

# **Analiza jezikovnih vprašanj, nastalih pri gradnji SIMflexa - oblikoslovnega in glasoslovnega slovarja za slovenski knjižni jezik**

Darinka Verdonik, mag. Matej Rojc, dr. Zdravko Kačič

Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko, Center za jezikovne tehnologije, Smetanova ul. 17, 2000 Maribor, Slovenija

## **Povzetek**

Članek predstavlja jezikovna vprašanja, ki so se postavljala ob zasnovi in gradnji oblikoslovnega in glasoslovnega slovarja za slovenski knjižni jezik (s kratico SIMflex), ki ju urejamo na Fakulteti za elektrotehniko, računalništvo in informatiko v Mariboru v Centru za jezikovne tehnologije in sta bila oktobra 2002 končana v prvi fazi. S takimi in podobnimi vprašanji se sreča slovenist pri gradnji jezikovnih virov za jezikovne tehnologije na oblikoslovni in glasoslovni ravni. Vprašanja analiziramo s stališča uporabnosti slovarjev pri razvoju jezikovnih tehnologij, čemur je SIMflex v prvi vrsti namenjen. Največ prostora namenjamo vprašanjem edninskih samostalnikov ter vrsti in določnosti pridevnikov. Ker podobnih razprav za slovenski jezik nismo zasledili, se nam pa zdijo pomembne za uspešno gradnjo virov, potrebnih za razvoj jezikovnih tehnologij, želimo s tem prispevati k analizi podobnih problemov.

**Ključne besede:** Siflex, SImlex, SIMflex, jezikovne tehnologije, jezikovni viri, sinteza govora, prepoznavna govora/besedila, besedna vrsta, oblikoslovje, glasoslovje.

## **Analysis of Linguistic Questions Arised while Building the SIMflex - Morfological and Phonetic Lexicon for Slovenian Language**

### **Abstract**

Article represents linguistic questions, which arised while building morphological and phonetic lexicon for Slovenian literary language (by short SIMflex). Those lexicons are built on Faculty of Electrical Engineering and Computer Science in Maribor, in Centre for Language Technologies, and in October 2002 they were finished in their's first stage. While analysing those questions we try to find the kind of solutions that we presume are the best for developing language technologies. The most space we destine to questions of uncountable nouns, and type and definiteness of adjectives. Since we didn't find similar papers for Slovenian language, but we think them necessary for developing language technologies, we want to contribute to analyses of similar problems.

**Key words:** Siflex, SImlex, SIMflex, language technologies, language resourses, speech synthesis, speech/text recognition, part of speech, morphology, phonology.

## 1 UVOD

Tehnološki razvoj v zadnjih desetletjih je omogočil, da so raziskovalne skupine po svetu in tudi pri nas (Inštitut Jožef Stefan, Fakulteta za elektrotehniko v Ljubljani, Fakulteta za elektrotehniko, računalništvo in informatiko v Mariboru, podjetje Amebis idr. (Kačič, 2002)) začele razvijati tehnologije za sintezo in prepoznavo človeškega govora, t. i. jezikovne tehnologije. S tem se je začela razvijati znanstvena veja, ki združuje na prvi pogled povsem nepovezani vеди - računalništvo in jezikoslovje. Prva prispeva znanje, s pomočjo katerega so/bodo tehnologije izvedene, druga prispeva znanje o ključnem predmetu - jeziku. Kolikor tesneje sta obe znanji združeni, toliko uspešnejša je pot proti cilju. Tako kot mora jezikoslovec, ki pomaga razvijati jezikovne tehnologije, kar se da dobro razumeti, kako deluje stroj ter kaj vse in kako se da z njim narediti, mora tudi inženir čim bolje poznati zgradbo jezika in način, kako ta funkcionira; idealno pa bi seveda bilo, ko bi en človek združeval oboje. Kakorkoli se zdi stroj, ki bi tekoče govoril in prepoznaval človeški govor, na videz utopičen cilj, se vse več ljudi trudi doseči ga, zgodovina pa nas uči, da se velikokrat na videz utopični projekti še prehitro uresničijo.

Za uspešen razvoj jezikovnih tehnologij so med drugim potrebni jezikovni viri, v katerih so popisani elementi posameznega jezika. Tradicionalna slovnica te elemente opisuje na več ravneh: glasoslovni, oblikoslovni, skladenjski, pomenski... Ker je prav strukturalizem, na katerem temelji tradicionalna slovnica, eden najbolj "tehničnih" med različnimi vidiki raziskovanja jezika, je tudi za razvoj jezikovnih tehnologij primeren popis elementov jezika po teh ravneh. Začne se pri osnovnih, glasoslovni in oblikoslovni ravni (Internet #4; Rojc, Kačič, [2] 2000; Šef, 2001; Jakopin, Bizjak, 1997), prvi poskusi tudi za slovenski jezik pa so že na ravni skladnje in semantike (Hajdinjak, M., Mihelič, F., 2002).

## 2 ZASNOVA SLOVARJEV SIMFLEX

Osnovni namen oblikoslovnega (Simlex) in glasoslovnega slovarja (Siflex) za slovenski knjižni jezik (z eno besedo Simflex), ki ju na Fakulteti za elektrotehniko, računalništvo in informatiko v Mariboru sestavljamo od leta 1999, je torej kar najširša uporabnost pri razvijanju jezikovnih tehnologij, tj. pri sintezi in prepoznavi govora/pisanega besedila. To dejstvo moramo imeti v mislih pri vsaki odločitvi glede vsebine slovarjev. Pri vsakem podatku, ki ga vnesemo, se moramo vprašati, v kakšno pomoč bo ta informacija računalniku. Ali je, na primer, podatek, da je neki pridevnik kakovostni in drugi vrstni, da je neki glagol naklonski in drugi glavni, koristen? Ali pri pregibnih jezikih, kot je slovenščina, koristi podatek o končnici, in če, kaj je bolje določiti za končnico v primerih, ko se osnova pri sklanjanju podaljšuje (recimo *nebo* - *nebes+a* ali *nebe+sa*)? In če gremo še korak nazaj, kakšne lastnosti pa sploh ima za računalnik koristen podatek? Takšna vprašanja so za jezikoslovca seveda izredno neljuba, saj slabo pozna delovanje stroja. Neljuba pa so tudi za inženirja, saj tehnologija še ni razvita do konca in težko je reči, kateri koraki bodo še potrebni na poti k zastavljenemu cilju. Prav zato je zasnova Simflexa kar se da široka: v njem skušamo popisati vse pomembne oblikoslovne in glasoslovne lastnosti slovenskega knjižnega besedja.

## 3 KAJ VSEBUJETA SIMLEX IN SIFLEX

Praden podrobneje spregovorimo o vnesenih informacijah in o dilemah, ki se pojavljajo ob takem delu, še nekaj besed o izboru besedja. V prvi fazi je v obeh slovarjih zajetih 30.000 najpogostejših besednih oblik, s katerimi je pokrit kar največji odstotek besedila. Izbor je bil narejen na besedilih v elektronski obliki (časopisni članki, dostopni teksti iz leposlovja), ki so skupaj štela okoli 31 milijonov besed (Rojc, Kačič, 2000 [2]). Iz izbranih 30.000 besed je bil ročno pripravljen nabor besed v slovarski obliki (leme) za nadaljnjo obdelavo in ob koncu prve faze je v oblikoslovnem slovarju popisanih okoli 20.000 lem in približno 800.000 besednih oblik, v glasoslovnem slovarju pa 170.000 enot. Slovarja sestavljamo vzporedno, in sicer se v glasoslovni slovar za vsako besedno obliko, ki se pojavi v oblikoslovnem slovarju, zabeleži pripadajoč fonetični prepis, torej je vsaki besedni obliki

pripisana ustrezna izgovarjava. Podatki se vnašajo s pomočjo sistema Morf (Rojc, Kačič, 2000 [1]), ki omogoča v prvi fazi kar se da avtomatsko vnašanje oblik in variant (skupaj z naglasi) ter določanje oblikoslovnih lastnosti, v drugi fazi pa naredi avtomatsko grafemsko-fonemsko pretvorbo vsakega gesla posebej, kar je osnova za glasoslovni slovar.

Potem ko je bilo izbrano najpogostejše besedje, se je spričo pestrosti oblikoslovnih in glasoslovnih variant v slovenskem knjižnem in še večje pestrosti v neknjižnem jeziku pojavilo vprašanje, katere variante zajeti. Ker tudi v SIMlexu (oblikoslovnem slovarju) označujemo vrsto in mesto naglasa, je teh variant še toliko več. Ali torej vnesti samo knjižne oblike, ali vključiti tudi kakšne pogovorne in narečne oblike in če, katere (vseh je vsekakor odločno preveč), ali vnesti tudi starinske in stilno zaznamovane besede in besedne oblike? Jezikovne pripomočke, ki bi ustrezno popisovali aktualen, živ jezik, kakršnega bi sami gotovo hoteli imeti popisane v slovarjih, za slovenščino zaradi majhnosti le stežka zagotavljamo. Tako smo imeli ob začetku gradnje SIMflexa na voljo le SSKJ, za odločanje o oblikah smo si od leta 2000 pri preverjanju lahko nekoliko pomagali s korpusom Nova beseda (Internet #2), delno tudi s korpusom FIDA (Internet #3) in od leta 2001 s pravopisnim slovarjem. Leta 2000 je izšla še nova slovnica (Toporišič, 2000), a tudi ta, kot bomo videli v nadaljevanju, ne popisuje vseh oblikoslovnih lastnosti slovenskega knjižnega jezika tako natančno, da bi lahko samo z njeno pomočjo rešili vsa jezikovna vprašanja. Medtem ko oba korpusa pomagata pri odločanju o oblikoslovnih variantah, ki so še v rabi, žal še vedno nimamo referenčnega govornega korpusa za slovenščino (tj. obsežne elektronske zbirke govornih besedil, ki bi zajemala vzorčni delež besedil slovenskega jezika), s katerim bi si lahko pomagali pri odločanju, katere naglasne variante so dovolj aktualne, da jih zabeležimo.

V SIMflexu so tako vnesene naslednje variante: za vsako geslo/lemo v oblikoslovnem slovarju so izpisane vse knjižne oblike in vse knjižne variante naglasov in naglasnih mest (vse ročno preverjeno po SP 2001 in SSKJ), pogovorne variante, ki so označene v slovarskem delu slovenskega pravopisa 2001, ne pa tudi narečne ali izrazito starinske variante ali besede, ki jih tudi v korpusih ni najti. Vnesenih je tudi nekaj najpogostejših lastnih imen, zlasti osebnih lastnih imen in pomembnejših zemljepisnih imen, ki so se pojavila v naboru. 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 samostalniki so osnovnemu naboru dodane oblike za ženski (*diplomat – diplomatka*) oziroma moški spol (*devica – devičnik*).

#### 4 OBLIKOSLOVNI SLOVAR - SIMLEX

Potem ko določimo razpon variant, je treba razmisliti o podatkih, ki bi bili v takšnih slovarjih uporabni. Pri tem se ob zgradbi oblikoslovnega slovarja pojavlja bistveno več vprašanj kot ob zgradbi glasoslovnega, zato najprej obširneje obravnavamo prvega.

Tudi v SIMlexu je vsem vnesenim besednim oblikam označen jakostni naglas - tonemski naglas ni označen. Vsem besednim oblikam je določena končnica in vsaki lemi predpona oziroma predpone, če je teh več (npr. *po+vz+peti*). Morfemi, iz katerih je sestavljena osnova, v prvi verziji SIMlexa niso posebej označeni (npr. pri samostalniki *car* je v roditeljski 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. Z razvojem tehnologij pa se že kaže potreba, da bi bili označeni vsi morfemi vsake besede, tako da bo to morebiti naloga ene od prihodnjih faz.

Ko govorimo o oblikoslovnem slovarju, je osnovna kategorija seveda besedna vrsta. Pri tem se razdelitev v SIMlexu nekoliko loči od razdelitve v tradicionalni slovnici: zaimke (samostalniške, pridevniške in prislovne) obravnavamo posebej, ne v okviru samostalniške in pridevniške besede ter prislova, prav tako je kot posebna besedna vrsta določen števniki, posamostaljeni pridevniki spadajo v

besedno vrsto samostalnik, povedkovnik pa je upoštevan le deloma. Podrobneje o tem, zakaj tako, v nadaljevanju.

#### 4.1 Samostalnik

Vsakemu samostalniku so določene naslednje kategorije:

- ali je občno ali lastno ime
- spol
- sklanjatev (moška, ženska, srednja)
- vrsta (I., II., III., IV.) in tip sklanjatve
- živost
- skloni in števila

Vrsta in tip sklanjatve v podobnih slovarjih (Internet #4) navadno nista označena - pri razvoju jezikovnih tehnologij pa je tudi takšna nadaljnja tipologizacija besed lahko v pomoč. Podspol živost je v SImlexu označen v tožilniku ednine pri samostalnikih, ki se sklanjajo po I. moški sklanjatvi. Pri besedah, ki lahko imajo kategorijo živosti ali ne, npr. avtomobili (*ford, audi...*) ali kadar je od pomena odvisno, ali ima beseda kategorijo živosti ali ne (*žerjav* - ptica ali naprava), je treba vnesti obe možnosti.

##### 4.1.1 Posamostaljeni pridevniki

Posamostaljeni pridevniki so vpisani kot samostalniki, ki se sklanjajo po IV. sklanjatvi. Za posamostaljene pridevnike srednjega spola pravi slovnica iz leta 1976 (Toporišič, 1976), da se sklanjajo po III. srednji sklanjatvi. V SImlexu je ta sklanjatev označena kot IV. srednja sklanjatev, tako kot predvideva zadnja slovenska slovnica (Toporišič, 2000).

##### 4.1.2 Edninski samostalniki

Pri gradnji oblikoslovnega slovarja za slovenski knjižni jezik pri samostalnikih naletimo na vprašanje edninskih samostalnikov: 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. Poglejmo, kaj o tem pravijo pomembnejše slovenske slovnice.

Avtorji slovenske slovnice iz leta 1956 govorijo o tem 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 (Toporišič, 2000) 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.” (ibid., 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*.” (ibid., 275) Skoraj dobesedno enako piše o vrstah samostalnika v slovnici iz leta 1976 (Toporišič, 1976), številskost pa se navaja pod drugimi inherentnimi lastnostmi samostalniške besede, in sicer piše le „številnost: tri števila, eno samo število (redko dve)“ (ibid., 211).

S temi navedbami si žal ni mogoče veliko pomagati. Smiselni možnosti sta: ali vse samostalnike, razen množinskih, vpisati v vseh treh številih, ali pa vsa pojmovna, snovna in skupna imena vpisati kot

samoedninska. Prva rešitev se zdi za prepoznavo govora/besedila praktična, toda kaj pa besede, kot so *divjad, drevje, zlato, kruh, usposobljenost...* Množinske in dvojinske oblike teh so slišati skrajno neživljenjske: *drevja, drevij, drevjem...*; *zlata, zlat, zlatom...*; *kruhi, kruhov, kruhom...* zveni celo pogovorno. Druga rešitev je še bolj neprimerna: nekatera snovna, pojmovna in tudi skupna imena se namreč že prav pogosto pojavljajo v množini ali dvojini (npr. *vina*), torej je prav, da se te oblike zabeležijo. Ker je merodajna raba jezika, je ta problem najbolje reševati s pomočjo korpusov za slovenski jezik, kjer se lahko za snovna, pojmovna in skupna imena preveri, ali so v slovenskih besedilih bile uporabljene kakšne množinske ali dvojinske oblike teh samostalnikov.

Vprašanje edninskosti samostalnikov se pojavi tudi pri lastnih imenih. Stvarna in zemljepisna imena se namreč običajno rabijo samo v ednini, če se nanašajo samo na en predmet oziroma kraj/področje (npr. *Večer, Delo, Maribor, Koper...*), kljub temu da jih prav lahko sklanjamo v vseh treh številih. Raba takšnih imen v množini ali dvojini bi bila najverjetnejša v umetnostni literaturi, v takih primerih pa je že vprašljivo, če še gre za lastno ime. V prvi verziji SIMlexa so zato takšna stvarna in zemljepisna imena vpisana samo v ednini, tista, ki se lahko nanašajo na več krajev ali stvari (npr. *Bistrica*), pa v vseh treh številih. Pri osebnih lastnih imenih te dileme ni.

## 4.2 Pridevnik

Za vsak pridevnik so v SIMlexu enako kot pri samostalniku izpisani vsi skloni in vsa števila, in sicer za vse tri spole. Ostale označene slovnične lastnosti so:

- podspol živost
- določnost/nedoločnost
- obrazilno in opisno stopnjevanje

### 4.2.1 Določnost/nedoločnost

Slovenska slovnica pri pridevniku omenja dve vrsti določnosti: "Glede na možnost izražanja nedoločnosti in določnosti z različnimi morfemi se pridevniki delijo na tri skupine: v prvi so tisti, ki imajo nedoločno in določno obliko, v drugi tisti s samo nedoločno obliko in v tretji oni s samo določno obliko: 1. *mlad - mladi, gozden - gozdni, uspel - uspeli, dvojen - dvojni*; 2. *bratov, sestrin, bukov*; 3. *slovenski, jelenji, lepši*. Pridevniki kot taki pa so pod 1. nedoločni, pod 2. in 3. določni." (Toporišič, 2000: 320) Kot navaja slovnica dalje, se glasovno določna in nedoločna oblika ločita le v imenovalniku in enako se glasečem tožilniku ednine za moški spol. Možnosti, kako označevati določnost v oblikoslovnem slovarju, je torej več: 1. pri pridevnikih se v vseh sklonih in spolih označi določnost glede na delitev v tri skupine, kot jo navaja slovnica (ibid.), 2. določnost/nedoločnost se označi samo tam, kjer se ta izraža glasovno, 3. določnost/nedoločnost se označi glede na stavek v slovnici: "Pridevniki kot taki..." (ibid.) Možnost 3 je bolj semantična kot oblikoslovna informacija, možnost 2 pa manj praktična in uporabna za razvoj jezikovnih tehnologij kot možnost 1. V SIMlexu je tako določnost označena samo tam, kjer se to loči po končnici. Glede na to odločitev nesklonljivim pridevnikom tipa *poceni*, ki ne ločijo določnosti/nedoločnosti glasovno, te kategorije ni smiselno označevati.

### 4.2.2 Stopnjevanje

Informacija o obrazilnem in opisnem stopnjevanju pridevnikov v SIMlexu obsega kvalifikator o vrsti stopnjevanja ter imenovalnik ednine osnovnika, primernika in presežnika oziroma obeh primernikov (*bolj, manj*) in presežnikov (*najbolj, najmanj*) pri opisnem stopnjevanju. Kot posebna lema so v prvi fazi obravnavani samo tisti primerniki in presežniki, ki so bili v osnovnem naboru. Kateri pridevnik se stopnjuje in kako, lahko natančneje preverjamo šele po izidu pravopisnega slovarja. V SIMlexu je vrsta stopnjevanja vedno vnesena tako, kot predvideva SP 2001, na podlagi primerov v korpusih pa so dodani: 1. pri večini pridevnikov, ki se stopnjujejo obrazilno, oblike za opisno stopnjevanje (npr. za moški spol pridevnika *prijazen* so izpisane stopnje *prijaznejši, najprijaznejši* in *bolj/manj/najbolj/najmanj prijazen*), 2. opisno ali obrazilno stopnjevanje ali oboje pri tistih pridevnikih, za katere je kakšna oblika primernika ali presežnika zabeležena v korpusih, v SP pa ni

zabeležena (npr. pri besedi *umetniški* – “vendar zato film ni bolj umetniški, ampak kvečjemu manj” (Internet #2)).

### 4.2.3 Vrsta pridevnika

Za posebno težavno in zelo povezano z določnostjo/nedoločnostjo se pokaže določanje vrste pridevnika. Poglejmo, kaj pravijo o tem pomembnejše slovenske slovnice.

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 in torej za nas najrelevantnejša 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 nastane 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*...). V slovnica za takšne primere žal ne najdemo zadovoljivega odgovora.

Ada Vidovič Muha (1981) določa vrsto pridevnika glede na besedno zvezo, v kateri se ta pojavi, in pridevnike v zgoraj navedenih besednih zvezah označi za vrstne. Z izpeljavo pa dokaže, da lahko celo 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)

Pri gradnji SImlexa smo imeli besede brez konteksta, torej bi morali s pomočjo korpusov predvideti najrazličnejše besedne zveze, v katerih bi se posamezen pridevnik lahko pojavil. Za skoraj vsak pridevnik bi tako morali predvideti dve ali celo tri vrste, s tem pa bi vnesli zmedo in nedoslednost. Za potrebe razvoja jezikovnih tehnologij je primernejša naslednja zelo poenostavljena delitev, ki se opira 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 *-o*, so določeni kot svojilni (npr. *sestrin*, *bratov*, *lipov*).

### 4.3 Glagol

V SIMlexu je za vsak glagol označeno, ali je glavni ali pomožni ali naklonski, določena sta glagolski vid in prehodnost. 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 in deležje na -e v osnovni obliki,
- vse tvorne sedanjske in velelniške oblike, oblike za preteklik, prihodnjik, predpreteklik, sedanji pogojniki in za pretekli pogojniki (v vseh osebah, številih in spolih),
- vse trpne oblike z deležnikom na -n/-t za sedanji, velelnik, preteklik, prihodnjik in sedanji pogojniki (v vseh spolih, številih in osebah).

#### 4.3.1 Vrste glagolov

Delitev na glavne, pomožne in naklonske glagole je običajna v podobnih slovarjih (Internet #1), tudi v mednarodnih projektih, kot je npr. Multext East (Internet #4), je zabeležena, zato jo je dobro vključiti v oblikoslovni slovar. V kakšno pomoč pa nam je lahko ta informacija pri razvoju jezikovnih tehnologij? Na prvi pogled se zdi najbolj smotno, če kot pomožni glagol označimo *biti*, saj z njim tvorimo čase in v stavku ne more biti brez dopolnila. Dopolnilo imajo tudi naklonski glagoli, vendar pa v slovenščini to ne more biti edina razlikovalna lastnost glede na polnopomenske (glavne) glagole, saj slovnica (Toporišič, 2000) našteva poleg naklonskih še trinajst vrst takih glagolov, pa še te niso nujno vse: "Vsega imamo torej 14 vrst pomožnih glagolov (in v precejšnji meri njim ustreznih drugih besednih vrst in zvez, zlasti povedkovnikov), morebiti pa jih je tudi še več." (ibid., 587) V jezikovni rabi pa se lahko prav vsi glagoli (tudi *biti*) uporabljajo tudi brez dopolnila (npr.: *Glagoli so besede, ki povedo, kaj kdo dela in kaj z njim je.*; *Smem?*). V SIMlexu je tako glagol *biti* vnesen kot glavni in kot pomožni glagol, za naklonske glagole pa so v prvi fazi označeni *morati*, *smeti*, *moči*, *hoteti*, *želeti* (primerjaj Toporišič, 2000: 401). Ali je ta rešitev ustrezna, se bo pokazalo z razvojem tehnologij in uporabo slovarja v praksi.

#### 4.3.2 Glagolska prehodnost

Slovenska slovnica loči več vrst prehodnosti (ibid., 354-355). Pri gradnji oblikoslovnega slovarja imamo možnosti, da (1.) kot prehodne označujemo samo direktno prehodne glagole ali (2.) vse glagole, ki lahko imajo ob sebi predmetno dopolnilo, ali pa (3.) ločeno označujemo obe vrsti prehodnosti. V SIMlexu je upoštevana možnost 2.

#### 4.3.3 Zložene glagolske oblike

Zložene glagolske oblike lahko označujemo po delih (posebej *sem* in posebej *šel*) ali kot eno enoto (*sem šel*). Pri oblikoslovnem označevanju besedila (kar je ena od faz označevanja besedila, potrebnih za razvoj tehnologije) se običajno označujejo po posameznih enotah, pri skladenjski analizi besedila pa je v pomoč, če računalnik že ima informacijo o celotni zloženi obliki, prav tako to pomaga pri prepoznavi govora/besedila. V SIMlexu so zato zložene glagolske oblike vpisane kot ena enota, ob tem pa so podatki vneseni tako, da lahko kasneje avtomatsko določimo potrebne oblikoslovne lastnosti vsake sestavne enote posebej.

#### 4.3.4 Trpnik

Prvi problem, povezan s trpnim glagolskim načinom, je, da lahko v slovenščini trpnik tvorimo: 1. z deležnikom na -n/-t, 2. s prostim morfemom *se*. Pri slednjem je treba ločiti, ali se s pomočjo *se* izraža trpnost ali pa gre za povratni glagol. Vnašanje te besedice tudi močno oteži avtomatsko generiranje glagolskih oblik, zato so v prvi verziji SIMlexa vnesene samo trpne oblike z deležnikom na -n/-t, ki

"ima širšo rabo" (Toporišič, 2000: 359). Pri povratnih glagolih, ki imajo ob sebi obvezno *se* (npr. *oglasiti se, ozirati se*), pa je bila ta besedica v prvi fazi ročno dodana.

Drugo vprašanje je, kateri glagoli lahko tvorijo trpnik in kateri ne. Ob primerih se je pokazalo, da ga tvorijo večinoma direktno prehodni glagoli (seveda ne vsi), med temi nedovršni glagoli redkeje. Po tem načelu in s pomočjo korpusov so trpne oblike vnesene v SIMlexu.

#### 4.4 Prislov

Delitev prislovov na vrste se v slovnici iz leta 1976 (Toporišič, 1976: 343-345) in zadnji (Toporišič, 2000) nekoliko razlikuje. Delitev v SIMlexu je povzeta po zadnji slovnici (ibid.), vendar je nekoliko poenostavljena: ločimo okoliščinske, časovne, lastnostne in vzročnostne prislove, slovnica pa dodatno razvršča prostorske in časovne prislove v skupen razred okoliščinskih, lastnostne in vzročnostne pa v skupen razred svojstvenostnih prislovov, poleg tega se vse štiri vrste, ki jih ločimo, delijo v podrazrede.

V SIMlexu je pri prislovih označen še izvor (samostalniški, pridevniški, glagolski, zaimenski ali števniki (ibid., 406)), stopnjevanje je vpisano po enakem načelu kot pri pridevniki (glej 4.2.2).

#### 4.5 Števniki in zaimek

Števniki in zaimki so v SIMlexu obravnavani kot posebna besedna vrsta, in ne v okviru pridevniške besede oziroma samostalniške in pridevniške besede ter prislova, kot je tradicija v jezikoslovju. Takšna delitev je bila izbrana, ker imajo oboji nekatere oblikoslovne posebnosti, poleg tega so enako razvrstitev uporabili v nekaterih mednarodnih projektih, npr. Multex-East (Internet #4). Pri zaimkih je informacija o pripadnosti besedni vrsti po tradicionalni slovnici ohranjena znotraj gesla s kvalifikatorji "samostalniški", "pridevniški" in "prislovni" zaimek. Opozoriti velja še na posebnost osebnih svojilnih pridevniških zaimkov, ki ločijo poleg "števila predmetov" (za prvo osebo *moj, moja, moji*) tudi "število lastnika" (*moj, najin, naš*), v tretji osebi ednine pa tudi "spol lastnika" (*njegov*, če je lastnik moškega, in *njen*, če je lastnica ženskega spola).

#### 4.6 Ostale besedne vrste

Od ostalih besednih vrst je najvprašljivejša kategorija povedkovnika. Čeprav v podobnih slovarjih (Internet #1; Internet #4) navadno ni upoštevana in je tudi s stališča jezikoslovja lahko vprašljiva, saj ima temelje drugod kot ostale besedne vrste, pa vseeno za silo rešuje vprašanje razvrstitve besed, kot so *treba, res, rad, vseč...*, ki v nobeno od ostalih besednih vrst ne spadajo povsem. Te in podobne besedice so tudi v SSKJ in SP 2001 velikokrat različno uvrščene (na primer besedico *treba* SSKJ uvršča med prislove, SP in slovnica (Toporišič, 2000) pa med povedkovnike). V SIMlexu so te besede označene kot povedkovnik in kot prislov ali katera druga najustreznejša besedna vrsta.

Tudi členki v SSKJ in SP 2001 niso usklajeno označeni: v SSKJ so velikokrat uvrščeni med prislove (npr. tudi členka *da, ne*). V takih primerih se SIMlex ravna po pravopisu 2001 in slovnici (Toporišič, 2000). Tako kot v pravopisu 2001 so označeni tudi prislovni zaimki - v SSKJ so še uvrščeni med prislove. Neuskklajeni so tudi primeri, kot npr. beseda *nekajkrat*: ta je po SSKJ prislov, enako po slovnici (Toporišič, 2000), po pravopisu 2001 pa kratnostni prislovni zaimek, kategorija, ki je slovnica ne omenja. Ti in podobni primeri kažejo, da bi bilo morda treba o statusu takšnih besed za potrebe jezikovnih tehnologij še enkrat premisliti.

## 5 GLASOSLOVNI SLOVAR - SIFLEX

Pri gradnji glasoslovnega slovarja za slovenski knjižni jezik je veliko manj jezikovnih dilem. Ker so v slovarjih SIMflex naglasne variante zajete že z oblikoslovnim slovarjem (kjer sta prav tako označena vrsta in mesto naglasa), so v glasoslovnem slovarju dodane samo knjižne izgovorne variante, in sicer:

1. posebne glasovne zveze (po SP, paragrafi 688 do 704), npr. če prideta skupaj črki *t* in *s*, sta predvidena izgovora z obema glasovoma ali z zlitim *c* – recimo za besedo *odsek* *O t - s \E k* in *O - ts \E k*; 2. premene po zvonečnosti (če se beseda konča na zvoneči soglasnik, se izgovori nezvoneči par, recimo samostalnik *breg* se izgovori *b r /e: k*); 3. 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*).

Fonetični simboli so usklajeni z abecedo SAMPA (Speech Assessment Methods Phonetic Alphabet) za slovenski jezik (Zemljak et al., 2002). Vsaka beseda je zlogovana.

## 6 ZAKLJUČEK

V članku predstavljamo jezikovna vprašanja, ki so se pojavljajo ob zasnovi in gradnji oblikoslovnega in glasoslovnega slovarja za slovenski knjižni jezik (SIMflex), ki ju od leta 1999 urejamo na Fakulteti za elektrotehniko, računalništvo in informatiko v Mariboru v Centru za jezikovne tehnologije. Slovarja sta zaključena v prvi fazi in po koncu te obsega SIMlex okoli 20.000 lem in 800.000 besednih oblik, SIFlex pa približno 170.000 enot. Delo seveda nadaljujemo z novim naborom besed.

Vsi jezikovni problemi so obravnavani s stališča, kako izbrati takšne rešitve, ki bodo najustreznejše glede na osnovni namen slovarjev: uporabo pri jezikovnih tehnologijah. Podrobnejšo oceno sedanje zasnove pa bomo lahko naredili ob uporabi slovarjev. Pri tem nikakor ne pričakujemo, da se bodo vse rešitve pokazale za najboljše, saj na tem področju še ni veliko izkušenj, iz katerih bi se lahko učili. Za slovenščino npr. nismo našli podobnih razprav drugih avtorjev, menimo pa, da so nujne za uspešno delo na tem področju. Če predpostavljamo, da imajo statistične metode pri razvoju jezikovnih tehnologij mejo uspešnosti, ki je ne bodo mogle preseči brez pomoči jezikovnih virov, je lahko uspešnost tehnologije odvisna prav od načina, kako je jezik v jezikovnih virih popisan. Ker je slednje predvsem delo jezikoslovca, bo prej ali slej pri razvoju jezikovnih tehnologij, če naj bo uspešen, potrebno širše sodelovanje jezikovne stroke.

## 7 ZAHVALA

Pri vnosu podatkov za SIMflex so sodelovale Alenka Januš, Branka Meolic, Tanja Šenveter, Barbara Volčjak in mag. Melita Zemljak, slednja je sodelovala tudi pri zasnovi osnovne zgradbe obeh slovarjev. Podjetje ČZP Večer je prispevalo besedilni korpus, tj. zbirko člankov dnevnega časopisa Večer od leta 1998 do 2000 v elektronski obliki.

## Literatura

- Bajec, A., Kolarič, R., Rupel, M. (1956). Slovenska slovnica, Ljubljana.
- Hajdinjak, M., Mihelič, F. (2002). Semantična analiza vremenskih napovedi. V: Informacijska družba IS'2002: Jezikovne tehnologije, 10.
- Internet #1 (04. 12. 2002). [www.cis.uni-muenchen.de/projects/CISLEX.html](http://www.cis.uni-muenchen.de/projects/CISLEX.html).
- Internet #2 (04. 12. 2002). [http://bos.zrc-sazu/s\\_beseda.html](http://bos.zrc-sazu/s_beseda.html).
- Internet #3 (04. 12. 2002). <http://www.fida.net/slo/>.
- Internet #4 (04. 12. 2002). <http://nl.ijs.si/ME/V2/msd/html/>.
- Jakopin, P., Bizjak, A. (1997). O strojno podprtem oblikoslovnem označevanju slovenskega besedila. V: Slavistična revija, 3-4, 513.
- Kačič, Z. (2002). Pomen združevanja raziskovalnih potencialov pri preseganju jezikovnih pregrad v okviru jezikovnih tehnologij naslednjih generacij. V: Informacijska družba IS'2002: Jezikovne tehnologije, 111.
- Rojc, M., Kačič, Z. [1]. (2000). A Computational Platform for development of Morphologic nad Phonetic lexica. V: Second International Conference on Language Resources and Evaluation, 277.
- Rojc, M., Kačič, Z. [2]. (2000). Design of Optimal Slovenian Speech Corpus for Use in the Concatenative Speech Synthesis System. V: Second International Conference on Language Resources and Evaluation, 321.
- Rojc, M., Kačič, Z., Verdonik, D. (2002). Design and Implementation of the Slovenian Phonetic and Morphology Lexicons for the Use in Spoken Language Applications. V: Third International Conference on Language Resources and Evaluation, 1296.
- Šef, T. (2001). Analiza besedila v postopku sinteze slovenskega govora. Doktorsko delo, FRI.
- Toporišič, J. (1976). Slovenska slovnica. Založba Obzorja, Maribor.
- Toporišič, J. (2000). Slovenska slovnica. Založba Obzorja, Maribor.
- Verdonik, D., Rojc, M., Kačič, Z. (2002). Zasnova in izgradnja oblikoslovnega in glasoslovnega slovarja za slovenski knjižni jezik. V: Informacijska družba IS'2002: Jezikovne tehnologije, 44.
- Vidovič Muha, A. (1981). Pomenske skupine nekakovostnih izpeljanih pridevnikov. V: Slavistična revija, 1, 19.
- Zemljak, M., Kačič, Z., Dobrišek, S., Gros, J., Weiss, P. (2002). Računalniški simbolni fonetični zapis slovenskega govora. V: Slavistična revija, 2, 159.