Oznake: slovarska baza in slovar

Iztok Kosem

Abstract
This paper discusses the various possibilities concerning dictionary labels that have been brought about by recent developments in lexicography, especially with regards to the content of the planned dictionary of contemporary Slovenian. First, key decisions concerning dictionary labels are presented. This is followed by a discussion on different possible approaches to labelling, including the use of automatic methods to identify label candidates. The main part of the paper focuses on labels in the proposed dictionary of contemporary Slovenian, with some options considered and suggestions provided on how to improve and optimise the labelling process and the subsequent visualisation of labels in the dictionary.

Keywords: labels, Slovenian, automatic methods, dictionary, user

Ključne besede: oznake, slovenščina, avtomatske metode, slovar, uporabnik

1 UVOD

Oznake so različni tipi slovarskih pojasnil, ki uporabnike opozarjajo, da ima beseda, besedna zveza, njen pomen itd. določene slovnične omejitve, da se nanaša na določeno časovno obdobje, tip besedila, regionalne posebnosti, da se uporablja na določenem strokovnem področju, da izraža določen odnos do vsebine ali udeležencev ipd.

V sodobni leksikografiji smo s korpusi dobili možnost zelo podrobne analize realne jezikovne rabe, z novimi slovarskimi mediji pa možnost prikaza večje količine informacij, ki jih lahko uporabniku predstavimo na različne načine. Vse to pomeni, da lahko informacije, povezane z oznakami, določamo sistematično za več elementov slovarske mikrostrukture, npr. za posamezne kolokatorje ali nize kolokatorjev, pa tudi v različnih oblikah, npr. v obliki opozoril, daljših pojasnil ipd. Pomembna sprememba v sodobni leksikografiji se je zgodila tudi v razmerju slovarska baza – slovar, saj sodobne slovarske baze vsebujejo veliko več informacij kot iz njih izpeljani slovarji, poleg tega pa lahko ena sama slovarska baza služi za izdelavo različnih slovarjev. Slovarska baza tako vsebuje informacije, relevantne za različne tipe slovarskih uporabnikov. To je pomembno tudi z vidika načina določanja oznak in njihovega prikaza v slovarju, saj je treba pri načrtovanju novih slovarjev – zlasti slovarjev, ki se jezikovnega opisa lotevajo povsem na novo, kot je predlagani slovar sodobnega slovenskega jezika (SSSJ) – ob upoštevanju potreb vseh možnih uporabnikov natančno opredeliti razmerje med slovarsko bazo in slovarjem.

V prispevku najprej predstavimo ključne odločitve glede uporabe oznak v slovarju, nato različne načine določanja oznak, vključno z možnostjo vključevanja avtomatskih metod, ki jih omogočajo sodobne jezikovne tehnologije. V osrednjem delu se osredotočimo na oznake v SSSJ z vidika razmerja med slovarsko bazo in slovarjem ter predstavljamo nekaj možnosti in predlogov, kako izboljšati oz. optimizirati procesa določanja oznak in njihovo vizualizacijo v spletni različici slovarja.

2 OZNAKE

V zvezi z oznakami moramo pri snovanju slovarja sprejeti več odločitev. Najprej se moramo odločiti, kaj želimo v slovarju označevati. Največkrat se posebej označujejo slovnične, stilistične, področne, časovne, regionalne in registrske posebnosti, nekateri slovarji, zlasti slovarji za tuje govorce, označujejo tudi pragmatične posebnosti in frekvenco. Od naštetih vrst označevanja so samo frekvenčne oznake

in v določeni meri tudi časovne tiste, za katere se lahko naknadno, torej med izdelavo slovarja ali celo po njej, odločimo, da jih bomo dodali.

Po izbiri vrste označevanja moramo določiti oznake za vsako izbrano vrsto. Z leksikografskega vidika je bolj učinkovito, če imajo leksikografi na voljo omejen nabor oznak za posamezno vrsto, saj morajo pri analizi gradiva v programu za izdelavo slovarjev samo izbrati ustrezno oznako iz nabora in se jim ni treba ukvarjati s poimenovanjem ali razmišljanjem, ali uporabiti obstoječo oznako ali oblikovati novo. Vendar je pri marsikateri oznaki nabor zelo težko vnaprej povsem omejiti (gl. razdelek 2), sploh pri izdelavi povsem novega slovarja. Pri določenih vrstah oznak je težava tudi stopenjskost oz. razlike med stopnjami, ki jih opredeljujejo posamezne oznake, npr. pri registrskih *formalno, manj formalno, zelo formalno, neformalno* ipd. Zelo pomembno je tudi vprašanje poimenovanja posamezne oznake, ki mora biti uporabniku jasno in razumljivo. Za SSKJ je bila recimo zaradi prevelike splošnosti in premajhne obvestilnosti kritizirana oznaka *ekspresivno*[1] (Müller 2009). Poleg tega so raziskave (Rozman 2010) med šolskimi uporabniki Slovarja slovenskega knjižnega jezika (SSKJ) razkrile tudi napačno tolmačenje okrajšanih oznak, vendar so tovrstne težave dejansko bolj ali manj omejene na tiskane izdaje slovarjev, kjer so zaradi prostorskih omejitev pogosteje uporabljane okrajšave.[2]

Ne glede na medij pa mora biti v slovarju jasno opredeljen način prikazovanja oznak z vidika njihove umestitve v slovarsko geslo. Po eni strani se odločamo, kam oznako pozicionirati, pred element, ki ga označujemo, ali za njim. To je na ravni slovarske mikrostrukture različno: medtem ko oznake sledijo iztočnicam, stalnim zvezam in frazeološkim enotam, jih pri pomenih in podpomenih ponavadi najdemo pred razlagami, tj. na samem začetku. Mesto oznake v geslu pa določa tudi njen doseg oz. opredeljuje, katere dele gesla oznaka zajema. Za leksikografa to ne predstavlja bistvenih težav: oznaka za iztočnico (v zaglavju) velja za celotno geslo, oznaka pred pomenom za celoten pomen itd. Vendar pa, kot opozarjata Atkins in Rundell (2008: 231), se takšne dosledne rabe oznak slovarski uporabniki niti ne zavedajo. Ker uporabniki enojezičnih slovarjev največkrat ne preberejo celotnega gesla, ampak se osredotočijo le na določene dele, največkrat na razlago, zapis iztočnice, sinonime in zglede (gl. Harvey & Yuill 1997; Hartmann 1999; Kosem 2010; Verlinde in Binon 2010; Lorentzen in Theilgaard 2012), se lahko, če je oznaka v zaglavju in velja za celotno geslo, pri obsežnejših geslih zgodi, da jo uporabnik prezre. Idealne rešitve ni, vendar pa digitalni mediji vsekakor ponujajo več možnosti različnih vizualizacijskih rešitev (o tem več v razdelku 3).

1 S tega vidika je presenetljivo, da so se avtorji Osnutka koncepta za novi slovar slovenskega knjižnega jezika (Gliha Komac et al. 2015) odločili obdržati to oznako.

2 Okrajšave najdemo tudi v elektronskih različicah slovarjev, a gre praviloma za prvotno tiskane slovarje, prenesene na splet ali v kateri drugi digitalni medij, kar velja tudi za SSKJ, SSKJ2 in SNB.

Od medija neodvisna vizualizacijska težava je kopičenje oznak. V takšnem primeru je treba uporabnikom jasno sporočiti, ali gre za razmerje in-in oz. ali-ali. Poleg tega je daljši niz okrajšanih oznak za uporabnika še bolj zahteven za tolmačenje, saj mora poznati pomen vsake okrajšave v nizu:[3]

> **bêdro** -a stil. -ésa s (é, ẹ́) noga nad kolenom; *stegno*: smejal se je in se tolkel po bedrih / obirati kurje bedro
> // nav. mn., pog., šalj. *noga*: hlače mu kar opletajo po suhih bedresih

Dodatno zahtevnost vnašajo kombinacije okrajšanih oznak in ostalih elementov, recimo številke za osebo pri glagolih, ki jih uporabnik ne najde na seznamu razvezanih oznak:

> **bíti**[1] bíjem nedov., 3. mn. stil. bijó; bìl (í î)

Če so že takšni primeri rabe okrajšanih oznak potencialno zahtevni za uporabnika, gotovo podobno velja za gesla, v katerih se oznake pojavljajo pri več (pod)pomenih in dejansko imajo pomembno vlogo pri razlikovanju med posameznimi (pod)pomeni ali pri prepoznavanju ustreznega (pod)pomena:

> **beséda**[2] -e ž, rod. mn. stil. besedí (ẹ́)
> 2. ed. in mn. **misel, izražena z besedami:** /…/
> 3. nav. ed., ekspr. zagotovilo, obljuba: /…/
> ed., star. dogovor: /…/
> 4. **izražanje misli z govorjenjem:** /…/
> 5. ed. in mn. **govorni ali pisni nastop v javnosti:** /…/
> // nav. ed. **možnost, pravica do govorjenja, zlasti v javnosti:** /…/
> 6. ed., knjiž. **izmenjava mnenj, misli;** pogovor, govor: /…/
> 7. ed., nav. vznes. **sistem izraznih sredstev za govorno in pisno sporazumevanje;** jezik: /…/

Za snovanje uporabniku prijaznejših rešitev imamo več možnosti, od uporabe različnih oblik pisav, različnih barv ali barvnih odtenkov itd. Glede uporabe barv npr. raziskave kažejo, da barvne oznake uporabniki hitreje opazijo, pa tudi informacijo si bolje zapomnijo (Dziemianko 2015). Razlikovanje lahko vzpostavimo tudi z dodelitvijo stalnega prostora (na zaslonu) za določene tipe oznak v geslu (gl. razdelek 3) in podobnimi vizualizacijskimi rešitvami.

Pri oznakah se moramo odločiti tudi o tem, kdaj sploh uporabiti oznako. Gre za vprašanje, kateri način podajanja informacije, ki jo ponuja oznaka, je z vidika uporabnika najustreznejši pri danem pomenu, podpomenu ali katerem drugem delu

3 Vsi primeri na tej strani so iz spletne različice SSKJ2.

gesla. Gantar in Kosem (2013) navajata dve obliki označevanja: eksplicitno in implicitno. Eksplicitno označevanje je klasično označevanje z oznakami, pri katerem je oznaka jasno ločen (največkrat enobesedni) element slovarske mikrostrukture. Pri implicitnem označevanju pa informacija, ki bi jo sicer posredovala oznaka, postane del slovarske razlage, bodisi zaradi manjšega izpostavljanja, tesnejše povezanosti z razlago (zlasti pragmatika) ali zaradi mejnosti (terminološko – splošno). Implicitno označevanje je dejansko lahko učinkovitejše od eksplicitnega, zlasti za materne govorce, saj raziskave kažejo, da uporabniki skoraj vedno preberejo razlago, zelo redko pa ločeno podane slovnične informacije in informacije o rabi (Hartmann 1999; Kosem 2010). V razlagi oznaka postane sestavni del pojasnjevalne informacije o pomenu besede, kar si uporabniki lažje zapomnijo (Barnbrook 2002), medtem ko je pri eksplicitnem načinu podajanja ločena od ostalih delov gesla in jo morajo uporabniki tolmačiti skupaj z razlago, zgledom, stalno zvezo ali s katerim drugim elementom slovarskega gesla. Pri implicitnem označevanju ima leksikograf tudi možnost uporabe daljše ubeseditve, kar je pri informacijah o pragmatičnih in stilističnih posebnostih rabe besede zelo uporabno.

Obstaja tudi kombinacija eksplicitnega in implicitnega označevanja, pri katerem oznako oz. informacijo, ki jo posredujemo, predstavimo v obliki komentarja za zgledi oz. proti koncu pomenskega opisa oz. razlage. McCreary (2004) je pri poskusih s študenti leksikografije ugotovil, da je takšna oblika označevanja z leksikografskega vidika lažja kot pa vključevanje informacije v razlago. Rezultat je informacijsko razbremenjena razlaga, hkrati pa leksikograf dobi možnost, da v komentarju napiše nekoliko izčrpnejše pojasnilo.

Pri označevanju gre torej predvsem za ugotovitev rabe, ki zahteva oznako, izbiro ustreznega načina predstavitve tovrstne informacije in pri eksplicitnem označevanju še za ustrezno vizualizacijo oznake v slovarju. V samem procesu morajo leksikografi odgovoriti na vprašanje, ali sploh uporabiti oznako, katero oznako uporabiti, in zlasti pri implicitnem označevanju, kako jo ustrezno ubesediti. V nekaterih primerih se oznaka sprva zapiše eksplicitno, v končni verziji slovarja pa je prestavljena v razlago. Kaj se v takšnih primerih zgodi s prvotno oznako? Jo sploh še lahko najdemo? Kako lahko v slovarju poiščemo razlage s takšnimi informacijami? Pri omogočanju takšnih funkcionalnosti igra ključno vlogo načrtovanje tega, kako se bodo oznake in z njimi povezane informacije določale na ravni slovarske baze in slovarja oz. slovarjev.

3 METODE DOLOČANJA OZNAK

Preden se posvetimo oznakam v slovarski bazi in slovarju, je treba nameniti nekaj besed metodologiji določanja oznak med analizo korpusnega gradiva. Sodobna

leksikografska analiza ni več zgolj ročna; razvoj jezikovnih tehnologij je namreč leksikografom omogočil tudi uporabo avtomatskih načinov pridobivanja podatkov o jezikovni rabi. Številni tipi oznak, npr. slovnične, registrske, področne in regijske, so namreč povezani s porazdelitvijo besed v korpusnih besedilih in jih je načeloma mogoče avtomatično pridobiti iz korpusa (Rundell in Kilgarriff 2011). Seveda to ne pomeni, da so oznake avtomatično pripisane v slovarsko bazo, ampak gre zgolj za opozorila na potencialne oznake, ki v bazo oz. slovar preidejo le, če jih potrdijo leksikografi.

Avtomatično pridobivanje slovničnih oznak je bilo že preizkušeno tudi za slovenščino, in sicer v zaključni fazi izdelave Leksikalne baze za slovenščino (LBS), ko so bila pri testiranju postopka avtomatskega luščenja leksikalnih podatkov (ALLP; Kosem et al. 2012b; 2013a) pridobljena tudi opozorila o potencialnih slovničnih oznakah. Osnovo za tovrstne informacije so predstavljale gramatične relacije v slovnici besednih skic v orodju Sketch Engine, za katere je mogoče glede na poizvedbo (besedo v iztočnici) skladenjske odnose definirati glede na en sam element, ki izpostavlja en pojav v odnosu do vseh ostalih elementov v korpusu. Na ta način lahko pridobimo podatek o tem, ali nek pojav, kot npr. množinska oblika, tretjeosebna oblika ipd. statistično izstopa. Za vsako od potencialnih oznak je treba določiti statistično mejo, pri kateri se oznaka pripiše v avtomatsko izvoženo geslo.

Medtem ko je slovnične oznake mogoče na tak način avtomatično izluščiti iz kateregakoli korpusa, pa to ne velja za druge vrste oznak. Če namreč želimo pridobiti podatek o npr. področju ali registru rabe iztočnice oz. posameznega pomena, morajo biti besedila v korpusu že opremljena z ustrezno informacijo. Tako je določeno besedilo ali celo njegov del (npr. odstavek) lahko opredeljeno kot *športno, na internetnih forumih* itd. Do takšnega korpusa pridemo tako, da izdelamo podrobno taksonomijo, za vsako taksonomsko kategorijo izberemo učno množico dokumentov, katerim taksonomske kategorije pripišemo ročno, potem pa s strojnim učenjem označimo vsa besedila (ali njihove dele) v korpusu.

V zvezi z določanjem oznak se moramo odločiti tudi o tem, ali bomo leksikografom ponudili izdelan nabor oznak za posamezen tip ali pa bodo leksikografi imeli že vnaprej relativno svobodo pri ubeseditvi oznak. Slednji pristop je primeren zlasti za tiste tipe oznak, »kjer jezikovna raba niha in kaže različne pomenske, stilne, pragmatične in druge omejitve, ki jih je težko ustrezno zajeti z vnaprej določenimi kategorijami« (Gantar in Kosem 2013: 145). Kot primer lahko navedemo bazo DANTE (Atkins et al. 2010; Rundell in Atkins 2011), kjer so takšen pristop uporabili pri oblikovanju pragmatičnih oznak. Končni nabor je vseboval 511 pragmatičnih oznak, a se jih samo 92 pojavi več kot enkrat. Pregled oznak pokaže, da se oznake pogosto malenkostno razlikujejo

v ubeseditvi, posredujejo pa isto informacijo (npr. tako *emphasis, emphatic* in *emphatic use* opozarjajo, da gre za poudarek; podobno *expresses disapproval* in *disapproval* opozarjata, da gre za neodobravanje), kar je pri izdelavi slovarja na podlagi baze dobro poenotiti. Po drugi strani fleksibilnost pri ubeseditvi oznak ponuja določene prednosti, kot je npr. razkrivanje stopenjskosti oznak, npr. *expresses disapproval* ('izraža neodobravanje'), *can express disapproval* ('lahko izraža neodobravanje'), *often expressing disapproval* (»pogosto izraža neodobravanje«), *expresses strong disapproval* (»izraža močno neodobravanje«). Pri vnaprej določenem naboru oznak bi namreč težko predvideli vse takšne odtenke v rabi.

Nekoliko drugačna možnost določanja oznak je uvedba njihove hierarhične strukturiranosti, kot jo npr. za področne oznake priporočata Atkins in Rundell (2008), za slovenski prostor pa Kosem (2011). Bistvo takšnega pristopa je, da hierarhija oznak z nadrejenimi in podrejenimi oznakami leksikografu ponuja možnost izbire splošnejše oznake (ali oznak), kadar se težko opredeli za eno samo (bolj določno). Če se leksikograf odloči za bolj določno oznako (npr. *tenis*), pa to hkrati pomeni avtomatičen pripis tudi njej nadrejene oznake oz. več oznak (npr. *šport*). Takšen način lahko kombiniramo tudi z opisanim fleksibilnim pristopom, kjer lahko splošnejše oznake določimo vnaprej, ubeseditev določnejših pa prepustimo leksikografom. To v fazi finalizacije slovarja omogoča hitrejše poenotenje oznak.

4 OZNAKE V SLOVARSKI BAZI IN SLOVARJU SODOBNEGA SLOVENSKEGA JEZIKA

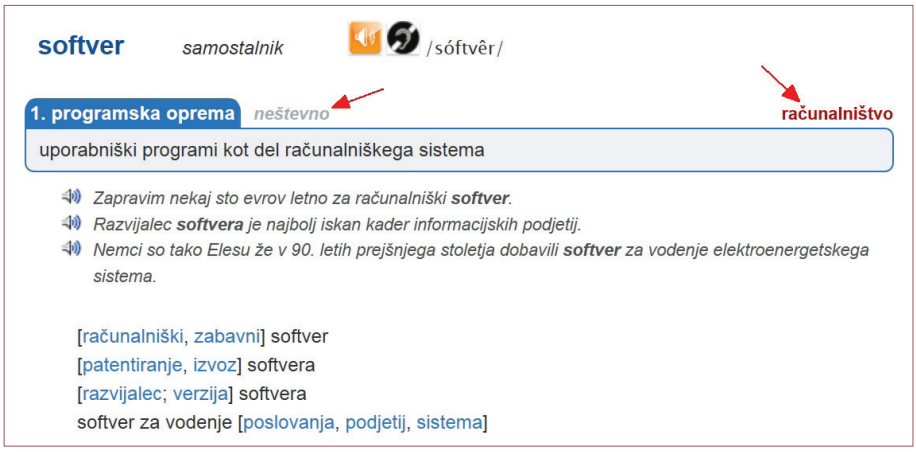
Prevladujoči medij sodobnih slovarjev je postal splet, ki ponuja možnost prikaza veliko večje količine informacij, tako tekstovnih kot multimedijskih, in omogoča različne povezave ter številne iskalne možnosti. Spletni slovarji »temeljijo na elektronskih slovarskih podatkovnih bazah, ki so strukturirane tako, da je podatke mogoče v čim večji meri pridobivati avtomatsko, jih urejati in povezovati z drugimi podatkovnimi bazami in uporabljati za nadaljnje jezikoslovne analize, hkrati pa jih izrabljati tudi v jezikovnotehnološke namene« (Gantar in Kosem 2013: 145). Ravno zaradi relevantnosti za raziskovalce in jezikovne tehnologe vsebujejo slovarske baze veliko več podatkov, kot jih vsebujejo na njej temelječi slovarji. Pri tem ne gre le za podatke, ki so primarno namenjeni računalniški obdelavi jezika, pač pa tudi za podatke, ki so koristni leksikografom pri izdelavi slovarskih gesel. O več slovarjih in ne o enem samem govorimo zato, ker je slovarska baza, zlasti slovarja večjega obsega, kot je SSSJ, lahko osnova številnim večjim in manjšim slovarjem, splošnim in specializiranim, za katere se predvideva, da bodo podatke v zvezi z jezikovno rabo vključevali različno.

Pri leksikografski analizi in izdelavi prvih različic gesel beležimo čim več z oznakami povezanih informacij. Sem sodi tudi avtomatsko luščenje potencialnih oznak, ki je opravljeno že pri izvozu podatkov iz korpusa. Pri beleženju oznak uporabljamo eksplicitno (oznaka) in implicitno metodo (razlaga oz. drugi del gesla), uporabljamo tudi t. i. skrite oznake, ki so eksplicitne oznake, a samo na ravni slovarske baze, saj jih uporabljamo pri omogočanju naprednejših iskanj po slovarju (glej razdelek 4.1). Leksikografi lahko skrite oznake uporabijo tudi v primerih, ko želijo opozoriti na informacijo, ki je sicer podana implicitno, in ko niso povsem prepričani o upravičenosti uporabe določene oznake.

Pri redakciji gesel se potrjujejo odločitve leksikografov, sprejete med analizo, in vnašajo morebitne spremembe, npr. eksplicitna oznaka postane skrita, informacija je posredovana implicitno. Na tej točki so tudi poenotene ubeseditve oznak. Med redakcijo je treba opravljati redne analize gesel z istimi oznakami ali z oznakami istega tipa, na podlagi ugotovitev pa se dopolnijo navodila za leksikografe in/ali izboljša postopek avtomatskega luščenja kandidatov za oznake, posledično pa se izboljša in pospeši analiza gradiva ter izdelava gesel.

Pri vizualizaciji oznak v slovarju se odločamo o oznakah, ki jih bomo na tak ali drugačen način vključili v slovar, in o tem, kakšno vlogo bomo vključenim oznakam namenili oz. kako, če sploh, jih bomo prikazali. Praviloma bodo v slovarju prikazane vse eksplicitne oznake v slovarski bazi, medtem ko bodo skrite oznake uporabljene večinoma zgolj za (naprednejša) iskanja (gl. 4.1). Če se slovarska baza uporabi za namene izdelave drugih slovarjev, pa se lahko zgodi, da se določene skrite oznake prikažejo tudi v slovarju ali pa se določena informacija v bazi preoblikuje v oznake, npr. informacija o frekvenci leme v oznake za skupine najpogostejših 1.000, 2.000 itd. besed v (pisnem) jeziku.

Čeprav je vizualizacija oznak zadnji korak v celotnem postopku, je eden najbolj ključnih. Že na začetku smo omenili nekaj pomembnih ugotovitev glede vizualizacije oznak, tako v zvezi z njihovo umestitvijo kot obliko (barvo ipd.). Elektronski medij nam ponuja številne možnosti, ki jih bo moral SSSJ čim bolj izkoristiti. Ena od možnosti prikaza oznak je predstavljena v Predlogu za izdelavo Slovarja sodobnega slovenskega jezika (Krek et al. 2013b: 34–35) in predvideva različne vizualizacijsek rešitve za različne tipe oznak (Slika 1). Prednost takšnega prikaza je, da ima vsak tip oznake točno določeno (prepoznavno) mesto in obliko (barvo, font itd.) v geslu oz. pri pomenu, ker predvidevamo, da bi to lahko uporabniku pomagalo informacijo hitreje opaziti in prepoznati ter jo zato tudi hitreje uzavestiti.



Slika 1: Prikaz slovničnih in področnih oznak (Krek et al. 2013b).

Pri snovanju vizualizacijskih rešitev je treba imeti v mislih različne pojavne oblike slovarja, od spletnega do slovarja na mobilnih napravah. Razlike v količini informacij, ki jih hkrati lahko prikažemo na različnih medijih, so namreč zelo različne in dejansko lahko govorimo o različnih tipih uporabnikov z vidika navad uporabnika medija.

V vsakem primeru morajo biti rešitve prikazovanja oznak in njihove ubeseditve preizkušene med slovarskimi uporabniki. Uporabniške študije se lahko začnejo izvajati že na začetku izdelave slovarja, saj se lahko uporabijo vzorčna gesla ali samo njihovi deli. Vpliv takšnih študij je zelo pomemben, saj rezultati vplivajo na vse korake v leksikografskem procesu, od določanja oznak do njihove vizualizacije.

4.1 Prednosti slovarske baze z vidika vključevanja oznak: nekaj možnosti

V tem razdelku predstavljamo nekaj možnosti, ki prikazujejo potencialno prednost uporabe ločenih korakov označevanja za slovarsko bazo in slovar za leksikografe in slovarske uporabnike. Naštete možnosti se bodo uporabile tudi pri izdelavi SSKJ.

4.2 Možnosti za leksikografe

Prednost slovarske baze kot podatkovnega vira, ki se razlikuje od dokončnega slovarja, je za leksikografe predvsem v bogastvu in raznolikosti informacij, ki jih

lahko vsebuje. Gre predvsem za podatke, ki so za dodajanje oznak dragoceni, bi jih pa pri tradicionalnem načinu določanja oznak prezrli oziroma bi jih upoštevali samo pri sprejemanju odločitev, ali oznako uporabiti ali ne. Vzemimo za primer slovnične oznake o številu, kot so *množina*, *v množini*, *navadno v množini*, *navadno v ednini*, *navadno v množini ali dvojini* itd.[4] Kot smo že omenili v razdelku 3, lahko opozorila o morebitnih slovničnih oznakah iz korpusa izvažamo avtomatsko, pri čemer za vsako oznako določimo določeno statistično mejo, ob kateri se opozorilo izpiše.[5] Vendar pa ob izpisu opozorila izgubimo informacijo o točnem odstotku pojavitev iztočnice v množini, ki bi bila koristna npr. za redaktorje pri potrjevanju geselskih člankov. Takšne informacije bomo pri izdelavi SSSJ vključili v bazo na ravni iztočnice in tudi na ravni posameznih kolokatorjev oz. skupin kolokatorjev ali celo struktur. Informacijo o pogostosti posameznih oblik iztočnice v korpusu Gigafida že imamo zapisano v Sloleksu (Dobrovoljc et al. 2015), tako da nam je v slovarsko bazo niti ni treba zapisati, ampak samo navedemo sklic na geslo v Sloleksu. Podatke o pogostosti pojavljanja oblik iztočnice s posameznim kolokatorjem pa lahko izvozimo med postopkom avtomatičnega pridobivanja podatkov iz korpusa. Takšni statistični podatki v slovarski bazi bodo zelo koristni pri posodabljanju gesel ob morebitnih povečanjih korpusa, saj bomo s pomočjo primerjave z novejšimi podatki lahko ugotovili morebitne spremembe v rabi iztočnice, njenih kolokatorjev in struktur ter posledično pomenov in oznake posodabljali oz. spreminjali veliko bolj sistematično. Dodatna prednost zapisovanja statističnih informacij v slovarsko bazo je tudi v tem, da lahko odstotkovni prag zelo znižamo ali celo izvozimo podatke za vse kolokatorje in strukture, potem pa se pri pripravi slovarja odločamo o mejnih vrednostih za vključitev posamezne oznake. Takšna rešitev je veliko boljša tudi za potencialne uporabnike slovarske baze, kot so npr. jezikoslovci, jezikovni tehnologi in drugi raziskovalci.

Ločena koraka analize in izdelave osnutkov gesel ter njihovega redakcijskega pregleda dajeta tudi možnost, da se poleg razlik v določanju oz. oblikovanju oznak vpelje različen pristop k določanju z oznakami povezanih informacij. Predstavljajmo si scenarij, da leksikograf pri analizi oblikuje razlago za določen pomen, ki vsebuje implicitno oznako, potem pa se redaktor odloči, da je bolje, da je oznaka predstavljena eksplicitno, kar seveda zahteva tudi preoblikovanje razlage. Del leksikografovega truda, vloženega v razlago, je tako izničen. Možna rešitev bi bila odprava implicitnega predstavljanja oznak oz. z njimi povezanih informacij v koraku analize in uvedba ločenega koraka (ali povečanje obsega nalog pri redakciji) za odločitve o dokončnem prikazu oznak v geslih. Pri analizi bi torej leksikografi oznake navajali samo eksplicitno, pri čemer bi morala uredniška ekipa slovarja pri vseh oznakah, ki bi jih lahko predstavili tudi implicitno, predvideti nekoliko

4 Navedeni primeri oznak so vzeti iz Leksikalne baze za slovenščino.

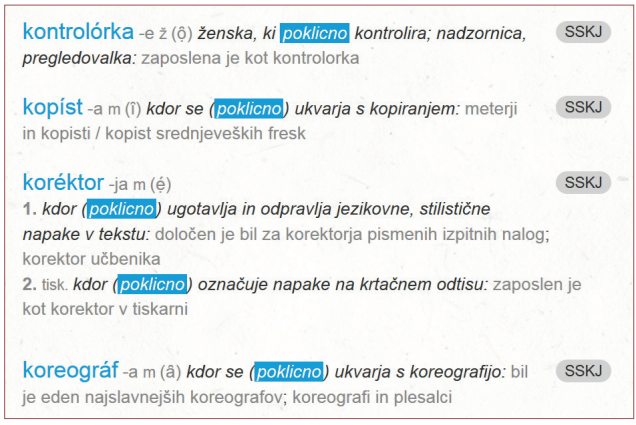
5 Mejne statistične vrednosti ne upoštevajo pogostosti določenega jezikovnega fenomena (npr. da je trpnik pri večini glagolov precej redkejši kot tvornik). Rundell in Kilgarriff (2011) svetujeta, da v takšnih primerih za kriterij vzamemo odstotkovno mero, npr. določen zgornji odstotek glagolov, ki se pogosto pojavljajo v trpniku.

večjo fleksibilnost pri ubeseditvi ali pa možnost uporabe krajše oblike oznake in daljšega pojasnila. Tako se tudi zmanjša možnost, da se implicitno predstavljena informacija o omejitvi rabe, pripisana med analizo, hkrati ne določi s skrito oznako in jo je kasneje v bazi težko najti.

V ločenem koraku ali med redakcijo bi se nato sprejemale odločitve o tem, kako najbolje predstaviti informacijo v oznaki. Takšna rešitev bi terjala tudi drugačen pristop k razlagam: med analizo bi se naredil neke vrste osnutek razlage za vsak (pod)pomen, lahko bi se uporabila tudi sinonimna razlaga ali kazalnik, med redakcijo pa bi se potem oblikovala dokončna razlaga.

4.3 Možnosti za slovarske uporabnike

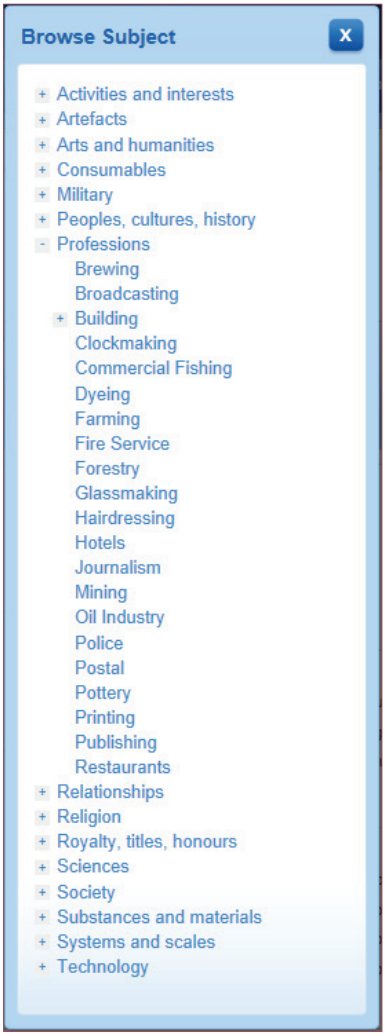
Pomembna prednost slovarske baze je predvsem v možnosti uporabe (določenih vrst) skritih oznak, ki so lahko zelo koristne za napredna iskanja. Pri tem lahko v bazo vključimo tudi tipe oznak, izdelane samo za slovarsko bazo. Prednosti takšne rešitve najlažje ponazorimo s primerom. Recimo, da želimo v slovarju poiskati vse iztočnice oz. pomene, ki pomenijo poklice. V obstoječih slovarjih, ki nimajo tako z informacijami bogatih slovarskih baz, moramo iskati določene besede v delih gesla, največkrat v razlagah. Tako lahko recimo pomene, ki se nanašajo na poklice, v SSKJ na portalu Fran poiščemo z iskanjem *poklicno* v razlagah (Slika 2).



Slika 2: Nekaj rezultatov za iskanje *poklicno* v razlagah SSKJ (portal Fran).

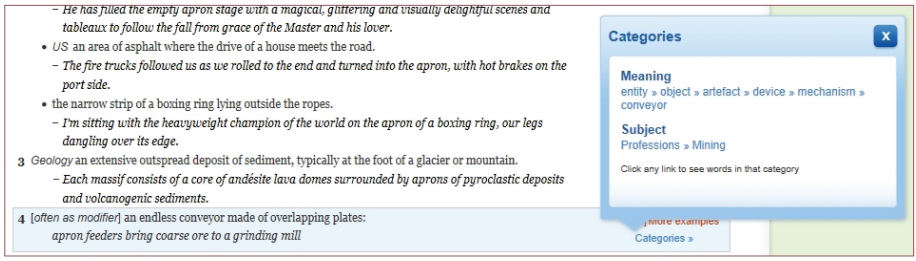
Težavi sta vsaj dve: po eni strani uporaba *poklicno* v razlagi ne pomeni nujno, da gre za poklic (npr. ***kri*** 3b *izraža socialno, poklicno pripadnost*), po drugi strani pa imamo iztočnice oz. pomene, pri katerih gre za poklic, a beseda *poklicno* v razlagi ni uporabljena (npr. ***prodajalec*** *kdor prodaja*).

V slovarski bazi to lahko rešimo tako, da pri vseh iztočnicah oz. njihovih pomenih in podpomenih, ki se nanašajo na poklice, uporabimo skrito oznako *poklic*, pa mogoče tudi z oznako določenega poklica. Tako lahko uporabnikom kasneje omogočimo, da na enostaven način poiščejo vse poklice ali vse iztočnice oz. pomene, povezane s posameznih poklicem. Podobno informacijo ima v slovarski bazi tudi angleški slovar Oxford[6] in jo je do nedavne prenove vmesnika ponujal pri naprednem iskanju (Slika 3), pa tudi v pojavnem oknu pri izbiri posameznega pomena (Slika 4).



Slika 3: Napredno iskanje po strokovnem področju v slovarju Oxford (prejšnja verzija vmesnika).

6 http://www.oxforddictionaries.com/ (dostop 8. 8. 2015).



Slika 4: Informacija o tematiki (Subject; v tem primeru o poklicu) pri določenem pomenu v slovarju Oxford (prejšnja verzija vmesnika).

Za naprednejše uporabnike, med katere štejemo npr. jezikoslovce in druge raziskovalce ter jezikovne tehnologe, bo dejansko bolj zanimiva slovarska baza kot sam slovar, saj njihove informacijske potrebe ponavadi presegajo informacije, ki jih ponuja slovar, poleg tega želijo čim večjo svobodo pri analizi in obdelavi jezikovnih podatkov. Za njihove potrebe lahko oznake v bazo dodamo tudi naknadno oz. neodvisno od leksikografske analize, npr. za označitev določenega besedišča, za katerega že imamo izdelan seznam. Na primer, za poučevanje in učenje slovenščine kot tujega jezika bi lahko uporabili oznake A1, B1, B2 itd. za označitev besedišča po ravneh skupnega evropskega referenčnega okvira (CEFR). Takšna informacija v slovarski bazi je potem koristna tudi za učitelje slovenščine kot drugega/tujega jezika, izdelovalce učnih gradiv, raziskovalce ipd., konec koncev pa tudi za leksikografe, ki bi se lotili izdelave slovarja za tujce.[7]

5 ZAKLJUČEK

V prispevku smo pokazali, katere korenite spremembe prinaša vpeljava slovarske baze, opremljene s številnimi informacijami, od katerih vse niso predvidene za prikazovanje v slovarju, tudi za postopke označevanja. Ob upoštevanju prednosti slovarske baze, kot so npr. fleksibilnost pri ubeseditvi oznak pri analizi korpusnega gradiva in možnost uporabe skritih oznak, nikakor ne smemo pozabiti na vlogo ustrezno pripravljenega korpusnega gradiva in raziskav med uporabniki, ki lahko še dodatno izboljša določanje oznak in njihovo ubeseditev ter vizualizacijo.

SSSJ bo izkoristil vse naštete prednosti za označevanje omejitev v rabi besed in podobnih informacij, saj bo le tako zadostil potrebam slovarskih uporabnikov, tudi naprednejših, navsezadnje pa tudi potrebam leksikografov, ki jim bo z uvedbo sodobnih postopkov določitve oznak in njihovega zapisa omogočil hitrejšo in objektivnejšo analizo korpusnega gradiva.

7 Glej prispevek Rozman et al. (2015) za diskusijo o tem, ali je potreben slovar za tujce ali pa bi njihove potrebe lahko zadovoljil tudi splošni slovar z določenimi dodanimi elementi.