Leksikalna baza za slovenščino: komu, zakaj in kako (naprej)?

Polona Gantar

Cobiss: 1.01

V prispevku so opisane smernice pri oblikovanju leksikalne baze za slovenščino, zlasti vprašanje različnih uporabnikov ter vrste in načina strukturiranja leksikalno-slovničnih podatkov v njej. Posebej so izpostavljene dileme, ki zadevajo določitev obsega in izbora leksikalnih enot ter razporeditev leksikalno-slovničnih podatkov ob upoštevanju predpostavke, da bodo podatki v leksikalni bazi za slovenščino namenjeni primarno spletnim aplikacijam in sodobnim elektronskim medijem.

Ključne besede: leksikalna podatkovna baza, uporabniška prijaznost, stavčne definicije, leksikografija, spletni slovarji

The Slovenian lexical database: For whom, why, and how (to proceed)?

This article describes the guidelines in the formation of the Slovenian lexical database, especially the issue of various users and the types and manners of structuring lexical and grammatical information in this database. Special emphasis is placed on questions dealing with the scope and selection of lexical units and the arrangement of lexical and grammatical information, while taking into account the premise that information in the lexical database is primarily intended for web applications and modern electronic media.

Keywords: lexicography, lexical database, user friendliness, sentence definitions, online dictionaries

1 Namen

Leksikalna baza za slovenščino (LBS)[1] se izdeluje z namenom, da bo vsebovala zadostno količino relevantnih na korpusu temelječih podatkov o slovenskem besedišču, ki se jih tradicionalno pričakuje od priročnikov tipa *slovar*. Predvsem torej, kaj neka leksikalna enota pomeni in v kakšnem besedilnem in situacijskem kontekstu jo govorci običajno uporabljamo. Na podlagi kompleksnega opisa leksikalnih enot v LBS naj bi bilo mogoče izdelati različne končne opise z različnimi kombinacijami vsebovanih podatkov in v različnih končnih izdelkih, pri čemer je mogoče izpostaviti zlasti splošni enojezični in šolski slovar, slovar za učenje slovenščine kot tujega jezika, dvojezične slovarje in priročnike, ki se osredotočajo samo na določen tip leksikalno-slovničnih podatkov, npr. kolokacijski in sinonimni slovar, slovar

[1] Dostopna na: http://www.slovenscina.eu/Vsebine/Sl/Domov/Domov.aspx.

večbesednih izrazov, frazeologije ipd. V tem smislu je LBS namenjena predvsem leksikografom in jezikovnim analitikom ter še ne predstavlja končnega izdelka.

V procesu oblikovanja leksikalne baze pa smo začeli razmišljati o LBS kot o bazi jezikovnih podatkov, ki bi jo bilo mogoče povezati z drugimi bazami jezikovnih podatkov in jo v obliki večpredstavnega spletnega jezikovnega portala ponuditi neposredno uporabnikom v čim bolj prijazni obliki ter na različnih stopnjah informativnosti in zahtevnosti, pri tem pa izkoristiti možnosti, ki jih ponuja spletni medij. V končni fazi bi to pomenilo, da mora uporabnik v iskalniku zgolj identificirati svoj jezikovni (pravopisni, slovarski, slovnični) problem, sistem pa mu na podlagi različnih baz strukturiranih jezikovnih podatkov ponudi ustrezen in zanesljiv odgovor.

1.1 Preprostost, strokovnost in doseganje *avtoritete* priročnika

»Obstajata dve poti k učinkovitejši rabi slovarjev:
prva je, da radikalno izboljšamo slovarje,
druga je, da radikalno izboljšamo uporabnike.«[2]
(Atkins – Varantola 2008: 337)

Pri zasnovi LBS smo veliko razmišljali o tem, kako si uporabnik dejansko predstavlja zanesljiv odgovor ter v kolikšni meri je pomembna in na kakšnih podlagah se vzpostavlja avtoriteta priročnika.[3] Osnovna dilema ostaja med (a) preprosto informacijo, ki je na račun nujno potrebne kratkosti manj podrobna, s tem pa tudi manj obremenjena s terminologijo in manj poučna, in (b) daljšimi opisi, podkrepljenimi z jezikoslovno analizo problema, s strokovno literaturo in pripadajočo terminologijo. Ob tem se neizogibno zastavljata vprašanji, katere strategije uporabiti za doseganje uporabniške prijaznosti, ne da bi se pri tem morali odreči strokovnosti, in kakšen je dejansko »eleganten in učinkovit pomenski opis«, kot ga denimo omenjata Čermák (2009: 26) in Rundell (2010), ko govorita o uporabniški prijaznosti, učinkovitosti in zanesljivosti slovarske informacije.

Usmeritev k spletni končni obliki LBS (pri čemer so knjižni izdelki vedno lahko njen neposredni produkt) je pomembno vplivala na spremembe v oblikovanju in strukturiranju podatkov v bazi. Vprašanja, ki so bila tradicionalno povezana z iskanjem najboljših rešitev pri logičnem urejanju podatkov, kjer je moral uporabnik obvladati tudi logiko notranje strukturiranosti gesla (abecedna ureditev, vsebina gesla, oblika iztočnice, gnezdenje, terminologija ipd.), so postala bolj ali manj tehnične narave: abecedna ureditev ni več relevantna, poznavanje logične ureditve gesla v smislu razumevanja besednovrstnih konverzij, osnovne oblike, gnezdenja ipd. ni več potrebno, in tudi ne omejitev količine informacij z vidika porabe prostora. Nastala pa je vrsta novih vprašanj, povezanih z značilnostmi in možnostmi spletnega medija, med drugim:

(a) opredelitev obsega leksikalne enote oz. segmenta besed, ki predstavlja za uporabnika potencialni slovarski problem;

[2] »There are two direct routes to more effective dictionary use: the first is to radically improve the dictionary; the second is to radically improve the users.« (Prevod P. G.)

[3] To vprašanje velja v prvi vrsti za pravopis, vendar pa je pomembno tudi za slovarske informacije, zlasti pri stopnji pomenske členitve in pri pomenskih opisih.

(b) strukturiranost leksikalno-slovničnih podatkov v različnih stopnjah zahtevnosti oz. informativnosti;
(c) organizacija spletne strani z vidika navigacije in iskanja podatkov (hitra dostopnost in učinkovitost dane informacije) ter povezava tako na druge podatkovne vire znotraj spletnega portala kot na druge razpoložljive spletne jezikovne vire;
(č) način prikaza besedila na strani (stalne in opcijske rubrike) ter medbesedilne povezave (hiperpovezave, pasice, pomoč ipd.);
(d) oblikovne možnosti elektronskega besedila s stališča multimedijskosti (vključitev slik, zvoka, videa, izvažanje in tiskanje podatkov, dodajanje zaznamkov ipd.);
(e) možnost povezav z drugimi bazami, kot so npr. Wikipedija, Wordnet, FrameNet, in s spletom nasploh.

Poleg za človeškega uporabnika je bila LBS že od samega začetka predvidena tudi za namene računalniške obdelave naravnega jezika (RONJ**)**, konkretneje za izboljšanje razčlenjevalnika[4] in označevalnika[5] za slovenščino, računamo pa tudi, da bo na podlagi LBS mogoče narediti prve poskuse samodejnega razdvoumljanja pomenov slovenskih leksemov. Temu namenu so podrejeni podatki o (a) besednovrstni strukturiranosti pomensko relevantnih besednih zvez (ali stavčnih fraz, npr. pridevnik + samostalnik; samostalnik + samostalnik v rodilniku) – t. i. skladenjske strukture in stavčni vzorci (pri glagolskih iztočnicah), in (b) beleženje udeležencev s t. i. semantičnimi tipi v stavčno strukturiranih pomenskih shemah. Semantični tipi udeležencev so skupaj s pomenskimi indikatorji (neposrednimi nadpomenkami ali sinonimi) tudi kandidati za dopolnjevanje slovenske ontološke mreže sloWNet[6] (Fišer 2009).

1.2 Ciljni uporabniki

Glede na izhodiščno dvonamenskost LBS sta tudi potencialna uporabnika LBS dva: človek in računalnik. Človeškemu uporabniku so v prvi vrsti namenjene pomenske informacije: pomenska členitev (z oblikovanjem pomenskega menija), opis pomena s pomenskim indikatorjem in pomensko shemo, ki predstavlja izhodišče za oblikovanje stavčno strukturiranih razlag (Gantar – Krek 2009), razlaga ter kolokacije in korpusni zgledi. Vse druge informacije – skladenjske strukture in stavčni vzorci, vključno s semantičnimi tipi udeležencev v stavčno strukturirani razlagi pri posameznih pomenih glagolov ter pri nekaterih pomenih pridevnikov in samostalnikov – so namenjene primarno RONJ in slovničnim analizam.

---

[4] Rezultate skladenjske razčlenitve za poljubno besedilo v slovenščini je mogoče preveriti na spletnem servisu projektne strani http://razclenjevalnik.slovenscina.eu/ (dostop 29. 9. 2011).

[5] Program je bil razvit pri projektih Jezikoslovno označevanje slovenščine (http://nl.ijs.si/jos/) in Sporazumevanje v slovenskem jeziku, njegovo delovanje pa je mogoče preizkusiti na: http://oznacevalnik.slovenscina.eu (dostop 29. 9. 2011).

[6] Več o projektu na http://lojze.lugos.si/~darja/slownet.html (dostop 29. 9. 2011).

1.2.1 Združljivost različnih profilov uporabnikov

Človeški uporabnik LBS zajema tri profile: (a) splošnega uporabnika, pri čemer se predvideva raven zahtevnosti in obvestilnosti na ravni srednješolske oz. gimnazijske izobrazbe (prim. SSKJ 1, Uvod: XI),[7] kar glede vsebine informacij ustreza tudi (b) šolskemu uporabniku in (c) vsaj v nekaterih segmentih tudi učencu slovenščine kot tujega jezika. Tujejezični in šolski uporabnik, ali bolje, vsebina leksikalno-slovničnih podatkov, ki so namenjeni enemu ali drugemu, zahteva upoštevanje določenih specifik in razlik. Šolskemu uporabniku smo se želeli približati z izborom specializiranega besedišča (gl. 1.3.1),[8] obenem pa smo prilagodili tudi pomenske opise v obliki pomenskih indikatorjev in pomenskih shem, ki predstavljajo izhodišče za oblikovanje razlag stavčnega tipa. Poleg splošnih načel smo pri oblikovanju pomenskih indikatorjev težili k razumljivosti in kratkosti, hkrati pa naj bi pomenski indikatorji kot gradniki pomenskih menijev vzpostavljali zadostno mero pomenske razločevalnosti glede na druge pomene besede v iztočnici. Poleg tega smo menili, da so za razumevanje pomena z vidika uporabnikov, ki se slovenščino učijo kot tuji jezik, najprimernejši pomenski opisi v obliki *stavčnih definicij*, ki najbolj naravno (npr. z izkazano tipično skladenjsko realizacijo, kot je denimo pri nekaterih pridevnikih povedna ali primarno prilastkova raba, povratnosvojilnost pri glagolih ipd.) vključujejo podatke o besedilnem okolju, v katerem se realizira pomen. Poleg tipične skladenjske rabe gre predvsem za izpostavitev udeležencev, razmerij med njimi in okoliščin (tudi zunajjezikovnih), ki so potrebne za razumevanje posameznega pomena. Kljub omenjenim premislekom pa ostaja odprto vprašanje, ali je omenjeni trojni profil uporabnika dejansko združljiv tudi v enem samem slovarskem izdelku. V nadaljevanju projekta želimo zato konkretne rešitve, zlasti berljivost stavčnih razlag in obvestilnost pomenskih menijev, preveriti pri različnih ciljnih skupinah uporabnikov.

1.3 Vrsta in strukturiranje leksikalno-slovničnih podatkov

1.3.1 Viri

Primarni nabor iztočnic v LBS izhaja s seznama 5000 najpogostejših lem v korpusu FidaPLUS, ker pa je bil v okviru projekta SSJ zgrajen nov milijardni referenčni korpus Gigafida (Logar Berginc – Šuster 2009; Logar Berginc – Krek 2010), smo v nadaljevanju izdelave LBS podatke pridobivali iz novega korpusa. Poleg tega smo z namenom približati se šolskemu uporabniku na podlagi korpusa osnovno- in

[7] Dejstvo, da se 60 % populacije po končani srednji šoli ali gimnaziji vpiše na fakulteto, narekuje potrebo po novi definiciji splošnega uporabnika, pri čemer izobrazbeno izhodišče v smislu večje ali manjše zahtevnosti oz. preprostosti slovarja ni bistveno – vsaj v našem primeru ne, saj smo si prizadevali za preproste pomenske opise in zmanjšanje slovničnega in slovarskega metajezika ne glede na izobrazbeno lestvico potencialnega uporabnika. Pri zasnovi baze ali slovarja je smiselno končnega uporabnika opredeliti predvsem glede na to, ali gre za otroka ali za odraslega uporabnika in ali so podatki in njihov opis namenjeni rojenemu govorcu ali učencu tujega jezika (Atkins 2008: 37).

[8] LBS bo ob zaključku aktivnosti (junij 2012) vsebovala 2500 iztočnic, od tega jih bo približno 500 vključenih na podlagi učbeniškega geslovnika.

srednješolskih učbeniških besedil izdelali geslovnik s približno 1000 lemami, kjer smo poleg frekvence upoštevali še večpomenskost (strokovni izrazi, ki imajo tudi splošni pomen, in strokovni izrazi, ki prehajajo v splošni jezik), splošno rabljene aktualne prevzete besede, pa tudi besede, ki so po našem mnenju za šolskega uporabnika zanimive z vidika učnih vsebin, nove predmetnosti (neregistrirani izrazi v SSKJ) in generacijske pripadnosti.

Izhajajoč iz vsaj deloma različnih potreb predvidenih skupin ciljnih uporabnikov, je smiselno v LBS s širokim spektrom uporabnosti zajeti čim več leksikalno in slovnično relevantnih podatkov. Z vidika sodobne leksikografije to ni več mogoče brez obsežnih besedilnih korpusov, hkrati pa obsežne količine podatkov poleg izstopajočih frekventnih pojavov pokažejo tudi jezikovno variantnost in posebnosti v vsej njihovi razsežnosti. To dejstvo neizogibno vodi v iskanje odločitev, kaj od obrobnega je poleg tipičnega za uporabnika prav tako zanimivo/pomembno, ne nazadnje tudi z vidika pojavov, ki se v jeziku šele uveljavljajo oz. se uveljavljajo zgolj v specifičnih jezikovnih situacijah. Odločitve glede tega morajo med drugim upoštevati dejstvo, da npr. orodje Sketch Engine (SkE),[9] ki omogoča hitrejše pridobivanje relevantnih podatkov iz korpusa, določenih relacij bodisi zaradi zapletenosti slovnice besednih skic, ki so pogoj za generiranje kolokacijskega obnašanja besed, bodisi zaradi nefrekventnosti določenega pojava/posebnosti ne izpostavi. Ena izmed rešitev tega problema je sprotno izboljševanje slovnice besednih skic na podlagi povratnih informacij iz baze ter preizkušanje slovnic in orodij, ki so bili izdelani za druge jezike.[10] Druga rešitev je odločitev, da ostaja temeljni vir za pridobivanje leksikalno-slovničnih podatkov v LBS ročna analiza najmanj 150 do 300 konkordanc. Na podlagi ročne analize konkordanc leksikograf izdela osnovno pomensko sliko besede (določi osnovne pomene in podpomene), oblikuje pomenski meni, registrira tipični besedilni kontekst za posamezne pomene, udeležence (oz. prehodnost pri pridevnikih in samostalnikih), stalne zveze in frazeološke enote. Ko je na tej podlagi izdelana osnovna pomenska slika konkretne leme v iztočnici, je s pomočjo orodja Sketch Engine oz. aplikacije Besedne skice za slovenščino (Krek – Kilgarriff 2006) izdelan kolokabilni del geselske strukture (kolokacije in pripadajoče skladenjske strukture), s pomočjo aplikacije GDEX, ki je bila v okviru projekta prilagojena posebej za slovenščino (Kosem idr. 2011), pa so izbrani dobri korpusni zgledi.

[9] SkE ⟨http://www.sketchengine.co.uk/⟩ (dostop 29. 9. 2011) je le eno od – sicer že solidno standardiziranih – orodij za luščenje leksikografskih podatkov iz korpusa. Jezikovno oz. slovensko specifične parametre za luščenje podatkov bomo pri izdelavi LBS testirali pri poskusu samodejne izdelave gesel, kjer bomo kot učno množico uporabili v bazi že strukturirane leksikalno-slovnične podatke.

[10] Na korpusu FidaPLUS smo preizkusili slovaško varianto slovnice besednih skic, ki jo je za potrebe izdelave Slovarja sodobnega slovaškega jezika (Slovník súčasného slovenského jazyka (A–G)) izdelal Vladimír Benko, v prihodnje pa nameravamo preizkusiti tudi sistem avtomatskega luščenja relevantnih leksikografskih podatkov iz korpusa, ki ga uporabljajo na Inštitutu za nemški jezik v Mannheimu in ki ga je razvil Cyril Belica.

Primer 1: Glagolsko geslo s pripadajočimi pomenskimi in kolokacijskimi podatki

POPOPRATI *glagol*
1. dodati poper
2. popestriti

1. indikator dodati poper
pomenska shema če ČLOVEK *popopra* JED ali ŽIVILO, ji doda poper, s čimer dobi poseben, nekoliko pekoč okus
kolokacije
[rahlo, obilno] popoprati
popoprati [jed, meso, zrezke]
razširjene kolokacije
popoprati s [črnim, belim, zmletim, mletim] poprom
popoprati z [grobo, sveže] zmletim poprom
skladenjske zveze
popoprati po okusu
2. indikator popestriti
pomenska shema če ČLOVEK popopra IZJAVO, DOGODEK ali VZDUŠJE, jo s pripombo ali dejanjem zaostri ali naredi bolj zanimivo

Dodatnih gradivnih virov pri izdelavi LBS ne predvidevamo, se pa pri razbiranju pomena in oblikovanju pomenskega opisa stalnih zvez, ki so v rabi na specializiranih področjih in hkrati del splošnega jezika, avtorji zatekajo tudi k spletnemu iskanju informacij. Take zveze je namreč pogosto težko pomensko opisati zgolj na podlagi konkordanc, saj zahtevajo specializirano védenje, hkrati pa mora biti njihov opis preprost, namenjen splošnemu uporabniku in ne strokovnjaku. Glede na to, da LBS ni zasnovana kot terminološka baza in da hkrati predvidevamo njeno objavo znotraj širšega jezikovnega portala, se ponuja možnost napotitve uporabnika na relevanten vir v obliki spletnih povezav, npr. *Wikipedija*, *islovar* ipd. Za zdaj se v primeru terminoloških stalnih zvez, ki jih uvrščamo pod posamezne pomene ali od pomena neodvisno, odločamo le za navedbo ustreznega področja rabe (t. i. področne oznake), ki pa je lahko kombinirana z razlago v pomenski shemi:

Primer 2: Umestitev stalnih zvez v samostalniško geslo

GREDA *samostalnik*
1. indikator del vrta
pomenska shema *greda* je del vrta ali njive, v katerem so v vrsti posajene rastline
SZ-pomen topla greda
SZ-pomen zaprta greda
pomenska shema *topla greda* je umetno narejen prostor, ki zagotavlja toplotne razmere, v katerih je mogoče gojiti ali prezimovati rastline
kolokacije
[prenosna] topla greda
[sejati, posejati, posaditi] v toplo gredo

SZ-geslo topla greda
oznaka ekologija
pomenska shema *topla greda* je rezultat procesa, pri katerem se toplotno sevanje, ki prihaja v ozračje z Zemlje, vrača nazaj in povzroča višjo temperaturo, kot bi bila, če bi Zemljino površje ogrevalo le sonce
kolokacije
[učinek] tople grede
[povzročati] toplo gredo

1.3.2 Vrste leksikalnih enot v leksikalni bazi za slovenščino

Glede na možnosti spletne postavitve LBS je za uporabnika poznavanje logične urejenosti podatkov znotraj gesla manj pomembno. Pomembno pa je pri vključevanju podatkov določiti vrsto in obseg leksikalne enote, na katero so podatki pripeti. V LBS obravnavamo kot leksikalno enoto (a) vsak pomen in podpomen besede v iztočnici ter (b) stalne zveze in (c) frazeološke enote. Te enote v LBS predvidevajo pomenski opis, so lahko opredeljene glede na področje rabe, stil in besedilni kontekst (s t. i. oznakami) ter imajo evidentirano tipično besedilno okolje.

Poleg tega kot samostojne enote v LBS obravnavamo tudi t. i. *skladenjske zveze*, ki so ustaljeni večbesedni delci jezika, za katere je značilno, da izkazujejo relativno pomensko prozornost (pomenskega opisa zato zanje ne predvidevamo) in strukturno trdnost ob relativno spremenljivem besedilnem okolju oz. oblikovno napovedljivem vezljivostnem mestu, npr. *pod vplivom česa*, *v skladu s čim/kom*, *v času* (*česa*), *v barvi* (*česa*) ipd.

Vrste leksikalnih enot v LBS in struktura leksikalno-slovničnih podatkov, ki jih predvidevajo

Leksikalna enota	**Pomenski opis**	**Sintagmatika**	**Oblike rabe**	**RONJ**
pomen ali podpomen	– indikator – pomenska shema	– kolokacije – razširjene kolokacije	–besednovrstna konverzija* – restrikcije (ustaljenost v določenem številu, skladenjskem položaju ipd.)	– oblikovanje ontologij – formalizacija besednozvezne strukture, npr. Prid + Sam – stavčni vzorci – formalizacija udeležencev v stavčni razlagi
stalna zveza	– indikator/ razlaga	– kolokacije – variante	– oblike rabe	
skladenjska zveza		- kolokacije	- oblike rabe – pretvorbe	
frazeološka enota	– indikator/ razlaga	– kolokacije	– oblike rabe – pretvorbe	

* Besednovrstno konverzijo (nominalizacija, adjektivizacija ipd.) in homonimijo (prekrivnost celotne paradigme znotraj iste besedne vrste) obravnavamo kot samostojne (pod)pomene in ne kot samostojne iztočnice ali podiztočnice.

1.3.3 Obseg leksikalne enote

Razmerja med posameznimi vrstami leksikalnih enot – v LBS zlasti med kolokacijami, razširjenimi kolokacijami in stalnimi zvezami, med razširjenimi kolokacijami in skladenjskimi zvezami, med kolokacijami in stalnimi zvezami in ne nazadnje med stalnimi zvezami in frazeološkimi enotami – so večkrat zabrisana, zato so potrebna čim bolj jasna načela, ki leksikografom omogočajo čim bolj enotne odločitve. Pri snovanju LBS smo se pri določanju vsebinskih in formalnih parametrov za prepoznavanje zgoraj omenjenih leksikalnih enot opirali tako na teoretična spoznanja kot na praktične izkušnje pri analizi korpusa (Gantar idr. 2009; 2009a). Izkazalo se je, da posamezne besede izkazujejo bolj ali manj obsežne kolokabilne nize. Teoretično je mogoče predvidevati, da kolokabilno zaprti nizi napovedujejo pomensko in strukturno trdnost zveze, torej potencialne stalne zveze ali frazeološke enote, in obratno: bolj odprt oz. obsežen kot je kolokabilni niz, več možnosti je, da gre za tipično besedilno okolje besede, tj. za kolokacijo in ne za leksikalizirano (stalno ali frazeološko) zvezo. V praksi pa se je izkazalo, da je obsežnost kolokabilnega niza glede na prepoznavanje samostojnih leksikalnih enot relativna, npr. *šola* v pomenu ‘ustanova’ kolocira z besedami kot *osnovna*, *višja*, *srednja*, *visoka* ipd. in hkrati skupaj z omenjenimi pridevniki tvori samostojne leksikalne enote, ki potrebujejo lastni pomenski opis: *osnovna šola*, *srednja šola*, *visoka šola* itd. Rešitev, ki smo jo glede tega sprejeli v LBS, je, da navedemo celotni kolokacijski niz pri ustreznem pomenu samostalnika, hkrati pa še samostojne stalne zveze, ki jih pomensko opišemo in jim določimo njihovo lastno kolokabilno okolje, če obstaja, npr. [*vpisati se*, *hoditi*] *v osnovno šolo*, [*končati*, *obiskovati*] *osnovno šolo*, [*devetletna*, *osemletna*] *osnovna šola* itd. To pomeni, da bo uporabniku podatek na voljo v obliki kolokacije in stalne zveze s pomenskim opisom. V nadaljevanju projekta želimo določiti in preveriti predvsem mehanizme samodejnega prepoznavanja leksikalno relevantnih besednih zvez, in sicer z upoštevanjem že registriranih in formaliziranih skladenjskih struktur, ki se tipično pojavljajo pri posameznih besednih vrstah, s testiranjem različnih statističnih vrednosti medbesedne povezovalnosti in z izboljšavami slovnice besednih skic v orodju Sketch Engine.

Strategije pri ločevanju kolokacij od stalnih zvez se pri različnih besednih vrstah razlikujejo. Pri pridevnikih predvidevamo večje število stalnih zvez in manj samostojnih pomenov. Pri samostalnikih registriramo stalne zveze pod posameznimi pomeni ali pa od pomena neodvisno, če tvorijo pomensko samostojne leksikalne enote. Pri glagolih stalnih zvez ne beležimo zaradi možnosti različnih funkcijskih realizacij (konverzij oz. transformacij), ki jih omogočajo glagolske zveze. Zveze glagola z ustaljeno besednozvezno kombinacijo (tipično predloga in samostalnika) beležimo bodisi pri ustrezni samostalniški iztočnici (gl. primere spodaj), pri frazeoloških enotah (če presodimo, da gre za pomensko in/ali strukturno samosvojo enoto, ki potrebuje lasten pomenski opis) ali pri skladenjskih zvezah (brez pomenskega opisa), saj menimo, da nastopajo predvsem kot niz različic ob sicer trdnem besednozveznem jedru, npr.

Primer 3: Skladenjske zveze pri samostalniških iztočnicah

[začiniti, soliti, popoprati, sladkati ...] po okusu
[prebijati se, živeti, shajati, životariti, preživeti ...] iz meseca v mesec
[padati, prileteti, spustiti se] pod kotom [x] stopinj
[gibati se, krožiti, vrteti, masirati, nadaljevati] v smeri urnega kazalca

Z vidika tujejezičnega uporabnika (ki se uči slovenščino kot tuji jezik) ostaja odprto vprašanje, ali je tudi omenjene skladenjske zveze smiselno pomensko opisati, saj je merilo pomenske prozornosti, ki pravi, da je »pomen zveze več kot vsota pomenov njenih delov« (Atkins – Rundell 2008: 167), vezano na občutek rojenega govorca, pomisleke glede razumljivosti takih zvez pa vzbuja tudi dejstvo, da niso vedno neposredno prevedljive v tuji jezik, npr. češ. *po čase/začas*[11] → sln. *čez čas/sčasoma*; ang. *in (less than) no time* → sln. *v hipu/kot bi trenil/takoj*;[12] nem. *höchste Zeit* → sln. *skrajni čas*.[13]

Možnost podrejanja stalnih zvez posameznemu pomenu besede v iztočnici (gl. zgoraj primer *greda*) med drugim predvideva iskanje – in posledično posredovanje tega podatka uporabniku – pomenske sorodnosti stalne zveze kot celote ali njenih sestavin s katerim od pomenov iztočnice. Ker je tako pomensko povezanost težko identificirati in ker se je v praksi pokazalo, da so odločitve slovaropiscev pogostokrat različne ali celo nasprotujoče si, ostaja vprašanje smiselnosti pomenskega podrejanja stalnih zvez sploh. Iz istih razlogov smo se odločili, da pomensko ne podrejamo frazeoloških enot, čeprav je v nekaterih primerih pomenska povezava katere od sestavin frazeološke enote s katerim od registriranih pomenov očitna (gl. v nadaljevanju primer samostalnika *oblak*). Prepoznavanje razlik med stalnimi zvezami in frazeološkimi enotami je za spletno postavitev LBS za uporabnika manj pomembno, saj tako ene kot druge predvidevajo enak tip podatkov: pomenski opis (in pomensko členitev), registracijo različic in pretvorbenih možnosti, evidentiranje tipičnega besedilnega okolja in predstavitev s korpusnimi zgledi.

V primerjavi s posameznimi pomeni besede v iztočnici stalne zveze in frazeološke enote ne predvidevajo strukturne analize, kar je zlasti pomembno za frazeološke enote, za katere velja, da imajo anomalno strukturno in pomensko zgradbo,[14] zato njihove formalne sestave ne beležimo v obliki skladenjskih struktur. Odprto pri tem ostaja vprašanje, ali je ta praksa z vidika RONJ dejansko sprejemljiva tudi za stalne zveze, ki so v primerjavi s frazeološkimi enotami pogosto strukturirane kot običajne samostalniške, pridevniške in prislovne zveze, npr. Sam + Prid: *arhivsko vino*, *varovalna barva*, *biotska raznovrstnost/raznolikost*; Sam + Sam: *avtomobil bomba* itd.

[11] Primer za češčino je povzet po: *Slovník české frazeologie a idiomatiky: výrazy neslovesné*, Praha: Academia, 1988.

[12] Primer za angleščino s slovenskim ustreznikom je povzet po: *Veliki angleško slovenski slovar* 2: *L–Z*, Ljubljana: DZS, 2006.

[13] Primer za nemščino s slovenskim ustreznikom je povzet po: Doris Debenjak idr., *Veliki nemško-slovenski slovar*, Ljubljana: DZS, 1993.

[14] »[...] frazem ali idiom je enkratna zveza najmanj dveh prvin, od katerih vsaj ena funkcionira v konkretni zvezi na drugačen način kot v drugih zvezah oz. se kot taka pojavlja zgolj v konkretnem izrazu« (Čermák 1985: 177). Prevod P. G.

V nadaljevanju bomo pri izdelavi LBS preverjali tudi različne načine navajanja pretvorbenih možnosti, ki jih izkazuje dejanska raba frazeoloških enot (tj. razmerje med osnovno obliko enote in tipičnimi variantnimi oblikami in pretvorbami), in možnosti opisa pomenskih odtenkov, ki jih različne stavčne funkcije frazeoloških enot prinašajo s seboj, čeprav gre v večini primerov le za variantne oz. pretvorbene možnosti izhodiščnih samostalniških ali glagolskih zvez, npr. *trn v peti*: *biti čigav trn v peti*, *biti trn v peti koga*, *biti trn v peti za koga*, *biti trn v čigavi peti*, *biti komu trn v peti* itd.

Posebej smo se ustavili ob vprašanju, kako obravnavati razmerje med metaforičnim pomenom besede in pomenom frazeološke enote kot celote. Glede na to, da se za pomensko podrejanje frazeoloških enot besednim pomenom nismo odločili, se zastavlja vprašanje, ali določene zveze obravnavati kot kolokacije pri ustreznem (navadno metaforičnem) pomenu besede ali kot frazeološke enote s samostojnim pomenskim opisom in predvidljivim besedilnim okoljem. Primer za to je npr. samostalnik *oblak*, pri katerem je eden od pomenov ‘nekaj nerealnega ali oddaljenega, česar si ljudje želijo ali o čemer sanjarijo’, ki se v istem pomenu pojavlja tudi v bolj ali manj ustaljenih zvezah, kot denimo: [*živeti*, *plavati*, *biti*] *v oblakih*; *spustiti se z oblakov* (*na trdna/realna tla/na zemljo*), hkrati pa tudi zunaj njih, kot denimo v zgledu *S prijatelji boste sanjali o novih načrtih*, *a bo za zdaj vse ostalo v oblakih*.

1.3.4 Stopnja in način pomenske členitve

Leksikalne enote členimo v LBS pomensko zelo podrobno na podlagi priporočila, da je to smiselno predvsem v fazi oblikovanja podatkovne baze (Atkins – Rundell 2008: 268). S podrobno razdeljenimi pomeni v podatkovni bazi je namreč slovaropiscem na voljo celoten spekter možnosti, iz katerih lahko izpeljejo več različnih slovarjev. V praksi to pomeni, da temelji pomenska členitev na različnih kolokatorjih, zlasti pri pridevnikih, npr.

SMUČARSKI *pridevnik*
1 namenjen smučanju in smučarjem
kolokacije
smučarska [vozovnica, karta]
smučarski [center, tečaj]
smučarska [šola]
1.1 o opremi za smučanje
kolokacije
smučarski [čevelj]
smučarske [palice]
2 o športu
kolokacije
smučarski [skoki, poleti]
smučarski [reprezentant]

Smiselnost podrobne pomenske členitve je torej predvsem v možnosti poznejše sinteze in v uporabnosti za različne končne izdelke in/ali uporabnike, se pa kljub temu

poraja dilema, ali je podrobna pomenska členitev enako smiselna pri vseh besednih vrstah. Iz zgornjega primera je razvidno, da bistvo pomena dejansko temelji na samostalniškem jedru, npr. *smučarske palice*, s tem pa se vsaj v strukturi leksikalne baze znova zastavlja vprašanje razmejevanja med kolokacijami in stalnimi zvezami.

(a) Osamosvojitev pragmatične informacije ali vključitev v pomenski opis?
Vse leksikalne enote v LBS predvidevajo pomenski opis, in sicer na dveh ravneh: (a) s t. i. pomenskimi indikatorji, ki so primarno namenjeni oblikovanju pomenskega menija (sinonimi in neposredne nadpomenke, s katerimi dosežemo hitro navigacijo po geslu in možnost hitre in ustrezne identifikacije »pomenskega« problema), ter (b) v obliki stavčno strukturirane pomenske sheme pri glagolih in nekaterih pomenih pridevnikov in samostalnikov ali (b1) v obliki razlage, zlasti pri (nevezljivih) samostalnikih in stalnih zvezah. Ključni pomenski opis, ki mora zadovoljiti trojni profil potencialnega uporabnika, je stavčno strukturirana pomenska shema oz. razlaga. Bistvo stavčne razlage je med drugim vključitev pragmatičnih pomenskih sestavin (podčrtano), ki so nujno potrebne za ustrezno razumevanje in tvorjenje besedila, npr.

> CRKNITI *glagol*
> 1 umreti
> če rečemo, naj ČLOVEK *crkne*, na zelo grob način povemo, da nam je vseeno, če umre, ali da mu to celo privoščimo
> 1.1 poginiti
> če rečemo, da je ŽIVAL *crknila*, na grob način povemo, da je poginila
> 1.2 oveneti
> če rečemo, da RASTLINA *crkne*, na grob način povemo, da oveni, navadno zato, ker zanjo ne skrbimo ali ker nima ustreznih razmer za uspevanje

Sprva smo za pragmatične informacije predvideli samostojni element <pragmatika>, in sicer z namenom, da bi bilo mogoče iz celotne baze naknadno samodejno pridobiti pragmatične informacije, povezane s pomenom posamezne leksikalne enote. V praksi pa se je pokazalo, da so posamezni pragmatični segmenti znotraj pomenskih razlag težko opredeljivi, ker so sestavni del celotnega pomenskega opisa. Hkrati se je tudi pokazalo, da leksikografi posameznih elementov znotraj pomenskih opisov ne prepoznavajo kot pragmatične (ali jih prepoznavajo zelo različno), in ne nazadnje so precej slabi tudi rezultati anket, ki merijo razumevanje zlasti slovničnih in pragmatičnih informacij v obliki okrajšav, torej v obliki metainformacij, ločeno od pomenskega opisa (Rozman idr. 2010). V spodnjih zgledih so podčrtani pragmatični deli razlag, ki jih avtorji niso vključili v samostojni element <pragmatika>:

> če ČLOVEK *benti* nad drugim ČLOVEKOM, DOGAJANJEM ali obstoječimi RAZMERAMI, izraža negodovanje ali nestrinjanje, navadno tako, da uporablja kletvice in žaljive besede

če ČLOVEK *blebeta*, veliko govori, navadno nepremišljeno ali o nepomembnih STVAREH

brezbarven OBRAZ ali del obraza je bled, navadno zaradi negativnih čustev, kot sta jeza ali strah

Zaradi omenjenih ugotovitev smo opustili ločevanje pragmatičnih elementov znotraj pomenske sheme/razlage, hkrati pa smo uvedli element <oznaka>, v katerem imamo možnost eksplicitno opredeliti lastnosti pomena, kot so specifičen govorni položaj, npr. *v neformalni/formalni situaciji*, odnos govorca do vsebine sporočila, npr. *odklonilno*, *slabšalno*, *kot grožnja* ipd. Trenutno se seznam možnih oznak pragmatičnega tipa še dopolnjuje na podlagi korpusne analize konkordanc.

1.3.5 Skladenjske informacije v LBS in omejitve v rabi

Korpusna analiza in tudi tipična skladenjska razmerja, ki jih zapolnjujejo kolokacije, kažejo, da so izbire med možnimi skladenjskimi realizacijami (tj. tistimi, ki jih omogoča slovenska slovnica) pogostokrat omejene bodisi (a) zgolj na določene izbire, kar je zlasti pogosto pri frazeoloških enotah, da so (b) bolj tipične v kateri od možnih skladenjskih realizacij (npr. pasivizacija, raba pridevnika v povedkovem določilu, omejenost v določeni osebi, spolu, številu ipd.) ali pa (c) posamezne možnosti govorci izkoriščamo bolj ali manj enakovredno.

(a) Implicitnost ali eksplicitnost skladenjskih in drugih slovničnih omejitev v rabi

Tovrstne informacije, zlasti ko gre za omejene ali celo nerealizirane sicer možne izbire, so po našem mnenju zelo pomembne za šolskega uporabnika in za učenca slovenščine kot tujega jezika, pa tudi za RONJ, zato predvidevamo dve možnosti njihovega eksplicitnega beleženja:

(a) znotraj skladenjskih struktur v elementu restrikcija (poudarjeno), npr.

ARGUMENT *samostalnik*
struktura: gbz**<r>brezosebno</r>** SBZ2
kolokacija [zmanjkuje, zmanjka] argumentov

CIVILIST *samostalnik*
struktura: gbz na SBZ4**<r>navadno v množini</r>**
kolokacija: [streljati] na civiliste

DVOUMEN *pridevnik*
struktura: Kol-rbz PBZ2**<r>v samostalniški rabi</r>**
kolokacija: [nekaj, veliko] dvoumnega

in (b) v obliki tipičnih stavčnih vzorcev (pri glagolih), ki poleg prototipične skladenjske realizacije, razvidne iz stavčne razlage, izkazujejo še druge tipične možnosti, npr.

BRITI *glagol*
če se MOŠKI *brije* |ali| če si MOŠKI brije BRADO ali BRKE, si s PRIPOMOČKOM odstranjuje dlake
- briti se
- kdo se brije
- briti si kaj
- briti se s čim

BENTITI *glagol*
če ČLOVEK *benti* nad drugim ČLOVEKOM, DOGAJANJEM ali obstoječimi RAZMERAMI, izraža negodovanje ali nestrinjanje, navadno tako, da uporablja kletvice in žaljive besede
- kdo benti
- bentiti nad čim/kom
- bentiti čez koga/kaj
- benti na koga/kaj
- bentiti, ker
- bentiti zaradi česa

Tretja možnost izražanja skladenjskih omejitev v LBS je implicitna, vključena v stavčno razlago pomena besede v iztočnici (podčrtano), kar pride do izraza zlasti pri pridevnikih in samostalnikih:

POZOREN *pridevnik*
1 ustrežljiv
če je ČLOVEK *pozoren* do drugega ČLOVEKA, je do njega ustrežljiv in mu izkazuje naklonjenost

NOTA *samostalnik*
1 značilnost
če neke LASTNOSTI dajejo IZDELKU, KRAJU ali DEJANJEM svojo *noto*, se v njem izražajo in ga delajo posebnega

2 Sklep

Leksikalna baza za slovenščino vsebuje kompleksno strukturo leksikalno-slovničnih podatkov na podlagi korpusne analize in je namenjena izdelavi različnih končnih izdelkov slovarskega tipa. Obenem je zasnovana kot baza podatkov, ki bo uporabniku dostopna v okviru širšega jezikovnega portala na spletu. Poleg slovarskih podatkov vsebuje tudi podatke, ki so primarno namenjeni računalniški obdelavi in

izboljšavi jezikovnotehnoloških aplikacij za slovenščino. Izraba spletnega medija je za razliko od klasičnih slovarjev v knjižni obliki izpostavila nove probleme pri vključevanju in strukturiranju leksikalno-slovničnih podatkov. Pri tem je v ospredju zagotavljanje relevantne, hitro dostopne in zanesljive informacije na uporabniku čim bolj prijazen način. Podatki v leksikalni bazi za slovenščino so namenjeni trem različnim profilom uporabnikov – splošnemu, šolskemu in učencu slovenščine kot tujega jezika –, hkrati pa so sprejete konkretne rešitve pri upoštevanju različnih potreb posameznega profila, med drugim z vključitvijo učbeniškega geslovnika, z oblikovanjem razlag stavčnega tipa in z navajanjem tipičnih skladenjskih uresničitev in omejitev v njihovi rabi. Pri nadaljnjem delu za leksikalno bazo bomo z vidika treh tipov uporabnikov s pomočjo anket preverili učinkovitost pomenskih opisov, uporabnost pomenskih izbir in ustrezno razmejitev informacij glede na predvidene tipe v bazi opisanih leksikalnih enot.

Literatura

Atkins 2008 = Sue Atkins, Theoretical Lexicography and its Relation to Dictionary-making, v: *Practical lexicography: a reader*, ur. Thierry Fontenelle, Oxford: Oxford University Press, 2008, 31–50.

Atkins – Rundel 2008 = Sue Atkins – Michael Rundell, *The Oxford Guide to Practical Lexicography*, Oxford: Oxford University Press, 2008.

Atkins – Varantola 2008 = Sue Atkins – Krista Varantola, Monitoring Dictionary Use, v: *Practical lexicography: a reader*, ur. Thierry Fontenelle, Oxford: Oxford University Press, 2008, 337–375.

Čermák 1985 = František Čermák, Frazeologie a idiomatika, v: František Čermák – Josef Filipec: *Česká lexikologie*, Praha: Academia, 1985, 166–248.

Čermák 2009 = František Čermák, Leksikografovi zapiski o korpusnem slovarju, *Jezik in slovstvo* 54 (2009), št. 3–4, 25–42.

Fišer 2009 = Darja Fišer, sloWNET – slovenski semantični leksikon, v: *Infrastruktura slovenščine in slovenistike*, Ljubljana: Znanstvena založba Filozofske fakultete, 2009 (Obdobja 28), 145–149.

Gantar idr. 2009 = Polona Gantar idr., *Specifikacije za izdelavo leksikalne baze za slovenščino: standard za izdelavo posamezne leksikalne enote v leksikalni bazi*, Projekt »Sporazumevanje v slovenskem jeziku« ESS in MŠŠ, 2009 ⟨http://www.slovenscina.eu/Media/Kazalniki/Kazalnik6/SSJ_Kazalnik_6_Specifikacije-leksikalna-baza_v1.pdf⟩.

Gantar idr. 2009a = Polona Gantar idr., *Specifikacije za izdelavo leksikalne baze za slovenščino: opis analize referenčnega korpusa*, Projekt »Sporazumevanje v slovenskem jeziku« ESS in MŠŠ, 2009 ⟨http://www.slovenscina.eu/Media/Kazalniki/Kazalnik5/SSJ_Kazalnik_5_Specifikacije-opis-analize-korpusa_v1.pdf⟩.

Gantar – Krek 2009 = Polona Gantar – Simon Krek, Drugačen pogled na slovarske definicije: opisati, pojasniti, razložiti?, v: *Infrastruktura slovenščine in slove-*

nistike, Ljubljana: Znanstvena založba Filozofske fakultete, 2009 (Obdobja 28), 151–159.

Kosem idr. 2011 = Iztok Kosem – Miloš Husak – Diana McCarthy, GDEX for Slovene, *Proceedings of the 2nd international conference on electronic lexicography*, eLEX2011 ⟨http://www.trojina.si/elex2011/elex2011_proceedings.pdf⟩.

Krek – Kilgarriff 2006 = Simon Krek – Adam Kilgarriff, Slovene Word Sketches, v: *Jezikovne tehnologije* 5, ur. Tomaž Erjavec – Jerneja Žganec Gros, Ljubljana: Inštitut Jožef Stefan, 2006, 62–65.

Logar Berginc – Krek 2010 = Nataša Logar Berginc – Simon Krek, New Slovene corpora within the Communication in Slovene project, v: *Abstract: International Conference SLAVICORP, Corpora of Slavic Languages, 22–24 November 2010*, 8.

Logar Berginc – Šuster 2009 = Nataša Logar Berginc – Simon Šuster, Gradnja novega korpusa slovenščine, *Jezik in slovstvo* 54 (2009), št. 3–4, 57–68.

Rozman idr. 2010 = Tadeja Rozman idr., *Nova didaktika poučevanja slovenskega jezika: sporazumevanje v slovenskem jeziku*, Ljubljana: Ministrstvo za šolstvo in šport – Amebis, 2010.

Rundell 2010 = Michael Rundell, Defining Elegance, v: *A Way with Words: Recent Advances in Lexical Theory and Analysis, A Festschrift for Patrick Hanks*, ur. Gilles-Maurice de Schryver, Kampala: Menha Publishers, 2010 (Linguistics Series).

SSKJ 1 = *Slovar slovenskega knjižnega jezika* 1: *A–H*, Ljubljana: SAZU – ZRC SAZU, Inštitut za slovenski jezik – Državna založba Slovenije, 1970.

The Slovenian lexical database: For whom, why, and how (to proceed)?

Summary

The Slovenian lexical database contains a complex structure of lexical and grammatical information based on corpus analysis and is intended for the production of various dictionary-type final products. At the same time, it is designed as a database that will be accessible to users as part of a broader linguistic portal on the web. In addition to lexicographic information, it also contains information primarily intended for computer processing and for improving language technology applications for Slovenian. In comparison to traditional dictionaries in book format, the use of web-based media has presented new issues in the inclusion and structuring of lexical and grammatical information. Here the emphasis is on ensuring relevant, quickly accessible, and reliable information in the most user-friendly manner. The information in the Slovenian lexical database is intended for three different user profiles: general users, students, and those learning Slovenian as a foreign language. At the same time, concrete solutions have been adopted for taking into account different needs of individual profiles, among other things including a textbook glossary, designing sentence-type definitions, and citing typical syntactic realizations and limitations in their use. Further work on the lexical database from the perspective of the three types of users will involve using a survey to check the effectiveness and applicability of the semantic descriptions and the suitable demarcation of information with regard to anticipated types of lexical units described in the database.