SPLETNI PORTAL *SLOGOVNI PRIROČNIK*: LUŠČENJE IN PRIKAZ PODATKOV O JEZIKOVNI RABI

Kaja Dobrovoljc
Trojina, zavod za uporabno slovenistiko, Ljubljana

Simon Krek
Institut »Jožef Štefan«, Ljubljana

UDK 811.163.6''271.1'354:004.738.5

V prispevku predstavljamo proces luščenja in prikazovanja korpusnih podatkov, kakršen je bil vzpostavljen pri pripravi demonstracijskih gesel na spletnem portalu *Slogovni priročnik*. Kot most med nevtralnimi korpusnimi podatki in vizualizacijo normativnih podatkov na portalu služi leksikon besednih oblik, njihovo pretakanje iz leksikona na portal pa usmerja mehanizem kratkega odgovora, ki omogoča, da se podatki na portalu avtomatsko prilagajajo spremembam v jeziku oz. referenčnem korpusu.

spletni portal, jezikovni priročniki, standardizacija, pravopis, jezikovne tehnologije, luščenje podatkov

The paper presents the process of corpus data extraction and representation for the purpose of creating the *Style Guide* web portal for Slovene. The neutral corpus data and information about language codification are merged within a lexicon of inflected forms and subsequently visualised through the 'short answer' system that enables the portal data to automatically adapt to any changes in the language or its reference corpus.

web portal, language reference books, standardisation, normative guide, language technologies, data extraction

1 Kratka predstavitev portala

Spletni portal Slogovni priročnik[1] je jezikovni priročnik, ki govorcem slovenščine s sopostavljanjem trenutno veljavnega pravopisnega standarda in korpusnih podatkov na izčrpen in razumljiv način pomaga pri reševanju raznovrstnih jezikovnih zadreg, s katerimi se srečujejo pri tvorbi besedil. Zasnova, zgradba in funkcionalnosti portala so podrobneje predstavljeni v Krek 2012, enem od kazalnikov krovnega projekta Sporazumevanje v slovenskem jeziku (Bizjak Končar idr. 2011) ter ločenem prispevku pričujočega simpozija, zato jih v nadaljevanju predstavljamo zgolj na kratko.

Vsaka izmed približno 700 obravnavanih zadreg je opredeljena s specifično kodo kategorije, ki določa njeno hierarhično umestitev v pripadajočo jezikovno ravnino in ustreza enemu odgovoru na portalu.[2] Vsak odgovor nato sestavljajo tri ločene rubrike: rubriki *Na dolgo in široko* ter *Za navdušence* skupaj tvorita t. i. dolgi (pod)odgovor, ki uporabniku ponuja laično razlago obravnavane jezikovne problematike, rubrika *Kratko in jedrnato* (kratki odgovor) pa s kombinacijo grafa in njegovega kratkega opisa splošno razlago posamezne jezikovne zadrege v dolgem odgovoru dopolnjuje s pojasnili o rabi in normativni ustreznosti njene konkretne leksikalne reali-

1 http://slogovni.slovenscina.eu

2 Trenutno je možen ogled 15 demonstracijskih gesel z zgledi, druge jezikovne zadrege pa so skupaj s pripisano kategorijo in tipičnimi zgledi navedene v ontološko urejenem kazalu portala.

zacije, ki jo je vstopni sistem pred tem prepoznal kot relevantno za uporabnika.

Za pridobitev izčrpnega nabora problematičnih primerov znotraj posamezne kategorije, ki omogoča, da portal uporabniku v kratkem odgovoru vedno ponudi konkreten odgovor na konkretno jezikovno vprašanje, je torej nujna temeljita predhodna analiza jezikovne rabe, s čimer označujemo sodobno slovenščino, kakršno prikazuje uravnotežen nabor avtentičnih besedil v referenčnem besedilnem korpusu. V nadaljevanju prispevka tako podrobneje predstavljamo luščenje, analizo, uvoz in prikazovanje korpusnih podatkov na omenjenem spletnem portalu ter vlogo, ki jo v tem procesu igrata njegovi osrednji podatkovni zbirki: korpus Gigafida v obsegu milijarde besed[3] in leksikon besednih oblik Sloleks.[4]

2 Strojno luščenje korpusnih podatkov

V prvi fazi luščenja korpusnih podatkov iz korpusa Gigafida smo podrobneje opredelili oblikoslovne, skladenjske oziroma druge lastnosti posamezne jezikovne zadrege in pripravili ustrezna navodila za strojno luščenje in želeni izpis podatkov. Različni tipi jezikovnih zadreg so glede na zahtevnost, obseg in pričakovano stopnjo korpusnega šuma zahtevali različno zasnovana navodila, pri vseh pa smo skušali v čim večji meri upoštevati dve poglavitni načeli: navodila za luščenje naj bodo pripravljena brez vnaprejšnjega sklepanja o dejanski jezikovni rabi (tudi če to pomeni večjo količino korpusnega šuma) ter naj se, če je le mogoče, opirajo zgolj na alfanumerično obliko pojavnic, ne pa na njihove strojno pripisane metapodatke o oblikoskladenjskih oznakah ali lemi.

Za ponazoritev procesa pridobivanja in analize korpusnih podatkov vzemimo kategoriji C1a3a (Sklanjanje moških samostalnikov z neobstojnim samoglasnikom: slovenska lastna imena, npr. Klemen – Klemna/Klemena, Sajovic – Sajovica/Sajovca) in C1a3b (Sklanjanje moških samostalnikov z neobstojnim samoglasnikom: tuja lastna imena, npr. Russell – Russella/Russlla, Clinton – Clintona/Clintna).

Za pridobitev seznama vseh lastnih imen, ki bi lahko bila relevantna za omenjeni zadregi, opis iskanih korpusnih pojavnic razdelimo na tri dele – osnovo, variabilni del in obrazilo[5] – ter jih izluščimo iz korpusa za nadaljnjo obdelavo. Na ta način lahko primerjamo variantne oblike, vključno s podatki o variabilnosti pri posamezni obliki v oblikoslovni paradigmi. Najbolj zanimive za naš namen so tiste kombinacije, ki imajo distribucijo variabilnega dela najbolj razpršeno po obeh možnostih.

V drugi fazi tako dobljeni spisek razdelimo na ločene podsezname, ki predstavljajo vsak svojo kombinacijo (izpuščenega ali obdržanega) samoglasnika in enega ali dveh soglasnikov (npr. -en/-n-, -ek/-k-, -ic/-c-, -ell/-ll-), ter podatke o frekvenci vseh obrazil določene osnove in posameznega variabilnega dela združimo. Če spisek vseh pojavnic s statističnimi podatki iz korpusa razporedimo po izračunu, ki proti vrhu potiska oblike z najbolj razpršeno distribucijo in največ pojavitvami,[6] dobimo za podseznam -en/-n- naslednji spisek (najvišje vrednosti si sledijo po vrsticah od zgoraj navzdol, navajamo prvih petnajst):

3 http://www.gigafida.net; več v Logar Berginc idr. 2012.

4 http://www.slovenscina.eu/sloleks; več v Arhar 2009; Krek, Erjavec 2009.

5 V danem primeru je osnova opredeljena kot niz črk z veliko začetnico (npr. Klem-), variabilni del kot katerikoli enojni ali podvojeni soglasnik, pred katerim lahko stoji poljuben samoglasnik (npr. -en- ali -n-), in obrazilo kot dokončen nabor končnic v paradigmi imen, ki se sklanjajo po prvi moški sklanjatvi (npr. -a, -u, -om/-em).

6 Pri statistični obdelavi podatkov, ki je namenjena zgolj razvrščanju, ne pa tudi selekciji podatkov, smo glede na naravo jezikovne zadrege poleg frekvence posameznih variant in njihovih medsebojnih razmerij upoštevali tudi dolžino korena, frekvenco strojno pripisanih ali ugibanih lem, frekvence oblik s specifičnimi obrazili ipd.

Tabela 1: Korpusni spisek potencialnih imen z neobstojnim polglasnikom pred črko n

Osnova	(Ugibana) lema	Število pojavitev v korpusu Gigafida		Rezultat
		osnova -**en**- obrazilo *npr. Klemena*	osnova -**n**- obrazilo *npr. Klemna*	
Klem	Klemen	1843	3839	0,46
Lor	Loren	908	505	0,29
Berg	Bergen	208	375	0,25
Niels	Nielsen	164	120	0,25
Test	Testen	501	2326	0,24
Robb	Robben	163	333	0,24
Natlač	Natlačen	223	147	0,23
Gold	Golden	37	29	0,21
Gall	Gallen	105	148	0,20
Ols	Olsen	112	64	0,20
Bid	Biden	102	117	0,20
Bjorndal	Bjorndalen	112	163	0,20
Franz	Franzen	117	114	0,19
Jens	Jensen	138	60	0,19
Patt	Patten	85	113	0,19

3 Ročna analiza korpusnih podatkov

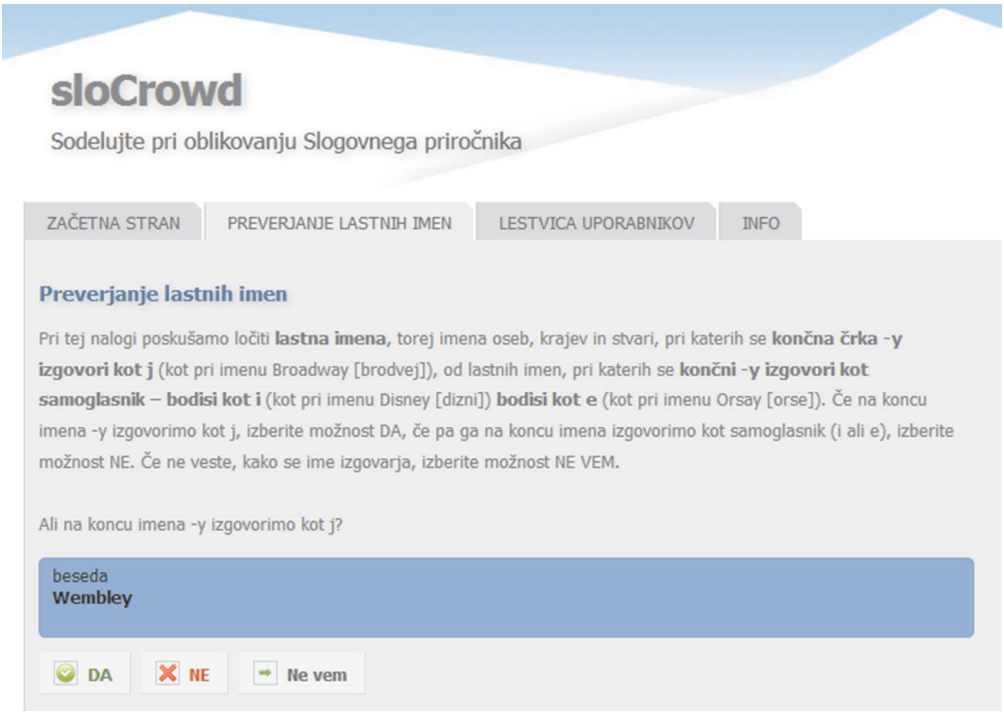
Pri jezikovnih zadregah z ozko opredeljenimi oblikoslovnimi lastnostmi tako dobljeni spiski ne potrebujejo nadaljnje ročne analize[7] in so pripravljeni za neposreden uvoz v leksikon besednih oblik (gl. poglavje 4), pri večini avtomatsko generiranih izpisov pa je pred zaključkom analize nujna še faza ročne analize, bodisi z namenom odstranjevanja korpusnega šuma oz. validacije podatkov bodisi z namenom pripisovanja informacije o normativni zaznamovanosti posameznih variant.

Za ta namen smo uporabili prosto dostopno spletno aplikacijo sloCrowd,[8] ki z izkoriščanjem moči množic (angl. crowdsourcing) omogoča preprosto ročno obdelavo obsežnih korpusnih spiskov. Kot ugotavljajo avtorji aplikacije, se s prenosom bremena verifikacije na širšo množico zmanjša čas verifikacije, dobljeni rezultati pa so lahko celo bolj zanesljivi, saj se o posameznem primeru odloča večje število uporabnikov (Tavčar, Fišer, Erjavec 2012: 198). Sistem zahteva predhodno prijavo uporabnikov in vključuje posebni mehanizem, ki z naključnim umeščanjem primerov iz vnaprej pripravljene zlate množice preverja njihovo zanesljivost.

Za ponazoritev analize s pomočjo izkoriščanja moči množic vzemimo korpusni spisek za kategorijo C1a3f (Oblikoslovje > Samostalniki > Moške sklanjatve > Samostalniki na samoglasnik > Sklanjanje moških imen, ki se končajo na -y). S strojnim luščenjem korpusnih podatkov dobimo seznam imen (tj. osnov, ki se začnejo z veliko začetnico in končajo s črko y), ki se v korpusu pojavljajo tako z obrazili s podaljšavo z -j-

7 Tak je denimo problem izbire podaljšave pri pregibanju akronimov (A3c3a), pri katerem se z luščenjem nizov velikih črk, vezaja in omejenega nabora končnic pojavlja zelo malo korpusnega šuma.

8 http://dis.ijs.si/ales/slocrowd; SloCrowd je posebna izpeljava izvorne aplikacije sloWCrowd (Tavčar, Fišer, Erjavec 2012), ki je bila razvita za namen popravljanja ročno zgrajenega semantičnega leksikona za slovenščino sloWNet (Fišer 2009) in je dostopna na http://nl.ijs.si/slowcrowd.



Slika 1: Primer naloge v spletni aplikaciji sloCrowd

(npr. Harryja, Sydneyja, Playboyja) kot brez nje (Harrya, Sydneya, Playboya). Imena, ki se končajo na sklop soglasnika in črke y, ne potrebujejo nadaljnje obdelave, saj imajo zaradi predvidljivega izgovora črke y jasno določljivo standardno (pregibanje s podaljšavo, npr. Henryja) in nestandardno paradigmo (brez podaljšave, npr. Henrya).[9] Pri imenih, ki se končajo na sklop samoglasnika in črke y, pa normativna zaznamovanost ni avtomatsko določljiva, saj moramo poznati izgovor tujega imena. Spisek takih imen smo zato uvozili v sloCrowd in uporabnike[10] prosili, da s preprostim klikanjem določijo pravilen izgovor končnega sklopa (slika 1).

Naloga je bila zaključena, ko je vsako izmed imen pregledalo pet različnih uporabnikov, pri čemer so imena z manj kot štirimi enakimi odgovori pregledali in dokončno potrdili še zanesljivi tretji odločevalci. Tem so bile sicer zaupane tudi druge oblike ročne analize, denimo lematizacija ali odstranjevanje šuma s pregledovanjem konkretnih korpusnih konkordanc.

Podoben delovni proces smo uporabili tudi pri pripravi večine drugih demonstracijskih gesel, ki poleg avtomatskega luščenja korpusnih podatkov potrebujejo še določeno mero ročne jezikoslovne analize.[11] Tako dobljeni izčiščeni korpusni spiski so po eni strani namenjeni piscem dolgih odgovorov, ki v zgoščen in pregleden opis obravnavane problematike umeščajo ustrezne korpusne zglede, po drugi strani pa so namenjeni predvsem

9 Izjema so enozložna imena, pri katerih se končni -y izgovori kot [aj], npr. Fry, Sly, Sky.

10 Pri poskusni uporabi orodja sloCrowd v okviru izdelave demonstracijskih gesel Slogovnega priročnika je sodelovalo okoli 100 študentov Oddelka za prevajalstvo Univerze v Ljubljani, ki so do oddaje prispevka skupaj pregledali nekaj več kot 8000 primerov (pri čemer je vsak primer potrdilo od 3 do 5 uporabnikov, odvisno od narave naloge).

11 Seznam vseh nalog je objavljen na vstopni strani aplikacije sloCrowd.

uvozu v leksikon besednih oblik Sloleks in njihovi posledični vizualizaciji v obliki kratkega odgovora.

4 Leksikon besednih oblik

Za delovanje portala je ključna povezava med spiskom jezikovnih zadreg, podatki o pogostosti pojavljanja oblik v korpusu Gigafida in podatki o njihovi normativni zaznamovanosti. Vse naštete informacije so združene v leksikonu besednih oblik, saj sam korpus Gigafida normativnih podatkov ne vsebuje – obdelan je s statističnim označevalnikom in lematizatorjem (Grčar, Krek, Dobrovoljc 2012), ki posameznim pojavnicam pripisuje osnovno obliko besede in jim določa jezikoslovne lastnosti, ti podatki pa se v krožnem procesu prenašajo v leksikon.

Format Lexical Markup Framework (LMF),[12] ki je bil uporabljen pri sestavljanju leksikona, omogoča, da vsaki leksikonski enoti oz. obliki pripišemo poljubno število dodatnih informacij. Posamezna enota v leksikonu besednih oblik Sloleks, ki je sicer v obliki spletnega slovarja prosto dostopen na projektni spletni strani,[13] postane del portala Slogovni priročnik šele takrat, ko ji poleg običajnih atributov, ki opredeljujejo oblikoslovne lastnosti in zapis oblik, pripišemo kodo kategorije iz nabora jezikovnih zadreg (vrednost atributa *SPSP*), podatek o normativni zaznamovanosti (vrednost atributa *norma*) in podatek o tipu variantne oblike (vrednost atributa *tip*),[14] sicer je za portal nevidna oz. nerelevantna. Primer tako strukturirane leksikonske enote prikazuje slika 2.

```
<LexicalEntry id="LE_S_Matija" xmlns:d="urn:LEKSIKON_SSJ">
     <feat att="besedna_vrsta" val="samostalnik" />
     <feat att="vrsta" val="lastno_ime" />
     <feat att="spol" val="moški" />
     <feat att="SPSP" val="C1a2a" />
     <Lemma>
          feat att="zapis_oblike" val="Matija" />
     </Lemma>
<…>
     <WordForm>
     <feat att="število" val="ednina" />
     <feat att="sklon" val="rodilnik" />
          <FormRepresentation>
               <feat att="zapis_oblike" val="Matija" />
               <feat att="msd" val="Slmer" />
               <feat att="SPSP" val="C1a2a" />
               <feat att="norma" val="variantno" />
               <feat att="tip" val="C1a2a_s_1" />
               <feat att="pogostnost" val="858" />
          </FormRepresentation>
          <FormRepresentation>
               <feat att="zapis_oblike" val="Matije" />
               <feat att="msd" val="Slmer" />
               <feat att="SPSP" val="C1a2a" />
               <feat att="norma" val="variantno" />
               <feat att="tip" val="C1a2a_s_2" />
               <feat att="pogostnost" val="4018" />
          </FormRepresentation>
     </WordForm>
<…>
</LexicalEntry>
```

Slika 2: Primer zapisa variantnih oblik *Matija* in *Matije* v rodilniku ednine moškega imena Matija v formatu LMF[15]

Leksikon besednih oblik je kot vmesna točka med korpusom in portalom uporabljen predvsem pri zadregah, ki spadajo na področje oblikoslovja, besedotvorja, pravopisa in (ob načrtovani vključitvi podatkov o izgovarjavi) pravorečja, manj pa pri odgovarjanju na vprašanja o skladnji, besedišču in besedilu. Pri teh tematskih sklopih prikazovanje statističnih podatkov o rabi namreč ni smiselno ali pa se ti iz korpusa generirajo neposredno (kot denimo pri že obstoječih odgovorih glede izbire predlogov v/na, s/z oz. tekom/med).

12 Format LMF je od leta 2008 tudi standard ISO za zapis strojno berljivih leksikalnih podatkov. Več o prilagoditvi formata za oblikoslovne bogate jezike, kot je slovenščina, v Krek, Erjavec 2009.

13 http://www.slovenscina.eu/sloleks

14 Atribut *norma* lahko zavzame vrednosti *nestandardno* (za oblike, ki niso v skladu s trenutnim pravopisnim standardom), *variantno* (za več oblik, ki so v skladu s trenutnim pravopisnim standardom) ali *nejasno* (za oblike, kjer norma ni jasno določljiva zaradi neskladij med pravopisnimi pravili in pravopisnim slovarjem). Odsotnost normativne oznake pomeni, da je oblika standardna, torej normativno nezaznamovana. Atribut *tip* je namenjen ločevanju med dvema ali več oblikoslovnimi variantami znotraj kategorije in je sestavljen iz oznake kategorije, podčrtaja, okrajšane normativne oznake in številke, npr. C1a3a_s_1 za Klemna/Jemna, Klemnu/Jemnu, Klemnom/Jemnom itn., C1a3a_s_2 za Klemena, Klemenu, Klemenom itn. ter C1a3a_n_1 za Jemena, Jemena, Jemenu itn.

15 Kategorija C1a2a: Oblikoslovje > Samostalniki > Samostalniki na samoglasnik > Sklanjanje moških samostalnikov, ki se končajo na črko -a.

Prikazovanje tako strukturiranih leksikonskih podatkov v odgovorih Slogovnega priročnika usmerja mehanizem kratkega odgovora, ki ga podrobneje predstavljamo v nadaljevanju.

5 Prikazovanje korpusnih podatkov

Kratki odgovor, ki se navzven prikazuje kot kombinacija grafa in njegovega kratkega opisa, je zapisan v formatu XML in ga sestavljajo trije deli. Prvi (na sliki 3 je označen z elementom <tipi>) definira vse možne variante znotraj kategorije in identifikacijsko oznako (tip), s katero so te označene v leksikonu. Ta legenda variantnih tipov ima zgolj informativni namen in za razliko od drugega in tretjega dela ni namenjena strojni obdelavi.

Drugi del (na sliki 3 ga uvaja element <tabela>) z določanjem pogojev glede pripisane kategorije, tipa, norme ipd. zagotavlja ustrezen izpis podatkov iz leksikona na grafu in v besedilu nad njim ter obenem opredeljuje dodatne povezave z oblikami na spletni konkordančnik korpusa Gigafida oz. na njihovo paradigmo v spletnem vmesniku leksikona Sloleks, kadar je to smiselno. Normativni podatki se na grafu prikazujejo barvno, in sicer modra barva označuje standardne oblike, siva nestandardne, rumena pa oblike, pri katerih norma ni jasno določljiva.

Tretji del vsebuje besedilo, ki se na portalu izpisuje nad grafom. Ker ni nujno, da se v korpusu za vsak primer pojavljajo vse možne variante znotraj kategorije, oziroma se lahko razmerja med variantami s časom spreminjajo, je v tem delu opisano statistično stanje pri vseh možnih kombinacijah standardnih ali nestandardnih oblik za obravnavano kategorijo. Slika 3 (v prvem izmed niza elementov <tekst>) tako prikazuje prvega izmed trinajstih možnih kratkih odgovorov za kategorijo C1a1g (Sklanjanje angleških in

```
<odgovor_Kratko id="C1a1g">
      <tipi>
            <tip kateri="C1a1g_s_1" opis="podaljšava_da,nemi-e_ne" />
            <tip kateri="C1a1g_s_2" opis="podaljšava_ne,nemi-e_ne" />
            <tip kateri="C1a1g_n_1" opis="podaljšava_da,nemi-e_da" />
            <tip kateri="C1a1g_n_2" opis="podaljšava_ne,nemi-e_da" />
      </tipi>
      <tabela iskanje="1234">
            <!-- obrazilo: "Shakespearom" -->
            <beseda katera="1" tip="obrazilo" povezava="gigafida">
                  <pogoj att="norma" val="variantno"/>
                  <pogoj att="tip" val="C1a1g_s_2"/>
            </beseda>
            <!-- obrazilo: "Shakespearjem" -->
            <beseda katera="3" tip="obrazilo" povezava="gigafida">
                  <pogoj att="norma" val="nestandardno"/>
                  <pogoj att="tip" val="C1a1g_n_2"/>
            </beseda>
            <!-- obrazilo: "Shakespearejem" -->
            <beseda katera="4" tip="obrazilo" povezava="gigafida">
                  <pogoj att="norma" val="nestandardno"/>
                  <pogoj att="tip" val="C1a1g_n_1"/>
            </beseda>
            <!-- obrazilo: "Shakespeare" -->
            <beseda katera="5" tip="lema" povezava="gigafida"/>
      </tabela>

<!-- varianta 1: ŠTIRJE, standardno12, nestandardno34 -->
<tekst var="S00.S00.N00.N00" graf="1234">Na grafu si lahko ogledate podatke o rabi
oblik <beseda katera="1"/>, <beseda katera="2"/>, <beseda katera="3"/> in <beseda
katera="4"/> lastnega imena <beseda katera="5"/> v korpusu Gigafida. Obliki,
zapisani z modro, sta ustrezni, sivi pa nista skladni s trenutnim pravopisnim
standardom.</tekst>
<…>
</odgovor_Kratko>
```

Slika 3: Primer kratkega odgovora v formatu XML

francoskih moških imen, ki se končajo na govorjeni [r]).

To pomeni, da je kratki odgovor zasnovan kot univerzalni mehanizem, ki omogoča dinamičen proces prikazovanja podatkov na portalu, saj se lahko nenehno prilagaja novemu stanju v jeziku, kakršno je zabeleženo v kombiniranem sistemu virov besedilni korpus – leksikon besednih oblik – portal Slogovni priročnik. Ko se posodobi in poveča besedilni korpus, se statistični podatki iz korpusa prelijejo v leksikon besednih oblik, na portalu pa so ti podatki uporabljeni za vizualizacijo brez potrebe po poseganju v sistem. Besedila v kratkem odgovoru torej ni treba vedno znova ročno spreminjati in prilagajati novim stanjem, temveč sistem sam izbere pravi odgovor za prikaz glede na stanje, ki ga najde v (redno posodobljenem) leksikonu besednih oblik (Krek 2012: 228).

6 Sklep

Predstavljeni postopek luščenja in analize korpusnih podatkov omogoča, da uporabnikom spletnega portala Slogovni priročnik ponudimo odgovor za tiste leksikonske enote, ki jim v določeni jezikovni zadregi povzročajo največ težav, ne pa naključno izbranih in v večini primerov tudi že desetletja ponavljajočih se primerov, kar je bila praksa dosedanjih pravopisnih priročnikov. Z vnosom teh enot v leksikon besednih oblik se med milijardnim korpusom, osrednjo leksikonsko bazo in spletnim portalom obenem vzpostavi neprekinjen krog, ki omogoča, da se prikaz korpusnih podatkov na portalu avtomatizirano prilagaja spremembam v jeziku.

Literatura

ARHAR, Špela, 2009: Učni korpus SSJ in leksikon besednih oblik za slovenščino. *Jezik in slovstvo* 54/3–4. 43–56. http://www.jezikinslovstvo.com/pdf/2009-03-04-Razprave-Spela-Arhar.pdf

BIZJAK KONČAR, Aleksandra, DOBROVOLJC, Helena, DOBROVOLJC, Kaja, LOGAR BERGINC, Nataša, KOCJANČIČ, Polonca, KREK, Simon, ROZMAN, Tadeja, 2011: *Slogovni priročnik: sporazumevanje v slovenskem jeziku: kazalnik 17 – Standard za korpusno analizo težav pri tvorbi besedil*. http://www.slovenscina.eu/Media/Kazalniki/Kazalnik17/Kazalnik_17_Slogovni_prirocnik_SSJ.pdf

FIŠER, Darja, 2009: Pristopi za avtomatizirano gradnjo semantičnih zbirk. Nina Ledinek, Mojca Žagar Karer, Marjeta Humar (ur.): *Terminologija in sodobna terminografija*. Ljubljana: Založba ZRC, ZRC SAZU. 357–370.

GRČAR, Miha, KREK, Simon, DOBROVOLJC, Kaja, 2012: Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. Tomaž Erjavec, Jerneja Žganec Gros (ur.): *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. 89–94.

KREK, Simon, 2012: Spletni portal Slogovni priročnik. Boža Krakar Vogel (ur.): *Slavistika v regijah – Koper: Zbornik Slavističnega društva Slovenije 23*. Ljubljana: Zveza društev Slavistično društvo Slovenije. 225–231.

KREK, Simon, ERJAVEC, Tomaž, 2009: Standardised Encoding of Morphological Lexica for Slavic Languages. Volodymyr Shyrokov, Ludmila Dimitrova (ur.): *Organization and development of digital lexical resources: proceedings*. *MONDILEX Second Open Workshop, Kyiv, Ukraine*. Kijev: National Academy of Sciences of Ukraine, Ukrainian Lingua-Information Fund. 24–29.

LOGAR BERGINC, Nataša, GRČAR, Miha, BRAKUS, Marko, ERJAVEC, Tomaž, ARHAR HOLDT, Špela, KREK, Simon, 2012: *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko, Fakulteta za družbene vede.

TAVČAR, Aleš, FIŠER, Darja, ERJAVEC, Tomaž, 2012: sloWCrowd: orodje za popravljanje wordneta z izkoriščanjem moči množic. Tomaž Erjavec, Jerneja Žganec Gros (ur.): *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. 197–202.