Špela Arhar
Trojina, zavod za uporabno slovenistiko

UDK 811.163.6'322.2

UČNI KORPUS SSJ IN LEKSIKON BESEDNIH OBLIK ZA SLOVENŠČINO

Glavni namen prispevka je predstavitev priprave učnega korpusa ter leksikona besednih oblik za slovenščino. 400.000 besed obsegajoči korpus SSJ predvideva štirinivojsko označenost: lematizacijo, označenost na oblikoskladenjski ter skladenjski ravni ter označenost lastnih imen. Vse oznake bodo ročno pregledane. Skupaj s korpusom JOS100k tvori korpus SSJ polmilijonski učni korpus za učenje statističnih modelov za npr. oblikoskladenjsko označevanje ter skladenjsko razčlenjevanje slovenščine. Leksikon besednih oblik bo prinašal okvirno 100.000 leksikonskih enot, vsebujočih oblikoslovne paradigme posameznih besed z naborom informacij, prekrivnih s sistemom oblikoskladenjskega označevanja JOS. Predvidena je vključitev informacij o (besedotvorni) povezanosti leksikonskih enot, v primeru v jezikovni rabi izpričane oblikovne variantnosti pa bodo vključeni tudi podatki o pogostnosti oblik ter njihovi trenutni opredeljenosti v normativnih virih. Vključitev večbesednih enot je predvidena na ravni večbesednih lastnih imen ter oblik, ki se variantno pišejo skupaj oziroma narazen.

Ključne besede: učni korpus, označevanje korpusa, oblikoskladenjsko označevanje, skladenjsko označevanje, označevanje lastnih imen, leksikon besednih oblik

1 Uvod

Prispevek predstavlja aktivnosti projekta *Sporazumevanje v slovenskem jeziku* (SSJ),[1] ki so vezane na avtomatsko označevanje slovenščine. V središču interesa

[1] »Operacijo delno financira Evropska unija iz Evropskega socialnega sklada ter Ministrstvo za šolstvo in šport. Operacija se izvaja v okviru Operativnega programa razvoja človeških virov za obdobje 2007–2013, razvojne prioritete: razvoj človeških virov in vseživljenjskega učenja; prednostne usmeritve: izboljšanje kakovosti in učinkovitosti sistemov izobraževanja in usposabljanja 2007–2013.« <www.slovenscina.eu>.

projekta sta nadgradnja oblikoskladenjskega označevanja ter zasnova skladenjskega označevanja, v zvezi s čimer se projektni cilji združujejo v dve temeljni aktivnosti: priprava učnega korpusa, ki bo osnova za učenje statističnih označevalnikov ter skladenjskih razčlenjevalnikov, ter zasnova in gradnja leksikona besednih oblik, tj. podatkovne zbirke z naborom informacij o oblikoslovnih lastnostih besedišča.

Izboljšane metode označevanja bodo v sklopu projekta uporabljene pri označevanju novega referenčnega korpusa za slovenščino, vsi rezultati projektnih aktivnosti pa bodo po zaključku projekta prosto dostopni za uporabo.[2]

2 Označevanje korpusnih besedil

Kljub nekaterim pomislekom[3] se je s širitvijo korpusnega jezikoslovja na področja morfološko bogatejših jezikov označevanje korpusnih besedil – vsaj npr. lematizacija, tj. pripis osnovne oblike besedam – izkazalo za neobhoden korak pri gradnji korpusov. Obenem so nova spoznanja o možnostih obdelave označenih besedil botrovala razvoju programske opreme drugega korpusnojezikoslovnega vala: razvijalci programov sledijo specifičnim raziskovalnim potrebam, programi so ciljno usmerjeni v pridobivanje, organizacijo in prikaz samo tiste vrste podatkov, ki jih uporabnik za določeno nalogo potrebuje.[4]

František Čermák npr. v razpravi, prvič objavljeni leta 1995, že piše o **zunanjih** ter **notranjih korpusnih podatkih.** O zunanjih podatkih govorimo pred vključitvijo v korpus (gre torej za neoznačena pisna ter transkribirana govorjena besedila), po vključitvi ter obdelavi besedil pa govorimo o notranjih podatkih:

> Tako dostopni in strojno berljivi notranji podatki v samem računalniku so takšne vrste in lastnosti, kakršne tvorci korpusa glede na zamišljeni cilj in uporabo dodajo. Kakorkoli je to tudi mogoče, praktično noben korpus danes ne daje na razpolago samo podatkov v obliki preprostih linearnih besednih verig. /…/ Stopenj tovrstnega označevanja je lahko toliko, kolikor velika je potreba in kolikor jih je mogoče računalniško (programsko) uspešno vključiti in uveljaviti /./ (Čermák 1995 v Gorjanc in Krek 2005: 148.)

Problematika označevanja korpusnih besedil je danes tako vezana predvsem na poskuse izboljšave kakovosti označevanja: ko govorimo o označevanju

[2] Pod licenco *Creative Commons* – <http://creativecommons.si/licence>.

[3] V raziskovalni praksi, imenovani *popolni korpusni pristop*, se temeljna premisa po nadomeščanju jezikoslovčeve intuicije z analizo avtentičnih jezikovnih podatkov odraža v odločitvi za delo z golimi, jezikoslovno neoznačenimi korpusnimi besedili – kakršno koli pripisovanje kategorij nekorpusnega izvora besedam v korpusu ima namreč aprioren vpliv na kasnejšo interpretacijo korpusnih podatkov (Tognini Bognelli 2001 v Teubert in Krishnamurty 2007). Sistematičen pregled argumentov za in proti označevanju korpusov prinaša npr. McEnery et al. 2006: 29–32.

[4] Tipičen predstavnik tovrstnega programa je denimo program Besedne skice oziroma *Word Sketches* <http://www.sketchengine.co.uk/>, ki na podlagi oblikoskladenjsko označenega korpusa ter nabora slovničnih pravil prikazuje različne vrste kolokacijskih podatkov za leksikografske potrebe (Krek in Kilgarriff 2006).

obsežnih korpusov (trenutni referenčni korpus *FidaPLUS*[5] npr. obsega več kot 620 milijonov besed), je avtomatizacija seveda edina možna izbira, avtomatski označevalniki pa niso (in – vsaj zaenkrat – ne morejo biti) neoporečno zanesljivi. V primeru da niso uzaveščene, lahko označevalne napake vodijo v neustrezno interpretacijo korpusnih podatkov, kar pomeni, da je izboljšava označevanja še toliko bolj bistvena, če je za označevani korpus predviden širok in raznovrsten nabor uporabnikov oziroma namenov rabe.

Poleg seznanjanja korpusnih uporabnikov z omejeno natančnostjo avtomatskega označevanja je lahko rešitev omenjenega problema le neprekinjeno izboljševanje označevalnih metod, pri čemer je nujno upoštevati, da so označena besedila ključnega pomena ne le za korpusno jezikoslovje, ampak tudi za obdelavo naravnega jezika ter razvoj jezikovnih tehnologij za izbrani jezik.

Projekt SSJ z aktivnostmi, vezanimi na izboljšavo označevanja (korpusnih) besedil, v večji meri nadaljuje delo, začeto v sklopu projekta *Jezikoslovno označevanje slovenščine* (JOS).[6] V nadaljevanju prispevka je na kratko predstavljeno trenutno stanje na področju oblikoskladenjskega ter skladenjskega označevanja in označevanja lastnih imen za slovenščino.[7]

3 Učni korpus SSJ

Učni korpus (ang. *training corpus*) je jezikovni vir, ki se uporablja za učenje statističnih modelov za izbrane namene, npr. razvoj statističnih označevalnikov ali skladenjskih razčlenjevalnikov. Ker je ključno, da so podatki za statistično učenje najvišje možne zanesljivosti, so oznake v učnih korpusih običajno v celoti ročno pripisane oziroma pregledane.

Učni korpus SSJ prinaša 400.000 besed, besedila bodo po zaključku projektnih aktivnosti štirinivojsko označena: korpus bo lematiziran, označen na oblikoskladenjski ter skladenjski ravni, označena pa bodo tudi lastna imena. Korpus SSJ je zasnovan na način, da bo skupaj s korpusom JOS100k (gl. naslednje poglavje) tvoril uravnoteženo celoto – končni rezultat gradnje bo torej polmilijonski učni korpus za učenje statističnih modelov za avtomatsko obdelavo slovenskih besedil.

3.1 Oblikoskladenjsko označevanje

Večji doprinos na področju oblikoskladenjskega označevanja slovenščine predstavljajo rezultati projekta JOS, in sicer revidirani nabor sistema oznak,

[5] <http://www.fidaplus.net> – več o korpusu v Arhar in Gorjanc 2007.

[6] Projektna dokumentacija ter rezultati projekta so na voljo na strani <http://nl.ijs.si/jos/>.

[7] Osredotočamo se na teme, ki so neposredno povezane z opisovanimi aktivnostmi, številni projekti označevanja slovenščine zato na tem mestu ostajajo ob strani, npr. izdelava odvisnostnih drevesnic (Erjavec in Ledinek 2006; Ledinek 2007), izdelava *WordNeta* kot del priprav na semantično označevanje (Fišer 2005; Fišer in Erjavec 2008), označevanje diskurza (Verdonik 2008; Verdonik et al. 2008), označevanje govornih zbirk (Žganec Gros et al. 2000) in še mnogi drugi.

korpusa JOS100k in JOS1M ter metode izboljšave označevanja s kombiniranjem rezultatov dveh označevalnikov. Vse našteto je bilo v literaturi že podrobneje predstavljeno, zato na tem mestu sledi le povzetek.

Oblikoskladenjske oznake JOS (priprava označevalnega sistema je opisana v Arhar in Ledinek 2008) so bile v letu 2008 zainteresirani javnosti ponujene kot predlog označevalnega standarda za slovenščino. Iz korpusa Fida+X[8] sta bila pripravljena dva korpusa, 100.000 besed obsegajoči JOS100k, ki predstavlja začetek sistematične priprave večnivojsko označenega učnega korpusa za slovenščino, ter milijon besed obsegajoči JOS1M. Oba korpusa sta bila preoznačena z novimi oblikoskladenjskimi oznakami, ki so bile pri prvem korpusu v celoti, pri drugem pa delno ročno pregledane (Erjavec in Krek 2008).

Na osnovi ročnih oznak korpusa JOS100k je bil naučen statistični označevalnik TnT, s katerim je bil korpus ponovno označen. Avtomatsko pripisane oznake, tako s strani označevalnika TnT kot izvorne oznake Amebisovega označevalnika,[9] so bile primerjane z ročnimi, pri čemer se je za prvi označevalnik izkazala 86,6% natančnost, za drugega pa 85,7% (ibid.: 52).

Nadaljnje raziskave so potekale na področju združevanja rezultatov obeh avtomatskih označevalnikov, z izhodiščno idejo, da je v primerih, kjer označevalnika besedo označita različno, mogoče statistično predvidevati, kateri od označevalnikov je označil pravilno – ter v nadaljevanju upoštevati ustrezno oznako. Statistični modul je bil naučen na podatkih korpusa JOS100k, trenutni rezultati izpričujejo do 79,73% natančnost napovedovanja, katera od dveh oznak je ustrezna (Rupnik in Grčar 2008).

V sklopu projekta SSJ že poteka ročno pregledovanje oblikoskladenjskih oznak učnega korpusa SSJ, torej novih 400.000 besed. 25 študentov pregleduje približno 10.000 besed obsegajoče datoteke, v začetni fazi po štirje označevalci na eno datoteko. Po končanem označevanju bodo oznake primerjane med sabo ter po potrebi problematična mesta ponovno pregledana. Pregled oznak oblikoskladenjskega nivoja naj bi bil predvidoma končan do jeseni 2009.

3.2 Skladenjsko označevanje

V sklopu projekta JOS je bil zasnovan sistem skladenjskega označevanja slovenščine. Sistem prinaša 10 oznak različnih tipov (za označevanje bodisi besednozveznih, stavčnočlenskih ali nadstavčnih razmerij), ki se pri označevanju kombinirajo z oblikoskladenjskimi oznakami.[10] Sistem označevanja je bil

[8] Korpus je nastal s pretvorbo referenčnega korpusa *FidaPLUS* v format xml, v zvezi s čimer so bile potrebne določene prilagoditve (gl. Erjavec in Krek 2008: 50).

[9] Korpus *FidaPLUS* je bil označen na podjetju Amebis (Arhar in Gorjanc 2007: 101–102), oznake so bile v korpusu Fida+X ohranjene, nato pa ob izdelavi podkorpusov JOS pretvorjene v oznake nabora JOS.

[10] Za natančnejšo predstavitev sistema označevanja na tem mestu žal ni prostora, ustrezna literatura je v pripravi (Ledinek in Erjavec 2009).

preizkušen in delno revidiran ob označevanju dela korpusa JOS100k, tj. 500 povedi, ki predstavljajo zlati standard za nadaljnje ročno označevanje (v nadaljevanju: pilotni korpus). Na osnovi označevalnih izkušenj so bila pripravljena navodila za označevalce, še prej pa programsko orodje, ki omogoča enostavno označevanje ter pregledovanje izbranih povedi.

V času pisanja prispevka poteka ročno skladenjsko označevanje celotnega korpusa JOS100k, kar skupaj z že označenimi 500 pomeni približno 6.100 povedi. Študentje označujejo korpus v paketih po 200 povedi, vsak paket po dva označevalca.[11] Ob označevanju poteka evalvacija označevalnega sistema, na osnovi katere se tako sistem kot navodila nadgrajujejo.

Na osnovi do sedaj označenih besedil bo v kratkem poskusno naučen izbrani statistični skladenjski razčlenjevalnik, s katerim bo označen preostanek korpusa. V primeru dovolj kvalitetnih rezultatov se bo avtomatsko označevanje uporabilo kot predpriprava ročnemu označevanju, s čimer se bo delo ustrezno pohitrilo.[12]

3.3 Prepoznava lastnih imen

V sklopu projekta SSJ se bodo v učnem korpusu pripisovale tri osnovne kategorije lastnih imen: *osebno*, *zemljepisno* ter *stvarno lastno ime*.[13] Za razliko od sistemov prepoznave lastnih imen (ang. *named entitiy recognition*, *NER*), ki v označevanje vključujejo tudi drugovrstne kategorije (npr. *datum*, *čas, odstotek*, *valuta*),[14] kategorije na tem mestu ostajajo primarno lastnoimenske, pri čemer je izhodiščna klasifikacija (osnova je *Slovenski pravopis* 2003: §25–160) nadgrajena z rezultati analize označevanj korpusnih besedil.

Sistem za prepoznavo lastnih imen bo deloval na osnovi metode prikritih modelov Markova (ang. *hidden Markov model*, *HMM*).[15] Poleg verjetnosti lastnoimenskih kategorij, ki bodo izračunane na osnovi podatkov iz učnega korpusa, bodo v obravnavo zajete oblikoskladenjske oznake besed v povedi ter oznaka vrste zapisa besede (npr. velika začetnica na začetku povedi ali znotraj nje, same male črke, same velike črke). Obenem bodo uporabljeni podatki o lastnih imenih iz leksikona

[11] Zaželeno bi bilo označevanje s strani treh označevalcev, vendar je zaradi pomanjkanja sredstev tretji označevalec vključen le na mestih, ki se izkažejo za problematična.

[12] Groba ocena koordinatorke označevanja Nanike Holz je, da označevalci v enem tednu označijo v povprečju 100 povedi.

[13] V času pisanja članka so kategorije preverjene na 500 stavkih pilotnega korpusa; ob označevanju večje količine besedil bodo dodatno vrednotene in po potrebi spremenjene.

[14] Nabor in razdelanost kategorij sta odvisna od namena sistema v razvoju; pri razvijanju zahtevnejših sistemov za luščenje ter prikaz informacij iz besedil so pogosto dodane še druge kategorije, gl. npr. Sekinejev označevalni sistem <http://nlp.cs.nyu.edu/ene/> ali sistem BBN <http://www.ldc.upenn.edu/Catalog/docs/LDC2005T33/BBN-Types-Subtypes.html>.

[15] Markovski proces je stohastični proces, ki obravnava verjetnost prihodnjega stanja izključno glede na trenutno stanje, brez upoštevanja preteklosti. Za prikrite modele Markova je značilno, da zaporedje stanj, skozi katere napreduje model, ni poznano.

besednih oblik (gl. naslednje poglavje), indikator za določanje lastnoimenskosti pa bo tudi sobesedilo obravnavane besede.[16]

4 Leksikon besednih oblik

Projekt SSJ predvideva do začetka leta 2011 gradnjo leksikona besednih oblik, za katerega so bile decembra 2008 pripravljene specifikacije standardov za izdelavo.[17]

Leksikon bo zajemal približno 100.000 enot, izbranih glede na naslednje kriterije: pri odprtih besednih vrstah glede na pogostnost v korpusih – učnem ter referenčnem korpusu, pri zaprtih (predlog, veznik, členek, zaimek) bodo vključene vse evidentirane enote. V leksikon bodo vključeni zahtevnejši primeri iz oblikoslovja (npr. primeri z variantnim pregibanjem, kot denimo *okvir – okvirja/ okvira*) in oblikovne variante zapisa večbesednih enot skupaj oziroma narazen, in sicer ne glede na pogostnost v korpusih.

Nadaljevanje poglavja je namenjeno predstavitvi nekaterih odločitev v zvezi z zasnovo leksikona.

4.1 Format zapisa leksikonske enote

Leksikon je zasnovan skladno s standardom ISO *Lexical Markup Framework,*[18] zapis je v formatu XML.

Osnova leksikona je enobesedna enota (o večbesednih enotah več v **4.3.1**), ki prinaša oblikoslovno paradigmo za obravnavano besedo ter vnaprej določen nabor informacij za vsako od navedenih oblik. Informacije so dveh tipov, na eni strani se v leksikonu pripisujejo vrednosti in kategorije, skladne s sistemom oblikoskladenjskega označevanja JOS, v posameznih primerih pa so predvidene tudi druge oblike informacij ter povezav med enotami (o teh v nadaljevanju). Oba tipa podatkov sta v leksikonski strukturi med seboj jasno ločljiva.

---

[16] Gre za metodo t. i. eksterne oziroma interne evidence. »Pri interni evidenci gre za tipične lastnosti in zgradbo imen (*d. o. o., d. d., Inštitut X, X vrh, X jezero*). K njej pripisujemo tudi začetnice imen (*George W. Bush*). Eksterna evidenca pa uporablja sobesedila oziroma okolice lastnih imen (*Prof. XY, Gospa XY ali XY ml.*).« (Arčan in Vintar 2006: 151.)

[17] Pričujoče poglavje se v veliki meri opira na omenjeni dokument, iz specifikacij so vzeti tudi vsi navedeni primeri.

[18] <http://www.lexicalmarkupframework.org>.

4.2 Informacije v leksikonski enoti

Tipična leksikonska enota prinaša opredelitev osnovne oblike ter nabor posameznih besednih oblik z ustrezajočimi atributi in vrednostmi, lahko pa tudi povezavo na druge leksikonske enote <Related Form> ter, kadar je potrebno razdvoumljanje enakopisnih oblik, informacijo o pomenu enote <Sense>.

4.2.1 Lema in besedne oblike

V nadaljevanju sledi za primer uvodni del leksikonskega zapisa za samostalnik *igranje.* Začetek predstavlja opredelitev z identifikacijsko oznako (*id*), ki vsebuje lemo z oznako njene besednovrstne opredelitve (*S_igranje*). Sledi opis lastnosti, ki veljajo za celotno paradigmo (v obravnavanem primeru npr. opredelitev besedne vrste, občnoimenskosti ter spola). Na naslednjem mestu je zapis leme <Lemma>, za tem pa si po vrsti sledijo besedne oblike <WordForm> z vsemi podatki – na prvem mestu je vedno zapis oblike, pri obravnavi samostalnikov sledi opredelitev števila ter sklona:

```
<LexicalEntry  id= ”LE_S_igranje”>
      <feat att=”besedna_vrsta” val=”samostalnik”/>
      <feat att=”vrsta” val=”občni”/>
      <feat att=”spol” val=”srednji”/>
      <Lemma>
         <feat att=”zapis_oblike” val=”igranje”/>
      </Lemma>
      <WordForm>
         <feat att=”zapis_oblike” val=”igranje”/>
         <feat att=”število” val=”ednina”/>
         <feat att=”sklon” val=”imenovalnik”/>
      </WordForm>
      <WordForm>
         <feat att=”zapis_oblike” val=”igranja”/>
         <feat att=”število” val=”ednina”/>
         <feat att=”sklon” val=”rodilnik”/>
      </WordForm>
      …
```

Primer 1: Igranje *(besedne oblike).*

4.2.2 Povezave na druge leksikonske enote

Enote v leksikonu so povezane med sabo v primeru, da izpričujejo katero od vnaprej določenih (oblikovno transparentnih) besedotvornih povezav. Predvidene povezave, ki so v vseh primerih obojestranske, se označujejo, kadar obe besedi izkazujeta zadostno stopnjo pogostnosti v korpusnih virih. Nabor povezav prikazuje spodnja tabela, primeri so podani v ležečem tisku:

izpeljani pridevnik (na -ov/-ev, -in) *česnov*	↔	izvorni samostalnik *česen*
glagolnik (na -ev, -nje) *čakanje*	↔	izvorni glagol *čakati*
občni samostalnik *loka*	↔	prekrivno lastno ime *Loka*
izpeljani samostalnik na -ost *veselost*	↔	izvorni pridevnik *vesel*
deležje na -e, -č, -aje *sede*	↔	glagol *sedeti*
deležnik na –č, deležnik stanja na -l, -n, -t *sedeč*	↔	glagol *sedeti*
izpeljani prislov *temeljito*	↔	izvorni pridevnik *temeljit*
elativ *pretesen, pretesno*	↔	izvorni pridevnik oziroma prislov *tesen, tesno*
okrajšava *dr.*	↔	neokrajšana različica *doktor*

Tabela 1: *Povezave med leksikonskimi enotami.*

Povezava z drugo besedo <RelatedForm> je navedena po opisu besednih oblik, navaja se s pomočjo identifikacijske oznake slednje. Za v prejšnjem poglavju predstavljen primer *igranje* sledi zapis povezave na glagol:

```
<RelatedForm>
    <feat att=”idref” val=”LE_G_igrati”/>
</RelatedForm>
```

Primer 2: Igranje *(povezava z glagolom).*

4.2.3 Razdvoumljanje oblik

Informacija o pomenu obravnavane enote je v leksikonu predvidena le na mestih, kjer je potrebno razlikovanje med enakopisnima besednima oblikama enake besedne vrste, in sicer zgolj v primeru, da razlikovanje ni mogoče na osnovi drugih razlikovalnih informacij. Razdvoumljanje z navedbo pomena <Sense> poteka tako predvsem v primerih enakopisnih glagolskih nedoločnikov, ki imajo del glagolske paradigme drugačen (npr. *izvesti – izvezem/izvedem, stati – stanem/stojim, odpeti – odpojem/odpnem* itd.). Obliko zapisa prikazuje spodnji primer za *stati – stanem*:

```
<Sense>
    <Definition>
<feat att=«text« val=«imeti določeno kupno ali prodajno ceno«/>
    </Definition>
</Sense>
```

Primer 3: Stati *(razdvoumljanje pomena).*

Če enakopisni samostalniški par izkazuje dva spola (npr. *prst – prst*), sta samostalnika vključena kot ločeni leksikonski enoti. Označevanje živosti je predvideno pri ustrezni sklonski obliki samostalnika, kakor prikazuje spodnji primer za samostalnik *koder*. Dodatne pomenske opredelitve za tovrstne primere torej niso predvidene.

```
<WordForm>
        <feat att=«zapis_oblike« val=«koder«/>
        <feat att=«število« val=«ednina«/>
        <feat att=«sklon« val=«tožilnik«/>
        <feat att=«živost« val=«ne«/>
</WordForm>
 <WordForm>
        <feat att=«zapis_oblike« val=«kodra«/>
        <feat att=«število« val=«ednina«/>
        <feat att=«sklon« val=«tožilnik«/>
        <feat att=«živost« val=«da«/>
 </WordForm>
```

Primer 4: Koder *(označevanje živosti).*

4.3 Lastna imena

V leksikon bodo vključena tudi lastna imena, ki izpričujejo dovolj visoko pogostnost v korpusnih virih. Zapis leme za te primere je z veliko začetnico. Pri vključevanju lastnih imen v leksikon se upoštevajo naslednje smernice: vsi priimki imajo po dve leksikonski enoti, eno po ničti ženski ter drugo po moški sklanjatvi. Po dva vnosa imajo tudi imena, ki so enakopisna za ženski ter moški spol (npr. *Saša*), prav tako imajo po dva vnosa primeri, ki se pojavljajo bodisi v srednjem ali ženskem spolu (npr. *Japonska – Japonsko*).

Kot lastna imena se v leksikon vključujejo tudi kratična imena, npr. zemljepisnega tipa, imena strank, podjetij, organizacij, zakonov itd. (npr. *EU*, *SDS*, *OŠ*, *ZDR*). Kratična imena, ki so prešla v samostalnike (npr. *Sazu*, *Nama*), so v leksikon vključena kot samostalniki.

4.3.1 Večbesedna lastna imena

V sklopu projekta SSJ bodo v leksikon zajete le večbesedne enote, namenjene bodisi identifikaciji variantnosti zapisa skupaj oziroma narazen – o teh več v naslednjem poglavju – bodisi prepoznavi večbesednih lastnih imen.

Večbesedne leksikonske enote v izogib podvajanju podatkov prinašajo zgolj referenco na identifikacijske oznake enobesednih komponent <ListOfComponents>. Poleg tega večbesedna enota vsebuje zapis osnovne oblike besedne zveze <Lemma>,

znotraj identifikacijske oznake zveze pa še vzorec, ki opredeljuje skladenjske lastnosti obravnavane besedne zveze (*mwePattern*). Zapis lastnoimenske besedne zveze ponazarja naslednji primer:

```
<LexicalEntry id=”LE_Škofja_Loka” mwePattern=”LI_dve_1_VV_PP_P2”>
    <feat att=”status” val=”lastno_ime”/>
    <Lemma>
        <feat att=”zapis_oblike” val=”Škofja Loka”/>
    </Lemma>
    <ListOfComponents>
        <Component entry=”LE_P_škofji”/>
        <Component entry=”LE_S_loka”/>
    </ListOfComponents>
</LexicalEntry>
```

Primer 5: Škofja Loka *(večbesedno lastno ime).*

Vzorci večbesednih enot prinašajo naslednje podatke o besedni zvezi: zaporedno mesto posamezne besede v zvezi, opredelitev zapisa posamezne besede z veliko oziroma malo začetnico ter opredelitev, ali se beseda kot del zveze pregiba enako kot njen enobesedni leksikonski referent. Omogočen je vnos še dveh neobveznih informacij: opredelitev ločila, ki ločuje oziroma povezuje elemente zveze (npr. presledek, vezaj), ter opredelitev jedra besedne zveze.

Vzorci v leksikonu obstajajo kot vnaprej definirane enote posebne vrste, ob vnosu besednih zvez pa se za povezavo na ustrezni vzorec uporablja skrajšana oblika vzorčnega zapisa. V primeru *Škofja Loka* uporabljeni zapis vzorca *LI_dve_1_VV_PP_P2* tako predstavlja: lastno ime, ki vsebuje dve besedi, zaporedna številka vzorca je 1, obe besedi se zapisujeta z veliko začetnico, obe besedi izkazujeta v zvezi podedovano pregibanje (tj. se pregibata enako kot enobesedni leksikonski referent), besedi ločuje presledek, jedro zveze pa je druga od obeh besed.

Sledi še primer zapisa obravnavanega vzorca kot posebne leksikonske enote, kjer so zgoraj opisane informacije (*zaporedje*, *pisava*, *pregibnost*) podane za vsak element besedne zveze <MWELex> posebej:

```
<MWEPattern id=”LI_dve_1_VV_PP_P2”>
    <MWENode>
<MWELex>
        <feat att=”zaporedje” val=”1”/>
        <feat att=”pisava” val=”velika_začetnica”/>
        <feat att=”pregibnost” val=”podedovana”/>
</MWELex>
<MWELex>
        <feat att=”zaporedje” val=”2”/>
```

```
                <feat att="pisava" val="velika_začetnica"/>
                <feat att="pregibnost"  val="podedovana"/>
                <feat att="ločilo" val="presledek"/>
                <feat att="jedro" val="da"/>
            </MWELex>
        </MWENode>
    </MWEPattern>
```

Primer 6: LI_dve_1_VV_PP_P2 *(vzorec).*

4.4 Normativni podatki

Za leksikon je predvidena najširša mogoča namenskost, med drugim tudi pri aplikacijah, vezanih na slogovni priročnik in pedagoško slovnico (ki sta prav tako projektni aktivnosti). K primerom, ki izražajo variantnost na oblikovni ravni, je zato smotrno vključevati tako podatke o pogostnosti posamezne variante v rabi kot tudi trenutno aktualno normativno opredelitev posamezne od variant.

Za normativno opredelitev so predvidene tri oznake, *variantno*, kadar sta obe navedeni oblikovni varianti v skladu s trenutno normo, ter *nestandardno* oziroma *standardno* za oblike, ki jih normativni priročniki kot take določajo. Podatek o pogostnosti je pri vsaki od oblik vključen kot število pojavitev v referenčnem korpusu.[19] Sledi primer zapisa za variantno obliko rodilnika ednine samostalnika *okvir*:

```
<WordForm>
    <feat att="število" val="ednina"/>
    <feat att="sklon" val="rodilnik"/>
    <FormRepresentation>
    <feat att="zapisOblike" val="okvirja"/>
    <feat att="norma" val="variantno"/>
    <feat att="pogostnost" val="2251"/>
    </FormRepresentation>
    <FormRepresentation>
    <feat att="zapisOblike" val="okvira"/>
            <feat att="norma" val="variantno"/>
            <feat att="pogostnost" val="1841"/>
    </FormRepresentation>
</WordForm>
```

Primer 7: Okvir *(oblikovne variante).*

Beleženje variantnosti (s podatkom o pogostnosti v rabi) je v leksikonu predvidena

[19] V vseh primerih v tem prispevki so prikazani podatki o pogostnosti iz korpusa *FidaPLUS*, normativni podatki so iz *Slovenskega pravopisa* (2003).

še na nekaterih mestih. Kratična imena, ki so pisana s samimi velikimi črkami, imajo v sklopu ene leksikonske enote trojno paradigmo, z vezajem, brez vezaja ter po ničti sklanjatvi:

```
<WordForm>
    <feat att=”število” val=”ednina”/>
    <feat att=”sklon” val=”rodilnik”/>
    <FormRepresentation>
        <feat att=”zapis_oblike” val=”BTC-ja”/>
        <feat att=”norma” val=”variantno”/>
        <feat att=”pogostnost” val=”421”/>
    </FormRepresentation>
    <FormRepresentation>
        <feat att=”zapis_oblike” val=”BTCja”/>
        <feat att=”norma” val=”variantno”/>
        <feat att=”pogostnost” val=”3”/>
    </FormRepresentation>
    <FormRepresentation>
        <feat att=”zapis_oblike” val=”BTC”/>
        <feat att=”norma” val=”variantno”/>
        <feat att=”pogostnost” val=”337”/>
    </FormRepresentation>
</WordForm>
```

Primer 8: BTC *(oblikovne variante)*.

Variantnost se beleži tudi pri primerih večbesednih enot, za katere raba izpričuje več različic na ravni zapisa skupaj oziroma narazen (npr. *čimbolj – čim bolj, naprimer – na primer*). Podobno kot večbesedna lastna imena tudi narazen pisane različice vsebujejo povezavo na ustrezni vzorec, ki pa v tem primeru vsebuje le tri tipe informacij: opredelitev zaporednega mesta posamezne besede v zvezi, opredelitev, ali se beseda kot del zveze pregiba enako kot njen enobesedni leksikonski referent, ter opredelitev ločila, ki ločuje oziroma povezuje elemente zveze. Spodnji primer podaja vnos zveze *na primer*, ki vsebuje povezavo na različico *naprimer*:

```
<LexicalEntry id=”LE_na_primer” mwePattern=”SLG_dve_2_PNP”>
    <Lemma>
        <feat att=”zapis_oblike” val=”na primer”/>
    </Lemma>
    <ListOfComponents>
        <Component entry=”LE_D_na”/>
        <Component entry=”LE_S_primer”/>
    </ListOfComponents>
    <RelatedForm>
<feat att=”status” val=”skupaj_narazen”/>
        <feat att=”idref” val=”LE_R_naprimer”/>
    </RelatedForm>
</LexicalEntry>
```

Primer 9: Na primer *(zapis skupaj – narazen)*.

5 Zaključek

V prispevku predstavljene aktivnosti se neposredno vpenjajo v mrežo ostalih ciljev projekta SSJ. Učni korpus kot vir za razvoj statističnih modelov za označevanje ter oblikovni leksikon kot podpora (predvsem oblikoskladenjskemu) označevanju ter prepoznavi lastnih imen neposredno vplivata na kakovost označenosti slovenskih besedil. Nadgrajena oziroma na novo razvita označevalna orodja bodo uporabljena za označevanje novega referenčnega korpusa, ta pa bo osnova za pripravo drugih temeljnih jezikovnih virov – v sklopu projekta predvsem gradnje leksikalne podatkovne baze ter priprave slogovnega priročnika ter pedagoške korpusne slovnice.

Literatura

Arčan, Mihael, in Vintar, Špela, 2006: Avtomatično prepoznavanje lastnih imen. Erjavec, Tomaž, in Žganec Gros, Jerneja (ur.): *Jezikovne tehnologije 5.* Ljubljana: IJS. 150–155.

Arhar, Špela, in Ledinek, Nina, 2008: Oblikoskladenjske oznake JOS: revizija in nadgradnja nabora oznak za avtomatsko skladenjsko označevanje slovenščine. Erjavec, Tomaž, in Žganec Gros, Jerneja (ur.): *Jezikovne tehnologije 6.* Ljubljana: IJS. 54–59.

Arhar, Špela, in Gorjanc, Vojko, 2007: Korpus FidaPLUS: Nova generacija slovenskega referenčnega korpusa. *Jezik in slovstvo* 52/2. 95–110.

Čermák, František, 2005: Jezikovni korpus: sredstvo in vir spoznanj. Gorjanc, Vojko, in Krek, Simon (ur.): *Študije o korpusnem jezikoslovju.* Ljubljana: Krtina. 137–171.

Erjavec, Tomaž, Holozan, Peter, Krek, Simon, Pivec, Matej, Rigač, Simon, Rozman, Simon, in Velušček, Aleš, 2008: *Specifikacije za leksikon besednih oblik – projekt »Sporazumevanje v slovenskem jeziku«.* Kamnik: neobjavljeno projektno poročilo.

Erjavec, Tomaž, in Krek, Simon, 2008: Oblikoskladenjske specifikacije in označeni korpusi JOS. Erjavec, Tomaž, in Žganec Gros, Jerneja (ur.): *Jezikovne tehnologije 6.* Ljubljana: IJS. 49–53.

Erjavec, Tomaž, in Ledinek, Nina, 2006: Slovenska odvisnostna drevesnica: prvi rezultati. Erjavec, Tomaž, in Žganec Gros, Jerneja (ur.): *Jezikovne tehnologije 5.* Ljubljana: IJS. 162–167.

Fišer, Darja, in Erjavec, Tomaž, 2008: Predstavitev in analiza slovenskega wordneta. Erjavec, Tomaž, in Žganec Gros, Jerneja (ur.): *Jezikovne tehnologije 6.* Ljubljana: IJS. 37–42.

Fišer, Darja, 2005: Pristopi k izdelavi leksikalnih podatkovnih zbirk. *Jezik in slovstvo* 50/6. 17–32.

Krek, Simon, in Kilgarriff, Adam, 2006: Slovene Word Sketches. Erjavec, Tomaž, in Gros, Jerneja (ur.): *Jezikovne tehnologije 5.* Ljubljana: IJS. 62–65.

Ledinek, Nina, in Erjavec, Tomaž, 2009: Odvisnostno površinskoskladenjsko označevanje slovenščine: specifikacije in označeni korpusi. *Infrastruktura slovenščine in slovenistike (Obdobja 28)*. Prispevek je v pripravi.

Ledinek, Nina, 2007: Slovenska odvisnostna drevesnica v raziskavah o induktivnem odvisnostnem označevanju. *Jezik in slovstvo* 52/1. 3–16.

McEnery, Tony, Xiao, Richard, in Tono, Yukio, 2006: *Corpus-Based Language Studies.* London: Routledge.

Rupnik, Jan, Grčar, Miha, in Erjavec, Tomaž, 2008: Improving morphosyntactic tagging of Slovene by tagger combination. Erjavec, Tomaž, in Žganec Gros, Jerneja (ur.): *Jezikovne tehnologije 6.* Ljubljana: IJS. 110–115.

SAZU in Inštitut Frana Ramovša ZRC SAZU, 2003: *Slovenski pravopis: elektronska izdaja.* Ljubljana: Založba ZRC.

Tognini-Bonelli, Elena, 2007: The corpus-driven approach. Teubert, Wolfgang, in Krishnamurty, Ramesh (ur.): *Corpus Linguistics: Critical Concepts in Linguistics*. London, New York: Routledge. 74–92.

Verdonik, Darinka, 2008: Označevanje vrste diskurzivnih označevalcev. Erjavec, Tomaž, in Žganec Gros, Jerneja (ur.): *Jezikovne tehnologije 6.* Ljubljana: IJS. 25–28.

Verdonik, Darinka, Žgank, Andrej, in Pisanski Peterlin, Agnes, 2008: Validacija označevanja diskurzivnih označevalcev v korpusih Turdis-2 in BNSIint. Erjavec, Tomaž, in Žganec Gros, Jerneja (ur.): *Jezikovne tehnologije 6.* Ljubljana: IJS. 29–32.

Žganec Gros, Jerneja, Mihelič, France, in Dobrišek, Simon, 2000: Govorne tehnologije: pridobivanje in pregled govornih zbirk za slovenski jezik. *Jezik in slovstvo* 48/3–4. 47–59.

Spletne strani

Besedne skice (Word Sketches): <http://www.sketchengine.co.uk/>. (Dostop 21. 6. 2009.)
Creative Commons: <http://creativecommons.si/licence>. (Dostop 20. 6. 2009.)
Jezikoslovno označevanje slovenščine: <http://nl.ijs.si/jos/>. (Dostop 21. 6. 2009.)
Korpus slovenskega jezika FidaPLUS: <http://www.fidaplus.net>. (Dostop 20. 6. 2009.)
Lexical Markup Framework: <http://www.lexicalmarkupframework.org>. (Dostop 20. 6. 2009.)
Označevalni sistem BBN: <http://www.ldc.upenn.edu/Catalog/docs/LDC2005T33/BBN-Types-Subtypes.html>. (Dostop 20. 6. 2009.)
Sekinejeva hierarhija označevanja lastnih imen: <http://nlp.cs.nyu.edu/ene/>. (Dostop 20. 6. 2009.)
Sporazumevanje v slovenskem jeziku: <www.slovenscina.eu>. (Dostop 21. 6. 2009.)