Leksikografski proces pri izdelavi spletnega slovarja sodobnega slovenskega jezika

Polona Gantar, Iztok Kosem in Simon Krek

Abstract
This paper describes each stage in the compilation of a database that is to be used as a basis for an online dictionary of contemporary Slovenian and in developing Slovenian language technologies. A proposal for a procedure for archiving different versions of entries, as well as different versions of the entire database during the compilation process, is also presented. Furthermore, we describe how to include detecting lexical change (the continuous updating of headwords) and dictionary users in the process. This is a topical issue in electronic lexicography, but one that still leaves many questions unanswered.

Keywords: dictionary-making process, automatic data extraction, online dictionary, detecting lexical change, gradual dictionary compilation

Ključne besede: leksikografski proces, avtomatsko luščenje podatkov, spletni slovar, detektiranje pomenskih sprememb, postopna izdelava slovarja

1 UVOD

Izdelava slovarskih priročnikov je v digitalni dobi po pričakovanju pogojena s sodobnim načinom življenja, ki ga na področju dostopa do najrazličnejših informacij prek spleta in mobilnih naprav, kot so pametni telefoni in tablice, določajo zanesljivost, hitra in brezplačna dostopnost ter prilagodljivost vsebin, če izpostavimo samo tri najbolj odločujoče (Müller-Spitzer et al. 2011). V takih okoliščinah se leksikografi in založniki upravičeno sprašujejo, kako kompleksne slovarske opise izdelati kvalitetno, vendar hkrati v čim krajšem času in s čim manjšim finančnim vložkom ter kako jih ohranjati aktualne in zanimive za jezikovno skupnost, ki so ji v prvi vrsti namenjeni. Povsem jasno je, kot ugotavljajo poznavalci leksikografske prakse zadnjih 10, 15 let (prim. Krek 2011; Rundell 2014),[1] da tiskani slovarji, kljub temu da danes še sobivajo z elektronskimi in spletnimi, v bližnji, še manj pa daljnji prihodnosti, niso več realnost. Zato se zdi načrtovanje procesa izdelave slovarja – še posebno v situaciji, kot je slovenska, kjer na korpusu temelječi jezikovni opisi sodobne slovenščine niso na voljo, izdelava novega slovarja slovenskega knjižnega jezika (NSSKJ)[2] pa konceptualno in izvedbeno sledi tiskani logiki –, še toliko bolj pomembno in za jezikovno skupnost ključno.

V Predlogu za izdelavo slovarja sodobnega slovenskega jezika (SSSJ; Krek et al. 2013b: 52–60) je predstavljen postopek izdelave slovarja v posameznih fazah, ki omogočajo postopno objavo slovarskih informacij glede na stopnjo leksikografske obdelave in količino podatkov, ki jih imajo gesla v posameznih fazah leksikografskega opisa. Hkrati je opisan tudi postopek sprotnega posodabljanja slovarskih gesel (ibid.: 46) in določitev prioritetnega zaporedja njihove obdelave (ibid.: 45). V pričujočem prispevku želimo posamezne faze podrobneje razčleniti in se osredotočiti na načrtovanje leksikografskega procesa, ki bo kos izzivom na dolgi rok in bo učinkovito izrabljal možnosti informacijsko-komunikacijskega znanja – konkretno jezikovnotehnoloških orodij – tako v metodološkem smislu pri pridobivanju in obdelavi jezikovnih podatkov kot pri posredovanju informacij uporabnikom. Ker je proces leksikografskega opisa za spletno zasnovane na korpusu temelječe slovarje še relativno neopisan,[3] se želimo v prispevku dotakniti tudi vse pomembnejšega vključevanja jezikovne

1 Na temo prihodnosti leksikografije je bila na konferenci Electronic Lexicography in the 21st Century (eLex, Bled, 10.–12. november 2011) organizirana okrogla miza z naslovom: *Will there still be dictionaries in 2020?* Posnetek je dostopen na: http://videolectures.net/elex2011_bled/ (dostop 27. 7. 2015).

2 Slovar že nastaja na Inštitutu za slovenski jezik Frana Ramovša (ISJFR). Ker gre za koncept, ki ima na področju sodobnih jezikovnih opisov edini finančno podporo na nacionalni ravni, je problem izvedbe, ki je konceptualno vezana na tiskani format, statičnost in odsotnost jezikovnotehnološke podprtosti, dejansko problem celotne jezikovne skupnosti tako v smislu smotrnosti porabe davkoplačevalskega denarja kot v smislu rezultata, ki ne upošteva leksikografskih in jezikovnotehnoloških trendov pri jezikovnem opisu.

3 To je tudi eden od razlogov, da je bil opis in načrtovanje leksikografskega procesa pri slovarjih, zasnovanih za splet, tema ene od delavnic evropske pobude European Network of e-Lexicography (ENeL) julija 2014 v Bolzanu. Prispevki so dostopni na: http://www.elexicography.eu/working-groups/working-group-3/wg3-meetings/wg3-bolzano-meeting/ (dostop 27. 7. 2015).

skupnosti v leksikografski proces in se opredeliti do problema, ki ga prinaša sprotno objavljanje slovarskih informacij, namreč vzpostavitev postopka arhiviranja in dostopanja do posameznih različic slovarske baze.

2 FAZE V LEKSIKOGRAFSKEM PROCESU

Leksikografski proces kot detajlno načrtovan proces izdelave slovarja je ena ključnih organizacijsko-izvedbenih nalog, ki vplivajo tako na organizacijo in sestavo leksikografskega tima, kot na finančno in časovno izvedbo projekta. Kot izpostavljata Tiberius in Krek (2014), obstajajo v literaturi predvsem opisi leksikografskega procesa pri izdelavi tiskanih slovarjev (prim. Dubois 1990; Landau 1984; Zgusta 1971), kjer se načeloma izpostavljajo tri zaporedne faze, tj. faza načrtovanja, faza pisanja in faza publiciranja. Možnost izrabe računalnika (tj. strojne obdelave jezikovnih podatkov), pojav interneta in količina ter medsebojni preplet različnih tako jezikovnih kot jezikovno povezanih informacij so nujno vplivali na način izdelave in objave leksikografskih vsebin. Leksikografski proces pri izdelavi nestatičnih spletnih slovarjev, kot ugotavlja Klosa (2013: 4), tako v splošnem predvideva šest faz, ki pa ne potekajo nujno v linearnem zaporedju, ampak se med seboj lahko prekrivajo in dopolnjujejo (Klosa 2013; Tiberius in Schoonheim 2015), in sicer: pripravljalna faza, faza pridobivanja podatkov, faza računalniške priprave podatkov, faza računalniške obdelave podatkov, analitična faza in faza priprave za spletno objavo.

2.1 Opisi leksikografskega procesa v predlogih za izdelavo novega slovarja slovenskega jezika

Na področju slovenske leksikografije se je v zadnjih petih letih, dejansko pa šele od objave do pred kratkim edinega javno predstavljenega Predloga za izdelavo SSSJ (Krek et al. 2013b), začelo resneje govoriti o tem, da slovenska jezikovna skupnost potrebuje slovar sodobne slovenščine, hkrati pa tudi o tem, v kolikšni meri naj bi se leksikografska praksa pri izdelavi slovarja v digitalni dobi še zanašala na leksikografsko tradicijo (in jezikovne opise) Slovarja slovenskega knjižnega jezika (SSKJ) (Gantar 2014) ter kako tako obsežen projekt izpeljati v čim krajšem času. V kontekstu leksikografskega procesa, ki je tesno povezan z vsebino, izvedbo in medijem, za katerega je slovar zasnovan, nas bo zato zanimalo, v kolikšni meri se predstavljeni leksikografski koncepti ukvarjajo z leksikografskim procesom, koliko je ta premišljen in izvedljiv in koliko se pri tem upošteva dejstvo, da bo novi slovar namenjen predvsem bodočim uporabnikom, tj. uporabnikom, ki bodo v slovarju iskali informacije npr. čez 10, 15 let.

Preden se osredotočimo na opis posameznih faz pri izdelavi SSSJ, kot ga predvideva v ta namen oblikovan konzorcij v okviru Centra za jezikovne vire in tehnologije,[4] si poglejmo, kako je zamišljena izdelava NSSKJ, katerega izdelava naj bi trajala vsaj 20 let,[5] in še prej, kakšno vlogo ima v metodološkem smislu in z vidika vključevanja gradiva v NSSKJ Sprotni slovar slovenskega jezika, ki se spogleduje s konceptom »slovarja v izdelavi«[6] (angl. *dictionary under construction*, prim. Klosa 2013: 3), kar predstavlja v slovenski leksikografiji novost.

2.1.1 Sprotni slovar slovenskega jezika

V uvodu v Sprotni slovar slovenskega jezika lahko beremo, da gre za rastoči slovar, ki pa ima v času nastajanja in objave zgolj informativno naravo.[7] Njegova izhodiščna različica vključuje besedišče, ki je v obstoječih slovarjih še neregistrirano, potrjuje pa ga korpusno gradivo. Geslovnik se postopoma dopolnjuje z besedjem, ki so ga uporabniki iskali, a ne našli, v slovarjih na inštitutski spletni strani http://bos.zrc-sazu.si, poleg tega naj bi zajemal tudi besede, ki jih obstoječi korpusi slovenščine še ne prinašajo, raba pa je bila registrirana v drugih, zlasti elektronskih virih.[8] V zvezi s procesom in metodologijo izdelave slovarja v uvodu izvemo še, da se bo na podoben način geslovnik dopolnjeval tudi v prihodnje in da bo »novo besedje slovarju predvidoma dodajano vsakih šest mesecev«.[9] Slovar sicer omenja »izhodiščno različico«, ne predvideva pa načina arhiviranja in dokumentiranja starejših različic ter dostopa do njih, negotov pa je tudi status vključevanja v NSSKJ, ki je v uvodu opredeljen takole: »Ali bodo posamezne iztočnice dejansko prešle v normativne ali razlagalne slovarje in bodo opisane natančneje, pa bo pokazal čas.« Kljub dobrodošli novosti, ki jo obeta slovar v naslovu, lahko ugotovimo, da v metodološkem smislu, tj. z vidika postopnega dodajanja slovarskih informacij, ne prinaša v leksikografsko prakso nič novega. Kot je mogoče razbrati iz uvoda, se gesla v celoti dodajajo na novo, žal pa tudi ni mogoče preizkusiti postopka dostopanja do posameznih različic.

4 http://www.cjvt.si/projekti/ (dostop 27. 7. 2015).

5 Prim. odzive v medijih ob objavi Osnutka, npr. http://www.24ur.com/novice/slovenija/na-nov-slovar-slovenskega-knjiznega-jezika-bomo-cakali-se-leta.html (dostop 27. 7. 2015).

6 Morda ustreznejši slovenski izraz, ki označuje ta tip spletnega slovarja, je »nikoli dokončani slovar«.

7 Avtor in strokovni pregledovalci se pri izboru in opisu besedišča sklicujejo izključno na informativnost in izrecno zanikajo kakršnokoli normativnost. Zdi se torej, da bo šele premislek uredniške ekipe NSSKJ uporabnika seznanil s primernostjo, ustreznostjo oz. neustreznostjo besed. Ob povabilu uporabnikom, naj predlagajo domače ustreznike, ki jih uporabljajo ali bi jih želeli (!) uporabljati – s čimer naj bi se krepilo »zavedanje govorcev o njihovem vplivu na jezikovno rabo, preko katere lahko posledično sodelujejo pri normiranju slovenskega besedja« (Krvina 2014: 91) – pa se dejansko zastavlja vprašanje razumevanja jezikovne norme in njenega določevalca: ali o njej torej odloča skupina ljudi s pozicije moči, kot si to zamišlja ISJFR pod okriljem SAZU, ki v smislu prijaznosti in ljudskosti občasno ponudi ta občutek moči tudi izbranim jezikovnim uporabnikom, ali pa je norma – torej tisti del jezika, ki ga jezikovna skupnost želi standardizirati – neločljivo povezana s tem, kako jezikovna skupnost jezik dejansko uporablja, kar pomeni, da je izhodišče standarda v ustrezno analizirani in interpretirani jezikovni rabi celotne jezikovne skupnosti? V zvezi s tem prim. prispevek Gorjanc et al. (2015).

8 Avtorji teh virov posebej ne navajajo.

9 Kot je razvidno iz kolofona, je od leta 2014 (in tudi v času pisanja tega prispevka) na spletnem portalu Fran (http://www.fran.si/132/sssj-sprotni-slovar-slovenskega-jezika, dostop 27. 7. 2015) še vedno na voljo le različica 1.0, zadnja sprememba pa je datirana z 2. 10. 2014.

2.1.2 Osnutek za NSSKJ

Proces izdelave NSSKJ je v kontekstu pregleda leksikografskih procesov potrebno omeniti predvsem zato, ker naj bi vključeval tri pomembne elemente v t. i. predredakcijski fazi, ki se prekrivajo s predvidenimi postopki v posameznih fazah pri izdelavi SSSJ (Krek et al. 2013b), in sicer (a) avtomatsko luščenje podatkov iz korpusa (b) izdelavo orodja za prepoznavanje pomenskih in slovničnih sprememb ter (c) nadgradnjo obstoječih korpusov.

Proces izdelave NSSKJ predvideva dve fazi, t. i. predredakcijsko fazo in fazo redakcije. Predredakcijska faza vključuje oblikovanje geslovnika, na podlagi katerega bo izdelan nabor iztočnic za uvrstitev v slovar. Avtorji predvidevajo, da bodo v okviru predpriprave slovarskim podatkom v slovarski bazi samodejno dodatni podatki, pridobljeni iz korpusov, kot tudi podatki iz obstoječih slovarjev[10] in jezikovnih zbirk v lasti ISJFR.[11] Med drugim naj bi bili iz korpusov avtomatsko izluščeni zapis iztočnice, besednovrstna opredelitev, podatek o pogostosti leme in njenih posameznih oblik, skladenjski podatki s pripadajočimi kolokacijami in stavčnimi zgledi ter nekatera slovnična opozorila ter podatki o jezikovni rabi, npr. o pisanju skupaj in narazen. Na podlagi nadaljnje analize teh podatkov pa bo izdelana celostna tipologija slovarske obravnave. Ker zahteva avtomatski postopek luščenja omenjenih podatkov iz korpusa natančno razdelane odločitve glede interpretacije podatkov v razmerju korpus – leksikon besednih oblik – slovar, saj so lematizacija in oblikoskladenjski podatki prilagojeni označevanju korpusa, njihov prenos v slovar pa zato ne more biti neposreden, nas bi zanimalo, kako bo postopek avtomatizacije pri izdelavi NSSKJ dejansko izpeljan, vključno z opisom procesa avtomatizacije, saj le ta, kot se je pokazalo pri avtomatizaciji postopkov pri izdelavi dela LBS, nikakor ni trivialen. Glede na to, da omenjeni postopki na ISJFR po vsej verjetnosti še niso bili preizkušeni (oz. po vednosti avtorjev niso bili predstavljeni v strokovni literaturi, na strokovnih posvetih ali osrednjih leksikografskih konferencah po Evropi), tudi ni mogoče pričakovati rezultatov evalvacije ali natančnejše predstavitve celotnega avtomatizacijskega postopka. Če so se avtorji NSSKJ odločili, da bodo pri postopku avtomatizacije uporabili metodologijo, ki je bila uporabljena pri izdelavi LBS, je treba še poudariti, da je luščenje podatkov iz korpusa, kot je bilo izpeljano pri izdelavi LBS, sledilo zelo jasnim metodološkim izhodiščem in je bilo prilagojeno organizaciji podatkov v slovarju, ki se v marsičem razlikuje od koncepta NSSKJ.[12]

10 Glede na trditev, da bodo podatki iz obstoječih slovarjev vključeni v kar največji meri, se pod vprašaj postavlja trditev, da bo slovar »izdelan popolnoma na novo« (NSSKJ: 1, 2), posledično pa tudi, ali lahko pričakujemo dejansko opis sodobne slovenščine ali ponovno adaptacijo SSKJ, ki je nastal na povsem drugačnem gradivu, v povsem drugem času in z drugačno metodologijo.

11 Glede na tip podatkov, ki bodo pridobljeni samodejno, predvidevamo, da bo uporabljen identični postopek avtomatizacije, kot je bil uporabljen pri izdelavi Leksikalne baze za slovenščino (LBS; Kosem et al. 2013; Kosem et al. 2013a). Ustrezni citati v Osnutku za NSSKJ sicer niso navedeni.

12 Sem denimo sodi drugačna obravnava besednovrstne konverzije in homonimije nasproti večpomenskosti, obravnava iztočnice v odnosu do pomena, sistem označevanja, ki temelji na komunikacijski obvestilnosti ipd.

Na podlagi tega lahko sklepamo, da so avtorji NSSKJ zamisel o izvedbi avtomatizacije in prilagoditvi na drugače zasnovan slovar bodisi prepustili kasnejšemu času, pri čemer to vpliva na prilagajanje celotnega leksikografskega procesa, ali pa se bodo temu postopku pri dejanski izvedbi enostavno odrekli. Če si sposodimo misel iz uvoda v Sprotni slovar slovenskega jezika, lahko rečemo, da bo o vključenosti in izvedljivosti postopka avtomatizacije pri izdelavi NSSKJ dejansko pokazal šele čas.

Poleg avtomatsko izluščenih podatkov se v procesu izdelave NSSKJ predvideva tudi izdelava orodja, ki bi prepoznavalo morebitne pomenske in slovnične spremembe (NSSKJ: 78). Na ta način naj bi se po besedah avtorjev skrajšal čas izdelave slovarja, saj naj bi redaktorji dobili izbrane podatke vnaprej pripravljene za delo v leksikografskem programu, hkrati pa se v nasprotju z namenom avtomatizacije (prim. Kosem et al. 2013) predvideva, da se »pri redakcijskem delu vsi podatki preverjajo, kot da bi jih bilo treba zbrati in obdelati povsem na novo« (NSSKJ: 78).

Naslednji pomemben element, ki ga izpostavljajo avtorji NSSKJ v procesu izdelave slovarja, je razvoj orodja za avtomatsko detektiranje pomenskih in slovničnih sprememb v jeziku. Glede na to, da gre za temo, ki je v sodobni leksikografski metodologiji zelo aktualna, bi pričakovali, da gre za orodje, ki je na slovenskem gradivu že preverjeno, rezultati pa opisani v katerem od znanstvenih prispevkov. Znanje, ki ga ima na tem področju ISJFR, bi bilo namreč mogoče vsaj posredno uporabiti tudi za druge jezike, posledično pa bi se sodobna slovenska leksikografija uvrstila na pomembno mesto znotraj evropske leksikografske prakse, kjer se avtomatska detekcija pomenskih sprememb šele uveljavlja (prim. Cook et al. 2014).

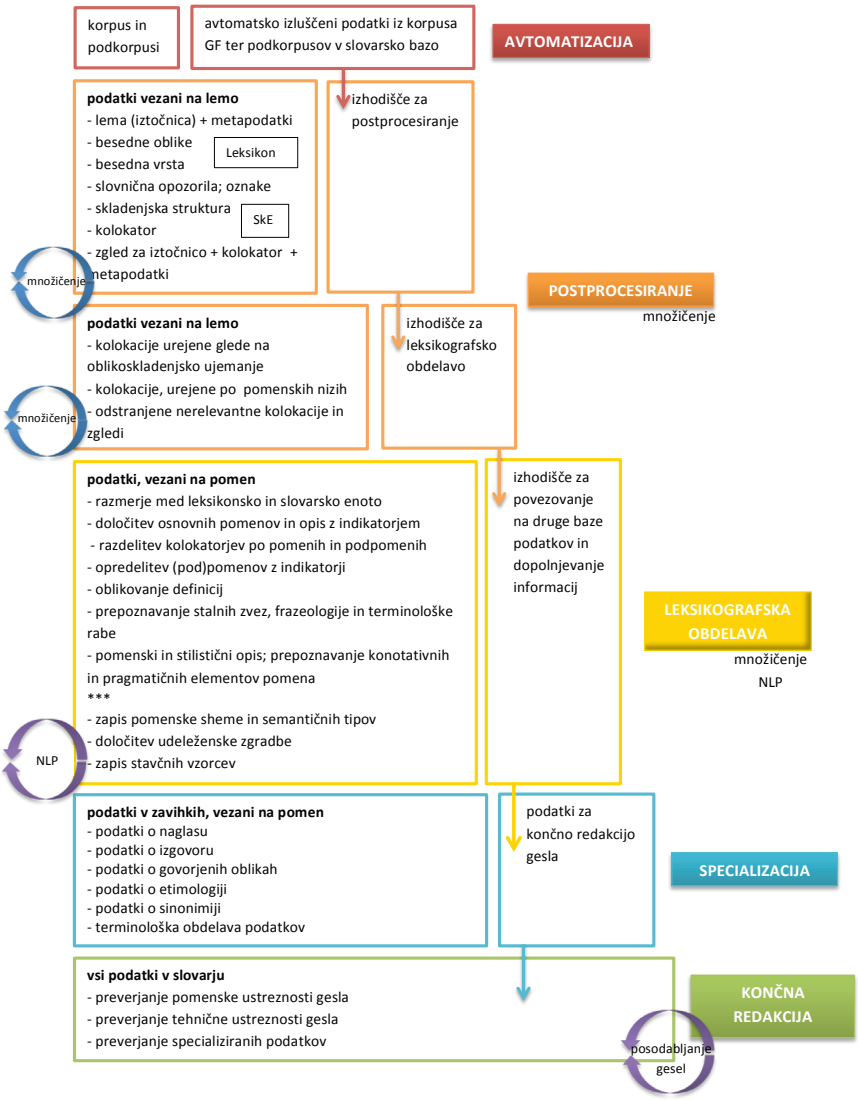
Sprotno posodabljanje korpusa je pri snovanju sodobnega slovarja, katerega uporabna vrednost ne poteče hkrati z njegovo objavo, logični postopek znotraj leksikografskega procesa. Ker pa naloga nikakor ni trivialna, bi od avtorjev osnutka za izdelavo NSSKJ pričakovali natančnejša pojasnila o tem, na kakšen način se bodo obstoječi korpusi posodabljali, do kakšne mere se bo nadgrajevala oz. spreminjala taksonomija v korpus vključenih besedil, kako bodo za nova besedila urejene avtorske pravice, kakšna bo dostopnost nadgrajenega korpusa itd. Za razliko od predvidenega postopka nadgrajevanja korpusa Gigafida pri izdelavi SSSJ, ki je opisan v pričujoči monografiji (Logar et al. 2015), namreč v leksikografskem procesu za izdelavo NSSKJ ni – razen pojasnila, da bodo »vse spremembe, ki bodo nastale med posameznimi posodobitvami slovarja, dokumentirane« ter da bo »zaradi sklicevanja na posamezne različice slovarja na voljo tudi ogled starejših različic« (NSSKJ: opombi 4 in 32), povedano nič konkretnega.

2.2 Leksikografski proces pri izdelavi SSSJ

Posamezne faze v izdelavi SSSJ so predvidene že v Predlogu (Krek et al. 2013b: 52), na tem mestu pa jih želimo razčleniti bolj podrobno, tudi v skladu z izkušnjami, ki smo jih pridobili pri nadgradnji procesa avtomatizacije, pri evalvaciji avtomatskega postopka in na podlagi izkušenj, ki jih imajo na tem področju slovarji, ki so primarno zasnovani za splet in predvidevajo postopno objavo, pri čemer so podatki na voljo uporabniku že v času nastajanja, predvideno pa je tudi sprotno posodabljanje in možnost nadgrajevanja – gre torej za slovarje, ki so na nek način nenehno v izdelavi, zaradi česar se zanje v angleščini uvaja izraz *online dictionaries under construction* (prim. Klosa 2013).

Izdelava slovarskih gesel je v okviru leksikografskega procesa pri izdelavi SSSJ predvidena v petih fazah (Slika 1) na način, ki omogoča, da so podatki uporabniku na voljo že v času izdelovanja slovarja, tj. že po prvi fazi, in ne šele po zaključku slovarskega dela. Prednosti, ki jih ima tak postopek, odtehtajo večjo zapletenost procesa, v katerem morajo biti posamezne faze zelo natančno določene, opravila leksikografov in drugih sodelavcev pa usklajena in vnaprej predvidena. Večstopenjskost izdelave gesel omogoča tudi učinkovitejšo in bolj ekonomično delitev dela. V prvi fazi tako večino dela opravi računalnik, človeško delo pa je vključeno šele v kasnejših fazah, kjer se znanje leksikografov izrablja le za specifične leksikografske postopke, ki zahtevajo izkušnje in strokovno usposobljenost. Predhodna delitev nabora iztočnic v težavnostne stopnje omogoča tudi postopno seznanjanje manj izkušenih leksikografov s postopki pomenskega členjenja in oblikovanja pomenskih razlag. Ker je za določena rutinska dela, kot je npr. odstranjevanje neustreznih in nerelevantnih podatkov ter razvrščanje kolokacij in zgledov pod ustrezne pomene, predvideno množičenje (gl. prispevek Fišer et al. 2015), je mogoče racionalizirati stroške človeških virov, predvsem pa pospešiti čas izdelave gesel.

Pri postopku izdelave slovarja v posameznih fazah kot tudi pri dostopanju do posameznih različic je zelo pomembno, da uporabnik takoj in jasno prepozna, v kateri fazi se nahaja geslo. V ta namen smo v spletni postavitvi predvideli podatek z datumom stanja gesla, ki se generira skladno s spremembami, ki potekajo v celotnem procesu geselske izdelave. Na ta način zagotovimo dvoje: prvič, uporabnik ima možnost navajanja reference na slovarsko geslo, ki velja za določeno stanje tega gesla, kar je zlasti pomembno pri citiranju podatkov v raziskovalne in izobraževalne namene, in drugič, s tem uporabniku nakažemo, kaj lahko od slovarskih podatkov pričakuje v smislu njihove količine, urejenosti in stopnje zanesljivosti.



Slika 1: Faze v leksikografskem procesu pri izdelavi SSSJ.

2.2.1 Prva faza: Avtomatizacija

Prva faza v izdelavi gesla je v celoti namenjena avtomatskemu luščenju leksikalnih podatkov iz korpusa Gigafida (Logar Berginc et al. 2012), ki predstavlja osnovno frekvenčno listo. Ta bo pred procesom luščenja dopolnjena z natančno in kompleksno statistično obravnavo podatkov iz korpusov Kres, Gos

(Verdonik in Zwitter Vitez 2011) in drugih prosto dostopnih korpusov. Poleg tega je za zajetje specializiranega besedišča predvidena izgradnja specializiranih podkorpusov in dopolnitev nekaterih že obstoječih, npr. učbeniškega podkorpusa (več o tem v Logar 2015 ter Vintar in Logar 2015), kot tudi podrobno tematsko označevanje strokovnih besedil že na ravni korpusnih metaoznak, ki se prek avtomatsko izluščenih podatkov prenesejo tudi v slovarsko bazo (prim. Gantar in Kosem 2013; Kosem 2015).

Upoštevajoč zgradbo geselskega članka, kot jo predvidevamo pri izdelavi SSSJ (prim. Klemenc et al. 2015), so avtomatsko izluščeni podatki naslednji:

- **Lema** v osnovni obliki, kot jo določa oblikoskladenjska označitev v korpusu Gigafida in leksikonu besednih oblik Sloleks (prim. Dobrovoljc et al. 2015) ter pripadajoče odvisne oblike (v samostojnem zavihku).
- Podatek o **frekvenci** leme v korpusu oz. podkorpusu.
- **Besedna vrsta** leme, kot jo določa oblikoskladenjska oznaka v korpusu in leksikonu besednih oblik Sloleks.
- Določena **slovnična opozorila**, ki se nanašajo na tipično skladenjsko ali besedilno obnašanje leme v korpusu, kot je denimo sopojavljanje z lastnimi imeni ali količinskimi izrazi, izstopajoča možnost tretjeosebne rabe glagola ali nastopanje v *se*-glagolskih ali citatnih zgradbah. Ti podatki, ki se iz korpusa pridobivajo s pomočjo kombinacije direktiv CONSTRUCTION+UNARY v orodju Sketch Engine (Kilgarriff et al. 2004), se bodisi v slovarski bazi generirajo kot opozorila leksikografom pri nadaljnji obdelavi leme v t. i. pomenski fazi oz. se lahko na ravni leme v slovar prepisujejo kot slovnične oznake, npr. *pogosto zanikano*, *pogosto v 3. os. ednine*, *pogosto z lastnim imenom* ipd. (prim. Kosem 2015).
- **Skladenjske strukture**, ki so bile predhodno registrirane na podlagi ročne analize besednih skic pri izdelavi LBS in na podlagi katerih je bila izdelana izpopolnjena različica slovničnih relacij v orodju SkE (Krek 2012). Ta različica slovnice besednih skic je namenjena izključno avtomatskemu luščenju kolokacijskih podatkov iz korpusa in ne ročnemu pregledovanju leksikografov, saj je na podlagi tako pridobljene besedne skice mogoče izluščiti bolj podrobne kolokacijske podatke, za analizo katerih bi leksikograf potreboval neprimerno več časa, zato bi zanj pomenila prej oviro kot pomoč v procesu pomenskega členjenja izhodiščne leme. Nova različica slovnice besednih skic tako vključuje tudi direktive, kot so CONSTRUCTION, COLLOC in SEPARATEPAGE, ki omogočajo luščenje vezljivostnih vzorcev pri glagolih, prepoznavanje elementov, ki na podlagi predvidenih za slovar relevantnih slovarskih enot spadajo v kategorijo t. i. skladenjskih zvez, kot so npr. zveze predlog-samostalnik-predlog: *v primerjavi z*, *v odnosu do*, *za razliko od* ipd., ter prikazovanju

relacij s tremi elementi (direktiva TRINARY) na novi spletni strani, kar omogoča uvedbo natančnejših relacij s predlogi, ki v prejšnji različici slovnice besednih skic niso bile upoštevane. Naprimer za relacijo glagol *pomesti* + predlog v tožilniku (koga-kaj_g4) dobimo samostojne stolpce s kolokatorji za različne predloge, ki se vežejo s tožilnikom: pomesti pod [preprogo, predpražnik, tepih], *pomesti na [smetišnico, smetišče, kup, tla], pomesti v [koš, kot]* ipd.

- **Kolokatorje**, ki se pojavljajo v posamezni skladenjski strukturi ob obravnavani lemi in tvorijo potencialne kolokacije pa tudi skladenjske in stalne zveze. Prepoznavanje zadnjih kot tudi uvrstitev v določen pomen je naloga leksikografa v pomenski fazi leksikografskega procesa.
- **Korpusne zglede**, ki vsebujejo lemo in kolokator v določeni skladenjski strukturi, za kar smo uporabili za slovenščino prilagojeno in v dveh korakih izpopolnjeno funkcijo orodja GDEX za izločanje čim bolj optimalnih korpusnih zgledov (Kosem et al. 2013), ki predstavljajo kandidate za vključitev v slovar (z možnimi prilagoditvami) in so kot taki torej potencialni slovarski zgledi.

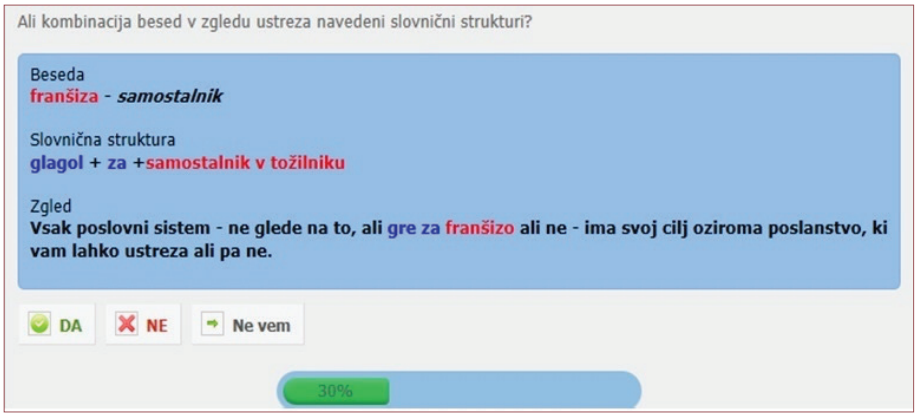
Postopek avtomatizacije, kjer smo s pomočjo v ta namen posebej prilagojene API skripte, ki vsebuje opise vseh relevantnih slovničnih relacij, izluščili zgoraj navedene podatke in jih avtomatsko prenesli v slovarsko bazo v programu iLex (Erlandsen 2004), kjer so bili pripravljeni za nadaljnjo obdelavo, je bil za slovenščino že preizkušen pri izdelavi LBS (Kosem et al. 2013; 2013a) in ovrednoten z vidika prekrivnosti izluščenih podatkov glede na ročno izdelavo gesel, v postopku izdelave kolokacijskega slovarja za slovenščino pa še nadgrajen in izboljšan. V nadgradnjo sodi predvsem avtomatsko odstranjevanje kolokatorjev, ki ponudijo same enake zglede, in postavitev leme in/ali kolokatorja pri izpisu v slovarsko bazo v ustrezen sklon, spol in število. Poleg tega smo ob predhodno natančno določenih parametrih za luščenje kolokatorjev, ki smo jih izdelali ločeno za različne frekvenčne skupine in posamezne besedne vrste (za podrobnosti gl. Kosem et al. 2013), v novem poskusu kolokatorje izluščili na podlagi združenega podatka o jakosti (angl. *salience*) in frekvenci kolokatorja, ki omogoča primerjavo z ročnim izborom kolokatorjev na podlagi besednih skic in v končni fazi omogoča izbor kolokatorjev, ki so za določeno lemo najrelevantnejši.

2.2.2 Druga faza: Postprocesiranje in odstranjevanje napak

Druga faza v procesu izdelave slovarja je namenjena (a) postprocesiranju, ki vključuje tudi čim bolj avtomatsko odstranjevanje napak oz. nerelevantnih avtomatsko izluščenih podatkov z možnostjo uporabe množičenja, (b) dodajanju metaoznak, ki omogočajo povezovanje podatkov znotraj slovarske baze in združevanje

z drugimi slovarskimi bazami (npr. izhodiščno slovarsko bazo in bazo za izdelavo kolokacijskega slovarja, slovarja sinonimov ipd.) ter (c) čiščenju nerelevantnih podatkov, ki so pri avtomatskem luščenju običajno posledica napačne lematizacije ali korpusnega šuma. V postopku postprocesiranja je avtomatsko izluščene podatke mogoče dodatno urediti, npr. postaviti izluščene kolokatorje v ustrezen spol in sklon glede na lemo, kot ga zahteva podstavna skladenjska struktura, ter vzpostaviti kolokacijske nize, v katerih so kolokatorji znotraj enega niza pomensko povezani. V postopku postprocesiranja je za namene združevanja različnih podatkovnih baz potrebno posebej označiti posamezne elemente znotraj kolokacij, npr. predloge, veznike, prosti morfem *se/si* pri glagolu ipd. ter odstraniti napačno zapisane strukture, ki so se npr. pri združevanju avtomatsko in ročno izdelanih gesel pokazale kot posledica ročnega zapisovanja. Dodajanje metaoznak posameznim elementom geselske zgradbe, npr. identifikacijska oznaka iz leksikona besednih oblik Sloleks pri lemi in kolokatorju, ter dodajanje opozoril o statusu leme, npr. o njeni avtomatski ali ročni izdelavi, je namenjena predvsem leksikografom in združevanju podatkov iz različnih baz.

Za odstranjevanje nerelevantnih kolokacij, ki se zaradi lematizacijskih napak in korpusnega šuma pojavijo pri avtomatskem luščenju podatkov iz korpusa, smo v tej fazi predvideli tudi možnost uporabe množičenja, pri katerem v posebej za to oblikovani nalogi uporabnike sprašujemo, ali uporabljena kombinacija v avtomatsko izluščenem zgledu ustreza predvideni skladenjski strukturi, kot prikazuje Slika 2:



Slika 2: Naloga v procesu množičenja za odstranjevanje avtomatsko izluščenih nerelevantnih kolokacij in pripadajočih zgledov iz LBS

Proces čiščenja avtomatsko izluščenih podatkov iz korpusa, za katerega smo uporabili orodje SlowCrowd[13] (Tavčar et al. 2012), ki se učinkovito izrablja tudi za

13 http://nl.ijs.si/slowcrowd/index.php?project=slowcrowdmain (dostop 27. 7. 2015).

čiščenje slovenske različice wordneta SloWNet (Fišer 2009), je bil preizkušen pri izdelavi LBS (Kosem et al. 2013a), prvi rezultati pa so pokazali, da je uporaba množičenja pri čiščenju podatkov zanesljiva in da lahko občutno skrajša čas v tej fazi leksikografskega procesa.

2.2.3 Tretja faza: Leksikografska obdelava podatkov

V naslednji fazi, ki je namenjena leksikografski obdelavi podatkov, zaradi česar je strokovno in organizacijsko najbolj zahtevna in hkrati časovno najobsežnejša, so naloge osredotočene na analizo podatkov z vidika pomena, tj. pomenske členitve in pomenskega opisa, posledično pa zadevajo tudi prepoznavanje slovničnih in skladenjskih, normativnih ter stilističnih lastnosti besed oz. njihovih pomenov.

V tej fazi ima leksikograf na voljo avtomatsko izluščene in izčiščene podatke za posamezno lemo, ki ji je avtomatsko pripisana besednovrstna oznaka, ki ustreza morfosintaktični oznaki v leksikonu, zato je njegova prva naloga prepoznavanje simetričnosti med t. i. leksikonsko in slovarsko enoto. Naloga nikakor ni trivialna, učinkovitost in enotnost leksikografov pri odločanju v posameznih primerih pa je povezana z natančnimi navodili, ki vključujejo vse možne situacije in predpostavljajo enotne rešitve pri obravnavi homonimije ter besednovrstne konverzije – tj. v skladu z dogovorjenimi merili slovarskega koncepta (prim. Gantar 2015; K. Dobrovoljc 2015). Konkretno je pri izdelavi SSSJ razmerje med leksikonsko in slovarsko iztočnico privzeto simetrično, morebitne posebnosti v slovarski bazi pa leksikograf označuje s posebnimi vnaprej izdelanimi strojno berljivimi restriktorji.

Osnovna naloga, ki jo v tej fazi opravi leksikograf, je pomenska razčlenitev leme in oblikovanje pomenskih razlag za prednostno določena gesla. Glede na zgradbo gesla, ki jo povzemamo po LBS, leksikograf izdela pomenski meni s pomočjo pomenskih indikatorjev, v katerem je prikazana pomenska zgradba gesla in razmerja med pomeni in podpomeni, ter t. i. pomensko shemo s prikazom tipičnega vezljivostnega vzorca za posamezni (pod)pomen pri glagolskih ter nekaterih samostalniških in pridevniških iztočnicah. Na tej stopnji je pomembno tudi dodajanje podatkov, ki so v slovarski bazi namenjeni računalniški obdelavi jezika, kamor sodi (z uporabo sofisticiranih avtomatskih postopkov) luščenje stavčnih vzorcev, prepoznavanje semantičnih tipov po vzoru Corpus Pattern Analysis (Hanks 2004; Hanks in Pustejovsky 2005) in pripisovanje udeleženskih vlog (angl. *semantic role labeling*).

V tej fazi leksikograf identificira tudi stalne besedne zveze, med katerimi posebej opozori na terminološke, ki potrebujejo obravnavo z vidika stroke, na katero se nanašajo, ter registrira in pomensko opiše frazeološke enote.

Leksikografsko delo je v tej fazi organizirano glede na stopnjo težavnosti gesla in vnaprej pripravljenih predlog za določene pomenske skupine gesel. Za učinkovito organizacijo dela je pomembna razdelitev nalog med (a) izkušene leksikografe, ki opravijo osnovno pomensko razčlenitev in izdelajo pomenske razlage, prepoznajo kompleksnejše slovnične in skladenjske vzorce, stilistične ter pragmatične posebnosti rabe ipd., med (b) leksikografe, ki so specializirani za redakcijo frazeoloških enot, opis slovničnih in skladenjskih lastnosti posameznega (pod)pomena, normativnih podatkov, ter (c) med leksikografe, ki se z leksikografskimi nalogami šele seznanjajo in opravljajo manj zahtevna leksikografska opravila, kot je preverjanje ustrezne razvrstitve kolokacij pod posamezne pomene (na podlagi rezultatov množičenja) in skladenjske strukture, urejanje pomenskih kolokacijskih nizov, prepoznavanje besedilnega okolja pri stalnih zvezah in frazeoloških enotah.

Za del nalog, ki predstavljajo rutinska leksikografska opravila in ne zahtevajo poglobljenega leksikografskega znanja, je predvidena naloga v okviru množičenja, kjer uporabniki razvrščajo avtomatsko izluščene zglede, ki vsebujejo kolokacijo, ki ustreza določeni skladenjski strukturi, pod ustrezni vnaprej določeni (pod)pomen.[14] S to nalogo želimo poleg razvrščanja kolokatorjev v že pomensko razčlenjeno geslo dobiti tudi povratno informacijo o ustrezni pomenski razčlenitvi ter detektirati pomenske opise, ki znotraj širše jezikovne skupnosti nimajo zadostne potrditve.

Po končani tretji fazi, v kateri imajo uporabniki na voljo že večino slovarsko relevantnih informacij, vezanih na pomen, je slovarsko geslo pripravljeno za dodajanje informacij, ki jih v spletno zasnovanem slovarju predvidevamo v drugih zavihkih/rubrikah, čemur je namenjena naslednja faza v leksikografskem procesu.

2.2.4 Četrta faza: Dodajanje specializiranih jezikovnih podatkov

Delo v četrti fazi postopno izdelanega slovarskega gesla je namenjeno dodajanju t. i. zunajbaznih podatkov in dopolnjevanju na spletnem portalu že prikazanih podatkov, kjer se predvideva specializirano znanje jezikoslovcev in strokovnjakov drugih področij, zlasti terminologov pa tudi jezikoslovcev, zadolženih za standardizacijo in reševanje normativnih vprašanj (gl. Popič et al. 2015). Podatki, ki se posameznim delom slovarskega gesla dodajajo v tej fazi, so naslednji:

- podatki o **naglasu** v povezavi z leksikonom besednih oblik in v skladu z že omenjenim prepoznavanjem simetričnosti leksikonske in slovarske

14 Naloga je podrobneje predstavljena v prispevku Fišer et al. (2015).

enote, kamor sodi tudi prepoznavanje prekrivnosti oz. neprekrivnosti naglasne in spregatvene paradigme kot enega izmed meril za obravnavo homonimije v odnosu do večpomenskosti;

- podatki o **izgovoru** na podlagi korpusa Gos in na podlagi za to določenih vnaprej izdelanih parametrov, kamor sodi označitev naglasnega mesta, izgovor odvisnih oblik in zapis izgovora, ki iz zapisa iztočnice ni predvidljiv (Jurgec 2015);
- podatki o **govorjenih oblikah** in posebnostih posameznih oblik na ravni pomena, kot jih je mogoče pridobiti iz govornega korpusa Gos (Verdonik 2015);
- podatki o **etimologiji,** natančneje o izvoru besede in njenih sorodnih oblikah v različnih jezikih ter podatek o starinskih oblikah ali zapisih besede v slovenskem jeziku glede na časovno umeščenost besedila, v katerem se je konkretna oblika pojavila;
- podatki o **sinonimiji** na podlagi analize kolokatorjev v funkciji Sketch Difference v orodju Sketch Engine in z vključitvijo podatkov iz slovenske različice wordneta SloWNet ter
- **terminološka obdelava podatkov.** V ta namen bo organizirana mreža strokovnjakov za posamezna strokovna področja in vzpostavljena spletna platforma, ki bo omogočala spremljanje in usklajevanje dela.

2.2.5 Peta faza: Končna redakcija gesla

Zadnja faza v leksikografskem procesu pri izdelavi SSSJ je namenjena končni redakciji celotnega gesla in medsebojnemu usklajevanju podatkov, ki jih vključujejo posamezni slovarski zavihki. Leksikografova naloga je ugotavljanje konsistentnosti podatkov glede na leksikografski koncept, strukturo gesla, pa tudi ugotavljanje skladnosti podatkov z izpričanim realnim jezikovnim stanjem. Leksikograf lahko zato pred zaključno objavo posamezno geslo uredi, dopolni ali pa vrne v katero od predhodnih faz, če npr. ugotovi nedoslednosti v pomenski členitvi besede, pri opisu stalnih zvez in frazeoloških enot ali pomanjkljive podatke v segmentu terminološke obdelave.

Pred zaključkom te faze je opravljen tudi postopek avtomatskega detektiranja pomenskih sprememb, ki vrnejo obravnavo besede v fazo pomenskega členjenja, prepoznavanja večbesednih leksikalnih enot, izrabe moči množic in ponovne končne redakcije. Čeprav torej predstavlja peta faza zaključen leksikografski proces, se slovarski opis na tej točki ponovno vrača v avtomatski izvoz podatkov, ki jih v zvezi s posameznimi pomeni narekuje leksikalni razvoj slovenščine, izpričan v nadgradnji korpusnih virov.

3 SPROTNO POSODABLJANJE SLOVARSKE BAZE

V celotnem postopku izdelave gesel in njihove predstavitve uporabnikom ima zelo pomembno vlogo slovarska baza, ki po eni strani predstavlja vir vseh slovarskih informacij, po drugi pa tudi arhiv vseh sprejetih odločitev v posameznih fazah leksikografskega procesa. Ker so faze izdelave slovarja jasno razmejene, je treba z vidika slovarske baze zagotoviti, da je v slovarskem orodju vzpostavljen delotok, ki lahko redaktorjem in leksikografom v vsakem trenutku postreže z informacijo, v kateri fazi je posamezno geslo. Dodatno raven kompleksnosti pri načrtovanju slovarske baze prinašata dve povezani odločitvi: sprotno posodabljanje slovarja in predvidena možnost, da so gesla na voljo uporabnikom po vsaki dokončani fazi.

Ko govorimo o sprotnem posodabljanju, imamo v mislih posodabljanje obstoječih gesel v slovarski bazi, ki so že šla skozi vse faze leksikografskega procesa, deloma pa tudi izdelavo povsem novih gesel, sploh tistih, ki so bila obravnavana prednostno (npr. neologizmi). Slednja morajo biti namreč v bazi opremljena s posebnim opozorilom o svoji pomembnosti, ki jih ločuje od drugih gesel, tudi zato, da se lahko pri posodobitvi slovarja uporabnike nanje opozori. Podobno velja tudi za posodabljanje obstoječih že dokončanih gesel, kjer gre lahko bodisi za dodajanje novih podatkov (npr. pomenov, kolokacij, frazeoloških enot) na podlagi analize novega gradiva (npr. spremljevalnega korpusa ali dolgoročno gledano nove verzije referenčnega korpusa) bodisi za popravljanje obstoječih (npr. popravljanje odkritih napak), vendar pa je tu za razliko od prednostnih gesel pomembnejši časovni vidik, torej kdaj so bile nove informacije dodane (in na podlagi katerega vira).

Pri sprotnem posodabljanju je treba posebej obravnavati primere, ko v slovarsko bazo dodajamo nove podatke, ki nadomestijo stare samo na ravni prikazovanja uporabnikom. Tak primer so naprimer slovarski zgledi, kjer se lahko na neki točki odločimo za zamenjavo obstoječih slovarskih zgledov z novimi (prim. Klein in Geyken 2010; Lemnitzer et al. 2015). Zaradi tega je treba v slovarsko bazo pri zgledih (in drugih mikrostrukturnih slovarskih elementih) zapisati opozorilo o tem, ali jih v slovarju prikazujemo ali ne. Na ta način v bazi vedno ohranjamo tudi vse predhodne zglede, uporabnikom pa so v določeni fazi vidni le tisti, ki so glede na vsebino gesla najbolj aktualni.

Možnost, da so gesla na voljo uporabnikom po posamezni fazi leksikografskega procesa, sama na sebi v bazi ne zahteva dodatnih informacij, razen tistih, ki se nanašajo na delotok; dobro je le iz takšnega postopka izločiti obstoječa slovarska gesla, ki jih posodabljamo z novimi informacijami, saj bi kombinacija leksikografsko pregledanih in nepregledanih podatkov v geslu lahko zmotila uporabnike. Sicer so za objavo po fazah precej bolj relevantne vizualizacijske

rešitve v slovarju, ki pa vseeno potrebujejo tudi določeno informacijo (npr. datum objave in različica).

Ključno je torej vzpostaviti postopek, ki leksikografom prikaže jasno sliko o tem, v katerih fazi izdelave se geslo nahaja, kdaj je bilo dodano v slovar, kdaj so bili geslu dodani novi podatki (kdaj je nastala nova različica[15]). Po našem mnenju lahko tak postopek zagotovi učinkovito in pregledno leksikografsko delo in omogoči jasno in razumljivo predstavljanje slovarskih podatkov uporabnikom.

4 OBRAVNAVANJE POSAMEZNIH RAZLIČIC IN VIZUALIZACIJA PODATKOV

V tem delu se posvetimo trem vprašanjem, relevantnim za posodabljanje slovarja s predlaganim leksikografskim procesom: Kako pogosto posodabljati slovar? Kako jasno ločiti nedokončana gesla oz. njihove informacije od končanih? Kako obravnavati posamezne različice gesel in slovarja kot celote?

Tuje prakse pri pogostosti posodabljanja spletnega slovarja kažejo dva pristopa: redni nekajmesečni intervali ali sproti, čim nastanejo nova gesla. Prvi pristop uporabljajo slovarji, kot je npr. Oxford English Dictionary (OED),[16] kjer slovar posodobijo vsake štiri mesece in imajo posebno stran,[17] namenjeno objavam v zvezi s posodobitvami. Podoben način posodabljanja uporablja tudi Macmillan English Dictionary (MED),[18] ki pa ne podaja ločenih informacij o tem, kdaj je posodobitev opravljena,[19] ampak uporabnike opozori na izbrane nove besede v rubriki New Words na naslovni strani.

Drugi pristop, torej takojšnjo objavo dokončanih gesel, zasledimo v Velikem slovarju poljskega jezika[20] (Żmigrodzki 2014) in v Slovarju sodobnega nizozemskega jezika[21] (Tiberius in Schoonheim 2015). Poudariti velja, da gre pri omenjenih slovarjih za projekta, pri katerih se slovarja delata na novo in sta torej sproti nastajajoča slovarja v pravem pomenu besede. Tako je motivacija za čim prejšnjo objavo leksikografskih vsebin zaradi uporabnikov, pa najbrž tudi financerjev, toliko

15 Pomembno je ločiti med različicami gesel, ki nastanejo ob večjih spremembah (npr. po končanju posamezne faze ali pri posodobitvi obstoječega slovarskega gesla z novimi podatki), in med različicami, ki jih sproti beleži program za izdelavo slovarjev. Ta namreč beleži vse sprotne spremembe, ki jih pri izdelavi gesla vnaša leksikograf, in tako omogoča primerjavo dveh različic, vpogled v izbrisane podatke, njihovo obnovitev ipd. Zato je dobro dosledno uporabljati terminologijo, ki jasno ločuje med tema procesoma.

16 http://www.oed.com/ (dostop 27. 7. 2015).

17 http://public.oed.com/the-oed-today/recent-updates-to-the-oed/ (dostop 27. 7. 2015).

18 http://www.macmillandictionary.com/ (dostop 27. 7. 2015).

19 Na spletni strani s pogostimi vprašanji in odgovori: (http://www.macmillandictionary.com/faq.html (dostop 27. 7. 2015). je podana informacija, da slovar posodobijo večkrat na leto.

20 Wielki słownik języka polskiego: http://www.wsjp.pl/ (dostop 27. 7. 2015).

21 Algemeen Nederlands Woordenboek: http://anw.inl.nl/show?page=search1 (dostop 27. 7. 2015).

večja kot pri slovarjih, ki večinoma samo dodajajo nove besede ali posodabljajo obstoječo vsebino. S tega vidika je predlog v konceptu NSSKJ, ki predvideva posodabljanje slovarja enkrat letno (NSSKJ: 3), premalo ambiciozen in premalo uporabniško naravnan – pričakovali bi namreč, da bi slovarskim uporabnikom, za katere vemo, da že več kot dvajset let čakajo na nov opis slovenskega jezika, čim hitreje ponudili rezultate leksikografskega dela.

Logiki sprotnega objavljanja sledi tudi leksikografski proces pri predlaganem SSSJ, v katerem predvidevamo sprotno objavo gesel v vseh fazah izdelave. V slovenskem prostoru sicer že obstajajo slovarji, ki uporabljajo podobno prakso, npr. iSlovar,[22] ki ločuje štiri stopnje dokončanosti gesla: »predlog« (predlagal urednik ali uporabnik), »pregledano« (pregledal urednik), »strokovno pregledano« (pregledala in uredila strokovna skupina) in »urejeno« (pregledala slovaropisna skupina; gre za končno redakcijo). Opozoriti je treba, da to ne pomeni, da se gesla dodajajo vsak dan ali celo vsako uro, ampak gre za paketne posodobitve (torej več gesel hkrati) v precej pogostih intervalih, ki zagotavljajo boljšo preglednost tako za leksikografe kot za uporabnike.

Z opozorili o dokončanosti gesla tudi poskrbimo za ločevanje nedokončanih gesel od dokončanih. V Predlogu za izdelavo SSSJ (Krek et al. 2013b: 52–60) je predvideno, da se stanje gesla prikaže z barvnimi pikami (od rdeče do zelene), hkrati pa se navede tudi datum zadnje posodobitve gesla (Krek et al. 2013b: 27). Datum je razločevalni element tudi v primeru, da geslo ostane v isti fazi (npr. pri posodobitvi dokončanega slovarskega gesla). Končna oblikovalska rešitev bo mogoče drugačna od prikazane v Predlogu, a bo v vsakem primeru morala vključevati vsaj ti dve informaciji. Poleg tega bo, podobno kot pri OED, na posebni (pod)strani slovarja dokumentirana vsaka posodobitev, npr. s seznami iztočnic in njihovim statusom, izpostavljene bodo tudi morebitne sistematične spremembe.

Ena izmed odločitev glede sprotnega objavljanja se nanaša tudi na to, kaj po spremembah narediti s prejšnjimi različicami gesel. To je sicer manj problematično v primeru objavljanja posameznih faz, saj je za uporabnika v vsakem primeru najbolj relevantna različica gesla, ki vsebuje največjo količino informacij. Na simpoziju o angleškem slovarju OED leta 2014 je bila prav to ena od perečih tem, saj so se nekateri udeleženci pritožili, da po posodobitvi gesel nimajo več vpogleda v prejšnje različice. Njihov argument je bil, da npr. s posodobitvijo razlag izgubimo informacijo o tem, kako so na določen pomen oz. rabo besede v jezikovni skupnosti gledali v obdobju izdelave prvotne različice gesla. Tak argument je seveda povsem legitimen, vendar pa velja spomniti, da gre v tem primeru za historični slovar, pri katerem je diahroni pogled na rabo jezika ključnega pomena.

22 http://www.islovar.org (dostop 27. 7. 2015).

Pri izdelavi SSSJ nameravamo o možnosti primerjanja različnih prejšnjih in zadnje različice slovarja razmisliti skladno z uporabniškimi raziskavami, kjer bi se zdelo smiselno o taki možnosti razmisliti, če bi uporabniki izrazili to potrebo.

Že sedaj predviden način dostopa do starejših različic gesel bo za raziskovalce, pa tudi za namene strojne obdelave jezika, omogočen z rednim objavljanjem novih različic prosto dostopne slovarske baze, ki bodo dejansko usklajena s posodobitvami spletne različice slovarja, razen v primerih, ko bo prišlo do sprememb v podatkih, ki so relevantni samo za slovarsko bazo. Pomemben del takšnih objav bo podrobna dokumentacija, kjer bodo opisane ne le spremembe na ravni slovarskih gesel, ampak tudi vse vsebinske in tehnične spremembe v slovarski bazi, npr. novi tipi skritih oznak, spremembe na ravni DTD (angl. *Document Type Definition*) itd.

5 ZAKLJUČEK

Leksikografski proces pri izdelavi slovarja, ki predvideva sprotno objavljanje gesel v posamezni fazi, njihovo posodabljanje in možnost dostopa do posameznih različic izdelanega gesla, je kompleksen postopek, ki zahteva vnaprej predvideno in natančno izdelano strategijo, ki vpliva na dejansko izvedljivost celotnega postopka in organizacijo leksikografskega dela. Leksikografski proces, kot ga predlagamo v prispevku, temelji v izhodišču na avtomatskem luščenju osnovnih leksikalnih podatkov, ki se v naslednjih fazah dopolnjujejo, hkrati pa se izločajo podatki, ki so bodisi napačni ali pa za uporabnika nerelevantni. Pri tem je pomembno razlikovati med slovarsko bazo in podatki v njej, ki so namenjeni tekoči izvedbi sprotnega posodabljanja in uporabniku niso vidni, in podatki, ki jih ima uporabnik na voljo v posameznih različicah. Za izvedbo predlaganega procesa je posebej pomembno, da je delotok natančno razdelan in opisan, kar leksikografom in redaktorjem omogoča kontinuirano in konsistentno izvedbo, uporabniku pa je ves čas pred očmi informacija, v kateri fazi se geslo nahaja, ter posledično, kateri podatki so mu v določeni fazi na voljo, hkrati pa tudi, kako dostopati do posameznih različic, če ga to zanima za raziskovalne namene, za nadaljnje strojno procesiranje ali izrabo podatkov v pedagoške namene. V zvezi z leksikografskim procesom, ki ga predlagamo, je pomembno tudi poudariti, da je izdelan izključno za objavo na spletu in za natančno predvideno geselsko zgradbo, notranjo organizacijo ter vrsto leksikalnogramatičnih podatkov, ki se dopolnjujejo tako znotraj geselske zgradbe (v zavihku pomen, npr. pomenski meni, kolokacije, skladenjske strukture, stavčni vzorci, zgledi pri posameznih leksikalnih enotah, tj. pomenih, stalnih zvezah in frazeoloških enotah) ter s podatki v drugih zavihkih na spletni strani, tj. s podatki o besednih oblikah, govoru, normi, sinonimiji itd.