Zasnova in izgradnja oblikoslovnega in glasoslovnega slovarja za slovenski knjižni jezik

Darinka Verdonik, Matej Rojc, Zdravko Kačič, Bogomir Horvat

Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko,
Center za jezikovne tehnologije,
Smetanova ul. 17, 2000 Maribor, Slovenija
darinka.verdonik@guest.arnes.si

Povzetek

Članek predstavlja jezikoslovni vidik sestavljanja oblikoslovnega in glasoslovnega slovarja za slovenski knjižni jezik (SImlex in SIflex), ki ju urejamo na Fakulteti za elektrotehniko, računalništvo in informatiko v Mariboru v Centru za jezikovne tehnologije. Ob tem opozarjamo na važnejša jezikovna vprašanja, ki so se pojavila ob delu, in izbrane rešitve. Projekt se bliža koncu prve faze. Ko bo zaključena, bosta oblikoslovni in glasoslovni slovar vsebovala popis vseh knjižnih oblik in najvažnejših oblikoslovnih lastnosti okoli 20.000 najpogostejših lem ter fonetične prepise vsake oblike.

1. Uvod

Za izvedbo avtomatske prepoznave in sinteze govorjenega besedila so med drugim potrebni jezikovni viri, v katerih so popisani elementi posameznega jezika. Z jezikoslovnega stališča lahko te elemente opazujemo na več ravneh: glasoslovni, oblikoslovni, skladenjski, pomenski, pragmatični... Oblikoslovni in glasoslovni slovar, oba sestavljamo vzporedno, popisujeta elemente slovenskega knjižnega jezika na najnižjih, a temeljnih ravneh – glasoslovni in oblikoslovni. Za uspešno avtomatsko prepoznavo in sintezo govora bodo seveda potrebni še jezikovni viri, ki bodo predstavljali jezik na višjih ravneh, SImlex (oblikoslovni slovar slovenskega knjižnega jezika) in SIflex (glasoslovni slovar slovenskega knjižnega jezika) pa predstavljata dva od osnovnih virov, potrebnih za dosego ambiciozno zastavljenega cilja.

Naš namen v tem članku je predstaviti oba slovarja zlasti z jezikoslovnega stališča: poleg vsebinskega opisa predvsem opozarjamo na jezikovna vprašanja, ki so se pojavljala ob sestavi, in navajamo izbrane rešitve. S tem skušamo prispevati k analizi problemov, ki nastajajo pri takšnem delu. Tehnične rešitve samo omenjamo, saj so bile predstavljene že drugod (Rojc, Kačič, 2000 [1]; Rojc, Kačič, 2000 [2]; Rojc, Kačič, Verdonik, 2002). Ker sta slovarja v osnovi namenjena podpori razvoja jezikovnih tehnologij, so bile vse jezikovne rešitve izbrane tako, da bi kar najbolj ustrezale temu cilju.

Medtem ko podobno obsežnega glasoslovnega slovarja, kot je SIflex, za slovenščino še nimamo, je bilo na področju strojne obdelave oblikoslovnih lastnosti slovenščine več narejenega (Erjavec, 2001; Rojc, Kačič, [2] 2000; Šef, 2001). Prav tako je bilo več narejenega na področju oblikoslovnega označevanja besedil (Jakopin, Bizjak, 1997), kar pa je naslednja faza, pri kateri nam bo oblikoslovni slovar v veliko pomoč.

2. Zajeto besedje in variante

Izbor besed, s katerimi bi pokrili kar največji odstotek besedila, smo opravili na besedilih v elektronski obliki (časopisni članki, dostopni teksti iz leposlovja), ki so skupaj štela okoli 31 milijonov besed. Izmed teh smo izbrali 30.000 najpogostejših in te so bile osnovni nabor, na podlagi katerega smo ročno pripravili nabor besed v slovarski obliki za nadaljnjo obdelavo (Rojc, Kačič, 2000 [2]). Oblikoslovni in glasoslovni slovar sestavljamo s pomočjo sistema Morf (Rojc, Kačič, 2000 [1]), ki omogoča v prvi fazi kar se da avtomatsko vnašanje oblik in variant ter določanje oblikoslovnih lastnosti, v drugi fazi pa naredi avtomatsko grafemsko-fonemsko pretvorbo, ki je osnova za glasoslovni slovar.

V obeh slovarjih (glasoslovnem in oblikoslovnem) so za vsako geslo izpisane vse knjižne oblike in vse knjižne variante naglasov in naglasnih mest, pogovorne variante, ki so označene v slovenskem pravopisu 2001, ne pa tudi narečne ali izrazito starinske variante ali besede. Vnesenih je tudi nekaj najpogostejših lastnih imen, zlasti osebnih lastnih imen. Pri vseh glagolih, ki imajo dvojni naglas v nedoločniku, je vpisana tudi neknjižna naglasna varianta deležnika na *-l* in označena s kvalifikatorjem "pogovorno" (npr. za glagol *stopiti* so glagolske oblike, ki se tvorijo z deležnikom na *-l*, vpisane dvakrat: z oblikami *stop/ila -i -e -o*, tj. knjižno, in z oblikami *st/opila -i -e -o* z dodanim kvalifikatorjem "pogovorno" (znak "/" označuje ostrivec)). Pri samostalnikih so dodane oblike za ženski (*diplomat – diplomatka*) oziroma moški spol (*devica – devičnik*). Vse vnesene oblike in naglasi so ročno preverjeni po SSKJ in slovenskem pravopisu iz leta 2001.

3. Oblikoslovni slovar - SImlex

Oblikoslovni slovar smo začeli sestavljati leta 1999 in bo oktobra 2002 končan v prvi fazi. V tej smo določili vsebino slovarja: opravljen je bil izbor besedja, določili smo, katere podatke bomo vnesli za posamezne besedne vrste, ob vnašanju pa smo reševali zlasti jezikovna vprašanja, ki so se pojavljala ob delu in ki jih navajamo v

nadaljevanju. V tej fazi smo že izvedli tudi vsebinske popravke. Ko bo končana, bo SImlex obsegal 20.000 lem in približno 800.000 besednih oblik.

Zakodiranje informacij in njihova priprava za oblikoslovno označevanje besedil bosta sledila v naslednji fazi. SImlex namreč zajema več informacij o posamezni besedi, kot jih je bilo definiranih pri nekaterih standardih za oblikoslovno označevanje (npr. Multext (Erjavec, 2001)), in je namenjen za širšo uporabo pri sintezi in prepoznavi govora, oblikoslovnem označevanju...

V oblikoslovnem slovarju imajo vse vnesene besedne oblike označen jakostni naglas, tonemski naglas ni označen. Pri vseh besednih oblikah pregibnih besednih vrst je označena končnica, za vsako lemo je posebej določeno, ali ima beseda predpono ali ne in tudi, če ima več predpon (npr. *po+vz+peti*). Morfemov, iz katerih je sestavljena osnova, v prvi fazi nismo posebej označevali (npr. pri samostalniku *car* je v rodilniku ednine ločena končnica *-a*, ne pa tudi pripona *-j- – carj+a*), prav tako ni označeno, če je osnova zložena iz več korenov (tj. pri zloženkah in sklopih), tudi medpone niso določene.

Osnovna kategorija je seveda besedna vrsta, ločimo pa naslednje:

3.1. Samostalnik

Vsakemu samostalniku so določene naslednje kategorije:
- ali je občno ali lastno ime
- spol
- sklanjatev (moška, ženska, srednja)
- vrsta sklanjatve (I., II., III., IV.)
- živost/neživost je označena v tožilniku ednine samo pri samostalnikih, ki se sklanjajo po I. moški sklanjatvi
- skloni in števila

Pri tem so za vsak kriterij upoštevane vse variante, npr. beseda *mercedes* je vpisana kot občno in kot lastno ime, kot občno ime ima vneseni obliki za živost in neživost. Le pri nekaterih zemljepisnih imenih nismo vpisali te variante, če je beseda običajno rabljena kot občno ime (npr. beseda *vas* je vnesena samo kot občno ime, čeprav je lahko tudi lastno, tj. *Vas* pri Kočevju). Če pa je lahko beseda v naboru samo zemljepisno lastno ime, npr. *Moskva*, je seveda tako označena. Pri moških osebnih lastnih imenih, ki v pogovornem jeziku pri sklanjanju podaljšujejo osnovo s *-t* (npr. *Vojko -a* in pog. *-ta*), je zabeležena tudi ta pogovorna varianta (tako kot v SP 2001).

Vpisani so tudi posamostaljeni pridevniki, in sicer kot samostalniki, ki se sklanjajo po IV. sklanjatvi. To velja tudi za posamostaljene pridevnike srednjega spola: za te pravi slovnica iz leta 1976 (Toporišič, 1976), da se sklanjajo po III. srednji sklanjatvi. Sami smo se odločili, da je za naše namene ustreznejša oznaka IV. srednja sklanjatev, poleg tega enako predvideva zadnja slovenska slovnica (Toporišič, 2000).

3.1.1. Edninski samostalniki

Pri vnašanju samostalnikov se je pojavilo vprašanje, ali vpisati vsem samostalnikom, razen množinskim, vsa tri števila, ali pa pri pojmovnih (*lepota*), snovnih (*moka*) in skupnih (*drevje*) imenih, ki se običajno rabijo le v ednini, vpisati samo ednino.

Ta pojav omenjajo že avtorji slovenske slovnice iz leta 1956, in sicer v okviru števila. Ugotavljajo, da "/k/adar splošno govorimo o snoveh, jih rabimo le v ednini. Kadar imamo v mislih določene dele kake snovi, moramo to posebej izraziti, n. pr.: pet hlebov kruha..." (Bajec et al., 1956: 87) Dalje pravijo, da "/n/ekatera snovna imena rabimo tudi v množini, a v nekoliko drugačnem pomenu; včasih mislimo na /.../ različne vrste..." (ibid.). Nič pa ne omenjajo pojmovnih in skupnih imen.

V najnovejši slovenski slovnici je ta problem omenjen na dveh mestih: pri obravnavanju številskosti pregibnih besed in pri vrstah samostalniških besed. O številskosti pravi tako: "Številskost je zmožnost besede za pregibanja bodisi v vseh treh ali le v katerem izmed treh števil. Troštevilska je večina vseh pregibnih besed." (Toporišič, 2000: 271) V nadaljevanju za samostalnik našteva, da "so enoštevilska t. i. samomnožinska imena (možgani, vile, vrata)...", samoedninskih pa pri tem ne omenja. Pri vrsti samostalniških besed pravi slovenska slovnica tako: "Občna imena delimo na števna in neštevna, npr. *potok – lepota, železo, grmovje* (pojmovna, snovna, skupna). Razlika se lepo vidi v množini: *potoki* = 'več kot dva potoka' *proti lepote* = 'več vrst lepote'. Iz neštevnosti vodi prehod v števnost: *tri železa* = 'trije kosi železa'; tak prehod je lahko stilno opazen (pri Prešernu*: plesale lepote Ljubljane so cele*)." (Toporišič, 2000: 275) Skoraj dobesedno enako piše o vrstah samostalnika v slovnici iz leta 1976 (Toporišič, 1976), številskost pa navaja pod drugimi inherentnimi lastnostmi samostalniške besede, in sicer piše le „številnost: tri števila, eno samo število (redko dve)" (ibid., 211).

Ker se (vsaj nekatera) snovna, pojmovna in verjetno tudi skupna imena torej pojavljajo tudi v množini (in dvojini), je bilo očitno, da vseh teh imen ne moremo vpisati kot samoedninskih (npr. množinska oblika *vina* se rabi že prav pogosto). Naslednje vprašanje je bilo, ali pri vseh teh vpisati vsa tri števila (torej tudi pri samostalnikih, kot so *divjad, drevje, zlato, kruh, usposobljenost, varnost...*, ki jih tudi v korpusu nismo našli v nobenem sklonu dvojine ali množine). Odločili smo se, da pri takšnih samostalnikih vendarle vpišemo samo ednino, saj ni preveč verjetno, da bi v besedilih naleteli na oblike *drevij, drevjih, drevji, zlat, zlatih, varnostim, varnostih*, kot bi se glasile nekatere množinske in dvojinske oblike teh samostalnikov. Osnovni kriterij, kdaj vpisati vsa tri števila in kdaj samo ednino, je bil, ali se beseda pojavi v katerem sklonu dvojine ali množine v korpusu Nova beseda, deloma pa je bilo to prepuščeno tudi lastni presoji tistih, ki so sestavljali slovarja.

Z vprašanjem edninskosti samostalnikov smo se ukvarjali tudi pri lastnih imenih. Stvarna in zemljepisna imena, ki

se nanašajo samo na en predmet oziroma kraj/področje, se namreč običajno rabijo samo v ednini (npr. *Večer, Delo, Maribor, Koper*...). Zato smo sklenili, da v prvi fazi takšna stvarna in zemljepisna imena vpišemo kot edninska, tista, ki se lahko nanašajo na več krajev ali stvari (npr. *Bistrica*), pa kot troštevilna. Osebna lastna imena pa so vedno vpisana v vseh treh številih.

3.2. Pridevnik

Za vsak pridevnik so enako kot pri samostalniku izpisani vsi skloni in vsa števila, in sicer za vse tri spole. Ostale kategorije:

- Pri pridevnikih, ki v moškem spolu ločijo določno in nedoločno obliko, sta vneseni obe in ustrezno označeni. Za ženski in srednji spol kategorija določnosti ni označena. Kategorijo določnosti torej označujemo glede na zunanjo obliko, ki jo imajo pridevniki, in ne glede na pomenske lastnosti (Toporišič, 2000: 320). Tako smo se odločili v skladu z osnovnim namenom slovarja, tj. uporabo pri razvijanju jezikovnih tehnologij.
- Pri vseh oblikah za moški spol sta v tožilniku ednine označeni obliki za živost in neživost.
- Pri pridevnikih, ki imajo v pravopisu 2001 navedeno končnico za obrazilno stopnjevanje, sta izpisana primernik in presežnik v imenovalniku ednine za vse tri spole. Tisti primerniki in presežniki, ki so bili v naboru, so obravnavani kot posebna lema. Pri večini pridevnikov, ki se stopnjujejo obrazilno, so dodane tudi oblike za opisno stopnjevanje (npr. za moški spol pridevnika *prijazen* so izpisane stopnje *prijaznejši, najprijaznejši, bolj/manj/najbolj/najmanj prijazen*), čeprav v pravopisu to ni predvideno. Opisno stopnjevanje je vpisano pri vseh pridevnikih, za katere je označeno v pravopisu 2001, poleg tega pa še pri tistih, za katere smo našli kakšno obliko primernika ali presežnika v korpusu Nova beseda (Nova beseda) (npr. pri besedi *umetniški* – "*vendar zato film ni bolj umetniški, ampak kvečjemu manj*").

3.2.1. Vrsta pridevnika

Za posebno težavno in zelo povezano z določnostjo/nedoločnostjo se je pokazalo določanje vrste pridevnika. Že slovnica štirih avtorjev (Bajec et al., 1956) navaja, da ločimo kakovostne (določajo kakovost in odgovarjajo na vprašanje kakšen), svojilne (določajo svojino in odgovarjajo na vprašanje čigav) in vrstne pridevnike (določajo vrsto in odgovarjajo na vprašanje kateri). Pridevniki, ki ločijo določno in nedoločno obliko (tj. kakovostni pridevniki), v določni obliki prav tako odgovarjajo na vprašanje kateri. Delitev na kakovostne, vrstne in svojilne pridevnike predvideva tudi slovnica iz leta 1976 (Toporišič, 1976).

Najnovejša slovenska slovnica pravi o vrsti pridevnikov naslednje: „Pridevnik v ožjem pomenu zaznamuje lastnost, in sicer kakovost (*mlad*) ali mero (*majhen*), vrsto (*jutranji, slovenski*) in svojino (*materin, očetov*); količino zaznamujejo števniki. Na splošno ločimo lastnostne kakovostne in merne (*mlad, majhen*), vrstne (*obči, slovenski, lipov*) in svojilne pridevnike (*očetov, materin, božji*). Deležniki (razen obeh opisnih na *-l* oz. *-n/-t*) se uvrščajo med lastnostne. Vprašalnice za te tri vrste so: *kakšen* ali *kolikšen*, *kateri* in *čigav*." (Toporišič, 2000: 320) V nadaljevanju navaja, da je „/v/elika posebnost lastnostnega pridevnika /.../ oblikoslovno izražanje kategorije določnosti." (ibid.)

Problem pri določanju vrste pridevnika je nastal pri besedah, kot so *avtobusen, magneten, kamnit, baročen*... Ti pridevniki namreč ločijo določno in nedoločno obliko, kar je značilno za lastnostne pridevnike, a v besednih zvezah, kot so *avtobusna postaja, magnetni zapis, kamnita ograja, baročni kip*, zaznamujejo prej vrsto kot kakovost (*avtobusna* proti *železniška postaja, lesena – kamnita ograja*...). Tudi Ada Vidovič Muha te (in podobne) pridevnike v teh besednih zvezah označi za vrstne. Prav tako z izpeljavo dokaže, da lahko tudi pridevniki, kot je *očetov*, v določenem kontekstu izražajo vrsto: „*(Ukradli so) (mi) očetovo uro --> (njihova kraja) (moje) očetove ure*. Drugostopenjska globinska pretvorba izkazuje vrstni pomen drugega pridevnika: *--> kraja ure, ki sem jo imel od očeta / ki jo je dal oče meni*." (Vidovič Muha, 1981: 27)

Ob upoštevanju konteksta bi potemtakem morali za skoraj vsak pridevnik predvideti dve ali celo tri vrste. Ker bi s tem vnesli zmedo in nedoslednost, smo sklenili, da je za naše potrebe najprimernejša naslednja, zelo poenostavljena delitev, ki se opira samo na obliko pridevnika:

- vsi pridevniki, ki ločijo določno in nedoločno obliko, so v določni in nedoločni obliki in v vseh spolih določeni kot kakovostni, kar ustreza oznaki lastnostni pridevniki po najnovejši slovnici (Toporišič, 2000) (npr. *avtobusen –i –a –o, kamnit –i –a -o*)
- vsi pridevniki, ki ne ločijo določne in nedoločne oblike in se v imenovalniku ednine moškega spola končajo na *-i*, so določeni kot vrstni (torej tudi *divji, božji* ipd.)
- vsi pridevniki, ki ne ločijo določne in nedoločne oblike in se v imenovalniku ednine moškega spola končajo na *-0*, so določeni kot svojilni (npr. *sestrin, bratov, lipov*)

3.3. Glagol

Za vsak glagol je označeno, ali je glavni ali pomožni ali naklonski, določena sta glagolski vid in prehodnost, in sicer so za prehodne označeni vsi glagoli, ki lahko imajo ob sebi predmetno dopolnilo, in ne samo direktno prehodni. Pri glagolih *biti, imeti, hoteti* sta označeni zanikana (*nisem, nimam, nočem*) in nezanikana (*sem, imam, hočem*) oblika.

Za vsak glagol so izpisani:

- nedoločnik
- namenilnik
- deležnik na -l
- deležnik na -n/-t
- glagolnik
- deležnik/deležje na -č

- deležnik/deležje na -ši
- deležje na -e
- vse tvorne sedanjiške oblike
- vse tvorne velelniške oblike
- vse tvorne oblike za preteklik
- vse tvorne oblike za prihodnjik
- vse tvorne oblike za predpreteklik
- vse tvorne oblike za sedanji pogojnik
- vse tvorne oblike za pretekli pogojnik
- vse trpne oblike z deležnikom na -n/-t za sedanjik
- vse trpne oblike z deležnikom na -n/-t za velelnik
- vse trpne oblike z deležnikom na -n/-t za preteklik
- vse trpne oblike z deležnikom na -n/-t za prihodnjik
- vse trpne oblike z deležnikom na -n/-t za sedanji pogojnik

Zložene glagolske oblike so torej vpisane kot ena enota, in ne po delih (npr. v zloženi obliki *sem šel* ni posebej določen *sem* kot pomožnik v prvi osebi ednine in posebej *šel* kot deležnik moškega spola ednine), vendar imajo določene vse podatke, s pomočjo katerih bomo lahko kasneje avtomatsko določili potrebne oblikoslovne lastnosti vsake sestavne enote posebej.

Tiste oblike, ki jih posamezen glagol ne tvori, so označene z zvezdico. Trpnik je vnesen pri večini direktno prehodnih glagolov, razen v redkih primerih, ko deležnik na *-n/-t* označuje samo stanje, in ne trpnosti (npr. *biti napit*, čeprav je glagol *napiti (koga s čim)* direktno prehoden). Deležniki sami (ne v zloženih glagolskih oblikah) so izpisani samo v imenovalniku ednine moškega spola, razen tistih, ki so se pojavili v naboru in so obdelani kot pridevniki (npr. *ubit, umrli*). Enako velja za glagolnike (v naboru so bili npr. *ugotovitev, ugrabitev, varčevanje*) in deležje na *-e*, ki je lahko tudi prislov (npr. *molče*).

3.4. Prislov

Pri prislovih ločimo štiri vrste: prostorske, časovne, lastnostne in vzročnostne prislove. Ta delitev je povzeta po slovenski slovnici (Toporišič, 2000), vendar je nekoliko poenostavljena. V slovnici spadajo prostorski in časovni prislovi v skupen razred okoliščinskih prislovov, lastnostni in vzročnostni prislovi spadajo v skupen razred svojstvenostnih prislovov, poleg tega se še vse štiri vrste, ki jih ločimo, delijo v podrazrede.

Po izvoru določamo prislove glede na to, ali so samostalniški, pridevniški ali glagolski.

Stopnjevanje je vpisano enako kot pri pridevnikih, torej obrazilno, če je tako določeno v pravopisu (2001), in pri teh prislovih običajno tudi opisno (npr. za prislov *močno* so vpisane oblike *močneje, močnejše, najmočneje, najmočnejše, bolj močno, manj močno, najbolj močno, najmanj močno*), samo opisno pa pri prislovih, za katere smo našli kakšno stopnjo primernika ali presežnika v korpusu Nova beseda. Opisno stopnjevanje prislovov v slovenskem pravopisu namreč ni označeno.

Poleg prislovov, ki so bili v naboru, so vpisani tudi prislovi, nastali iz obravnavanih pridevnikov. Pri teh je vpisana enaka vrsta stopnjevanja (oziroma nestopnjevanje) kot pri izvornem pridevniku.

3.5. Števnik

Števniki in zaimki so obravnavani posebej, in ne v okviru samostalniške in pridevniške besede, saj imajo oboji nekatere oblikoslovne posebnosti, poleg tega so enako razvrstitev uporabili v nekaterih mednarodnih projektih, npr. Multex-East (Erjavec, Holožan, 1996).

Do zdaj so vpisani samo glavni in vrstilni števniki od ena do dvajset, sto, tisoč, milijon, milijarda in bilijon. Ostale bomo avtomatsko generirali iz teh. Ločilni in množilni števniki bodo vneseni kasneje.

3.6. Zaimek

Vsem zaimkom je določeno, ali so samostalniški ali pridevniški in vrsta (osebni, oziralni, poljubnostni...). Izpisane in določene so oblike za različne sklone, spole, števila, osebe, naslonske oblike. Pri svojilnih zaimkih ločimo število svojine (npr. *moj* = ednina svojine), v tretji osebi ednine svojilnih zaimkov pa še spol svojine (*njegov* = moški spol svojine).

3.7. Členek

Besede so kot členek označene po kvalifikatorjih v slovenskem pravopisu 2001, in ne po SSKJ, kjer je ta besedna vrsta zelo pomanjkljivo označena. Določena je tudi vrsta členka, in sicer po slovenski slovnici (Toporišič, 2000) ločimo 14 vrst: členek čustvovanja, dodajalni, izvzemalni, mnenja/domneve, možnosti/verjetnosti, navezovalni, potrjevanja ali soglašanja, poudarni, presojevalni, spodbujevalni nikalni, spodbujevalni trdilni, vprašalni, členek zadržka ter zanikanja in nesoglašanja.

3.8. Povedkovnik

Kot povedkovnik so vnesene le najbolj značilne besede: *treba, res, rad, všeč*... Vse te besede so vnesene še kot prislov ali katera druga najustreznejša besedna vrsta, saj kategorija povedkovnika v slovarje, ki so nastajali v okviru različnih mednarodnih projektov, običajno ni vključena.

3.9. Veznik

Veznikom je določeno, ali so priredni ali podredni ter ali so enodelni ali večdelni.

3.10. Predlog

Predlogom je določeno, ali so pravi ali nepravi in s katerim sklonom se vežejo.

3.11. Medmet

Določena je vrsta: razpoloženjski, velelni, posnemovalni.

4. Glasoslovni slovar - SIflex

SIflex smo hkrati s SImlexom začeli sestavljati leta 1999 in bo ob dokončanju prve faze (oktobra 2002) vseboval okoli 170.000 enot. Tako obsežnega glasoslovnega slovarja za slovenščino še nimamo. Urejamo ga vzporedno z oblikoslovnim slovarjem: za vsako obliko, ki se pojavi v oblikoslovnem slovarju, v glasoslovnem slovarju s pomočjo avtomatske grafemsko-fonemske pretvorbe zapišemo ortografski zapis in ustrezen fonetični prepis ter nato oboje ročno pregledamo in vnesemo morebitne popravke ali variante. Tako ima vsaka besedna oblika pripisano ustrezno izgovarjavo. Takšen slovar bo v veliko pomoč pri izbiri pravilnega fonetičnega prepisa (npr. homografi). Orodje za avtomatsko grafemsko-fonemsko pretvorbo je podrobneje opisano v (Rojc, Kačič, Verdonik, 2002).

Fonetični simboli so usklajeni z abecedo SAMPA (Speech Assessment Methods Phonetic Alphabet) za slovenski jezik (Zemljak et al.). Vsaka beseda je zlogovana.

Pri fonetičnem prepisu upoštevamo variantne izgovarjave posebnih glasovnih zvez (po SP, paragrafi 688 do 704), npr. če prideta skupaj črki *t* in *s*, sta predvidena izgovora z obema glasovoma in z zlitim *c* – recimo za besedo *odsek* *O t - s \E k* in *O - ts \E k*). Upoštevani so premene po zvenečnosti (npr. če se beseda konča na zveneči soglasnik, se izgovori nezveneči par, recimo samostalnik *breg* se izgovori *b r /e: k*) in paragraf 656 v SP, po katerem je v nekaterih besedah (označene so v SSKJ) mogoč tudi izgovor z *l* poleg izgovora z *U* (npr. *kopalka* se lahko izgovori kot *k O - p /a: l - k a* ali kot *k O - p /a: U - k a*).

5. Zaključek

V članku smo predstavili oblikoslovni in glasoslovni slovar za slovenski knjižni jezik, ki ju od leta 1999 urejamo na Fakulteti za elektrotehniko, računalništvo in informatiko v Mariboru v Centru za jezikovne tehnologije, ter izpostavili nekatere jezikovne probleme, na katere smo pri tem naleteli. Slovarja bosta predvidoma v oktobru 2002 končana v prvi fazi. Ob njenem koncu bo oblikoslovni slovar vseboval približno 20.000 lem in okrog 800.000 besednih oblik, glasoslovni slovar pa okrog 170.000 enot, delo pa s tem ni končano, saj ju bomo v prihodnje dopolnjevali. Statistično vrednotenje v času pisanja članka še poteka. Slovarja bomo uporabili pri gradnji oblikoslovno označenega korpusa, morfološkega analizatorja, pri jezikovnem modeliranju, razpoznavi tekočega govora...

6. Zahvala

Pri vnosu podatkov za oba slovarja so sodelovale Alenka Januš, Branka Meolic, Tanja Šenveter, Barbara Volčjak in Melita Zemljak, slednja je sodelovala tudi pri zasnovi osnovne zgradbe obeh slovarjev. Iskrena hvala Mateju Rojcu za tehnično podporo, brez katere delo ne bi bilo mogoče. Podjetje ČZP Večer je prispevalo besedilni korpus, tj. zbirko člankov dnevnega časopisa Večer od leta 1998 do 2000 v elektronski obliki.

7. Literatura

Bajec, A., Kolarič, R., Rupel, M. 1956. Slovenska slovnica. Ljubljana.

Erjavec, T. (ur.). 2001. Specifications and Notation for MULTEXT-East Lexicon Encoding. http://nl.ijs.si/ME/V2/msd/html/.

Jakopin, P., Bizjak, A. 1997. O strojno podprtem oblikoslovnem označevanju slovenskega besedila. *Slavistična revija.* 3-4: 513-532.

Nova beseda. http://bos.zrc-sazu/s_beseda.html.

Rojc, M., Kačič, Z. [1]. 2000. A Computational Platform for development of Morphologic nad Phonetic lexica. In Gavrilidou, M., Carayannis, G., Markantonatou, S., Piperidis, S., Stainhaouer, G. (eds.), *Second International Conference on Language Resources and Evaluation*. Athens.

Rojc, M., Kačič, Z. [2]. 2000. Design of Optimal Slovenian Speech Corpus for Use in the Concatenative Speech Synthesis System. In Gavrilidou, M., Carayannis, G., Markantonatou, S., Piperidis, S., Stainhaouer, G. (eds.), *Second International Conference on Language Resources and Evaluation*. Athens.

Rojc, M., Kačič, Z., Verdonik, D. 2002. Design and Implementation of the Slovenian Phonetic and Morphology Lexicons for the Use in Spoken Language Applications. In Manuel Gonzalez Rodriguez, Carmen Paz Suarez Araujo (eds.), *Third International Conference on Language Resourses and Evaluation.* Les Palmas de Gran Canaria.

Šef, T. 2001. *Analiza besedila v postopku sinteze slovenskega govora*. Doktorsko delo, FRI.

Toporišič, J. 1976. Slovenska slovnica. Maribor: Založba Obzorja.

Toporišič, J. 2000. Slovenska slovnica. Maribor: Založba Obzorja.

Vidovič Muha, A. 1981. Pomenske skupine nekakovostnih izpeljanih pridevnikov. *Slavistična revija*, 1:19-39.

Zemljak, M., Kačič, Z., Dobrišek, S., Gros, J., Weiss, P. 2002. Računalniški simbolni fonetični zapis slovenskega govora. *Slavistična revija*. V tisku.