 a)

omejeno	[...]
na	E4
šest	ŠG4
milijonov	ŠG2

Slika 7: Večbesedni številnik kot dve ločeni enoti (primer 4 c)

Opazimo lahko, da uvaja združevanje večbesednih številk kot v primeru 4 a) čudno zaporedje oblikoslovnih oznak v model, ki si ga označevalnik ustvarja. Primer je razviden iz slike 6, kjer sledi oznaki E4 oznaka ŠG2 namesto ŠG4 kot na sliki 7.[8]

Primer 3 Večbesedna imena, npr. naslovi publikacij ali člankov (*Dom in svet*, *Slovar medicinske vede*) ali organizacij (*Policijska patrola*, *Mohorjeva knjižnica*) so v učnem korpusu obravnavana kot samostojne enote. Za strojno označevanje so najbolj problematične tiste zveze, v katerih je sklanjatvena »končnica« v sredini večbesedne enote (primer: *Slovar medicinske vede*, *Slovarja medicinske vede* itd.). Če so take večbesedne enote slovarju označevalnika neznane, jim označevalnik ne zna pravilno določiti oznake (npr. sklon), saj sodobni označevalniki pri neznanih besedah upoštevajo le končnico in začetek.

Kar se tiče segmentacije POSBesede, je ta v formatu urejevalnika Eva še označena, v formatu XML pa ne več. Ker pa je v formatu XML pri POSBesedi skoraj vedno stavek v eni vrstici, so lahko bila stavčna ločila ».«, »!«, »?« ter »...« naknadno označena z oznako SENT (ločilo na koncu stavka), kjer je bilo to ustrezno. Kjer se pika ni pojavila na koncu stavka, temveč je sledila številki ali kratici, kot v primerih 5 a) in 5 b), je bila označena kot ».«.

5 a) sredi 19. stoletja
5 b) tj. tista točka

[8] Oznaka za številnik je lahko bolj ali manj natančna (gl. 2.5); lahko vsebuje podatke o spolu, številu in sklonu ali pa tudi ne. O obvezni stopnji natančnosti v POSBesedi se njeni avtorji še niso dokončno odločili.

3.3 Tujejezično gradivo v korpusih

Tujejezično gradivo v POSBesedi ni oblikoslovno označeno. V urejevalniku Eva se vsaj oznaka za začetek in konec neslovenskih delov besedila še pojavi, v XML-različici POSBesede pa lahko spoznamo tuje gradivo le tako, da v njem besede nimajo oblikoslovnih oznak. Pri pretvorbi v vertikalni format je bilo tuje gradivo obdelano na dva različna načina:

1. Tuje gradivo je bilo izbrisano v učnem in v testnem korpusu. Tako nastajajo večje ali manjše vrzeli v besedilu, odvisno od tega, ali je tuje gradivo par besed ali več stavkov.
2. Vsaki tuji besedi je bila dodana oznaka `FM` (*foreign material*). S tujim gradivom vsebuje učni korpus 1.270.700 oznak z ločili oz. 1.034.082 brez; testni korpus vsebuje 214.626 oznak z ločili oz. 173.189 brez.

Če pri učenju upoštevamo tuje gradivo, je tovrstno gradivo mogoče prepoznati tudi pri strojnem označevanju. Strojno oblikoslovno označevanje, ki bi upoštevalo pojav tujejezičnega gradiva, do zdaj za slovenščino še ni bilo preizkušeno.

3.4 Označevalni slovar

Pri strojnem oblikoslovnem označevanju uporabljamo slovar besednih oblik, skupaj z njihovimi oznakami. Ta je lahko v vertikalnem formatu, v katerem vsaki besedni obliki sledijo oznake, ki so zanjo mogoče, npr. `prej A Sžd2 Sžp2` (gl. 2.4). Najenostavnejši označevalni slovar vsebuje za vpisane oblike le tiste oznake, ki so se pojavile v učnem korpusu. Če se npr. beseda *prej* v učnem korpusu pojavi samo kot prislov, je vpis v takem slovarju `prej A`. V tem primeru strojni označevalnik ne bi spoznal samostalniške rabe te oblike.

Število znanih oblik lahko povečamo, če obogatimo slovar z oznakami in lemami iz velikega splošnega slovarja. Pri enem izmed ocenjenih poskusov je bil označevalni slovar sestavljen iz naslednjih virov:

1. 82.397 besednih oblik s 105.127 oznak (brez lem) iz predelanega učnega korpusa (gl. tabelo 1);
2. oznake in leme vseh oblik 51.716 pravilnih samostalnikov, 18.128 pravilnih glagolov in 21.694 pridevnikov, izpeljanih iz Slovarja slovenskega knjižnega jezika, ter okoli 28.000 prislovov, ki nastajajo iz pridevnikov ali glagolov (pripravo opisuje Jakopin 2002: 25–32);
3. oznake in leme dodatnih 1.290 besednih oblik, npr. glagola *biti* in številnih zaimkov.

Skupaj vsebujeta že drugi in tretji vir več kot 3 milijone besednih oblik z lemami. Besedne oblike, ki se pojavijo v virih dvakrat, so bile vključene v obogaten slovar le enkrat, po možnosti skupaj z lemo. Lematizacijski slovar za strojni označevalnik namreč ne sme vsebovati dvojnih vpisov, v katerih bi bili besedna oblika in oblikoslovna oznaka enaki. To je lahko problem, če imata drugačni lemi. Npr. dve vrstici za obliko *avta*, ki je lahko oblika leme *avto* ali *avt* z enakimi oznakami, nista mogoči (gl. sliko 8). Namesto tega morajo v takih primerih biti seznami možnih osnovnih oblik napisani v eni vrstici (gl. sliko 9). To pa tudi pomeni, da oblikoslovni označevalnik pripiše

tistim oblikam kombinacijo dveh ali več lem (npr. `avt|avto`). Večpomenskost po označevanju torej še ni popolnoma rešena.

```
avta          Smd1 avt
avta          Smd1 avto
```

Slika 8: Nemogoči vrstici v lematizacijskem slovarju

```
avta          Smd1 avt|avto     [...]
```

Slika 9: Paralelni osnovni obliki v lematizacijskem slovarju

3.5 Rezultati učenja označevalnika TreeTagger

To poglavje povzame in razloži rezultate opravljenega učenja jezikovno neodvisnega označevalnika TreeTagger (Schmid 1994) z uporabo gradiv, opisanih v podpoglavjih 3.1 do 3.4. TreeTagger je v različici za Linux prosto dostopen za učne namene, raziskovanje in evalvacijo. Doslej še ni bil uporabljen za označevanje slovenskih besedil. V tem poglavju so povzeti rezultati petih poskusov učenja z uporabo privzetih nastavitev označevalnika.

Z uporabo izvirne POSBesede (gl. 3.1) in označevalnega slovarja iz učnega korpusa je bila točnost na testnem korpusu 79,99 % brez upoštevanja ločil in 83,88 % z ločili. Z uporabo predelane POSBesede (gl. tabelo 1) se je točnost povečala na 81,39 % (brez ločil), kar pomeni 1,4 % absolutne izboljšave. Z ločili je bila točnost 85,00 %. Na koncu je bil TreeTagger naučen še s predelanim korpusom in z uporabo obogatenega lematizacijskega slovarja (gl. 3.4). Uporaba obsežnejšega slovarja je spet povečala točnost, takrat do 83,43 % (brez ločili), kar pomeni 3,44 % absolutne izboljšave. Z ločili je bila točnost 86,65 %. Tabela 2 predstavlja rezultate teh poskusov.

V drugi poskusni seriji so bile upoštevane tujejezične besede, označene z oznako FM (gl. 3.3). Najprej je bila uporabljena izvirna POSBeseda, vključno s tujejezičnim gradivom. Čeprav je zaradi tega učni korpus večji, je naloga seveda bolj zahtevna, tako da je točnost padla na 79,50 % brez ločil oz. 83,45 % z ločili. S predelanim korpusom se je potem povečala tudi točnost tega označevalnika, in sicer na 80,60 % brez ločil (1,1 % absolutne izboljšave) oz. 84,34 % z ločili. Tako naučen označevalnik pravilno razpozna 490 izmed 1.591 oznak FM v testnem korpusu. Slovenskim besedam dodeli napačno oznako FM v 256 primerih. Tabela 3 predstavlja rezultate poskusne serije s tujejezičnim gradivom, pri kateri obogaten slovar še ni bil uporabljen.

Tabele 4 do 6 prikazujejo število in primere napačnih strojno določenih oznak glede na prejšnje, ročno preverjeno označevanje na celotnem testnem korpusu. Tabele se nanašajo na tri različice naučenega označevalnika TreeTagger, brez upoštevanja tujejezičnega gradiva; uporabljene oznake so iz nabora Jakopin/Bizjak (1997) (gl. 2.1). Vidimo, da v vseh različicah pride do številnih zamenjav sklonov samostalnikov, npr. med prvim in četrtim sklonom pri moških (Sme1/Sme4) ali med četrtim in šestim pri ženskih (Sže4/Sže6). Prav tako so pogosto napačno določene oznake nekaterih

Učni korpus	Izviren	Predelan	Predelan	Predelan, brez bibliografije
Označevalni slovar	Iz učnega korpusa	Iz učnega korpusa	Obogaten	Obogaten
Število oznak testnega korpusa	171.598	171.598	171.598	170.100
Število oznak z upoštevanjem ločil	212.950	212.950	212.950	210.435
Število uspešno dodeljenih oznak	137.260	139.656	143.160	142.129
Točnost (brez ločil)	79,99 %	81,39 %	83,43 %	83,56 %
Točnost (z ločili)	83,88 %	85,00 %	86,65 %	86,71 %

Tabela 2: Pregled rezultatov učenja TreeTaggerja

Učni korpus	Izviren	Predelan	Predelan, brez bibliografije
Število oznak testnega korpusa	173.189	173.189	171.169
Število oznak z upoštevanjem ločil	214.626	214.626	211.558
Število uspešno dodeljenih oznak	137.678	139.584	138.511
Točnost (brez ločil)	79,50 %	80,60 %	80,92 %
Točnost (z ločili)	83,45 %	84,34 %	84,56 %

Tabela 3: Pregled rezultatov z upoštevanjem tujejezičnega gradiva

slovničnih besed; primer so besede kot *pa*, *torej*, *saj*, *ne*, katerih uporaba je v gradivu dvoumna med členkom (Č, ČZ) in prirednim veznikom Vpr. Napake nastajajo tudi pri oznakah glagola *biti* in njegovih oblikah, ki so v gradivu dvoumne med rabo kot pomožni glagol (GP idr.) in glagol obstajanja (GO idr.). Tudi zamenjava med prislovom in pridevnikom, predvsem v prvem sklonu ednine srednjega spola, je pogosta.

Dobra novica je, da številne izmed najpogostejših zamenjav ne motijo lematizacije, saj ne pride pogosto do zamenjave med besednimi vrstami glagol, samostalnik in pridevnik. Lematizirana besedila, ustvarjena z najboljšo različico naučenega označevalnika, bi bila koristna pri leksikografskem delu. V slovenščini pa obstajajo celotne fraze, ki so glede na sklon dvoumne, kot npr. *lep avto* (prvi ali četrti sklon?) ali *pred novo hišo* (četrti ali šesti sklon?), pri katerih se ne moremo zanesti na strojno dodeljene oznake. Koristnost za nadaljno strojno sintaktično obdelavo je zaradi tega omejena.

V splošnem se rezultati še ne približajo točnosti označevalnikov za angleščino, ki uporabljajo veliko manjše nabore oznak. Najboljši rezultat pa je primerljiv s tistimi, ki so jih Džeroski idr. (2000) dosegli s štirimi označevalniki (gl. 2.5). Točnost najboljšega označevalnika iz njihovih poskusov pa je še vedno večja kot točnost TreeTaggerja pri tukaj predstavljenih poskusih. Razen razlik med načini delovanja označevalnikov so razlogi za to lahko razlike v gradivu, naboru oznak in težavnostni stopnji testnega gradiva.

Medtem ko so Džeroski idr. (2000) uporabili različna dela istega romana (*1984*) kot učni in testni korpus, so bili tu uporabljeni za učni korpus popolnoma drugi viri kot za testni korpus. Tako pride do razlike npr. pri točnosti razpoznavanja imen, ki jih lahko označevalnik označi kot navadne samostalnike, če jih še ni videl v učnih podatkih. TreeTagger se npr. na testnem korpusu pogosto moti pri imenih Angel in Bruno ter pri pridevnikih, ki nastajajo iz njih, saj vseh njihovih oblik še ni videl v učnem korpusu. V enem samem romanu pa se v različnih delih ne pojavi toliko različnih imen (gl. 2.4). Druga razlika je lahko skladnost (gl. 2.5) med učnim in testnim korpusom. Roman *1984* je označila (oz. preverila in popravila) samo ena oseba v razmeroma kratkem času, tako da

so nejasni primeri označeni na enak način. Pri POSBesedi pa oseba, ki je preverila in popravila testni korpus, ni označila tudi glavnega dela učnega korpusa (gl. 3.1). Že zaradi tega pride do neskladnosti, npr. pri izrazu *pol milijona*, kjer je en označevalec obliki *milijona* dal oznako za samostalnik, drugi pa za glavni številnik.

Testni korpus *Štirje časi* vsebuje tudi bibliografijo z 2.020 besedami, med njimi je veliko osebnih imen, življenjepisnih imen ter stvarnih imen kot so npr. naslovi publikacij (*Logično-filozofski traktat, Sanje o končni teoriji*) in imena založb (*Društvo Apokalipsa*, *Svetopisemska družba Slovenije*), pri katerih se je naučen TreeTagger velikokrat zmotil.

Primerjava z rezultati poskusov, ki jih predstavljata Erjavec/Džeroski (2004) za označevanje korpusa IJS-ELAN (gl. 2.1), zaradi uporabe drugačnih postopkov ni mogoča. H končni oceni natančnosti njunega označevalnika pripomore že dejstvo, da je bil korpus vnaprej označen, čeprav dvoumno.[9] Tudi učno gradivo je bilo izboljšano z dodajanjem ročno označenega izvlečka besedila, povzetega iz korpusa. Končna ocena natančnosti označevanja je bila izvedena ročno, torej brez predhodno označenega testnega korpusa.

	Korpus	Označevalnik	Primer	Št.
1	Vpr	Č	pa	872
2	Č	Vpr	pa	793
3	Sme1	Sme4	pogovor	788
4	Vpr	Vpo	kot	604
5	Sse1	Sse4	morje	394
6	Pse1	A	podobno	364
7	Sme4	Sme1	pogovor	357
8	Sse4	Sse1	morje	277
9	IOme1	Sme1	Angel	256
10	Sžp1	Sže2	misli	237
11	A	Pse1	podobno	234
12	E5	E4	po	233
13	Sže6	Sže4	žensko	232
14	GOce	GPce	je	216
15	Sže4	Sže1	resničnost	214
16	Č	Vpo	da	207
17	Sže1	Sže4	resničnost	206
18	ZKse1	ZKse4	to	190
19	Pse2	Pme2	prejšnjega	187
20	ZVR	ZRse1	kar	186
21	GPce	GOce	je	183
22	Sže2	Sžp4	roke	180
23	Sže2	Sžp1	misli	159
24	Pžp1	Pže2	črne	157
25	IOme1	IOme4	Schelling	152
26	Vpr	ČZ	ne	151
27	Sžp4	Sže2	roke	151
28	Pme1i	Pmp1	dobri	142
29	Pme2	Pse2	človeškega	142
30	GLmd	GLže	vedela	139

Tabela 4: Napačno dodeljene oznake na testnem korpusu (naučeno na izvirnem učnem korpusu)

	Korpus	Označevalnik	Primer	Št.
1	Sme1	Sme4	pogovor	770
2	Č	Vpr	pa	688
3	Vpr	Č	pa	525
4	Sse1	Sse4	morje	411
5	Sme4	Sme1	pogovor	358
6	Pse1	A	podobno	338
7	Vpr	A	tako	303
8	E5	E4	po	272
9	Sse4	Sse1	morje	263
10	IOme1	Sme1	Angel	259
11	Vpr	ČZ	ne	240
12	Sže6	Sže4	žensko	225
13	Sžp1	Sže2	misli	223
14	A	Pse1	podobno	219
15	GOce	GPce	je	213
16	Sže4	Sže1	resničnost	209
17	Č	Vpo	da	197
18	GPce	GOce	je	190
19	Sže1	Sže4	resničnost	186
20	ZVR	ZRse1	kar	184
21	Sže2	Sžp4	besede	184
22	Pse2	Pme2	prejšnjega	182
23	IOme1	IOme4	Schelling	159
24	ZKse1	ZKse4	to	158
25	Sže2	Sžp1	misli	158
26	Pžp1	Pže2	črne	149
27	Sžp4	Sže2	besede	149
28	Pme2	Pse2	človeškega	142
29	GLmd	GLže	vedela	136
30	KP	A	tj	130

Tabela 5: Napačno dodeljene oznake na testnem korpusu (naučeno na predelanem učnem korpusu)

[9] S tem postopkom je bila npr. besedna oblika *prej* označena s tremi oznakami (A, Sžd2, Sžm2; gl. 2.4.).

	Korpus	Označevalnik	Primer	Št.
1	Sme1	Sme4	pogovor	847
2	Č	Vpr	pa	688
3	Vpr	Č	pa	529
4	Sse1	Sse4	morje	408
5	Pse1	A	podobno	338
6	IOme1	Sme1	Angel	318
7	Sme4	Sme1	pogovor	317
8	Vpr	A	tako	305
9	Sžp1	Sže2	misli	273
10	A	Pse1	podobno	254
11	Sse4	Sse1	morje	253
12	Vpr	ČZ	ne	238
13	Pse2	Pme2	prejšnjega	225
14	GOce	GPce	je	212
15	Pme1i	Pmp1	dobri	203
16	Č	Vpo	da	197
17	ZVR	ZRse1	kar	184
18	Sže4	Sže1	resničnost	184
19	Pžp1	Pže2	črne	180
20	Sžp4	Sže2	besede	179
21	Sže6	Sže4	žensko	178
22	Sže1	Sže4	resničnost	178
23	GPce	GOce	je	175
24	E5	E4	po	170
25	Pmp2	Pžp2	rjavih	167
26	IOme1	IOme4	Schelling	159
27	Sže2	Sžp4	besede	159
28	ZKse1	ZKse4	to	156
29	Sže2	Sžp1	osebe	144
30	GLmd	GLže	vedela	139

Tabela 6: Napačno dodeljene oznake na testnem korpusu (naučeno na predelanem učnem korpusu; obogaten slovar)

4 Zaključek: *Quo vadis*?

100 milijonov besed iz XML-oblike korpusa Nova beseda (Jakopin/Lönneker 2004) je bilo označenih in lematiziranih s pomočjo najboljše različice naučenega TreeTaggerja. Slika 10 prikazuje izrezek tako označenega in lematiziranega članka časopisa *Delo* iz leta 2003, ki je del tega korpusa. Napačno dodeljeni oznaki sta napisani krepko. Razvidno je tudi, da lematizacijski slovar še ni popoln. Manjkajo še imena, svojilni zaimki in členki, ki pa jih je mogoče – z več ali manj težavami – dodati tudi naknadno. Drugi velik slovenski korpus, FIDA, še čaka na razdvoumljenje oznak, saj so v tem korpusu za vsako besedno obliko naštete vse možnosti oblikoslovnih oznak in lem (Erjavec 2003: 71), kar ima za posledico do neke mere nezanesljive rezultate statističnih izračunov pri leksikografskem delu (Gorjanc/Krek 2001: 42).

Pričujoča razprava prikazuje, da delo na strojnem označevanju slovenskih besedil še ne more biti zaključeno. Točnost oblikoslovnega označevanja, ki je okoli 85 % (z naborom 3.218 oznak), je že koristna pri leksikografskem delu, vendar se še vedno ne približa točnosti strojnega označevanja angleških korpusov, ki pa se ponavadi opravlja s precej manjšim naborom oznak.

Iz razprave tudi izhaja, da bi bila zelo potrebna navodila za ročno označevanje oz. popravljenje oznak slovenskih besedil. Taka navodila bi zmanjšala neskladnost med označevalci in izboljšala jezikoslovno dokumentacijo slovenskih označenih korpusov. Najboljše bi seveda bilo, če bi ta navodila upoštevala vse obstoječe nabore oznak za slovenščino in tudi pojasnila možne preslikave oznak med različnimi nabori oz. stopnjo možnih preslikav. Preslikava bi omogočila izboljšano izmenjavo in združevanje gradiva ter raziskovalnega dela. Lahko si predstavljamo zbirko tako dokumentiranega gradiva in programske opreme, ki bi bila prosto dostopna, vsaj za raziskovalne namene. Ta bi vsebovala tudi različne testne korpuse, tako da bi v prihodnosti lažje primerjali rezultate poskusov označevanja slovenskih besedil. Le tako bo v prihodnosti tudi slovenščina lahko izkoriščala »krepostni krog« (Calzolari 2004: 102), v katerem se

```
<s>
Po              E5              po
vstopu          Sme5            vstop
naše            ZSVapže2        -
države          Sže2            država
v               E4              v
Unijo           ISže4           -
bo              GFPce           biti
<lb/>
slovenščina     Sže1            slovenščina
postala         GLže            postati
eden            Sme1            eden
izmed           E2              izmed
uradnih         Pžp2            uraden
jezikov         Smp2            jezik
skupnosti       Sže2            skupnost
,               ,               -
je              GPce            biti
dejal           GLme            dejati
<lb/>
Rupel           IOme1           -
in              Vpr             in
še              Č               -
posebno         A               posebno
opozoril        GLme            opozoriti
na              E4              na
pomen           Sme4            pomen
učenja          Sse2            učenje
tujih           Pmp2            tuj
jezikov         Smp2            jezik
.               SENT            -
</s>
```

Slika 10: Izrezek označenega in lematiziranega članka iz časopisa *Delo*

besedilni korpusi in leksikografska ter jezikoslovno-tehnološka gradiva medsebojno obogatijo.

Zahvala

Opisani poskusi so bili omogočeni s pomočjo štipendije javne ustanove Ad Future na podlagi šestega javnega razpisa Ad Future (pogodbe o sofinanciranju 2004-006), dodeljene avtorici članka leta 2004 za sodelovanje v Laboratoriju za korpus slovenskega jezika Inštituta za slovenski jezik Frana Ramovša ZRC SAZU. Avtorica se lepo zahvaljuje Primožu Jakopinu (vodji Laboratorija) za svetovanje ter Miranu Željku za

lektoriranje besedila. Korpus Nova beseda je sestavil Primož Jakopin s pomočjo sodelavcev Inštituta za slovenski jezik Frana Ramovša ZRC SAZU, predvsem Aleksandre Bizjak in Helene Dobrovoljc. Izvirno POSBesedo in njun izbor oznak sta pripravila Primož Jakopin in Aleksandra Bizjak; pri označevanju je sodelovala tudi Lučka Uršič.

Literatura

Thorsten Brants, 2000a: Inter-Annotator Agreement for a German Newspaper Corpus. V *Second International Conference on Language Resources and Evaluation LREC-2000.* 1435–1439.

– – 2000b: TnT – A Statistical Part-of-Speech Tagger. V *Sixth Applied Natural Language Processing Conference ANLP-2000.* Seattle, WA: ACL. 224–231.

Chris Brew, Marc Moens, 1999: *Data-Intensive Linguistics.* Tehniško poročilo. Edinburgh: HCRC Language Technology Group, The University of Edinburgh. http://tangra.si.umich.edu/~radev/LNI-winter2004/resources/dilbook.ps [6. januar 2005]

Nicoletta Calzolari, 2004: Computational lexicons and corpora. Ur. V Piet Van Sterkenburg: *Linguistics Today – Facing a Greater Challenge.* Amsterdam/Philadelphia: John Benjamins. 98–107.

Doug Cutting, Julian Kupiec, Jan Pedersen, Penelope Sibun, 1992: A Practical Part-of-Speech Tagger. V *Proceedings of the Third Conference on Applied Natural Language Processing.* ACL. 133–141.

Sašo Džerovski, Tomaž Erjavec, Jakub Zavrel, 2000: Morphosyntactic Tagging of Slovene: Evaluating Taggers and Tagsets. V *Second International Conference on Language Resources and Evaluation LREC-2000*. 1099–1104.

Tomaž Erjavec, 1998: Oznake korpusa FIDA. *Uporabno jezikoslovje* 6. 85–95.

– – 2002a. *The IJS-ELAN Slovene-English Parallel Corpus. International Journal of Corpus Linguistics*, 7/1. 1–20.

– – 2002b: *Compiling and Using the IJS-ELAN Parallel Corpus. Informatica* 26. 299–307.

– – 2003: Označevanje korpusov. *Jezik in slovstvo* XLVIII/3–4. 61–76.

Tomaž Erjavec, Sašo Džerovski, 2004: Machine learning of morphosyntactic structure: Lemmatizing unknown Slovene words. *Applied Artificial Intelligence* 18. 17–41.

Vojko Gorjanc, Simon Krek, 2001: V *6th Conference on Computational Lexicography and Corpus Research »Computational Lexicography and New EU Languages«*. Birmingham: Centre for Corpus Linguistics. 41–47.

Primož Jakopin, 2002: *Entropija v slovenskih leposlovnih besedilih.* Ljubljana: Založba ZRC.

Primož Jakopin, Aleksandra Bizjak, 1997: O strojno podprtem oblikoslovnem označevanju slovenskega besedila. *Slavistična revija 45/3–4*. 513–532.

Primož Jakopin, Birte Lönneker, 2004: XML različica besedilnega korpusa Nova beseda. V *Znanstvene izdaje v elektronskem mediju. ZRC SAZU. Povzetki referatov.* Ljubljana: Inštitut za slovensko literaturo in literarne vede ZRC SAZU. 33–34.

Birte Lönneker, Primož Jakopin, 2004: Checking POSBeseda, a Part-of-Speech tagged Slovenian corpus. V *Zbornik 7. mednarodne multikonference IS 2004: Jezikovne tehnologije.* 48–55.

Wolfgang Menzel, 1995: Robust Processing of Natural Language. V Ipke WACHSMUTH, Claus-Rainer Rollinger, Wilfried Brauer (ur.): *KI-95: Advances in Artificial Intelligence.* Berlin: Springer. 19–34.

Helmut Schmid, 1994: Probabilistic Part-of-Speech Tagging Using Decision Trees. V *Proceedings of International Conference on New Methods in Language Processing*. Manchester. 44–49.

Irena Srdanivić – Erjavec, 2004: *Andersenove pravljice: oblikoslovna označitev in statistični opis besedila.* Seminarska naloga pri predmetu Besedilo in računalnik. Ljubljana: Filozofska fakulteta, Oddelek za splošno in primerjalno jezikoslovje.

Darinka Verdonik, Matej Rojc, 2004: Jezikovni viri projekta LC-STAR. V *Zbornik 7. mednarodne multikonference IS 2004: Jezikovne tehnologije.* 42–47.

Atro Voutilainen, 1999: An experiment on the upper bound of interjudge agreement: the case of tagging. V *9th Conference of the European Chapter of the Association for Computational Linguistics.* Bergen: ACL/University of Bergen. 204–208.

Ralph Weischedel, Marie Meteer, Richard Schwartz, Lance Ramshaw, Jeff Palmucci, 1993: Coping with Ambiguity and Unknown Words through Probabilistic Models. *Computational Linguistics* 19/2. 359–382.

Miran Željko, 2002: Pripomočki na spletu za prevajalce zakonodaje EU. V *Zbornik 5. mednarodne multikonference IS 2002: Jezikovne tehnologije.* 33–38.

Summary

The article deals with the currently possible methods to enrich Slovene corpora linguistically, especially in the area of part-of-speech tagging and lemmatization. While the procedure of part-of-speech tagging assigns a morphological tag to each word form, the lemmatization procedure provides each form with its entry form. The first part of the article presents existing tagsets for Slovene and explains tagging and lemmatization methods, and how their results are evaluated. The discussion includes, among other things, the fact that because of the rich Slovene morphology a training corpus for Slovene should be very large.

In evaluation there are two separate criteria, i.e., annotator agreement for manual tagging and precision for automated tagging. In the case of Slovene, the greatest weakness is the absence of a detailed tag description that would allow a uniform treatment of problematic cases. This absence reduces both agreement and precision.

In the second part of the paper the author presents the results of the experiments in training a robust automated tagger. The TreeTagger, which had not been used for Slovene before, tagged 100 million words of the Slovene *Nova beseda* corpus and lemmatized them. The training corpus used in this experiment is POSBeseda, a pre-tagged corpus containing over one million words. The article details the conversion, enhancement and use of POSBeseda for training the statistical tagger. It takes into account the presence of foreign-language material in corpora and gives an overview of the most common tagging mistakes made by the trained tagger.

The results do not reach the precision of automated taggers for English, which use much smaller tagsets. Displaying a precision of about 85%, the best results are nevertheless comparable to those of previous experiments conducted with other taggers trained for Slovene. The article compares the results wherever possible, and gives suggestions for future work.