Leksikon besednih oblik Sloleks in smernice njegovega razvoja

Kaja Dobrovoljc, Simon Krek in Tomaž Erjavec

Abstract

This paper presents Sloleks, the largest open-source machine-readable morphological lexicon of the Slovenian language to date. The development of Sloleks and the formal grammar behind it are first briefly presented, before a detailed presentation of the types and structure of the inflectional, derivational, grammatical and other information included in each lexical entry is provided, with special emphasis on its formal representation within the standardised XML LMF framework. Given that Sloleks is a strong candidate to be used in the compilation of a new dictionary of contemporary Slovenian, both as a source of morphological information as well as part of the language technologies behind it, the second part of the presentation explores the most important aspects of its future development, particularly the expansion of its entry list, the addition of pronunciation information, the normative categorisation of variants and a corpus-based re-evaluation of existing inflectional paradigms. An extensive usage-based open-source morphological lexicon of modern Slovenian, with a universally unified system of morphological description, will therefore have a long-term application for language technologies and other born-digital reference works for the Slovenian language.

Keywords: morphological lexicon, lexicon of inflected forms, machine readable dictionary, morphology, inflection, derivation, pronunciation, language standardisation

Ključne besede: oblikoslovni leksikon, leksikon besednih oblik, strojno berljivi slovar, oblikoslovje, pregibanje, besedotvorje, izgovor, jezikovna standardizacija

1 UVOD

Pri oblikoslovno bogatih jezikih, kot je slovenščina, je opis oblikoslovnih paradigem pri pregibnih besednih vrstah tradicionalno zelo pomemben. Že Bohoričeva slovnica (1584) skoraj polovico opisa namenja poglavju o pregibanju besed oz. etimologiji, kot to imenuje Bohorič. Oblikoslovne paradigme imajo podobno prominentno vlogo tudi v večini kasnejših slovenskih slovnic. Te se osredotočajo predvsem na sistemske vidike oblikoslovja, torej na oblikoslovne vzorce, ki jih konkretno ponazorijo z nekaj primeri, kar pomeni, da je eksplicitnih izpisov celotnih paradigem v slovnicah malo. Po drugi strani so v preddigitalni dobi slovarji, predvsem v obliki pravopisnih priročnikov, kasneje pa SSKJ, kot popisovalci besedišča tradicionalno vsebovali tudi podatke o pregibanju. V njih so morfološki opisi močno okrajšani, poleg iztočnice navadno omejeni na eno ali nekaj oblik paradigme, na podlagi katerih naj bi uporabniki sklepali na celotno oblikoslovno paradigmo. Tudi po prenosu tiskanih priročnikov v digitalno obliko so ti podatki ostali identični.

S pojavom računalnikov in razvojem področja procesiranja besedil v naravnih jezikih je bila kmalu izpostavljena potreba po dostopnosti strojno berljivih slovarjev in leksikonov besednih oblik (Atkins in Zampolli 1994). Za angleščino so bili prvi strojno berljivi slovarji za različne jezikovnotehnološke naloge izdelani že v šestdesetih letih prejšnjega stoletja (npr. Boguraev in Briscoe 1987), s splošno digitalizacijo jezikov v devetdesetih letih pa so začeli nastajati tudi oblikoslovni leksikoni za večino drugih evropskih jezikov.

Ker za računalniške potrebe ni dovolj, da je na voljo le vzorec ali nekaj oblik, so v teh leksikonih paradigme glede na razbremenjenost prostorskih omejitev tiskanega medija tipično izpisane v celoti in se nahajajo v formatu, ki je strojno berljiv. Oblikoslovni podatki, tradicionalno vsebovani v slovnicah in slovarjih in namenjeni uporabnikom knjižnih jezikovnih priročnikov, so torej z računalnikom kot novim »uporabnikom« dobili tudi novo področje aplikacije. Leksikoni morajo tako hkrati zadovoljiti jezikovnotehnološke potrebe v različnih računalniških aplikacijah, od črkovalnikov in slovničnih označevalnikov do razpoznavalnikov in sintetizatorjev govora, strojnih prevajalnikov ipd., po drugi strani pa je zaželeno, da so uporabni tudi kot samostojni, jezikovnim uporabnikom namenjeni oblikoslovni priročniki. Sodoben računalniški leksikon slovenskega jezika naj bi torej bil namenjen zadovoljevanju obeh potreb in zato organiziran drugače kot oblikoslovni podatki v tradicionalnih slovarjih in slovnicah ter tudi kot prvotni strojno berljivi leksikoni.

Sledenje tema dvema ciljema vsebinsko zasnovo leksikona postavlja pred dve nasprotujoči si tendenci: v jezikovnotehnoloških aplikacijah mora leksikon čim

bolj uspešno opredeljevati oblikoslovne lastnosti vseh besednih oblik, ki jih srečamo v realnih besedilih, vključno z govorjenimi besedili, in omogočiti preprosto strojno berljivost pripisanih podatkov. Pri tradicionalni slovarski rabi pa mora zagotoviti učinkovito podajanje pregibnih, izgovornih in besedotvornih informacij za človeškega uporabnika, vključno z normativnimi vidiki besedišča. V kontekstu rabe leksikona za potrebe izdelave bodočega slovarja sodobnega slovenskega jezika mora biti leksikon v tem smislu vsebinsko usklajen z obema poloma: po eni strani z oblikoslovnimi podatki v učnem korpusu, s pomočjo katerega se oblikoslovni označevalniki učijo strojno označevati besedilne korpuse, iz katerih pridobivamo podatke za slovar, po drugi strani pa mora biti leksikon usklajen z leksikografskimi podatki v slovarski bazi ali drugih povezanih podatkovnih zbirkah.

Glede zadovoljevanja tradicionalnih jezikovnopriročniških potreb je pri oblikovanju vsebine referenčnega oblikoslovnega leksikona za sodobni slovenski jezik ključna težava v tem, da so obstoječi referenčni jezikovni priročniki, torej slovnice (npr. Toporišič 2004), slovarji (npr. SSKJ2) in pravopisi (npr. SP 2001) glede obravnave oblikoslovnih podatkov neusklajeni, celo kontradiktorni (prim. Krek 2014), kar pomeni, da za izhodišče ni mogoče vzeti nobenega od omenjenih del, temveč je treba koncept vsebinsko oblikovati na novo. Ti priročniki v pretežni meri tudi niso nastajali na podlagi sodobnega gradiva, zato so razmeroma oddaljeni od jezikovne realnosti sodobne slovenščine, po kakršni poizvedujejo tako jezikovnopriročniški kot jezikovnotehnološki uporabniki.

Strojno berljivi leksikoni besednih oblik za slovenščino imajo razmeroma dolgo zgodovino. Na začetku devetdesetih let je podjetje Amebis začelo razvijati elektronski slovar slovenskega jezika ASES, ki vsebuje tudi eksplicirane oblikoslovne paradigme (Arhar in Holozan 2009). Baza tega slovarja oz. leksikona ni prosto dostopna, podatke, ki jih vsebuje, pa je mogoče najti v različnih izdelkih podjetja, kot so slovnični pregledovalnik Besana, strojni prevajalnik Presis, sistem za komunikacijo v naravnem jeziku itd. Kronološko gledano je bil prvi prosto dostopni računalniški leksikon za slovenski jezik izdelan v okviru projekta MULTEXT-East v devetdesetih letih in vsebuje prek 15.000 osnovnih oblik ali lem in njihove pregibne paradigme v tabelaričnem formatu (Erjavec et al. 1995).

V prvem desetletju tega stoletja so z razvojem govornih tehnologij, predvsem sinteze govora, postali pomembni tudi leksikoni, ki poleg oblikoslovnih podatkov vsebujejo informacije o izgovoru, denimo SIflex, SImlex (Rojc et al. 2002; Verdonik et al. 2002), LC-STAR (Verdonik et al. 2004; Verdonik in Rojc 2004), SI-PRON (Žganec Gros et al. 2006) itd. Precejšnjo težavo pri vseh omenjenih leksikonih s podatki o izgovoru predstavlja dejstvo, da niti eden ni prosto dostopen. Enako je z dostopnostjo računalniškega oblikoslovnega leksikona, ki je

bil v približno istem času izdelan na Inštitutu za slovenski jezik Frana Ramovša, vendar o njem ni na voljo nobenih podatkov razen navedbe o obstoju (Naglič et al. 2005: 36).

Nekoliko bolj specifičen je še leksikon, ki je na voljo v prosto dostopnem sistemu za strojno prevajanje Apertium in obsega nekaj nad 20.000 lem (Horvat in Vičič 2012; Vičič 2012). Čeprav ta v osnovi izhaja iz leksikona MULTEXT-East, je glede vsebine in formata nekoliko drugačen, ker je v precejšnji meri vezan na prevajalni sistem, zato ni uporaben kot splošni leksikon za slovenščino. V okviru projekta Sporazumevanje v slovenskem jeziku je nastal računalniški leksikon Sloleks (Dobrovoljc et al. 2013), ki ga tudi postavljamo v središče tega prispevka, saj glede na svoj obseg, odprto dostopnost in rabo v temeljnih jezikovnotehnoloških orodjih za slovenščino predstavlja smiselno izhodišče za nadaljnji razvoj referenčnega leksikona besednih oblik za slovenščino.

2 LEKSIKON BESEDNIH OBLIK SLOLEKS

V pričujočem razdelku podrobneje predstavimo vsebino leksikona besednih oblik Sloleks, format njegovega zapisa, nabor in organiziranost podatkov posamične leksikonske enote ter način njihovega prikazovanja v obstoječem spletnem vmesniku.

2.1 Vsebina leksikona

2.1.1 Geslovnik in paradigme

Leksikon besednih oblik Sloleks v trenutni različici (Dobrovoljc et al. 2013) vsebuje 100.805 gesel, pri čemer eno geslo ustreza eni lemi, njeni pregibni paradigmi in drugim oblikoslovnim podatkom. Nabor iztočnic oz. lem je bil izdelan na podlagi meril, določenih v specifikacijah za izdelavo leksikona besednih oblik (Erjavec et al. 2008), in sicer je bila v leksikon najprej zajeta večina lem ročno označenega učnega korpusa ssj500k (Krek et al. 2013a), vse leme leksikalno zamejenih besednih vrst (predlog, veznik, zaimek, členek) ter izbrani težji primeri iz oblikoslovja npr. tuja lastna imena, enakovidski glagoli s homonimnimi nedoločniki (npr. *stati*), moški samostalniki z živo in neživo obliko v tožilniku ednine (npr. *delfin*), težavni primeri z visoko stopnjo variantnosti (npr. *otrok*) itd. Preostala večina geselskih iztočnic je bila nato izbrana glede na pogostost rabe na podlagi seznama najpogostejših lem v takratnem referenčnem korpusu FidaPLUS v obsegu 620 milijonov besed (Arhar in Gorjanc 2007).

V drugi fazi izdelave leksikona so bile lemam pripisane pregibne oblike, s programom za polavtomatsko dodajanje oblikoslovnih paradigem, ki ga je razvilo podjetje Amebis d. o. o. za izdelavo podatkovne zbirke ASES (Arhar in Holozan 2009) in na njej temelječih orodij. Leksikon besednih oblik Sloleks vsebuje skoraj 2.800.000 pregibnih oblik, točno sestavo leksikona glede na število lem in oblik posameznih besednih vrst prikazujemo v Tabeli 1.

Tabela 1: Število lem in oblik v leksikonu besednih oblik Sloleks v1.2.

Besedna vrsta	Število lem	Število oblik
samostalniki	54.260	924.268
pridevniki	26.612	1.571.970
glagoli	10.242	260.826
prislovi	6.906	9.931
števniki	2.240	18.448
zaimki	169	6.182
predlogi	96	101
medmeti	85	85
okrajšave	70	70
členki	68	68
vezniki	54	54
večbesedne enote[1]	3	3
SKUPAJ	**100.805**	**2.792.006**

2.1.2 Sistem JOS

Slovnične informacije v leksikonu besednih oblik temeljijo na oblikoskladenjskih specifikacijah, razvitih v okviru projekta Jezikoslovno označevanje slovenščine (Erjavec in Krek 2008; JOS), z namenom označevanja besednih pojavnic v slovenskih besedilih. Sistem JOS je rezultat daljše tradicije razvoja računalniških slovnic za slovenščino, saj je usklajen s specifikacijami projektov MULTEXT (Ide in Véronis 1994) in nato MULTEXT-East, pri čemer so specifikacije MULTEXT-East 4.0 (Erjavec 2012), ki med skupno 12 jeziki pokrivajo večino slovanskih jezikov, za slovenščino identične specifikacijam JOS.

Specifikacije JOS definirajo dvanajst besednih vrst: samostalnik, pridevnik, glagol, prislov, zaimek, števnik, predlog, veznik, členek, medmet, okrajšavo in neuvrščeno, pri čemer se zadnja v leksikonu ne uporablja. Z izjemo členkov, medmetov

1 Večbesedne enote so bile v leksikon vključene kot del poskusnih gesel na portalu Slogovni priročnik.

in okrajšav so večini besednih vrst pripisane dodatne oblikoskladenjske lastnosti, toda ni nujno, da so vsem besednim oblikam posamezne besedne vrste vedno pripisane vse možne lastnosti. Nabor možnih kombinacij slovničnih lastnosti je opredeljen v obliki vnaprej definiranega seznama[2] 1.902 kombinacij besednovrstnih kategorij, oblikoskladenjskih lastnosti in njihovih vrednosti, navodila za njihovo pripisovanje pojavnicam v besedilih pa so podrobneje opisana v navodilih za označevanje učnega korpusa (Holozan et al. 2008).

Kot ponazarjajo Erjavec et al. (2015), so bile oblikoskladenjske specifikacije sistema JOS razvite predvsem za potrebe strojnega označevanja besedil, zato zaradi omejenih možnosti označevalnih orodij na nekaterih mestih odstopajo od uveljavljenih slovnic slovenskega jezika (Ledinek 2014: 34–48). Pri določanju besedne vrste se tako upošteva predvsem oblika besede, manj pa njena skladenjska vloga v besedilu. Tipičen primer so denimo deležniške oblike na -n, -t ali -č, ki so ne glede na skladenjsko vlogo vedno označene kot deležniški pridevniki, saj na trenutni stopnji razvoja oblikoskladenjskih označevalnikov ni mogoče pričakovati dovolj zanesljivega razdvoumljanja njihove prilastkovne ali povedkovne vloge. Podobne poenostavitve so bile upoštevane tudi pri določanju posameznih oblikoskladenjskih lastnosti, kjer je denimo lastnost oseba pripisana vsem glagolskim pojavnicam v sedanjiku (tudi če so te brezosebne, npr. *dežuje*), (ne)določnost pa vsem pridevnikom, čeprav svojilni pridevniki te lastnosti ne razlikujejo.

S sistemom JOS so bila implicitno, skozi proces označevanja učnega korpusa in izdelave leksikona besednih oblik, opredeljena tudi načela določanja osnovnih oblik (lem) pojavnic pregibnih besednih vrst. Ta se večinoma skladajo z lematizacijskimi načeli v obstoječih priročnikih za slovenščino, npr. imenovalniška oblika ednine pri samostalnikih, nedoločniška oblika pri glagolih, nestopnjevana nedoločna oblika moškega spola ednine pri pridevnikih in besednih številnikih ter nestopnjevana oblika pri prislovih, z nekaj sistemskimi izjemami.[3] Posebnost so zaimki, pri katerih je lema odvisna od vrste zaimka in njegovih lastnosti (npr. leme *vame, zame, čezme* itd. za navezne osebne zaimke, ki se pregibajo po številu, osebi oz. spolu, ali lema *se* za povratne osebne zaimke *sebe/se, sebi/se, sabo/seboj*).

2.2 Zapis

Pri zasnovi leksikona kot referenčne zbirke oblikoslovnih, besedotvornih in drugih sorodnih podatkov o slovenskem jeziku je poleg proste dostopnosti kot

2 http://nl.ijs.si/jos/msd/html-sl/msd.index.msds.html

3 Pri množinskih samostalnikih je kot lema denimo izbrana oblika v imenovalniku množine (*alimenti*) oz. edina možna oblika (*poštev*). Pri prislovu so primerniške in presežniške oblike *bolj*, *manj*, *prej*, *raje*, *več*, *večkrat* oz. *najbolj*, *najmanj*, *najprej*, *najraje*, *največ* oz. *največkrat* zaradi specifičnih skladenjskih vlog osamosvojene v svojo lemo.

predpogoja za splošno rabo pomembno tudi upoštevanje standardov, ki omogočajo fleksibilno strukturiranje vsebine in mednarodno primerljivost podatkov.[4] Leksikon besednih oblik Sloleks je tako izhodiščno zapisan v označevalnem jeziku XML v shemi Lexical Markup Language (LMF), mednarodnem standardu za zapis strojno berljivih leksikalnih podatkovnih zbirk za potrebe procesiranja naravnih jezikov (ISO 24613:2008), ki je bil razvit z namenom vzpostavitve skupnega modela za izdelavo in uporabo eno- in večjezičnih leksikalnih virov, upravljanja izmenjave podatkov med dvema ali več viri ter združevanja večjega števila posameznih virov v obsežnejše globalne elektronske vire (Francopoulo et al. 2006: 1).

Format LMF sestavljajo t. i. jedrni sklop (angl. *core package*) in njegove razširitve (angl. *extensions*). Jedrni sklop predstavlja strukturno ogrodje, ki opisuje nabor in hierarhijo univerzalnih podatkov v leksikalni podatkovni zbirki, kot so informacija o jeziku, imenu in dostopnosti vira, ter osnovni strukturi leksikalnih enot, razširitve jedrnega sklopa pa nato določajo način kombiniranja strukturnih gradnikov (iz jedrnega sklopa) z drugimi elementi (iz razširitev) za potrebe specifičnega leksikalnega vira, kot je denimo oblikoslovni leksikon.[5]

Prilagoditev formata LMF za zapis oblikoslovnih podatkov morfološko bogatih jezikov, kakršna je bila upoštevana tudi pri izdelavi leksikona besednih oblik Sloleks, je podrobneje opisana v Krek in Erjavec (2009), celoten nabor pričakovanih elementov, atributov in možnih vrednosti ter njihova hierarhična razporeditev pa sta opisana v pripadajoči shemi DTD (*Document Type Definition*), ki je namenjena validaciji podatkov v bazi, torej preverjanju skladnosti njene strukture in vsebine z izhodiščnimi načeli.

2.3 Struktura leksikonske enote

Osnovno enoto leksikona besednih oblik, v kateri so strukturirani podatki enega gesla, imenujemo leksikonska enota. Ena leksikonska enota ustreza eni osnovni obliki oz. lemi in njeni oblikoslovni paradigmi, torej naboru ene ali več pregibnih oblik s pripadajočimi slovničnimi lastnostmi. Vsaka leksikonska enota obvezno vsebuje podatek o lemi, besedni vrsti in vsaj eni besedni obliki, glede na besedno vrsto in druge značilnosti leme pa je tej lahko pripisano še poljubno število dodatnih pregibnih oblik ter slovničnih in drugih lastnosti. V nadaljevanju na kratko opišemo nabor vseh tipov in hierarhična razmerja oblikoslovnih podatkov v leksikonu besednih oblik Sloleks in ponazorimo njihov zapis v formatu XML LMF.

4 Prvi prosto dostopni leksikoni so bili v shranjeni v tabelaričnem formatu, ki ni najboljši format za zapis možnosti, da npr. obstaja več variantnih besednih oblik z več izgovori, ti pa so denimo na kompleksen način povezane z drugimi oblikami.

5 Razširitve pri tem določajo predvsem pričakovani nabor elementov v določenem tipu vira, njihovo število in hierarhično urejenost, ne pa njihove semantične vsebine, saj so standardizirana poimenovanja jezikoslovnih kategorij, kot so poimenovanja besednih vrst, lastnosti in vrednosti, določena v registru podatkovnih kategorij ISOcat (http://www.isocat.org/).

2.3.1 Iztočnica

Iztočnica oz. ključ leksikonske enote je opredeljena kot unikatni identifikator, po katerem se leksikonske enote ločijo med seboj, saj med njimi ni mogoče ločevati zgolj na podlagi iztočniške leme, ki se lahko v enakem zapisu pojavlja v več leksikonskih enotah različnih besednih vrst (npr. prislov in členek *ravno,* prislov in samostalnik *stran,* prislov in pridevnik *spet*) ali znotraj iste besedne vrste (npr. dovršni in nedovršni glagol *zlagati*, deležniški in splošni pridevnik *poročen*, ženski in moški samostalnik *prst*). Čeprav je iztočnica namenjena predvsem strojni obdelavi podatkov, ne pa neposrednemu prikazovanju uporabnikom, je v leksikonu zasnovana tako, da lahko iz nje razberemo podatek o besedni vrsti in lemi (govoreča šifra), npr. S_avtomobil. V primeru, da se znotraj iste besedne vrste pojavlja več enakih lem, je temu zapisu dodana še zaporedna številka, npr. G_vesti_1 (*vesti: vezem*) ali G_vesti_2 (*vesti: vedem*).[6]

```
<LexicalEntry id="LE_ebc318126ea71205d05cd0ce85f86362">
          <feat att="ključ" val="R_pazljivo"/>
```

Slika 1: Zapis iztočnice prislova *pazljivo* v formatu XML LMF.

2.4.2 Lema

Osrednji gradnik leksikonske enote, na katerega se pripenjajo vsi drugi oblikoslovni podatki, je lema. Lema predstavlja neonaglašeno osnovno, kanonično oz. citatno besedno obliko, pod katero so združene vse druge oblike z enakimi leksikalnimi in paradigmatskimi lastnostmi, običajno pa tudi z enakim pomenom oz. pomeni. Določanje osnovne oblike v leksikonu besednih oblik Sloleks sledi lematizacijskim načelom sistema JOS, ki so bila upoštevana tudi pri lematizaciji učnega korpusa (Holozan et al. 2008) ter razvoju programa za samodejno lematizacijo in oblikoskladenjsko označevanje slovenskih besedil (Grčar et al. 2012).

```
<Lemma>
     <feat att="zapis oblike" val="pazljivo"/>
</Lemma>
```

Slika 2: Zapis leme prislova *pazljivo* v formatu XML LMF.

6 Posebnost so moški in ženski priimki, ki se v leksikonu vedno pojavljajo v paru in jim je zato namesto številke pripisana informacija o spolu, npr. S_Novak_m (Novak: Novaka) in S_Novak_ž (Novak: Novak). Kadar se pri priimkih pojavlja prekrivni samostalnik z enako lemo in spolom, se tudi tem iztočnicam doda zaporedna številka, npr. S_Pavlica_ž_1 (za nesklonljivi ženski priimek Pavlica) in S_Pavlica_ž_2 (za sklonljivo žensko ime Pavlica).

2.4.3 Besedna vrsta in leksikalne lastnosti

Obvezna slovnična lastnost vsake leksikonske enote in temeljna uvrščevalna lastnost leme je podatek o besedni vrsti, poleg tega pa je večini lem na prvi ravni strukture leksikonske enote pripisana še ena ali več leksikalnih lastnosti. Leksikalne lastnosti so tiste slovnične lastnosti, ki veljajo za vse oblike v paradigmi in jih pripišemo na ravni leme, npr. vrsta (občno ali lastno ime) in spol pri samostalnikih, vrsta (glavni ali pomožni) in vid (dovršni, nedovršni ali dvovidski) pri glagolih, sklon pri predlogih itd. Po vzoru formalnih slovničnih opisov so leksikalne in druge slovnične lastnosti zabeležene v obliki parov lastnosti (npr. *spol* pri samostalnikih) in njihovih vrednosti (npr. *moški*, *ženski* ali *srednji*).

```
<feat att="besedna_vrsta" val="prislov"/>
<feat att="vrsta" val="splošni"/>
```

Slika 3: Zapis leksikalnih lastnosti (vrste) prislova *pazljivo* v formatu XML LMF.

2.4.4 Pregibna paradigma

Splošnim podatkom o identifikatorjih, lemi in leksikalnih lastnostih sledi izpis pregibne paradigme, ki jo sestavljajo ena ali več pregibnih oblik, njihove partikularne oblikoslovne lastnosti, podatki o pogostosti v referenčnem korpusu in normativne lastnosti morebitnih variantnih oblik.

2.4.4.1 Pregibne oblike

Vsaka leksikonska enota ima v paradigmi[7] vsaj eno obliko. V primeru nepregibnih besednih vrst je ta oblika običajno samo ena, v primeru pregibnih vrst pa je teh oblik več, njihov obseg pa je odvisen od besedne vrste, leksikalnih lastnosti in stopnje variantnosti v jezikovni rabi. Med pregibnimi besednimi vrstami so tako najkrajše paradigme stopnjevanih prislovov in nekaterih zaimkov, najdaljše pa so paradigme pridevnikov, ki se pregibajo po spolu, stopnji, številu, sklonu in določnosti ter v povprečju obsegajo 59 različnih oblik (prim. Tabela 1).

2.4.4.2 Pregibne lastnosti

Posameznim oblikam so v paradigmi pripisane tudi pregibne slovnične lastnosti. V primerjavi z leksikalnimi lastnostmi so pregibne lastnosti tiste, po katerih

7 Z izrazom pregibna paradigma označujemo vse pregibne oblike leme, kot jih določa sistem JOS, ne glede na to, ali so rezultat oblikospreminjevalnih (npr. sklanjanje) ali oblikotvornih (npr. stopnjevanje) procesov.

se oblike v paradigmi določene leme z določenimi leksikalnimi lastnostmi ločijo med seboj, zato jih pripišemo na ravni (abstrahiranih) slovničnih oblik, npr. sklon, število in živost pri samostalnikih; stopnja pri prislovih; oblika, oseba, število, spol in nikalnost pri glagolih itd. Nabor pregibnih lastnosti v leksikonu besednih oblik Sloleks temelji na sistemu JOS, pri čemer ni nujno, da se vse pregibne lastnosti neke besedne vrste uresničujejo pri vseh njenih lemah, temveč je njihov dejanski nabor odvisen od same leme ali njenih leksikalnih lastnosti.
Na ravni pregibnih lastnosti je vključen tudi podatek o sintetični preslikavi vseh slovničnih lastnosti besedne oblike v t. i. oblikoskladenjsko oznako, kakršna se uporablja pri strojnem slovničnem označevanju korpusnih besedil (gl. prispevek Erjavec et al. 2015).[8]

```
<WordForm>
     <feat att="stopnja" val="primernik"/>
     <feat att="msd" val="Rsr"/>
     /.../
</WordForm>
```

Slika 4: Zapis pregibnih lastnosti primerniške oblike prislova *pazljivo* v formatu XML LMF.

2.4.4.3 Izrazna variantnost

Kadar enemu naboru lastnosti (eni abstraktni slovnični obliki) ustreza več oblik neke leme, govorimo o dveh ali več izraznih variantah oblik (oblikovnih dvojnicah), med katerimi ločujemo s pripisom t. i. variantnih lastnosti. V obstoječi različici leksikona so to normativne variantne lastnosti, ki označujejo zaznamovanost oblike glede na obstoječi pravopisni standard (SP 2001). Oblike brez pripisane normativne zaznamovanosti so v skladu s standardom (npr. *grad: gradu* v dajalniku ednine), medtem ko oblike s pripisom *nestandardno* niso v skladu standardom (npr. *grad: gradi* v imenovalniku množine). V primeru variantnosti dveh ali več standardnih oblik je vsem oblikam pripisana lastnost *variantno* (npr. *grad: grada* ali *grad: gradu* v rodilniku ednine).

8 Vsem primerniškim oblikam prislovov je denimo pripisana oblikoskladenjska oznaka Rsr, saj v skladu z oblikoskladenjskimi specifikacijami JOS prva črka oznake prinaša podatek o besedni vrsti oblike (R: prislov), pri čemer nato pri prislovih druga črka označuje vrsto (s: splošni), tretja pa stopnjo (r: primernik).

```
<FormRepresentation>
     <feat att="zapis_oblike" val="pazljiveje"/>
     <feat att="norma" val="variantno"/>
     <feat att="pogostnost" val="97"/>
</FormRepresentation>
<FormRepresentation>
     <feat att="zapis_oblike" val="pazljivejše"/>
     <feat att="norma" val="variantno"/>
     <feat att="pogostnost" val="2"/>
</FormRepresentation>
```

Slika 5: Zapis variantnih primerniških oblik prislova *pazljivo* z normativnimi in korpusnimi podatki v formatu XML LMF.

2.4.4.4 Korpusni podatki

Posameznemu zapisu oblike v leksikonu besednih oblik Sloleks je pripisan tudi podatek o njeni pogostosti v referenčnem korpusu. Ta je iz korpusa pridobljen avtomatsko, in sicer s poizvedbo, kolikokrat se dana oblika z dano lemo in oblikoskladenjsko oznako pojavi v korpusu. Pri presoji zanesljivosti te informacije moramo upoštevati omejitve avtomatskega označevanja korpusnih besedil, saj pojavnicam v korpusu niso nujno pripisane prava lema in slovnične lastnosti. Trenutna natančnost strojnega lematizatorja in označevalnika za slovenščino (Grčar et al. 2012), s katerim je bil označen referenčni korpus Gigafida, je 91,34 %, pri čemer se natančnost za različne skupine lem in oblik precej razlikuje (ibid.: 92–94) in lahko vpliva na ustrezno interpretacijo korpusnih podatkov (Logar et al. 2015).

2.4.5 Povezane oblike

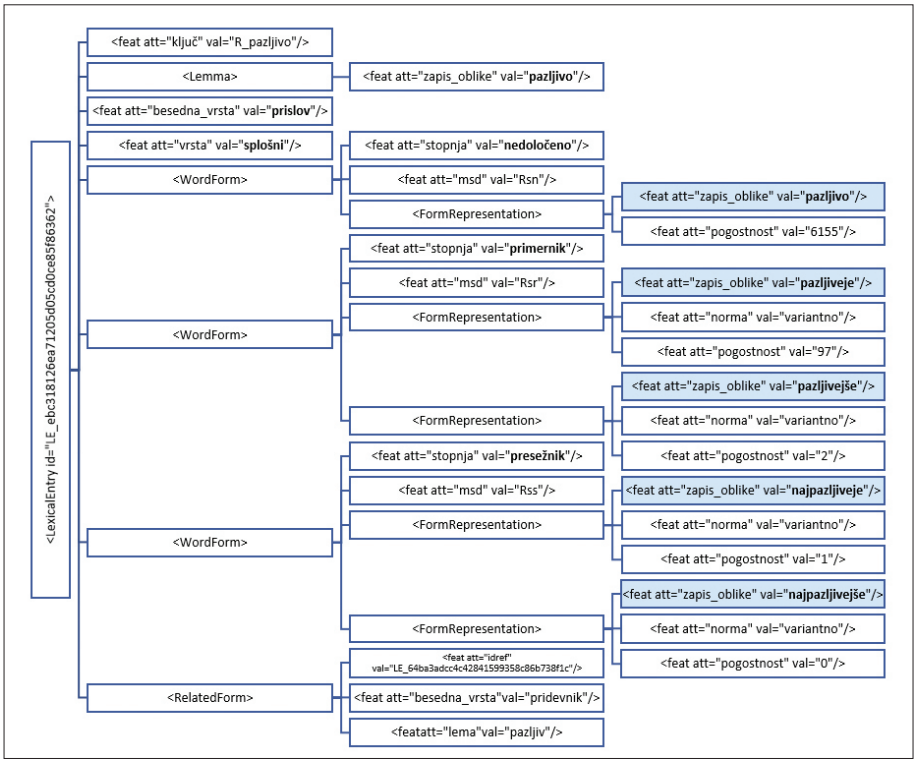
Poleg pregibnih oblikoslovnih lastnosti leme so v leksikonu besednih oblik Sloleks pri določenih skupinah lem beležene tudi informacije o njihovi besedotvorni povezanosti z drugimi lemami oz. leksikonskimi enotami. Trenutni nabor besedotvornih relacij v leksikonu besednih oblik Sloleks vključuje naslednje recipročne povezave: med samostalnikom in izpeljanim svojilnim pridevnikom (*kruh-kruhov*), med glagolom in izpeljanim glagolnikom (*briti--britje*), med pridevnikom in izpeljanim samostalnikom na -ost (*zarjavel-zarjavelost*), med glagolom in izpeljanim deležjem (*začeti-začenši*), med glagolom in izpeljanim deležnikom (*ujeti-ujet*), med pridevnikom in izpeljanim prislovom (*navihan-navihano*), med pridevnikom in izpeljanim elativom (*lep-prelep*),

med prislovom in izpeljanim elativom (*glasno-preglasno*) ter med lemo in njeno okrajšavo (*gospodična-gdč.*).

```
<RelatedForm>
     <feat att="idref" val="LE_64ba3adcc4c42841599358c8
     6b738f1c"/>
     <feat att="besedna_vrsta" val="pridevnik"/>
     <feat att="lema" val="pazljiv"/>
</RelatedForm>
```

Slika 6: Zapis povezanih iztočnic prislova *pazljivo* v formatu XML LMF.

V povzetek opisa strukture leksikonskih enot v leksikonu besednih oblik Sloleks na Sliki 7 prikazujemo celoten zapis leksikonske enote za prislov *pazljivo*, zaradi večje preglednosti tu v shematskem prikazu formata XML LMF.

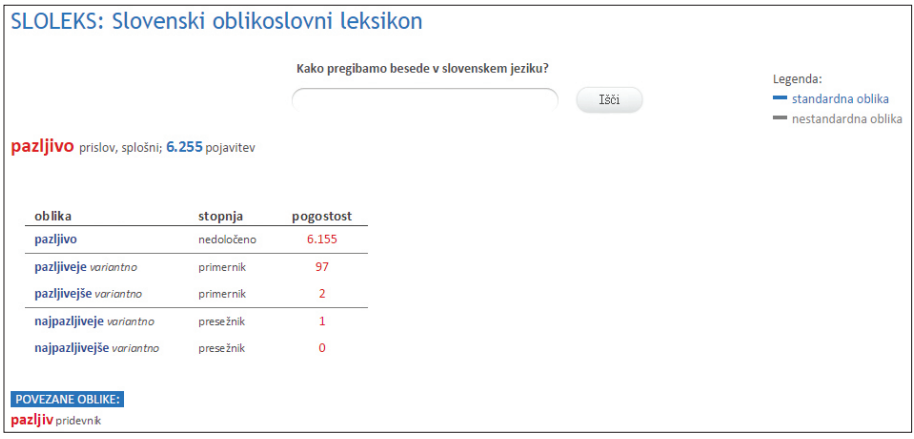


Slika 7: Prikaz strukture podatkov leksikonske enote v formatu XML LMF za prislov *pazljivo*.

2.5 Vizualizacija

Poleg uporabnosti v jezikovnotehnoloških aplikacijah strukturiran zapis informacij o oblikoslovnih lastnostih slovenskega besedišča prinaša številne prednosti tudi pri njegovi integraciji v druge priročnike ali samostojni vizualizaciji, saj nam omogoča poljubno določanje tako nabora prikazanih informacij kot tudi načina njihovega prikazovanja. Primer enega izmed načinov vizualizacije oblikoslovnega leksikona kot samostojnega oblikoslovnega priročnika je vizualizacija leksikona besednih oblik Sloleks na spletnem portalu projekta Sporazumevanje v slovenskem jeziku.[9]

Kot prikazuje posnetek prikaza leksikonske enote za prislov *pazljivo* (Slika 7) na Sliki 8, je podatek o lemi prikazan s poudarjeno pisavo, v isti vrstici pa mu sledita še podatka o besedni vrsti in vrsti ter skupna pogostost, ki v leksikonski enoti ni eksplicitno zapisana, a je izračunana na podlagi podatka o pogostosti posameznih pregibnih oblik. Sledi vizualno ločen prikaz pregibne paradigme z normativnimi in slovničnimi lastnostmi, pri čemer so posamezne kombinacije pregibnih lastnosti (slovnične oblike) ločene s črto. S klikom na podatek o pogostosti v referenčnem korpusu uporabnik dostopa do konkordanc v spletnem konkordančniku korpusa, ki jih na podlagi iskalnega pogoja v obliki dane kombinacije leme, oblike in oblikoskladenjske oznake mehanizem v korpusu poišče samodejno. Ob koncu gesla so prikazane tudi morebitne povezane leksikonske iztočnice z izpisano lemo in podatkom o besedni vrsti, ob kliku na lemo pa je uporabnik preusmerjen na ustrezno geslo.



Slika 8: Prikaz leksikonske enote za prislov *pazljivo* na spletnem portalu.

9 http://www.slovenscina.eu/sloleks (dostop 30. 6. 2015).

3 SMERNICE NADGRADNJE

3.1 Širitev geslovnika

Kot smo opisali v razdelku 2.1., leksikon besednih oblik Sloleks trenutno vsebuje približno 100.000 najpogostejših lem slovenskega besedišča in, v primerjavi z geslovniki drugih dostopnih oblikoslovnih priročnikov za slovenščino, ki so bodisi manjši po obsegu (Apertium, MULTEXT-East) ali niso nastajali na korpusni osnovi (SP 2001, SSKJ), pokriva doslej največji delež splošnega besedišča slovenskega jezika. Toda kljub temu je ob načrtovanju slovarskih in drugih jezikoslovnih opisov sodobne slovenščine na eni strani in upoštevanju naraščajočih in raznolikih potreb njenega strojnega procesiranja na drugi nujna tudi njegova nadaljnja širitev. To je smiselno načrtovati v obliki treh koncentričnih krogov, pri čemer vsak izmed krogov predstavlja temeljno izhodišče naslednjega, ni pa nujno, da so pri njihovi implementaciji uporabljena enaka vsebinska ali metodološka izhodišča.

V kontekstu prioritetne integracije oblikoslovnih podatkov v digitalno zasnovane priročniške opise sodobne slovenščine prvi koncentrični krog nadaljnjih širitev leksikona besednih oblik Sloleks predstavlja njegovo usklajevanje z geslovnikom slovarske baze, torej vključitev (manjkajočih) jedrnih leksikalnih enot slovenskega jezika, vključno z večbesednimi slovarskimi iztočnicami, variantnimi zapisi in drugimi lemami oz. oblikami, ki so oblikoslovno povezane z lemo neke slovarske iztočnice.

Drugi krog širitve geslovnika leksikona besednih oblik vključuje besedišče referenčnega korpusa slovenskega jezika. To glede na merila za zajem iztočnic namreč ni nujno tudi del slovarskega geslovnika, a je zaradi svoje pogostosti v rabi ključnega pomena za razvoj jezikovnih tehnologij, tudi tistih, ki se uporabljajo pri izdelavi slovarja, saj morajo lematizatorji, slovnični označevalniki in luščilniki leksikalnih podatkov poleg slovarske iztočnice natančno prepoznavati tudi besedišče v njeni okolici. V skladu s krepostnim krogom jezikoslovnega označevanja se s širitvijo leksikona izboljša jezikovni model označevalnika, z njim pa natančnost označevanja korpusa.

Primerjava prekrivnosti besednih oblik (različnic)[10] leksikona besednih oblik Sloleks in besedišča referenčnega korpusa Gigafida razkriva, da Sloleks vsebuje zgolj 43 % različnic, ki se v korpusu Gigafida pojavijo vsaj petkrat. Z višanjem frekvenčnega praga ta delež pričakovano narašča, vendarle pa leksikon besednih oblik pokriva zgolj 79 % od skupno 251.292 različnic, ki se v korpusu pojavijo najmanj 100-krat. Taka pogostost posamezne različnice (torej oblike, ne leme) v uravnoteženem in reprezentativnem korpusu jezika narekuje tudi potrebo po njenem formalnem opisu v ustrezni podatkovni bazi.

10 Pri tem smo namenoma primerjali zgolj besedne oblike z zapisom z malimi črkami, saj se nismo želeli opirati na strojno pripisan podatek o lemi ali posebnosti zapisovanja v korpusnih besedilih (npr. *slovenija*, *ljubljana* ipd.).

Podrobnejša analiza seznama najpogostejših različnic v korpusu Gigafida, ki jih v leksikonu besednih oblik Sloleks še ni, kaže, da bi to bazo v prihodnje veljalo dopolniti predvsem z naslednjimi skupinami besedišča:

- različnimi krajšavami (*p., s., j..; nan., dok., mr.; m2, cm3, a3; UV, MMS, VIP, SUV; VPS, SŽ* itd.),
- prevzetimi samostalniki (*city, miss, fax, art, dj, bluetooth, mac, facebook, prix, alias, maestro, college, gay, styling, fitness, volley, weekend, hiphop* itd.),
- nesklonljivimi prilastki (*turbo, online, anti, stereo, retro, audio, etno, latino, afro* itd.),
- nestandardnimi oblikami (*tud, kr, blo, brezveze, dobr, nevem, kao, jst, jap, tolk, nč, lahk, drgač, al, tm, zarad, mislm, pomoje, una, brezveze* itd.),
- medmeti (*živjo, bognedaj, jao, jp, hehe, he, hahaha, hahahaha, sviš, hehehe, khm* itd.),
- tujimi in domačimi lastnimi imeni (*obama, ilirika, evroliga, barca, clio, patria, beverly, pomurec, messi, airways, michel, svena, sarkozy, coca, evrovizija, titanik, čedad, wikipedia* itd.),
- zvrstno zaznamovanim besediščem (*škrinja, zalüblenih, mojoga, škürec, zadvečerek, špas* itd.),
- pa tudi z nekaterimi pogostimi domačimi oz. povsem podomačenimi besedami (*drugouvrščen, mimoidoči, prida, kapitalov, superpokal, štoparski, fotogalerija, tričetrt, bogve, drugoligaški, didžej, avtohiša, enoprostorec, osemvaljnik, supermodel, drska, preska, četrtinski, požarnik, klaviaturist, klientelizem, kapetanski, avtoprevozništvo, označba, predizbor, napak, prismučati, nezemljan, brezplačnik, evroobmočje, streljaj, dvetretjinski* itd.).

Za potrebe jezikovnih tehnologij je smiselno načrtovati tudi popis tujega besedišča, ki v slovenščino ni bilo prevzeto, a se v slovenskih besedilih pogosto pojavlja, denimo kot del tujih stvarnih lastnih imen (npr. *the, of, and* itd.).

Po geslovniku slovarske baze in pogostem besedišču referenčnega korpusa tretji krog razširitve geslovnika predvideva vključitev specializiranega besedišča za potrebe specifičnih jezikovnih priročnikov ali tehnoloških aplikacij, denimo tipično govorjenega besedišča, besedišča posameznih strokovnih področij, narečnega besedišča ali drugega zvrstno zaznamovanega besedišča. Za razliko od prvih dveh krogov, ki predstavljata univerzalno jedro opisa leksike nekega jezika, te tretje, domensko pogojene, širitve leksikona ni mogoče predvideti ali zagotoviti vnaprej. Ključnega pomena pa je, da je skupnosti omogočeno samostojno dopolnjevanje jedrnega leksikona, vključno z orodji in viri, ki jih za to potrebuje, začenši z odprtodostopno bazo pregibnih vzorcev.

3.2 Formalizacija oblikoslovnih vzorcev

Eno najpomembnejših vprašanj, povezanih tako s širitvijo kot reevalvacijo obstoječih oblikoslovnih leksikonov za slovenščino, je izdelava nabora strojno berljivih vzorcev pregibanja besed v slovenskem jeziku, ki bi omogočil validacijo pregibnih paradigem iztočnic v obstoječih priročnikih, pripisovanje paradigem novim lemam ter razvoj metod za njihovo samodejno prepoznavanje v besedilnih korpusih (prim. npr. Šnajder 2013 za hrvaščino).

Ob naboru oblikoslovnih leksikonov za slovenski jezik je mogoče sklepati, da v slovenskem prostoru že obstaja več tovrstnih zbirk pregibnih vzorcev, toda te širši raziskovalni skupnosti niso dostopne, načela njihovega oblikovanja, razvrščanja in usklajevanja z jezikovno rabo pa večinoma niso dokumentirana. Poskus implicitnega opisa vzorcev na podlagi luščenja pregibnostnih paradigem v večjih dostopnih podatkovnih zbirkah ali priročnikih, kot so SP2001, Apertium, pa tudi Sloleks (K. Dobrovoljc 2014), po drugi strani razkriva tudi nesistematičnosti in gradivne pomanjkljivosti, saj se v vseh virih pri naboru in razvrščanju vzorcev pojavljajo napake, nedoslednosti ali neskladnosti s sodobno jezikovno rabo.

To potrjuje, da je vsakršno načrtovanje izrabe ali nadgradnje obstoječih oblikoslovnih podatkovnih zbirk neločljivo povezano tudi z izdelavo aktualiziranega, prosto dostopnega seznama formaliziranih vzorcev pregibanja v slovenskem jeziku. Za razliko od tradicionalnih jezikoslovnih pristopov k opisu vzorcev pregibanja v slovenščini pa njihova jezikovnotehnološka namembnost narekuje upoštevanje spodaj opisanih načel ločene obravnave oblikovnih in izgovornih vzorcev, strojne berljivosti ter usklajenosti z jezikovno rabo.

3.2.1 Ločevanje oblikovnih in naglasnih vzorcev

Referenčni jezikovni priročniki za slovenščino tradicionalno pregibne izrazne lastnosti slovenskega besedišča opisujejo s hkratnim opazovanjem oblikovnih in izgovornih sprememb pri pregibanju, z razvrščanjem v t. i. sheme za dinamični naglas in oblikoslovje oz. sheme za tonemski naglas (prim. Rigler v SSKJ: XXXVIII–XLIX, LV–LVIII). Čeprav je tovrstno privzeto prikazovanje onaglašene paradigme delno vprašljivo tudi z vidika informativnosti za uporabnike (zlasti tuje, ki onaglaševanja niso vajeni), je predvsem z vidika uporabnosti v jezikovnotehnoloških aplikacijah smiselno pri opisu pregibnih lastnosti slovenskega jezika vzpostaviti ločnico med oblikoslovnimi vzorci v pisnem in naglasnimi vzorci v govorjenem jeziku. Z vidika strojnega procesiranja naravnih jezikov gre namreč

za dve ločeni ravni opisa (in procesiranja), ki sta medsebojno povezani, a izrazno neodvisni, saj se v jezikovni rabi vedno materializira le ena izmed obeh izraznih možnosti (zapis ali izgovor).

3.2.2 Strojna berljivost vzorcev

Drugi ključni vidik formalizacije pregibnih vzorcev je strojna berljivost njihovega zapisa, ki za razliko od diahrono ali pomensko motiviranih jezikoslovnih opisov vzorcev pregibanja v slovenščini zahteva skrajno formalistično pojmovanje osnove kot tistega nespremenljivega niza grafemov ali fonemov, ki je skupen vsem besednim oblikam v paradigmi, in obrazila kot tistega niza, ki je tej osnovi (krnu) dodan za tvorbo posamične oblike, ne glede, ali je dodan na koncu, začetku ali celo znotraj osnove. Strojno berljivi vzorec je tako zapisan v obliki algoritmičnih pravil, ki določajo načine krnjenja leme v osnovo in tvorjenja pregibnih oblik iz te osnove. Primer opisa pregibnih vzorcev za pregibanje po prvi ženski sklanjatvi, ki je prevedljiv v poljubni programski jezik, prikazuje Tabela 2.

To pomeni, da so tudi modifikacije ali posebnosti splošnih vzorcev, kot so premene, enoštevilske paradigme, raznospolske sklanjatve ipd., obravnavane kot samostojni pregibni vzorci. Z vidika njihove distribucije pa je nato smiselno ločevati med produktivnimi (sistemskimi) vzorci (ki se lahko povezujejo z odprto množico lem) in vzorci za izjeme (ki se povezujejo z zelo omejenim naborom lem). Tako produktivni vzorci kot vzorci za izjeme so formalizirani na enak način, informacija o njihovi omejeni distribuciji pa je lahko vzorcu pripisana v obliki neobvezujoče dodatne informacije ali v obliki restrikcije.[11]

Tabela 2: Primer opisa treh pregibnih vzorcev za samostalnike ženskega spola.

	vzorec ***lipa***		vzorec ***mravlja***		vzorec ***gora***	
osnova	lemi odstrani zadnjo črko	*lip*	lemi odstrani zadnjo črko	*mravlj*	lemi odstrani zadnjo črko	*gor*
Sozei	osnova + a	*lipa*	osnova + a	*mravlja*	osnova + a	*gora*
Sozer	osnova + e	*lipe*	osnova + e	*mravlje*	osnova + e	*gore*
Sozed	osnova + i	*lipi*	osnova + i	*mravlji*	osnova + i	*gori*
Sozet	osnova + o	*lipo*	osnova + o	*mravljo*	osnova + o	*goro*
Sozem	osnova + i	*lipi*	osnova + i	*mravlji*	osnova + i	*gori*

11 Na enak način lahko za potrebe različnih raziskav ali aplikacij vključimo tudi druge tipe metapodatkov, npr. o pričakovanih izraznih in slovničnih lastnostih lem, ki se pregibajo po določenem vzorcu, o medsebojni povezanosti vzorcev, o povezavah z naglasnimi vzorci ipd.

	vzorec ***lipa***		vzorec ***mravlja***		vzorec ***gora***	
Sozeo	osnova + o	*lipo*	osnova + o	*mravljo*	osnova + o	*goro*
Sozdi	osnova + i	*lipi*	osnova + i	*mravlji*	osnova + i	*gori*
Sozdr	osnova	*lip*	vrini *e* pred zadnji dve črki osnove	*mravelj*	osnova osnova + a	*gor* *gora*
Sozdd	osnova + ama	*lipama*	osnova + ama	*mravljama*	osnova + ama	*gorama*
Sozdt	osnova + i	*lipi*	osnova + i	*mravlji*	osnova + i	*gori*
Sozdm	osnova + ah	*lipah*	osnova + ah	*mravljah*	osnova + ah	*gorah*
Sozdo	osnova + ama	*lipama*	osnova + ama	*mravljama*	osnova + ama	*gorama*
Sozmi	osnova + e	*lipe*	osnova + e	*mravlje*	osnova + e	*gore*
Sozmr	osnova	*lip*	vrini *e* pred zadnji dve črki osnove	*mravelj*	osnova osnova + a	*gor* *gora*
Sozmd	osnova + am	*lipam*	osnova + am	*mravljam*	osnova + am	*goram*
Sozmt	osnova + e	*lipe*	osnova + e	*mravlje*	osnova + e	*gore*
Sozmm	osnova + ah	*lipah*	osnova + ah	*mravljah*	osnova + ah	*gorah*
Sozmo	osnova + ami	*lipami*	osnova + ami	*mravljami*	osnova + ami	*gorami*

Za ponazoritev razmerja med jezikoslovno in strojno motiviranimi opisi oblikoslovnih vzorcev v slovenščini vzemimo primer neonaglašenega pregibanja ženskih samostalnikov po prvi ženski sklanjatvi. Po Riglerjevih oblikoslovno-naglasnih shemah tej sklanjatvi ustreza pet shem (IAb1, IAb3, IB1b1, IB2a1 in IB2b1), pri čemer ima vsaka izmed shem enega ali več podtipov, oblikam v posameznih podtipih pa so lahko pripisane nadaljnje opombe z opisi ene ali več možnih modifikacij. Preslikavo razvezave tovrstnih oblikovno-naglasnih vzorcev v formalne vzorce pregibanja prikazujemo v Tabeli 3.[12] Vidimo lahko, da razmerje med obema načinoma opisa ni simetrično, saj imajo lahko posebnosti posameznih tipov drugačen formalni opis kot izhodiščni tip (npr. *ladja* drugače kot *lipa*, ali *mravlja* drugače kot *tabla*), prav tako pa ob ločevanju na oblikovne in naglasne vzorec nekateri naglasno razlikovalni tipi združijo v skupni oblikovni vzorec (npr. *gôra* in *stezà* v *steza*).

12 V primerjavo nismo vključili pridevniškega pregibanja (IAb1-podtip *désne*), tipa IAb1-agape in vzorcev za pregibanje po enem samem številu (*vile, grablje, nečke* ipd.).

Tabela 3: Primer preslikave Riglerjevih shem v formalizirane vzorce pregibanja po prvi ženski sklanjatvi.

Riglerjeve sheme		Formalni vzorci
1	shema IAb1, podtip *lípa*	vzorec ***lipa***
2	shema IAb1, podtip *lípa*, opomba 8, primer *óboj*	vzorec ***oboa***
3	shema IAb1, podtip *lípa*, opomba 8, primer *dékel*	vzorec ***dekla***
4	shema IAb1, podtip *lípa*, opomba 8, primer *grábelj*	vzorec ***mravlja***
5	shema IAb1, podtip *lípa*, opomba 8, primer *ládij*	vzorec ***ladja***
6	shema IAb1, podtip *lípa*, opomba 8, primer *kámer*	(vzorec *dekla*)
7	shema IAb1, podtip *lípa*, opomba 8, primer *zárij*	(vzorec *ladja*)
8	shema IAb1, podtip *lípa*, opomba 8, primer *mrávelj/márenj*	(vzorec *mravlja*)
9	shema IAb3, podtip *búkev*	vzorec ***bukev***
10	shema IAb3, podtip *cerkvé*	vzorec ***cerkev***
11	shema IB1b1, podtip *stezà*	vzorec ***steza***
12	shema IB1b1, podtip *stezà,* opomba 19	(vzorec *steza*)
13	shema IB1b1, podtip *stezà,* opomba 20	(vzorec *steza*)
14	shema IB2a1, podtip *gospá*	vzorec ***gospa***
15	shema IB2b1, podtip *gôra*	(vzorec *steza*)
16	shema IB2b1, podtip *nečké*	(vzorec *lipa*)

Kot smo izpostavili, bi morali za potrebe formalizacije na enak način opisati tudi vzorce za izjeme, ki se povezujejo zgolj s posameznimi lemami (npr. *ovca, mati, hči,* ipd.) in v obstoječih priročnikih niso omenjene kot del shem, temveč v geselskem zaglavju (npr. *ôvca -e stil. -é ž, rod. mn. ôvc in ovác (ó)* v SSKJ2). Njihov zapis v obliki samostojnega vzorca je poleg sistematičnosti smiseln tudi zato, ker tudi vzorci za izjeme izkazujejo določeno mero sistemskosti in se običajno povezujejo z več kot zgolj eno samo lemo (npr. *deska* tako kot *ovca*, *pramati* tako kot *mati* ipd.).

3.2.3 Skladnost z jezikovno rabo

Tretje ključno načelo, ki ga je treba upoštevati pri izdelavi nabora formaliziranih oblikoslovnih vzorcev za slovenščino, pa je upoštevanje oblikoslovnih tendenc v sodobni jezikovni rabi, o kakršni lahko sklepamo na podlagi uravnoteženih in reprezentativnih korpusov sodobne slovenščine. Z vidika jezikovnih tehnologij je upoštevanje jezikovne rabe pomembno zaradi zagotavljanja čim večje pokritosti pogosto rabljenega besedišča (ne glede na njegovo skladnost z obstoječo

kodifikacijsko normo), za jezikoslovje pa korpusni pristop k opisu oblikoslovja slovenskega jezika ponuja priložnost za aktualizacijo obstoječih opisov, ki niso nastali na tako obsežni gradivni osnovi.

Analiza rabe lahko prinaša nova spoznanja že na ravni opisa vzorcev pregibanja. Če za primer vzamemo zgolj zgoraj opisani nabor vzorcev za prvo žensko sklanjatev, analiza rabe v korpusu Gigafida denimo pokaže, da se domnevno sistemski podtip sheme IAb3, po katerem naj bi se samostalniki ženskega spola na *-ev* v rodilniku dvojine ali množine pojavljali tudi s končnico -á (*cérkev*: *cérkev* in *cerkvá*) pojavlja zgolj pri ponazoritveni lemi (gre torej za izjemo in ne pravilo), prav tako pa se v tožilniku dvojine v korpusu ne pojavlja navedena stilistična oblika *cerkvé*. Podobno analiza korpusa pod vprašaj postavlja tudi trditev, da se v rodilniku dvojine in množine med dva zvočnika *e* vriva samo v primeru, kadar je drugi zvočnik *r* (*kamra: kamer*), saj raba kaže, da do vrivanja *e* lahko prihaja tudi pri nekaterih drugih zvočniških sklopih (npr. *himna: himen*, *kolumna: kolumen*; *avla: avel*).

Še zlasti pa je analiza rabe pomembna z vidika oblikoslovnega razvrščanja leksike, torej povezovanja konkretnih lem s konkretnimi vzorci pregibanja. Za primer vzemimo obrazilno stopnjevanje prislovov, kjer slovenščina pozna dva sistemska vzorca: bodisi se prislovi ne stopnjujejo bodisi se stopnjujejo z variantnima končnicama *-ejše* oz. *-eje*, če so tvorjeni iz pridevnika in njihov pomen to dopušča (*zanimivo: zanimivejše/zanimiveje*). Analiza razvrščanja obeh sistemskih vzorcev v leksikonu besednih oblik Sloleks in slovarju Slovenskega pravopisa (K. Dobrovoljc 2014) kaže, da v obeh priročnikih prihaja do razhajanj z rabo, saj so nekateri v rabi stopnjevani prislovi v enem ali obeh priročnikih navedeni brez stopnjevanih oblik (npr. *smiselno, preudarno, poredko, enakovredno, korektno, športno* itd.) ali pa je stopnjevanost pripisana tistim prislovom, ki v rabi ne izkazujejo nobenega izmed variantnih obrazil (*arogantno, bistroumno, strahovito, zagonetno* itd.).

Še posebej pa mora biti skladnost s sodobno jezikovno rabo merilo pri določanju izjem. Pri stopnjevanju prislovov tako Slovenski pravopis za prislove *drago, ozko* in *težko* denimo navaja oblike *dražje, ožje* oz. *težje* (čeprav so v rabi tudi oblike *draže, ože* oz. *teže*), za prislov *kratko* obliki *krajše* in *kračje* (čeprav slednje v korpusu ni), za prislov *gladko* oblike *gladkeje, gladkejše, glaje* in *glajše* (čeprav se oblika *glaje* v referenčnem korpusu ne pojavi, oblika *glajše* pa samo enkrat) in tako dalje.

3.3 Dodajanje izgovora

Leksikon besednih oblik Sloleks trenutno ne vsebuje podatkov o izgovornih lastnostih vsebovanih besednih oblik, kar pomeni, da so zapisi osnovnih in

pregibnih oblik neonaglašeni. Z namenom celovitega opisa izraznih lastnosti sodobnega slovenskega besedišča je torej vključitev podatkov o izgovoru besednih oblik ena izmed prioritetnih nadgradenj obstoječe različice leksikona besednih oblik Sloleks. To je še toliko bolj pomembno z vidika govornih tehnologij, saj v slovenskem prostoru trenutno ne obstaja prosto dostopni leksikon, ki bi omogočal razvoj strojnih razpoznavalnikov in sintetizatorjev govora za potrebe najrazličnejših aplikacij, kot so samodejni podnaslavljalniki, bralniki za slepe in slabovidne, sistemi za komuniciranje v naravnem jeziku ipd.

3.3.1 Fonetični zapis

Podatek o izgovoru bi moral biti v leksikonu besednih oblik Sloleks beležen v obliki fonetične transkripcije v dogovorno izbrani strojno berljivi mednarodni fonetični abecedi (prim. Jurgec 2015), ki poleg zapisa glasov inherentno vsebuje tudi podatek o mestu in jakosti naglasa. Glede na specifične jezikovnopriročniške ali jezikovnotehnološke potrebe se referenčni fonetični zapis lahko na podlagi pravil pretvori tudi v druge fonetične abecede ali zapise onaglašenih oblik. Ti so lahko, ni pa nujno, v obliki samostojnih elementov zabeleženi tudi v samem leksikonu besednih oblik, pomembno pa je, da so ves čas usklajeni z izgovorom v referenčni izhodiščni fonetični transkripciji. Na enak način je mogoče vključevati tudi podatek o tonemskem naglasu, ki sicer glede na potrebe strojnih in človeških uporabnikov ni prioriteten (gl. Arhar et al. 2015).

3.3.2 Strukturna umestitev

Kot smo že opisali v razdelku 3.2.3, v leksikonu besednih oblik Sloleks izgovor obravnavamo kot osamosvojeno informacijo o glasovni podobi posamezne neonaglašene pisne oblike. Podatek o izgovoru tako pripisujemo na ravni obstoječega zapisa osnovne oz. pregibnih oblik. V primeru v slovenščini zelo pogoste izgovorne variantnosti je lahko eni neonaglašeni obliki pripisanih tudi več elementov s podatkom o izgovoru, med njimi pa tako kot pri variantnosti neonaglašenih oblik ločujemo z ustreznimi kvalifikatorji, ki nam omogočajo samodejni priklic izgovora posamezne ali vseh oblik v eni izmed variantnih izgovornih paradigem (glej razdelek 3.4). Na ta način obravnavamo vse tipe izgovorne variantnosti, ne glede na to, ali gre za glasovno (prevajalka: *prevajalka-pravajaụ̯ka*) oz. naglasno variantnost (agencija: *agéncija-agencíja*) vseh ali zgolj ene izmed pregibnih oblik v paradigmi.

3.3.3 Obravnava enakopisnic

Tudi po vključitvi informacij o izgovoru so leme z enakim zapisom in izgovorom ločene v več samostojnih leksikonskih enot, če izkazujejo različne izrazne lastnosti, tj. spadajo v različne besedne vrste (prislov in pridevnik *spet*), imajo različne leksikalne lastnosti (ženski in moški samostalnik *prst*) ali se drugače pregibajo (*vesti*: *vedem* in *vesti: vezem*). Po drugi strani v leksikonu besednih oblik Sloleks ne ločujemo med homonimnimi lemami s povsem prekrivnimi izraznimi, a različnimi pomenskimi lastnostmi (npr. moški samostalnik *bor*-drevo ali *bor*-element), zato take pare leksemov še naprej obravnavamo kot eno samo izrazno enoto besedišča, ki ji ustreza ena sama leksikonska enota (samostalnik moškega spola *bor*).[13] Ob dejstvu, da v leksikonu besednih oblik ne beležimo tonemskosti, enako velja tudi za pare pomensko različnih enakopisnic, ki se razlikujejo zgolj v tonemskem naglasu (pregibanje pridevnikov *bȗčen*-nanašajoč se na bučo in *búčen*-glasen opišemo v skupni leksikonski enoti splošnega pridevnika *bučen*).[14]

Po drugi strani pa vključevanje informacij o izgovoru spreminja način obravnave lem z enakim zapisom in drugačnim izgovorom, ki so bili doslej v primeru prekrivne leme, slovničnih lastnosti in neonaglašene pregibne paradigme obravnavani kot ena sama leksikonska enota, npr. *partija* (*partíja* in *pártija*), *častiti* (*částiti* in *častíti*) itd. Z uvedbo pomensko razločevalnih podatkov o izgovoru se oba leksema osamosvojita v samostojni leksikonski enoti (*S_partija_1* in *S_partija_2*), ob tem pa moramo upoštevati, da trenutni slovnični označevalniki za slovenščino ne vključujejo pomenskega razdvoumljanja oblikovno prekrivnih enakopisnic v kontekstu, zato so pojavnice tovrstnih parov označene z identično lemo in oblikoskladenjsko oznako. To pomeni, da je v leksikonu besednih oblik enakim oblikam ene in druge leme pripisana enaka korpusna pogostnost, pri luščenju informacij o besedilnem okolju, zgledov in drugih vrstah besedilnih podatkov pa med njima ni mogoče ločevati zgolj z avtomatskimi postopki.

3.4 Kategorizacija variantnosti

Oblikoslovna variantnost, tj. več izraznih možnosti iste slovnične oblike, je v slovenščini zelo pogosta in do nje prihaja na različnih ravneh: pri zapisu (*v naprej* ali *vnaprej*), izgovoru glasov (*/drsáuka/* ali */drsálka/*) ali naglaševanju leme (*upokójenec* ali *upokojênec*), kakor tudi pri izbiri oblikoslovne paradigme (*Luka: Luka, Luke* ali *Lukata*), zapisa ali izgovora pregibnih oblik (*college*: *collegea* ali *collega*) ter besedotvorju (*vanilija: vanilijev, vanilijin* ali *vanilin*).

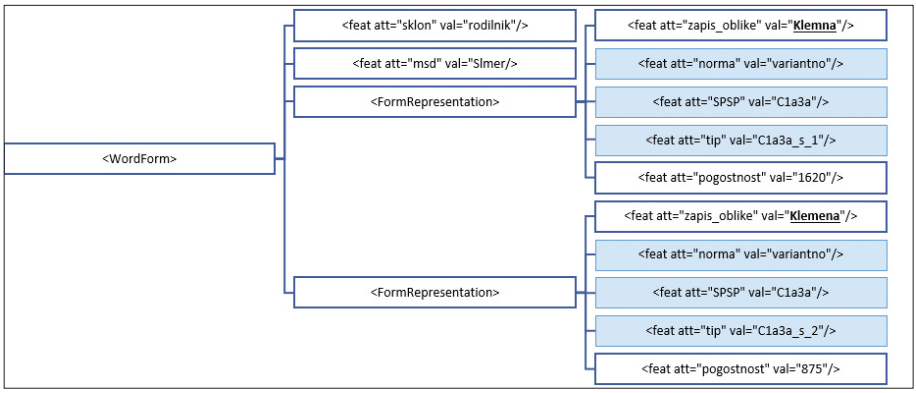
13 Za razmerje med leksikonsko in slovarsko iztočnico gl. prispevek K. Dobrovoljc (2015).

14 Za razumevanje preslikav med formalnooblikoslovno motiviranimi leksikonskimi enotami in pomensko motiviranimi slovarskimi enotami gl. Gantar (2015) in K. Dobrovoljc (2015).

Če bi v rabi izkazane oblikoslovne variante v leksikon besednih oblik vključevali zgolj kot niz več zapisovalnih ali izgovornih oblik z identičnimi slovničnimi lastnostmi, med njimi brez dodatnih pravil ne bi mogli sistematično ločevati, zato je z vidika različnih namenov uporabe oblikoslovnih leksikonov koristno, da variantam pripišemo tudi njihove razlikovalne (variantne) lastnosti oz. jih ustrezno tipologiziramo. Pri tem je treba poudariti, da tipologizacije ne smemo enačiti z normativno kvalifikacijo, saj prva označuje jezikovnosistemsko izbiro, druga pa njeno naknadno jezikoslovno interpretacijo, ki je družbenokonsenzualno pogojena in s tem bolj ali manj spremenljiva. Obe informaciji sta v leksikonu besednih oblik nepogrešljivi, saj tipologizacija omogoča usmerjen priklic posamičnih variantnih oblik ali celotnih variantnih paradigem ene ali več leksikonskih enot, podatek o njihovi normativni (ne)zaznamovanosti pa je ključen pri integraciji leksikona v jezikovne priročnike, navsezadnje pa informacijo o (ne)standardnosti posameznih variantnih izbir potrebujejo tudi jezikovnotehnološke aplikacije, ki generirajo besedila v slovenščini, kot so strojni prevajalniki, sintetizatorji govora ipd.

Dodajanje podatkov o tipih in normativni zaznamovanosti variantnih oblik smo v leksikonu besednih oblik Sloleks že preizkusili pri vzpostavitvi zasnove in delotoka spletnega portala Slogovni priročnik, namenjenega reševanju najpogostejših jezikovnih zadreg pri tvorbi besedil v slovenščini s sopostavljanjem informacije o veljavnem pravopisnem standardu in korpusnih podatkov (Krek 2012; Krek et al. 2013a; K. Dobrovoljc in Krek 2013). Zaledni mehanizem, ki uporabnikovo vprašanje poveže z ustrezno jezikovno zadrego in njenim pojasnilom ter vizualizira korpusne in normativne podatke za konkretno obliko ali paradigmo, po kateri sprašuje uporabnik, vse potrebne podatke črpa iz leksikona besednih oblik Sloleks, v katerem so leme ali pregibne oblike, povezane z eno izmed obravnavanih jezikovnih zadreg, ustrezno kategorizirane. Vsaki obliki (osnovni ali pregibni) so tako pripisani trije tipi kategorizacijskih podatkov: (i) kategorija jezikovne zadrege oz. variantne izbire, ki temelji na ontološko urejenem seznamu jezikovnih zadreg v slovenščini (H. Dobrovoljc in Krek 2011; Bizjak Končar et al. 2011), (ii) tip sistemske variante znotraj kategorije in (iii) njena normativna vrednost.

Primer take kategorizacije prikazuje izsek zapisa leksikonske enote za samostalnik *Klemen* na Sliki 9. Leksikonski enoti je že na prvem nivoju pripisan podatek, da se povezuje z jezikovno zadrego C1a3a (*Oblikoslovje* > *Samostalniki* > *Moške sklanjatve* > *Samostalniki z neobstojnim samoglasnikom* > *Slovenska lastna imena*), posameznim oblikam v paradigmi pa je poleg podatka, da spadajo v to kategorijo variantnosti v elementu z opredelitvijo tipa pripisan še podatek, kateri izmed obeh variant pripada (C1a3a_s_1 denimo označuje paradigmo z izpustom *e*, C1a3a_s_2 pa paradigmo brez izpusta), v elementu z opredelitvijo norme pa še ustrezni normativni kvalifikator *variantno,* ki označuje standardno dvojnico.



Slika 9: Del leksikonske enote *S_Klemen* s pripisom kategorije variantnega sklanjanja.

Tovrstna kategorizacija variantnosti v bazi nam tako po eni strani omogoča nadzorovan priklic podatkov posamezne leksikonske enote, kot je denimo seznam vseh jezikovnih zadreg, s katerimi se ta povezuje, ali ene oz. več oblik njene variantne paradigme, po drugi strani pa nam omogoča tudi samodejni priklic seznama vseh drugih lem, ki izkazujejo enako vrsto oblikovne, izgovorne ali besedotvorne variantnosti, npr. vseh slovenskih lastnih imen z neobstojnim samoglasnikom.

4 ZAKLJUČEK

Leksikon besednih oblik Sloleks z oblikoslovnimi, besedotvornimi, izgovornimi, normativnimi in drugimi podatki predstavlja vezni člen med različnimi jezikovnimi viri, ki jih predvideva Predlog za izdelavo slovarja sodobnega slovenskega jezika (Krek et al. 2013b). Predvsem so to različni jezikoslovno označeni korpusi slovenščine, od referenčnega, uravnoteženega, govornega do zgodovinskega in vseh ostalih. Na drugi strani so podatki iz leksikona neposredno uporabni in uporabljeni v priročniških virih, kot je (digitalni) slovar, spletni slogovni priročnik in drugi. Z enotno obravnavo morfologije slovenskega jezika leksikon v celotni označevalni in priročniški ekosistem prinaša konsistentno obravnavo pojavov, kar sicer v tem trenutku predstavlja enega ključnih primanjkljajev tako znotraj računalniškega procesiranja slovenščine kot tudi poučevanja slovenščine v celotnem izobraževalnem procesu, od osnovnih in srednjih šol do poučevanja slovenščine kot tujega jezika.

Obstoječi leksikon besednih oblik Sloleks predstavlja dobro osnovo za nadgradnjo, ki obsega predvsem obsežno širjenje geslovnika in oblik, vključno z

večbesednimi enotami, ter dodajanje informacij o izgovoru in normi. Vsi ti procesi morajo biti čim bolj avtomatizirani, kar obstoječe tehnologije že zagotavljajo, manjka pa strojno berljiv in na sodobni jezikovni rabi utemeljen popis pregibnih vzorcev za slovenščino. Nadgrajevanje je treba razumeti kot permanenten proces brez končne točke v času, saj v jezik vedno prihajajo nove besede, ki jih je treba tako opisati kot tudi zagotoviti njihovo strojno obdelavo v avtentičnih besedilih. Na ta način je koncept leksikona besednih oblik Sloleks tudi zastavljen. Najbolj ključen pri zagotavljanju širše rabe leksikona je dostop pod odprto licenco, saj šele ta zagotavlja, da se investicija v njegov razvoj tudi upraviči, v širši perspektivi pa slovenščini zagotavlja možnost za obstanek tudi v prihajajočem, vse bolj digitaliziranem svetu.