

## O STROJNO PODPRTEM OBLIKOSLOVNEM OZNAČEVANJU SLOVENSKEGA BESEDILA

Predstavljen je prvi slovenski oblikoslovni označevalnik, tako računalniški program kot tudi nabor oznak. Najprej je opisan nabor; obsega 4.797 oznak, kar odseva veliko pregibnost slovenskega jezika. V nadaljevanju je prikazano, kako deluje označevalnik. Ta poleg podpore ročnemu označevanju, pregledovanju in drugim opravilom vsebuje tudi avtomatični dvostopenjski razločevalnik. Ta skuša v zbirki že obdelanih besedil najti enolično označeno okolico obravnavane besede, če ne uspe, pa si pomaga še s statističnim razločevalnikom, ki sloni na pogostosti *n*-terčkov oznak. Doslej je bilo označenih 330.000 besed – štiri literarna dela in enomesečni vzorec elektronske izdaje časopisa Delo.

The first POS tagger for texts in the Slovenian language is presented. It includes the complete environment: the supporting software as well as the tagset, based on Slovenian grammar. The tagset consists of 4.797 tags, as the language is highly inflected. A description of the tagger follows; it includes a two-step disambiguator. The first step is based on the database of previously processed sentences, where a unambiguously tagged immediate neighbourhood of the observed word is being searched. It is followed by a probabilistic tagger, where frequencies of tag *n*-tuples up to level 5 are taken into consideration. So far 330.000 words have been tagged – four novels and a one-month sample of the leading Slovenian newspaper Delo, a selection from which is available in electronic form on Internet.

### Uvod

Oblikoslovno označevanje besedila je ena izmed stalnic kvantitativnega jezikoslovja, ki ji v zadnjem letu dni namenjajo večjo pozornost tudi jeziki srednje in vzhodne Evrope, kot jih ima navado označevati Evropska unija. Gre za prvo fazo pri razčlenjevanju besedila (angl. parsing), ki je nujni pogoj za nadaljnjo kvantitativno jezikoslovno obdelavo; področje se je razvilo predvsem zaradi potreb pri strojnem prevajanju in pa ob oblikovanju širše dostopnih (preko svetovnega spleta, angl. WWW) nacionalnih besedilnih korpusov (npr. BNC, British National Corpus in CNC, Czech National Corpus). Morfološko-sintaktični analizatorji, posebna programska oprema za tako označevanje besedila, ki vsebuje tudi seznam pravil, so bili najprej narejeni za angleški jezik (Dermatas 1995), od leta 1995 pa nastajajo tudi že za češčino, poljščino, madžarščino in bolgarščino (Hladka, Hajič 1995). Pri tem sta imeli vidno vlogo tudi Evropska unija s svojimi projekti in telesi MULTEXT East (Multilingual Text Tools and Corpora for Central and Eastern European Languages), EAGLES (Expert Advisory Group on Language Engineering Standards), ELRA (European Language Resources Agency), LE PAROLE (Language Engi-

\* <http://www.uni-lj.si/~ffjakopin> in <http://www.zrc-sazu.si/frisj/pj>

\*\* <http://www.zrc-sazu.si/frisj/aleks>

neering) in TELRI (Trans European Language Resources Infrastructure) (Teubert 1995) – pri zadnjem projektu sodelujeta tudi Inštitut za slovenski jezik ZRC SAZU in Inštitut Jožef Stefan – in multinacionalna družba Xerox, ki je leta 1993 odprla center za bazične jezikoslovne raziskave v Grenoblu (Chanod, Segond 1997). Ta je od leta 1996 svojo dejavnost s sedmih zahodnih jezikov razširil še na prej naštete štiri vzhodne.

Na Inštitutu za slovenski jezik smo se, tudi v okviru priprav na gradnjo Slovenskega nacionalnega korpusa, oblikoslovnega označevanja besedil lotili septembra 1996, po spletu zanimivih ugodnih okoliščin. Za poskusni vzorec smo izbrali roman *Pomladni dan* Cirila Kosmača. Ob njem je počasi rasel in se izpopolnjeval nabor oznak, ki je v pregibno bogatih jezikih, kakršen je slovenski, veliko obsežnejši, tudi več desetkrat, kot na primer v angleščini.

Označevanje je seveda potekalo s pomočjo računalnika, le da je bilo na začetku še zelo skromno podprto. Urejevalnik EVA, ki je bil ustrezno dopolnjen, je sproti gradil zbirko že uporabljenih besednih oblik in njihovih oznak. Decembra je bil roman do konca označen, januarja 1997 pa je sledila naslednja knjiga, Platonova *Država*. To delo smo izbrali zaradi sodelovanja v mednarodnem projektu TELRI, kjer v eni izmed delovnih skupin, namenjeni skupnim raziskavam, poteka raziskovanje vzporednih prevodov te knjige v dvanajstih jezikih ustanov soudeleženk (albanščini, angleščini, bolgarščini, češčini, francoščini, latvijščini, litovščini, madžarščini, nemščini, poljščini, slovaščini in slovenščini). Na posvetovanju »TELRI Workshop on Alignment and Exploitation of Texts«, ki je bilo v začetku februarja na Inštitutu Jožef Stefan, sva predstavila primerjavo oznak Kosmačevega romana in prve knjige Platonove Države (Jakopin, Bizjak 1997). Aprila je bila *Država* do konca označena, v maju pa je začel izhajati časopis Delo, ali boljše, kratek izbor iz vsake številke, tudi v elektronski obliki (pod klikljivim imenom »fax časopis« na svetovnospletni strani <http://www.delo.si>). Odločila sva se, da literarnima vzorcema dodava še publicistični jezik. Oblikoslovno sva označila 43 števil, od 6. maja do 17. junija 1997. Velikost vzorcev je navedena v tabeli 1.

Tabela 1: Izbrana besedila

| ime vzorca                    | obseg    | število povedi | število besed |
|-------------------------------|----------|----------------|---------------|
| 1. Ciril Kosmač: Pomladni dan | 176 str. | 5.922          | 61.565        |
| 2. Platon: Država             | 317 str. | 7.323          | 93.430        |
| 3. časopis Delo na internetu  | 43 dni   | 2.956          | 53.895        |
| Skupaj                        |          | 16.201         | 208.890       |

### Slovenski nabor oblikoslovnih oznak

Pri graditvi sistema oblikoslovnih oznak za izbrani jezik je treba najprej upoštevati slovnične zakonitosti in omejitve, ki jih postavlja način dela pri označevanju. Oznake morajo biti šifrirane tako, da je vanje mogoče spraviti vse značilnosti besedne vrste, potrebne za enolično razločevanje. Pri pripravi sistema

oblikoslovnih oznak za slovenski jezik smo izhajali iz Slovenske slovnice (Toporišič 1984). Slovnice značilnosti slovenskega jezika, ki smo jih uporabili pri sestavi oznak, smo nato primerjali z ustreznimi mednarodnimi zgledi, kot sta Brownov korpus (Francis, Kučera 1982) in korpus Penn Treebank (Marcus 1993). Kasnejši zgled je bil Kopernikov projekt MULTTEXT-East (<http://www.lpl.univ.aix.fr/projects/multext-east/>), pri katerem sodeluje tudi dr. Tomaž Erjavec z Inštituta Jožef Stefan; rezultati projekta še niso objavljeni. Koordinator tega projekta je dr. Jean Véronis iz »Laboratoire Parole et Langage« pri CNRS (Centre National de la Recherche Scientifique) in univerzi v kraju Aix-en-Provence, jeziki držav soudeleženk pa so bolgarski, češki, estonski, madžarski, romunski in slovenski. Oblikoslovni nabor, ki ga uporabljajo in ki ga je mogoče videti na svetovnospletni strani <http://nl.ijs.si/ME/Lexica/MorphSyn>, je zelo obsežen in pokriva vse oblikoslovne značilnosti naštetih jezikov. Za naše potrebe, kjer gre za označevanje v samo enem jeziku, večjih količin besedila in z ročnim preverjanjem, pa je žal manj primeren. Njegova največja pomanjkljivost so oznake, ki so po eni strani v angleščini in take, da jih jezikoslovec, ki opremlja novo ali preverja že označeno gradivo, ne more na hitro dešifrirati (oznake niso govoreče), po drugi strani pa je besedilo na zaslonu računalnika prikazano z vsako besedo in oznako v svoji vrstici, kar zmanjšuje njegovo preglednost.

Ponazorimo to s povedjo *Svet zunaj je bil videti mrzel še skozi zaprto okno* iz romana G. Orwella: 1984, označeno najprej z oznakami iz projekta MULTTEXT East (ME):

|         |        |           |     |
|---------|--------|-----------|-----|
| TOK     | Svet   | Ncmsn     | (1) |
| TOK     | zunaj  | Rgp       |     |
| TOK     | je     | Vcip3s-an |     |
| TOK     | bil    | Vcps-sma  |     |
| TOK     | videti | Vmn       |     |
| TOK     | mrzel  | Afpmsn    |     |
| TOK     | še     | Q         |     |
| TOK     | skozi  | Spsa      |     |
| TOK     | zaprto | Afpnsa    |     |
| TOK     | okno   | Ncnsa     |     |
| PTERM_P | .      | AAAA      |     |

in s slovenskimi oznakami, predlaganimi v prispevku:

|      |       |      |       |        |       |    |       |        |       |     |
|------|-------|------|-------|--------|-------|----|-------|--------|-------|-----|
| Svet | zunaj | je   | bil   | videti | mrzel | še | skozi | zaprto | okno. | (2) |
| Sme1 | A     | GPce | GLBme | GNE    | Pme1  | Č  | E4    | PTse4  | Sse4  |     |

Oznake iz povedi (1) so razložene v tabeli 2. Šifre, ki so bile uporabljene pri projektu ME, so v angleščini, zato je taka tudi njihova razlaga. Oznake iz povedi (2) pa so razložene v tabeli 3.

Tabela 2: Razlaga oznak (ME)

|           |                                                                |
|-----------|----------------------------------------------------------------|
| TOK       | Token                                                          |
| Ncmnsn    | Noun common masculine singular nominative                      |
| Rgp       | Adverb general positive                                        |
| Vcip3s-an | Verb copula indicative present third singular active no        |
| Vcps-sma  | Verb copula participle past singular masculine active          |
| Vmn       | Verb main infinitive                                           |
| Afpmsn    | Adjective qualificative positive masculine singular nominative |
| Q         | Particle                                                       |
| Spsa      | Adposition preposition simple accusative                       |
| Afpnsa    | Adjective qualificative positive neuter singular accusative    |
| Nnsa      | Noun common neuter singular accusative                         |
| PTERM_P   | Punctuation terminal point (AAAA)                              |

Tabela 3: Razlaga predlaganih oznak v povedi (2)

|           |                                                                  |
|-----------|------------------------------------------------------------------|
| Sme1      | samostalnik, moški spol, ednina, 1. sklon                        |
| A prislov |                                                                  |
| GPce      | glagol, pomožni, 3. oseba, ednina                                |
| GLBme     | glagol, opisni deležnik na -l (biti), moški spol, ednina         |
| GNE       | glagol, nedoločnik                                               |
| Pme1      | pridevnik, moški spol, ednina, 1. sklon                          |
| Č         | členek                                                           |
| E4        | predlog, 4. sklon                                                |
| PTse4     | pridevnik, deležnik stanja na -t, srednji spol, ednina, 4. sklon |
| Sse4      | samostalnik, srednji spol, ednina, 4. sklon                      |

Ker se projekt ME nanaša tako na indoevropske kot tudi na neindoevropske jezike, so v oznakah vključene vse oblikoslovne značilnosti teh jezikov. Posledica tega je, da se v oznakah pojavljajo tudi s pomišljajem označena prazna mesta, kadar kaka kategorija v jeziku ne nastopa: npr. na sedmem mestu pri Vcip3s-an. Slovenske oznake, grajene na oblikoslovnih značilnostih slovenskega jezika, takih, za druge potrebnih oblikoslovnih kategorij ne upoštevajo. Kljub temu pa so primerljive z mednarodnimi standardi. Slovenski nabor oznak upošteva vse obvezne besedne vrste in pregibne vzorce, navedene v priročniku za ocenjevanje leksike (Underwood, Navaretta 1997).

Slovenske oznake torej označujejo predvsem tiste oblikoslovne značilnosti, ki so bistvene za razločevanje večpomenskosti besednih oblik (angl. disambiguation). Najbolj nazoren primer težav pri razločevanju je recimo poved *Gori na gori gori*, kjer je besedna oblika *gori* lahko prislov (A), samostalnik (S) ali glagol (G), *na* pa predlog (E), glagol (G, velelnik od *dati*) ali medmet (M, npr. v povedi: *Na, pa smo tam*). Teoretično imamo v povedi torej 3 krat 3 krat 3 krat 3, se pravi 81 različnih možnosti označitve in med njimi 32 takih, kjer glagol nastopa natančno enkrat:

AEAG, AESG, AEGA, AEGS, AGAA AGAS, AGSA, AGSS, AMAG, AMSG, AMGA, AMGS, SEAG, SESG, SEGA, SEGS, SGAA, SGAS, SGSA SGSS, SMAG, SMSG, SMGA, SMGS, GEAA, GEAS, GESA, GESS, GMAA, GMAS, GMSA in GMSS.

Pri določanju števila kategorij za posamezno besedno vrsto pa sva poleg merila za razločevanje večpomenskosti besednih oblik upoštevala tudi načelo, da je pri označevanju treba ohraniti kar največ v obravnavanih besedilih shranjene semantične informacije. V ta namen so npr. lastna imena razvrščena v vrste: osebna imena, imena prebivalcev, veroslovna, živalska, zemljepisna, mitološka in stvarna imena.

V nadaljevanju je opisan nabor oznak, ki jih predlagava za slovenski jezik. Oznake vsebujejo možne kombinacije oblikoslovnih značilnosti za vsako posamezno besedno vrsto. Prva velika tiskana črka oznake označuje besedno vrsto, izjema so le kratice. Druge velike črke, ki sledijo, večinoma kažejo na vrste, enomestne šifre, pisane z malo začetnico ali števk, pa označujejo oblikoslovne kategorije.

Za lažje razumevanje je v tabeli 4 najprej naveden seznam enomestnih šifer, iz katerih so sestavljene oznake:

Tabela 4: Šifre oblikoslovnih kategorij

|          |               |               |              |
|----------|---------------|---------------|--------------|
| spol:    | m moški       | oseba:        | a prva       |
|          | ž ženski      |               | b druga      |
|          | s srednji     |               | c tretja     |
| število: | e ednina      | stopnjevanje: | j primernik  |
|          | d dvojina     |               | jj presežnik |
|          | p množina     | določnost:    | i            |
| sklon:   | 1 imenovalnik |               |              |
|          | 2 rodilnik    | razmerje:     | pr priredno  |
|          | 3 dajalnik    |               | po podredno  |
|          | 4 tožilnik    |               |              |
|          | 5 mestnik     |               |              |
|          | 6 orodnik     |               |              |

Oznake so sestavljene na osnovi pregibalnih vzorcev, zato jih bova predstavila po besednih vrstah. Najprej so navedene predmetnopomenske besedne vrste, nato slovnične, na koncu pa medmet in kratice. Poleg enomestnih šifer, ki so razložene v tabeli 4, nastopa v tabelah od 5 do 11 simbol  $\emptyset$ , kadar značilnost v oznaki ni posebej zaznamovana. Primer je stopnjevanje pri prislovih, kjer sta označeni samo primerniška in presežniška stopnja, osnovnik pa ni označen.

V tabeli 5 so razvrščena občna in lastna imena.

Tabela 5: SAMOSTALNIŠKA BESEDA

| besednovrstna oznaka | vrsta | spol  | število | sklon       | primer |             |
|----------------------|-------|-------|---------|-------------|--------|-------------|
| samostalnik          | S     | m,ž,s | e,d,p   | 1,2,3,4,5,6 | Sme1   | (dan)       |
| glagolnik            | SG    | m,ž,s | e,d,p   | 1,2,3,4,5,6 | SGse2  | (spoznanja) |
| osebna imena         | IO    | m,ž,s | e,d,p   | 1,2,3,4,5,6 | IOme1  | (Martin)    |
| imena prebivalcev    | IP    | m,ž,s | e,d,p   | 1,2,3,4,5,6 | IPme2  | (Čeha)      |
| veroslovna imena     | IV    | m,ž,s | e,d,p   | 1,2,3,4,5,6 | IVme3  | (Bogu)      |
| živalska imena       | IŽ    | m,ž,s | e,d,p   | 1,2,3,4,5,6 | IŽže3  | (Liski)     |
| zemljepisna imena    | IZ    | m,ž,s | e,d,p   | 1,2,3,4,5,6 | IZme2  | (Črnomlja)  |
| mitološka imena      | IM    | m,ž,s | e,d,p   | 1,2,3,4,5,6 | IMme5  | (Hadu)      |
| stvarna imena        | IS    | m,ž,s | e,d,p   | 1,2,3,4,5,6 | ISže2  | (Iliade)    |

Za postopek razločevanja samega bi zadostovala že razdelitev samostalnikov na občna in lastna imena, da pa bi uporabniku omogočili večjo svobodo pri iskanju, so lastna imena še naprej razčlenjena.

Kategorija živosti pri samostalnikih moškega spola ni upoštevana v oznakah, ker jo lahko dobimo iz slovarja besed in njihovih pregibalnih skupin.

Precejšnje težave so se pojavile pri označevanju večbesednih lastnih imen. Program in njegov slovar besed in oznak vsebujeta le enobesedna občna in lastna imena, zato so večbesedna imena razstavljena na posamezne besede, označene kot občna imena. Program označi ime *Zgodovina italijanskega srednjega veka* kot *Sže1 Pme2 Pme2 Sme2* in se tako za uporabnika ne ohrani več informacija o stvarnem imenu. Problem bo rešljiv le z večjim slovarjem večbesednih imen, v raziskavi pa smo ga za zdaj reševali ročno.

Tabela 6: GLAGOL

| besednovrstna oznaka        | vrsta  | oseba | spol  | število | sklon | primer |            |
|-----------------------------|--------|-------|-------|---------|-------|--------|------------|
| povedni sedanjik            | G      | a,b,c |       | e,d,p   |       | Gce    | (plava)    |
| pom. gl. biti sed.          | GP     | a,b,c |       | e,d,p   |       | GPce   | (je)       |
| pom. gl. biti prih.         | GFP    | a,b,c |       | e,d,p   |       | GFPcp  | (bodo)     |
| pom. gl. biti nik. obl.     | GZP    | a,b,c |       | e,d,p   |       | GZPae  | (nisem)    |
| glagol obstajanja biti      | GO     | a,b,c |       | e,d,p   |       | GOae   | (sem)      |
| gl. obst. biti nik. obl.    | GZO    | a,b,c |       | e,d,p   |       | GZOce  | (ni)       |
| gl. obst. biti prih.        | GFO    | a,b,c |       | e,d,p   |       | GFOce  | (bo)       |
| glagol tipa imeti nik. obl. | GZ     | a,b,c |       | e,d,p   |       | GZbe   | (nimaš)    |
| velelnik                    | GV     | a,b   |       | e,d,p   |       | GVbe   | (glej)     |
| opisni deležnik na -l       | GL     |       | m,ž,s | e,d,p   |       | GLže   | (obrisala) |
| opis. delež. gl. biti na -l | GLB    |       | m,ž,s | e,d,p   |       | GLBme  | (bil)      |
| trpni deležnik na -n/-t     | GN/GT  |       | m,ž,s | e,d,p   | 1,2   | GNme1  | (rojen)    |
| deležje na -č/-ši           | GČ/GŠI |       |       |         |       | GČ     | (loveč)    |
| nedoločnik                  | GNE    |       |       |         |       | GNE    | (povedati) |
| namenilnik                  | GNA    |       |       |         |       | GNA    | (gledat)   |
| pogojnik                    | GBI    |       |       |         |       | GBI    | (bi)       |
| prosti glagolski morfem     | Gmp    |       |       |         |       | Gmp    | (se)       |

V tabeli 6, v kateri je prikazan glagol, je vidno, da nekaterih oblikoslovnih značilnosti, ki so razvidne iz slovarja in jih od tam označenim besedam lahko tudi dodamo, pri oznakah nismo upoštevali (npr. glagolskega vida in načina).

Tabela 7: PRIDEVNIŠKA BESEDA

| besednovrstna oznaka             | vrsta  | spol  | število | sklon       | stopnjevanje | določnost | primer             |
|----------------------------------|--------|-------|---------|-------------|--------------|-----------|--------------------|
| pridevnik                        | P      | m,ž,s | e,d,p   | 1,2,3,4,5,6 | ø,j,jj       | ø,i       | Pmē1i (pomladni)   |
| opisni deležnik na -l            | PL     | m,ž,s | e,d,p   | 1,2,3,4,5,6 |              | ø,i       | PLmp4 (uspele)     |
| deležnik stanja na -n/-t         | PN/PT  | m,ž,s | e,d,p   | 1,2,3,4,5,6 |              | ø,i       | PNme4i (zgrešeni)  |
| deležnik na -č/-ši               | PČ/PSI | m,ž,s | e,d,p   | 1,2,3,4,5,6 |              | ø,i       | PČže2 (cvetoče)    |
| povedkovnik                      | PD     | m,ž,s | e,d,p   |             | ø,j,jj       |           | PDme (rad)         |
| svoj. prid. iz osebnih imen      | PIO    | m,ž,s | e,d,p   | 1,2,3,4,5,6 |              |           | PIOžp4 (Andrejeve) |
| svoj. prid. iz imen prebivalcev  | PIP    | m,ž,s | e,d,p   | 1,2,3,4,5,6 |              |           | PIPme1 (Brikin)    |
| svoj. prid. iz veroslovnih imen  | PIV    | m,ž,s | e,d,p   | 1,2,3,4,5,6 |              |           | PIVse4 (Kronovo)   |
| svoj. prid. iz zemljepisnih imen | PIZ    | m,ž,s | e,d,p   | 1,2,3,4,5,6 |              |           | PIZže5 (Krimski)   |
| svoj. prid. iz mitoloških imen   | PIM    | m,ž,s | e,d,p   | 1,2,3,4,5,6 |              |           | PIMme5 (Hadovem)   |
| svoj. prid. iz stvarnih imen     | PIS    | m,ž,s | e,d,p   | 1,2,3,4,5,6 |              |           | PISže5 (Mohorjevi) |

Lastnostni, vrstni in svojilni pridevniki niso posebej označeni, izjema so le svojilni pridevniki iz lastnih imen, ker se pišejo z veliko začetnico. Besede z veliko začetnico morajo imeti posebno oznako, ki jo program uporabi pri določanju koncev povedi.

V skupino pridevniške besede smo uvrstili tudi povedkovnik, čeprav je ta v Slovenski slovnici (Toporišič 1984) obravnavan kot samostojna besedna vrsta. Razlog za to odločitev so pravila v mednarodnih standardih, kjer se kaže usmerjenost k čim manjšemu številu paradigem. Za uporabnike pa informacijo o povedkovniku lahko ohranimo tako, da k oznaki P za pridevniško besedo dodamo še D, tako da je povedkovnik označen kot PD. V nekaterih primerih je določitev povedkovnika v vlogi povedkovega določila problematična, zato je bila v veliko pomoč literatura o povedkovem določilu in povedkovem prilastku (Orešnik 1996).

Sledita tabeli za zaimke in števniške. Zaimki in števniški so izvzeti iz pridevniške besede zaradi nekaterih oblikoslovnih posebnosti in ker tako razvrstitev priporočajo mednarodni standardi.

Tabela 8: ZAIMEK

| besednovrstna oznaka     | vrsta | oseba | spol    | število | spol  | število | sklon       | primer              |
|--------------------------|-------|-------|---------|---------|-------|---------|-------------|---------------------|
| osebni zaimek            | ZO    | a,b,c | m,ž,s,ø | e,d,p   |       |         | 1,2,3,4,5,6 | ZOcme5 (njem)       |
| osebni povratni zaimek   | ZOP   |       |         |         |       |         | 2,3,4,5,6   | ZOP2 (sebe)         |
| svojilni zaimek          | ZSV   | a,b,c | m,ž,s,ø | e,d,p   | m,ž,s | e,d,p   | 1,2,3,4,5,6 | ZSVaeme2 (mojega)   |
| povratni svojilni zaimek | ZSVP  |       |         |         | m,ž,s | e,d,p   | 2,3,4,5,6   | ZSVPme6 (svojim)    |
| vprašalni zaimek         | ZV    |       |         |         | m,ž,s | e,d,p   | 1,2,3,4,5,6 | ZVse1 (kaj)         |
| oziralni zaimek          | ZR    |       |         |         | m,ž,s | e,d,p   | 1,2,3,4,5,6 | ZRme2 (kakršnega)   |
| poljubnostni zaimek      | ZPO   |       |         |         | m,ž,s | e,d,p   | 1,2,3,4,5,6 | ZPOme1 (kdo)        |
| oziralni poljub. zaimek  | ZRPO  |       |         |         | m,ž,s | e,d,p   | 1,2,3,4,5,6 | ZRPOme6 (komerkoli) |
| nedoločni zaimek         | ZNE   |       |         |         | m,ž,s | e,d,p   | 1,2,3,4,5,6 | ZNEse1 (nekaj)      |
| drugostni zaimek         | ZD    |       |         |         | m,ž,s | e,d,p   | 1,2,3,4,5,6 | ZDme5 (drugem)      |
| totalni zaimek           | ZT    |       |         |         | m,ž,s | e,d,p   | 1,2,3,4,5,6 | ZTse1 (vse)         |
| nikalni zaimek           | ZNI   |       |         |         | m,ž,s | e,d,p   | 1,2,3,4,5,6 | ZNIme2 (nobenega)   |
| istostni zaimek          | ZI    |       |         |         | m,ž,s | e,d,p   | 1,2,3,4,5,6 | ZIme4 (isti)        |
| mnogostni zaimek         | ZM    |       |         |         | m,ž,s | e,d,p   | 1,2,3,4,5,6 | ZMse3 (marsičemu)   |
| kazalni zaimek           | ZK    |       |         |         | m,ž,s | e,d,p   | 1,2,3,4,5,6 | ZKmp1 (ti)          |
| poudarni zaimek          | ZPU   |       |         |         | m,ž,s | e,d,p   | 1,2,3,4,5,6 | ZPUže1 (sama)       |
| oz. zaimek kot veznik    | ZVR   |       |         |         |       |         |             | ZVR (ki)            |

Besednovrstna oznaka zaimek označuje vse zaimke, samostalniške, pridevniške in prislovne. Tako npr. oznaka ZR zajame samostalniške oziralne zaimke (kdor, kar), pridevniške (kakršen, kateri, čigar, kolikor) in prislovne oziralne zaimke (kjer, kadar, kakor, kolikor). Osebni in svojilni zaimki imajo tudi kategorijo osebe, svojilni zaimki pa imajo še spol in število kot inherentni slovnični kategoriji in spol in število, ki sta v skladenjski vlogi. Primer je *njegove* (knjige), ki ima oznako ZSVmežp1, kjer je moški spol ednine inherentna slovnična lastnost ( *on*, ne *ona* in *on* ne *oni*), ženski spol množine pa kaže na ujemanje z odnosnico.

V primeru, ko gre za oziralni zaimek *ki*, ki nastopa v funkciji veznika, sva se odločila, da mu dava posebno oznako, ZVR, pri čemer je V šifra za veznik. Tako uporabniki lahko najdejo poleg veznikov v besedilu tudi druge besedne vrste, ki nastopajo v vlogi veznika, kot je to npr. članek *ali*, ki ima oznako ČV.

Tabela 9: ŠTEVNIK

| besednovrstna oznaka | vrsta | spol  | število | sklon       | Primer                  |
|----------------------|-------|-------|---------|-------------|-------------------------|
| glavni števnik       | ŠG    | m,ž,s | e,d,p   | 1,2,3,4,5,6 | ŠGže3 (petintridesetim) |
| vrstilni števnik     | ŠV    | m,ž,s | e,d,p   | 1,2,3,4,5,6 | ŠVme2 (prvega)          |
| ločilni števnik      | ŠL    | m,ž,s | e,d,p   | 1,2,3,4,5,6 | ŠLse6 (dvojim)          |
| množilni števnik     | ŠM    | m,ž,s | e,d,p   | 1,2,3,4,5,6 | ŠMme5 (trojnih)         |
| nedoločni števnik    | ŠNE   |       |         |             | ŠNE (nekaj)             |
| število              | Š     |       |         |             | Š (77.500,00.–)         |

Števniki imajo oznako Š, kadar so v besedilu zapisani kot število (s števkami, pikami, vejicami, simboloma + in -), besednovrstne oznake pa imajo le, če so v besedilu zapisani z besedo.

Posebej se je treba ustaviti ob označevanju števnikov tipa pet(ero)/malo, kadar nastopajo v besedni zvezi, ki je v imenovalniku ali tožilniku. Tako kot v drugih slovanskih jezikih (Franks 1995, 93-205) sta tudi v slovenščini (Toporišič 1966, 159) imenovalnik in tožilnik izrazov, ki se v govoru končajo na 5-99 ali s čistimi stoticami ali tisočicami, količinska prislova s štetimi predmeti v množinskem rodilniku. Iz tega sledi, da je za števnike tipa pet(ero)/malo, ki so nepregibni, najustreznejša oznaka ŠG/ŠL/ŠNE. Ker pa je pri tvorbi oznak treba upoštevati tudi sistematičnost znotraj posamezne besednovrstne kategorije in vse oblikoslovne značilnosti, ki pripomorejo k natančnejšemu označevanju, so glavni in ločilni števniki dobili v imenovalniku in tožilniku oznako ŠG1/ŠG4/ŠL1/ŠL4.

V tabeli 10 so predstavljeni še prislov, nato slovnične besedne vrste in medmet.

Tabela 10: DRUGE BESEDNE VRSTE

| besednovrstna oznaka    | vrsta | razmerje | stopnjevanje | sklon     | primer       |
|-------------------------|-------|----------|--------------|-----------|--------------|
| prislov                 | A     |          | (ø,j,jj)     |           | A (resnično) |
| členek                  | Č     |          |              |           | Č (kar)      |
| zanikani členek         | ČZ    |          |              |           | ČZ (ne)      |
| členek v vezniški vlogi | ČV    |          |              |           | ČV (ali)     |
| predlog                 | E     |          |              | 2,3,4,5,6 | E2 (iz)      |
| veznik                  | V     | pr,po    |              |           | Vpr (in)     |
| medmet                  | M     |          |              |           | M (oh)       |

Prislovni izrazi, ki so skladenjsko zveza predlog in samostalnik, pridevnik, zaimek, števniki, prislov ali glagol, so označeni po svojih sestavnih delih.

Členek zanikanja je posebej označen, ker vpliva na sklon samostalnikov v stavku, prav tako tudi členek, ki nastopa v vlogi veznika.

Pri označevanju vzorca št. 3 *Delo na internetu* se je pokazalo, da je za kakovost označevanja treba uvesti tudi oznako za kratice in predvsem, da je treba uvesti posebno oznako za kratice, ki se pišejo z veliko začetnico in se pogosto tudi pregibajo v spolu, številu in sklonu.

Tabela 11: KRATICE

|                               | vrsta | spol    | število | sklon         | primer                                                       |
|-------------------------------|-------|---------|---------|---------------|--------------------------------------------------------------|
| kratica z malo začetnico      | K     |         |         |               | K (št.)                                                      |
| kratica z veliko začetnico    | KI    | ø,m,ž,s | ø,e,d,p | ø,1,2,3,4,5,6 | KI (ŠTUNFF)                                                  |
| naslov svetovnospletne strani | KURL  |         |         |               | KURL ( <a href="http://www.delo.si">http://www.delo.si</a> ) |

Če seštejemo in zmnožimo vse različne možnosti oznak za vse besedne vrste, dobimo število 4.797. Toliko različnih vrednosti ima lahko naš nabor. V vseh treh obravnavanih vzorcih (208.890 označenih besed) je bilo uporabljeno skupaj 1489 različnih oznak ali 31 %.

Nabor oznak, kot smo ga izdelali, je še vedno odprt. Ko se bo preizkušal na večjem vzorcu in na drugih besedilnih žanrih, bodo verjetno še potrebne prilagoditve in dopolnitve.

### Oznake in vzorec

Čeprav je naš vzorec razmeroma majhen, lahko iz njegovih oznak že vidimo nekaj značilnosti literarnega in publicističnega jezika. V tabeli 12 je za vsak vzorec posebej navedeno število besednih oznak po besednih vrstah.

Tabela 12: Skupno število oznak po besednih vrstah

|              | Pomladni dan | Država | Delo   | Skupaj  |
|--------------|--------------|--------|--------|---------|
| občna imena  | 10.173       | 17.488 | 14.078 | 41.739  |
| lastna imena | 1.206        | 668    | 4.269  | 6.143   |
| glagoli      | 20.269       | 21.498 | 9.297  | 51.064  |
| pridevniki   | 4.306        | 7.114  | 6.008  | 17.428  |
| zaimki       | 7.065        | 14.899 | 2.736  | 24.700  |
| števniki     | 399          | 1.363  | 3.188  | 4.950   |
| prislovi     | 3.625        | 5.254  | 1.908  | 10.787  |
| členki       | 3.590        | 5.536  | 1.557  | 10.683  |
| predlogi     | 5.269        | 8.215  | 6.038  | 19.522  |
| vezniki      | 5.377        | 11.368 | 3.398  | 20.143  |
| medmeti      | 274          | 18     | –      | 292     |
| kratice      | 12           | 9      | 1.418  | 1.439   |
| Skupaj besed | 61.565       | 93.430 | 53.895 | 208.890 |

Razmerja med posameznimi vzorci so veliko lepše vidna iz tabele 13, kjer so deleži oznak po besednih vrstah navedeni v odstotkih. V časopisnem jeziku je veliko več samostalnikov, lastnih imen, pridevnikov in kratic. Veliko manj je zaimkov, ki so bolj značilni za literarni jezik, medmetov, ki pogosto nastopajo v dialogu, pa v časopisnem vzorcu sploh ni bilo.

Tabela 13: Deleži oznak po besednih vrstah in skupaj, v odstotkih

|              | Pomladni dan | Država | Delo  | Celota |
|--------------|--------------|--------|-------|--------|
| občna imena  | 16.52        | 18.72  | 26.12 | 19.98  |
| lastna imena | 1.96         | 0.71   | 7.92  | 2.94   |
| glagoli      | 32.92        | 23.02  | 17.25 | 24.46  |
| pridevniki   | 6.99         | 7.61   | 11.15 | 8.34   |
| zaimki       | 11.48        | 15.95  | 5.08  | 11.82  |
| števniki     | 0.65         | 1.46   | 5.92  | 2.37   |
| prislovi     | 5.89         | 5.62   | 3.54  | 5.16   |
| členki       | 5.83         | 5.93   | 2.89  | 5.11   |
| predlogi     | 8.56         | 8.79   | 11.20 | 9.35   |
| vezniki      | 8.73         | 12.17  | 6.30  | 9.64   |
| medmeti      | 0.45         | –      | –     | 0.14   |
| kratice      | –            | –      | 2.63  | 0.69   |
| Skupaj       |              |        |       | 100.00 |

### **Strojni oblikoslovni označevalnik**

Oblikoslovno označevanje besedil (angl. part-of-speech tagging ali POS tagging) se je v večjem obsegu začelo šele na začetku devetdesetih let. Računalniki, dostopni akademskemu okolju, prej niso bili tako zmogljivi, da bi bili kos tovrstnim raziskavam. Če pomislimo, da iz 93.000 gesel v *Slovarju slovenskega knjižnega jezika* nastane preko 3.300.000 besednih oblik, da o imenih in redkeje rabljenih besedah (*Besedišče slovenskega jezika* obsega nadaljnjih 180.000 gesel) sploh ne govorimo, postane hitro jasno, da gre za zbirke, velike desetine megabajtov. Med njihovo obdelavo, razvrščanjem in podobnim pa je treba imeti še nekajkrat več prostora, po možnosti internega. Tudi večje hitrosti mikroprocesorjev, še vedno se podvojijo vsako leto in pol, so nadvse dobrodošle.

Kot na večini drugih področij kvantitativnega jezikoslovja so bili tudi tu prvi poskusi narejeni za angleški jezik. Najprej sta bila označena znani Brownov korpus (Francis in Kučera 1982) in korpus Penn Treebank, zbirka besedil s pensilvanske univerze (Marcus et al. 1993). Nabori oznak so bili zaradi majhne pregibnosti pri angleškem jeziku lahko skromni, pri pensilvanskem korpusu le 48 oznak (36 za besede in 12 za ločila), pri Brownovem 87 in npr. pri francoskem 88 (Chanod, Tapanainen 1993). Avtor se je z (ročnim) označevanjem angleškega besedila prvič srečal v delavnici seminarja COMPLEX, ki ga je septembra 1996 v Budimpešti vodil Gregory Grefenstette s Xeroxovega raziskovalnega centra v Grenoblu. Nabor oznak za slovenski jezik je, kot je videti iz prvega dela prispevka, za velikostni razred večji (4.797 oznak). Velik nabor oznak seveda znatno oteži izdelavo avtomatičnega označevalnika (Hajič, Hladka 1997) in omeji število metod, ki jih je mogoče uporabiti, po drugi strani pa nas, po besedah angleškega raziskovalca R. Garsida (Garside et al. 1987), bogato razčlenjen nabor oznak približa »idealu različnih oznak za vse razrede besed, ki imajo različen slovnični pomen«.

Kot je bilo omenjeno že v uvodu, je bil program za strojno podprto označevanje (označevalnik, angl. tagger) vgrajen v urejevalnik EVA, s čimer sta bili doseženi večja gibčnost (kolegi iz ekipe TELRI so mu rekli »interactive tagger«) in lažja uporaba. Najprej so bili dodani samo postopki za pomoč pri ročnem vnašanju oznak. V besedilu je program najprej poiskal in označil konce povedi (ta naloga zaradi imen in kratic ni povsem trivialna) in za vsako vrstico vrnil še eno prazno, potem pa je v eni delovni datoteki hranil besedilo, ki ga je bilo treba označiti, v drugi pa je sproti gradil in dopolnjeval slovar že obdelanih besed z oznakami. Pri vsaki še neoznačeni besedi je najprej pogledal v slovar, in če je besedo tam našel, je v obliki izbire (menija) na zaslonu pokazal vse že znane možnosti s frekvencami. Na sliki 1 je del povedi iz prve knjige Platonove *Države*.

Slika 1: Del povedi na računalniškem zaslonu med označevanjem

zvezi; pri tem\_\_ so\_\_ nejevoljni, kakor da\_ bi\_ bili\_ oropani  
 Sže5 E5 ZKse5 GPcp Pmp1 Vpo Vpo GBI GLBmp GNmp1  
 velikih stvari in\_ bi\_ bili nekoč imenitno živeli, zdaj pa le  
 Pžp2 Sžp2 Vpr GBI

| GLBmp | 141 |
|-------|-----|
| GLBsd | 2   |
| GLBžd | 4   |
| GLmp  | 1   |
| ?     |     |

Program se je ustavil na besedi *bili*. Beseda *bili* je bila že večkrat označena – 141-krat kot opisni deležnik na -l glagola biti, moški spol množina, 2-krat kot srednji spol dvojina, 4-krat kot dvojina ženskega spola, 1-krat samkrat pa kot opisni deležnik na -l, moški spol (glagol biti, bijem), množina. Opazimo tudi, da je program ob označevanju sproti dodajal mehke presledke (znaka za *in* in za *bi*), da so besede in oznake ostale ena pod drugo. Oseba, ki označuje, lahko izbere ponujeno najpogostejšo možnost (v izbiri so sicer razvrščene po abecedi) ali pa se odloči še za kako drugo oznako, ki je v izbiri ni. Kadar je imela beseda samo eno oznako, npr. samostalniki v prvem sklonu ednine ali glagoli v nedoločniku, jo je program označil sam in se pomaknil naprej. Iz slovarja besed in oznak so v tabeli 14 navedene besede, ki se začnejo z *žive*-.

Tabela 14: Besede na *žive*- v slovarju besed in oznak

žive · Gcp,1;Pže2,5;Pžp1,4;Pžp4,2;Smp4,2

živega · Pme2,2;Pse2,1;Sse2,3

živel · GLme,13

živela · GLže,12

živele · GLžp,2

živeli · GLmp,4

živem · Pme5,1;Pse5,2

živemu · Pme3,1;Sse3,1

živeti · GNE,18

živeče · PČmp4,1

živečih · PČmp2,1;Smp2,1

živečimi · PČsp6,2

Besedi vedno sledijo s podpičji ločene oznake, vsaka pa ima za seboj še z vejico razmejeno frekvenco. Besede v tem slovarju niso razvrščene po abecedi, ampak po kodah (č,š,ž so na koncu abecede), da je iskanje hitrejše.

Ko je količina s takim postopkom obdelanega gradiva naraščala, je postajal učni vzorec za avtomatični označevalnik, program, ki bi sam označeval besedilo, čedalje uporabnejši. Ko je bila obdelana in preverjena prva knjiga, Kosmačev *Pomladni dan*, ter prva knjiga Platonove *Države*, je nastal še dvostopenjski strojni označevalnik. Le-ta črpa informacijo iz že obdelanega gradiva in iz zbirke besed z oznakami, ki je na razpolago. Njegova prva stopnja sloni na slovarju že obdelanih povedi in deluje po načelu zgodovine:

*če ima neka beseda v slovarju besed in oznak (del kaže tabela 14) natančno eno (3) oznako, naj jo dobi, če pa ima več možnih oznak, je treba opazovati njeno okolico. Če ima kaka njena najmanj dve in ne več kot pet besed dolga okolica, pri čemer šteje tudi opazovana beseda, natančno en, iz slovarja označenih povedi razviden nabor oznak, podelimo besedi njeno oznako iz tega nabora, sicer pa pustimo besedo neoznačeno.*

Oglejmo si uporabo navodila (3) na kratkem primeru. Vzemimo poved iz črtice Cirila Kosmača *Kovač in hudič*, ki je izšla v koledarju Prešernove družbe za leto 1959:

Slika 2: Primer povedi, označene ročno

```

Minila je_ pomlad, minilo je_ poletje in_ prišla
GLže GPce Sže1 GLse GPce Sse1 Vpr GLže

je_ jesen, zlati čas_ vsega zlatega, čas_ šumečih vetrov
GPce Sže1 Pme1i Sme1 ZTse2 Sse2 Sme1 PČmp2 Smp2

in_ zrelih vonjav, čas_ velikih oblakov in_ nedosegljivih
Vpr Pžp2 Sžp2 Sme1 Pmp2 Smp2 Vpr Psp2

obzorij, čas_ sladkega in_ otožnega nemira.
Ssp2 Sme1 Pme2 Vpr Pme2 Sme2

```

Slika je verna kopija z računalniškega zaslona, le na vsaki dve vrstici je dodana še ena prazna, zaradi večje preglednosti. Če bi poved obdelali s prvo stopnjo obravnavanega označevalnika, bi dobili sliko 3.

Slika 3: Poved iz slike 2 po prvi stopnji označevalnika

Minila je\_\_ pomlad, minilo je\_\_ poletje in\_ prišla  
 GPce Sže1 GLse GPce Sse1 Vpr  
 je jesen, zlati čas vsega zlatega, čas šumečih vetrov  
 Smp2  
 in\_ zrelih vonjav, čas velikih oblakov in\_ nedosegljivih  
 Vpr Smp2 Vpr  
 obzorij, čas sladkega in\_ otožnega nemira.  
 Vpr Sme2

Program je označil 12 besed izmed 31 (39 %) in to pravilno. Beseda *minila* ima v slovarju označevalnika tri možne oznake (za vejicami so njihove frekvence): *GLmd,1*; *GLsp,1* in *GLže,12*, pred *je* pa se ni pojavila nikoli in je zato ostala neoznačena. *Je* ima pet možnih oznak: *Gce,2*; *GLme,2*; *GOce,287*; *GPce,7686* in *ZOče2,69*, pomlad pa dve: *Sže1,9* in *Sže4,3*. Dvojček *je pomlad* ima samo en nabor oznak: *GPce Sže1,1* (ki izvira iz konteksta *prihajala je pomlad*). *Minilo* je imelo samo eno oznako, *minilo je* tudi samo eno in od tod oznaka za *je*. Ostale označene besede so bile označene zato, ker so imele samo eno oznako. Nobena izmed besed *vetrov*, *oblakov* in *nemira* v slovarju označevalnika ni nastopila v rodilniku dvojine. Če bi, bi ostala na sliki 3 neoznačena. Prva stopnja označevalnika je razmeroma enostavno izvedljiva – potrebujemo le datoteko povedi z besedami in datoteko besed s kazalci v prvo datoteko. Obe skupaj, le zmerno komprimirani, zasedeta v pomnilniku manj prostora kot učno besedilo z oznakami.

Druga stopnja označevalnika sloni na predpostavki, da lahko zaporedje besed v besedilu smatramo za markovsko verigo  $n$ -tega reda, pojmu iz verjetnostnega računa. Pri markovski verigi drugega reda je vsak naslednji člen (dogodek) odvisen le od prejšnjega, pri markovski verigi  $n$ -tega reda pa od prejšnjih ( $n-1$ ) členov (dogodkov). Obravnavani označevalnik spet uporablja znane okolice, tokrat iz slovarja oznak, pri čemer je okolica spet dolga od dve do pet besed – se pravi da obsega opazovano besedo in od ene do štiri besede pred njo ali od ene do štiri besede za njo. Rekli bi lahko, da skuša označevalnik uganiti pravo oznako iz  $n$ -terčkov oznak, pri čemer gre  $n$  od 2 do 5. Za vsako okolico izračuna množico vseh možnih naborov oznak in pogleda, kateri od teh naborov se v slovarju  $n$ -terčkov (dvojčkov, trojčkov, četverčkov in peterčkov) dejansko pojavijo. Če je en sam, je pravi in iz njega program vzame oznako za opazovano besedo. Če ni, je treba pregledati še vse preostale možne okolice.

Poglejmo si (slika 4), kako se navedeno obnese na primeru, in podtaknimo drugi stopnji označevalnika poved iz slike 3 (ki je že šla čez prvo fazo).

Slika 4: Poved iz slike 3 po drugi stopnji označevalnika

Minila je\_\_ pomlad, minilo je\_\_ poletje in\_ prišla  
 GLže GPce Sže1 GLse GPce Sse1 Vpr GLže  
 je\_\_ jesen, zlati čas\_ vsega zlatega, čas šumečih vetrov  
 GPce Sme4 ZTse2 Sse2 Pmp2 Smp2  
 in\_ zrelih vonjav, čas velikih oblakov in\_ nedosegljivih  
 Vpr Pžp2 Sžp2 Pmp2 Smp2 Vpr Psp2  
 obzorij, čas sladkega in\_ otožnega nemira.  
 Ssp2 Vpr Pme2 Sme2

Število označenih besed se je dvignilo na 25, s tem da je beseda *čas* v drugi vrstici narobe označena (Sme4 namesto Sme1). Pravilno je torej označenih 24 besed od 31 ali 77%.

Poglejmo najprej, kako je prišlo do označitve prve besede, *Minila*, ki ima lahko v splošnem tri oznake: *GLmd,1*; *GLsp,1* in *GLže,12*, deležnik na -l, moški spol dvojine, srednji spol množine in ženski spol ednine, ki je tudi najpogostejši. Njena najmanjša okolica je še beseda *je*, ki je že označena in so torej tudi za to okolico teoretično možni trije nabori: (*GLmd GPce*), (*GLsp GPce*) in (*GLže GPce*). Ker se deležnik na -l in pomožni glagol ob njem vedno ujemata v številu, je označevalnik našel v svojem slovarju le tretjega in potem besedi *minila* dodelil oznako *GLže*.

Prva beseda, ki je označevalniku ni uspelo označiti, je *jesen*, druga beseda v drugi vrstici. Vzemimo njeno zadnjo okolico, ki bi še lahko prinesla odločitev, trojček *prišla je jesen*, besedo in dve pred njo. Ker sta prejšnji besedi že označeni, *jesen* pa ima le dve možni oznaki: *Sže1,2* in *Sže4,2*, prideta tudi za okolico v poštev dve možnosti: (*GLže, GPce, Sže1*) in (*GLže, GPce, Sže4*). Žal se izkaže, da sta obe možni; zgledi za prvo v slovarju označevalnika so: *prišla je noč*, *govorila je pravičnost* in *prihajala je pomlad*, za drugo, kjer je jesen v tožilniku, pa: *povesila je glavo*, *nagnila je glavo*, *pokazala je klop*, *imela je kavarno* in *pustila je šolo*.

Napačno označena beseda *čas* v drugi vrstici izvira iz še premajhnega učnega vzorca. Okolica, ki je dala slabo oznako, je trojček z opazovano besedo na začetku: *čas vsega zlatega*. Od vseh teoretičnih možnosti je označevalnik v slovarju n-terčkov oznak našel samo eno: *Sme4 ZTse2 Sse2*, ki izhaja iz vzorca časopisa *Delo*, iz besedne zveze (*povzročiti*) *kolaps vsega gospodarstva*. Tu je samostalnik *kolaps*, zgled opazovanemu času, v četrtem sklonu.

Primer, ki smo si ga ogledali, je eden takih, ki so za označevalnik manj ugodni. Dejansko se učinkovitost obeh stopenj označevalnika na novem besedilu podobnega žanra dvigne na precej čez 80 %.

Kako je videti označeno besedilo, je veliko bolj razvidno iz vsaj nekaj povedi dolgega odlomka. V tabeli 15 je dobre pol strani dolg odlomek iz Platonove *Države*, stran 36 in 37.

Tabela 15: Primer označenega besedila (odlomek iz Platonove Države, I. knjiga)

3.■ Po resnici in odkrito, pri Zevsu, ti bom povedal svoje  
 Š■ E5 Sze5 Vpr A E5 IVme5 ZObe3 GFPae GLme ZSVPse4  
 mnenje, dragi Sokrat.■ Pogosto se sestajamo močje istih let in  
 Sse4 Pme1 Iome1 ■ A Gmp Gap Smp1 ZISP2 Ssp2 Vpr  
 potrjujemo pravilnost starega pregovora.■<sup>4</sup> Ko se pogovarjamo,  
 Gap Sze4 Pme2 Sme2 ■Š Vpo Gmp Gap  
 skoraj vsi tarnajo in se z otožnostjo spominjajo mladostnih  
 A ZTmpl Gcp Vpr Gmp E6 Sze6 Gcp Pžp2  
 radosti, ljubezni, pitja in gostij ter vsega drugega, kar je s  
 Sžp2 Sžp2 Sse2 Vpr Sžp2 Vpr ZTse2 ZDse2 ZVR GOce E6  
 tem v zvezi; pri tem so nejevoljni, kakor da bi bili oropani  
 ZKse6 E5 Sze5 E5 ZKse5 GPcp Pmp1 Vpo Vpo GBI GLBmp GNmp1  
 velikih stvari in bi bili nekoč imenitno živeli, zdaj pa le še  
 Pžp2 Sžp2 Vpr GBI GLBmp A A GLmp A Vpr Č Č  
 životarili.■ Nekateri se tudi pritožujejo, da svojci z njimi -  
 GLmp ■ ZNEmpl Gmp Č Gcp Vpo Smp1 E6 ZOcmp6  
 ker so stari - grdo ravnaajo, in pri tem ubirajo žalostinke o  
 Vpo GPcp Pmp1 A Gcp Vpr E5 ZKse5 Gcp Sžp4 E5  
 nadlogah, ki jih je kriva starost.■ Meni se zdi, Sokrat, da ti  
 Sžp5 ZVR ZOčp2 GPce Pžel Szel ■ ZOae3 Gmp Gce Iome1 Vpo ZKmpl  
 ne obtožujejo pravega krivca, kajti ko bi tega bila kriva  
 ČZ Gcp Pme2 Sme2 Vpr Vpo GBI ZKse2 GLBže Pžel  
 starost, bi imeli jaz in vsi drugi moji starostni vrstniki iste  
 Szel GBI GLmp ZOae1 Vpr ZTmpl ZDmpl ZSVaemp1 Pmp1 Smp1 ZIžp4  
 težave.■ Naletel pa sem že na mnoge, ki se ne počutijo slabo,  
 Sžp4 ■ GLme Č GPae Č E4 Pmp4 ZVR Gmp ČZ Gcp A  
 tako se je nekoč namerilo, da sem bil ravno pri pesniku  
 ZK Gmp GPce A GLse Vpo GOae GLBme A E5 Sme5  
 Sofoklu, ko ga je nekdo vprašal: 'Kako je pri tebi, Sofokles, z  
 Iome5 Vpo ZOcm4 GPce ZNEm1 GLme ZV GOce E5 ZObe5 Iome1 E6  
 ljubeznijo?■ Ali še lahko občuješ z žensko?■' - 'Molči, človek!■'  
 Sze6 ■ Č Č A Gbe E6 Sze6 ■ GVbe Sme1 ■  
 je odvrnil pesnik.■ 'Vesel sem, da je to za mano.■ Tako mi je,  
 GPce GLme Sme1 ■ Pme1 GPae Vpo GOce ZKse1 E6 ZOae6■ ZKse1 ZOae3 GPce  
 kakor da bi pobegnili divjemu, pobesnelemu gospodarju.■ 'Te  
 Vpo Vpo GBI GLme Pme3 PLme3 Sme3 ■ ZKžp4  
 besede so mi že takrat ugajale in nič manj mi ne ugajajo danes.■  
 Sžp4 GPcp ZOae3 Č A GLžp Vpr ZNI Aj ZOae3 ČZ Gcp A ■

## Sklep

Prvi koraki pri oblikoslovnem označevanju slovenskih besedil so bili narejeni. Ob označevanju dveh knjig in časopisnega vzorca (skupaj 200.000 besed) sta postopno nastala tako že precej izbrušen nabor oznak kot tudi računalniško orodje, ki je s pomočjo že obdelanega besedila sposobno označiti nad 80 % novega gradiva in ki znatno olajša obdelavo ostanka in preverjanje celote. V preskusni fazi je še do-datek, ki išče oznake za neznane besede s pomočjo besednih oblik, generiranih iz Slovarja slovenskega knjižnega jezika.

Da bi učinkovitost strojnega označevalnika bistveno izboljšali, bi bilo treba povečati vzorec že obdelanega gradiva vsaj za pol velikostnega razreda, dodati še slovarje imen, predvsem osebnih, priimkov, krajevnih in, ne nazadnje, stvarnih imen. Nujna novost bo tudi seznam slovničnih pravil o besednih zvezah, s katerim bi okrepili drugo stopnjo označevalnika in se izognili napakam pri ročnem dopolnjevanju.

## VIRI IN LITERATURA

- Jean Pierre CHANOD, Frédérique SEGOND: RXRC, 1997: *La Petite Europe. La Tribune des Industries de la Langue* No. 23–24 (March). ISSN 1148–7666
- Jean Pierre CHANOD, Pasi TAPANAINEN, 1995: Creating a tagset, lexicon and guesser for a French tagger. *Texts to Tags: Issues in Multilingual Language Analysis – ACL SIG-DAT workshop*, University College Dublin. 58–64
- Evangelos DERMATAS, George KOKKINAKIS, 1995: Automatic stochastic tagging of natural language texts. *Computational Linguistics* 21/2. 137–163.
- Nelson W. FRANCIS, Henry KUČERA, 1982: *Frequency analysis of English usage. Lexicon and grammar*. Boston: Houghton Mifflin.
- Steven FRANKS, 1995: *Parameters of Slavic Morphosyntax*. Oxford: Oxford University Press.
- Roger GARSIDE et al., 1987: *The computational analysis of English. A corpus-based approach*. London: Longman.
- Jan HAJIČ, Barbora HLADKA: Probabilistic and Rule-Based Tagging of an Inflective Language – a Comparison. Technical Report No. 1, Institute of Formal and Applied Linguistics, Charles University, Prague (in preparation).
- Barbora HLADKA, Jan HAJIČ, 1995: A Simple Czech and English Probabilistic Tagger: a Comparison. *TELRI, Proceedings of the First European Seminar*. 191–196.
- Primož JAKOPIN, Aleksandra BIZJAK, 1997: Part-of-Speech Tagging in the Slovenian Translation of Plato's Republic. *TELRI Newsletter* 5 (april). 7–11.
- Mitchell P. MARCUS et al., 1993: Building a Large Annotated Corpus of English: the Penn Treebank. *Computational Linguistics* 19/2. 313–330
- Janez OREŠNIK, 1996: Nauk novejšje slovenistike o povedkovem prilastku. *SAZU Ljubljana, Razprave* II. 255–267.
- Wolfgang TEUBERT, 1995: What does TELRI mean? *TELRI Newsletter* 1 (September). 3.
- Jože TOPORIŠIČ, 1966: *Slovenski knjižni jezik* 2. Maribor: Obzorja.
- Jože TOPORIŠIČ, 1984: *Slovenska slovnica*. Maribor: Obzorja.
- Nancy UNDERWOOD, Costanza NAVARETTA, 1997: *A Draft Manual for the Validation of Lexica; Preliminary Report*. Paris: ELRA

## SUMMARY

Part-of-speech tagging has in the past year also spread into the domain of so-called Central & Eastern European languages. It is the first step in text parsing and a pre-requisite for further quantitative linguistic analysis, such as machine translation or setting up of national text corpora available on Internet; for instance BNC – British National Corpus and CNC, Czech National Corpus.

In the frame of preparations for the founding of Slovenian National Corpus the project of Slovenian POS tagger was set in motion in September 1996 by the authors of this article, a lingware specialist from the Faculty of Arts in Ljubljana and a linguist from the Fran Ramovš Institute of Slovenian Language at the Scientific Research Centre of the Slovenian Academy of Sciences and Arts.

The tagger and the tagset were developed using the novel *Pomladni dan* by the Slovenian 20th century writer Ciril Kosmač as a starting sample. Beside Slovenian grammar some other sources were considered while assembling the tagset, the tagsets of the Brown corpus, the Penn Treebank corpus and the tagset used in the frame of MULTEXT/East project. The main criteria were the legibility of the tags and the minimal size required for disambiguation. The tags had to be short, derived from Slovenian wording of linguistic terms and self-explanatory to such an extent that they would be not only machine-readable, but acceptable to human reader as well. The point is illuminated in the following sentence:

Seveda se lahko motijo. ('*Certainly they may be wrong.*')  
 Č Gmp A Gcp

The tag Č stands for particle, Gmp for separate verbal morpheme, A for adjective, and Gcp for verb, third person, plural.

The tags for verbs are given in table 1.

Table 1: VERB

| part-of-speech                       | Type   | Person | Gender | Number | Case | Example         |
|--------------------------------------|--------|--------|--------|--------|------|-----------------|
| main verb                            | G      | a,b,c  |        | e,d,p  |      | Gce (plava)     |
| auxiliary verb to be in present      | GP     | a,b,c  |        | e,d,p  |      | GPce (je)       |
| auxiliary verb to be in future       | GFP    | a,b,c  |        | e,d,p  |      | GFPcp (bodo)    |
| negative form of the aux. verb to be | GZP    | a,b,c  |        | e,d,p  |      | GZPae (nisem)   |
| verb to be                           | GO     | a,b,c  |        | e,d,p  |      | GOae (sem)      |
| negative form of the verb to be      | GZO    | a,b,c  |        | e,d,p  |      | GZOce (ni)      |
| verb to be in future                 | GFO    | a,b,c  |        | e,d,p  |      | GFOce (bo)      |
| negative form of the verb to have    | GZ     | a,b,c  |        | e,d,p  |      | GZbe (nimaš)    |
| imperative                           | GV     | a,b,c  |        | e,d,p  |      | GVbe (glej)     |
| participle ending in -l              | GL     |        | m,ž,s  | e,d,p  |      | GLže (obrisala) |
| participle of the verb to be         | GLB    |        | m,ž,s  | e,d,p  |      | GLBme (bil)     |
| participle ending in -n/-t           | GN/GT  |        | m,ž,s  | e,d,p  | 1,2  | GNme1 (rojen)   |
| participle ending in -č/-ši          | GČ/GŠI |        |        |        |      | GČ (loveč)      |
| infinitive                           | GNE    |        |        |        |      | GNE (povedati)  |
| supine                               | GNA    |        |        |        |      | GNA (gledat)    |
| conditional                          | GBI    |        |        |        |      | GBI (bi)        |
| separate verbal morpheme             | Gmp    |        |        |        |      | Gmp (se)        |

The tagger, or better, the software for the support of the tagging process, has developed over time from a simple tool used as an aid for manual tagging to a two-step disambiguator

in 1997. From the very beginning the tagger has been a part of text-editor EVA and therefore totally interactive. It is possible to tag just a single word or proceed from a selected place in text. The lexicon of words, their tags and tag frequencies is updated on the fly as another file, which can also be edited and adjusted when required. This file acts as a database, which is used for production of different statistical tables that give the necessary feedback for fine-tuning of the tagger and the tagset.

The disambiguator has two steps. The first is based on the history of tagged text: if a wordform in the lexicon has one and only one tag, it is the right one; if not, the neighbourhood of the word has to be examined. If there exists a neighbourhood, from 2 to 5 words long (including the word in question), that has a match in the history database, with one and only one set of tags, the word is given its tag from this set, otherwise it is left untagged. The second step is a probabilistic one, where the frequencies of all possible sets of tags for immediate tag neighbourhood (again 2 to 5 tags deep) are considered. If there is one and only one such set for any of the possible neighbourhoods, the corresponding tag is given to the observed word; otherwise the word is left untagged again. The table 2 shows a sentence (by Ciril Kosmač), after the two phases of disambiguator, as displayed on computer screen.

Figure 1: A disambiguated sentence

```

Minila je__ pomlad, minilo je__ poletje in_ prišla
GLže GPce Sže1 GLse GPce Sse1 Vpr GLže

je__ jesen, zlati čas_ vsega zlatega, čas šumečih vetrov
GPce Sme4 ZTse2 Sse2 Pmp2 Smp2

in_ zrelih vonjav, čas velikih oblakov in_ nedosegljivih
Vpr Pžp2 Šžp2 Pmp2 Smp2 Vpr Psp2

obzorij, čas sladkega in_ otožnega nemira.
Ssp2 Vpr Pme2 Sme2

```

(An attempt to translate the sentence might produce the following: *The spring has passed, the summer has passed, and the fall has come, the golden time of all golden, time of murmuring winds and ripe scents, time of huge clouds and unattainable horizons, time of sweet and gloomy unrest.*) In the sentence of table 2 there are 25 tagged words out of 31 with 1 error ( $24/31 = 77\%$  hit rate).

So far 330.000 words have been tagged, manually completed and verified – four novels and a one-month sample of the leading Slovenian newspaper Delo, a selection (about 5% of the full text) from which is available in electronic form on Internet. In table 3 the samples follow each other in the chronological order. Some other data about the first three samples (distribution of part-of-speech groups, for instance) are shown in the paper.

Table 2: The samples

| sample name                      | size    | sentences | words   |
|----------------------------------|---------|-----------|---------|
| 1. Ciril Kosmač: Pomladni dan    | 176 pp. | 5.922     | 61.565  |
| 2. Platon: Država                | 317 pp. | 7.323     | 93.430  |
| 3. Newspaper Delo (Internet)     | 43 days | 2.956     | 53.895  |
| 4. Ciril Kosmač: Prazna ptičnica | 139 pp. | 2.079     | 26.656  |
| 5. George Orwell: 1984           | 229 pp. | 6.684     | 90.760  |
| Total                            |         | 24.964    | 326.306 |

In the field of POS tagging of texts in the Slovenian language, the first steps have been made. A proven tagset, a disambiguator with a hit rate of 80%, and the supporting software, both incorporated into an own text editor for effective and comfortable use, have been established. A database of tagged samples (330.000 words) is now complemented by a lexicon of 3.300.000 wordforms with POS tags, based on 93.000 lemmas from the Dictionary of the Slovenian Literary Language.

The wish list for the future includes a bigger database (1 million tagged words) and a better disambiguator, with a hit rate of 90% or better.