

Specializirana leksika v splošnem slovarju

Špela Vintar

Abstract

This paper discusses theoretical and methodological issues related to specialised vocabulary in a dictionary of contemporary Slovenian. Key issues are addressed such as the role of terminology in a general dictionary, user requirements and needs, the complexity of the distinctions between general and specialised terms, as well as corpus composition and representativeness. We propose a model where lexical items are categorised into three levels of termhood, with each level of specialisation requiring a different strategy of lexicographical description. By illustrating possible relations between the proposed categories and the corpus-based methodology of candidate extraction, a working methodology is established for handling specialised units in a general dictionary.

Keywords: specialised vocabulary, general dictionary, terminology extraction, user requirements

Ključne besede: specializirana leksika, splošni slovar, luščenje terminologije, uporabniške zahteve

1 UVOD

V vsakdanjem življenju vsi uporabljamo specializirane izraze, saj se prav vsak od nas ukvarja ali srečuje s kako dejavnostjo ali življenjskim področjem, ki ga ne delijo vsi govorci slovenščine in ki zahteva določena znanja ali izkušnje – je torej specializirano. Ob tem smo kot materni govorci opremljeni z intuitivno zaznavo, da so nekateri izrazi bolj specializirani od drugih, svoj senzor terminološkosti pa znamo nasloniti tudi na konkretne in uporabne utemeljitve: »Te besede nihče ne razume« ali »Tako temu rečejo mehaniki«.

Prispevek predstavlja niz razmišljanj o vključevanju specializirane leksike v slovar sodobnega slovenskega jezika, vanje pa so zajeti različni vidiki od leksikografske tradicije pri nas in drugod, predvidevanj želja in potreb uporabnikov, zbiranju in izbiranju korpusnega gradiva pa vse do leksikografskega oziroma terminografskega opisa in predstavitve podatkov po meri uporabnikov.

Že na prvi pogled je jasno, da vzpostavitev celovitega metodološkega okvira za vključevanje specializirane leksike v slovar ni enostavna naloga, vendar bi pričakovali, da so na podobno temeljna vprašanja v preteklosti odgovarjale že leksikografske ekipe pomembnih splošnih enojezičnih slovarjev in da je njihova spoznanja razmeroma preprosto prenesti v naše okolje. A ko začnemo podrobneje pregledovati strokovno literaturo, je podrobnih in eksplicitnih metodoloških študij o specializirani leksiki v splošnih slovarjih presenetljivo malo, še posebej v primerjavi z bogatim naborom literature o metodologiji specializiranih in terminoloških slovarjev. O slovarskih konceptih in odločitvah v zvezi s specializirano leksiko nam tako še največ povedo sami slovarji in ponekod njihovi uvodi. Druga izrazitejša težava pri prenosu tujih izkušenj v slovensko okolje pa je specifičnost slovenske leksikografske tradicije na eni strani in aktualni družbeni ter jezikovnopolitični okvir na drugi strani, oboje namreč vpliva na pričakovanja uporabnikov slovarja in posledično na nabor funkcij, ki naj bi jih bodoči slovar opravljal. Iz tega izhaja osvobajajoče in hkrati zavezujoče prepričanje, da je metodologijo obravnave specializirane leksike v sodobnem slovarju slovenščine treba vzpostaviti na novo, z njenim preskušanjem v praksi pa naj bi se samodejno razvil tudi iterativni cikel metodoloških popravkov in izboljšav.

2 VLOGA SPECIALIZIRANE LEKSIKE V SPLOŠNEM SLOVARJU

Kot ugotavlja Boulanger (1996: 141) v svojem diahronem pregledu, je že od 19. stoletja naprej v leksikografiji opazna težnja po vse večjem vključevanju

strokovnega izrazja v splošne eno- in večjezične slovarje. Od razsvetljenstva naprej je vztrajno rasla vloga znanosti in tehnike v vsakdanjem življenju, prav tako se je zviševala raven izobrazbe in posledično vključenost t. i. tehnolektov v vernakularni jezik in jezikovne priročnike. Druga polovica 19. stoletja je bila tudi za slovenščino obdobje živahnega ustvarjanja strokovnega izrazja, pomembno znamenovano tudi s prevodi znanstvenih in strokovnih del iz pretežno nemščine, pa tudi drugih jezikov (Prunč 2009).

V 20. stoletju so se slovarji začeli odmikati od normativne vloge in se približevati uporabnikom, ti pa so od slovarjev velikega obsega pričakovali vse boljše zastopanost specializirane leksike. Landau (2001) ugotavlja, da so veliki splošni slovarji vse bolj podobni združevanju številnih specializiranih (LSP) slovarjev z nekdanjim splošnim, vzrok za to pa je po njegovem tudi dejstvo, da v segmentu specializirane leksike zaradi naglega razvoja znanosti in tehnologije nastaja bistveno več novega izrazja kot v splošnem besedišču. Utemeljitev za vse večje vključevanje specializirane leksike v slovarje je več, če jih povzamemo po Josse-
lin-Leray (2005):

- a) Leksikografska tradicija, ki – kot smo že omenili – vsaj zadnje stoletje in pol narekuje vse višji delež specializirane leksike v splošnih slovarjih (prim. razdelek 2.2).
- b) Vse intenzivnejše vstopanje specializirane leksike v vsakdanji diskurz, ki ga imenujemo tudi determinologizacija ali celo vulgarizacija terminologije. Pri laični rabi terminologije pogosto prihaja do odstopanj od prvotnega specializiranega pomena, pa tudi metaforizacije (npr. *anoreksija obrestnih mer*); za splošni slovar je pomembno, da opisuje tudi takšno rabo.
- c) Didaktična oz. pedagoška vloga splošnih slovarjev, ki je najprej temeljito preobrazila angleško slovaropisje predvsem zaradi potreb učencev angleščine kot tujega jezika, a v sodobnem globaliziranem svetu ta vloga vsaj deloma pritiče večini enojezičnih slovarjev.
- č) Težnja po izčrpnosti (angl. *comprehensiveness*), ki želi z ustvarjanjem enega vira za različne namene zadostiti čimveč uporabniškim potrebam, obenem pa je ob sodobnih informacijskih tehnologijah in primarnosti elektronske oblike izčrpnost bistveno lažje dosehati kot nekoč.
- d) Pričakovanja in potrebe uporabnikov, ti se namreč za tehnologijo in znanost vse bolj zanimajo in so vse bolj obveščeni.

Vsi navedeni razlogi obstajajo tudi v slovenskem prostoru. Če se pri leksikografski tradiciji omejimo le na SSKJ in njegovega načrtovanega naslednika NSSKJ, je SSKJ terminologiji namenil pomembno vlogo, pri tem pa ustrezno zamejil raven specializiranosti in aktualnost:

Terminologija je upoštevana nekako v obsegu srednje šole, zlasti če jo podpira publicistična ali poljudnoznanstvena raba. /.../ Terminološko gradivo je nastalo deloma z izpisovanjem poljudnoznanstvenih del, srednješolskih učbenikov in strokovnih slovarjev, deloma pa s prispevki okoli sto terminoloških svetovalcev. Od tega gradiva so bili za slovar odbrani samo termini, ki se rabijo v novejšem času. (SSKJ, Uvod: XVI–XVIII).

Zastopanost posameznih ved v SSKJ ni uravnotežena, prav tako pogostnost rabe ni bila poglaviti kriterij za vključevanje terminologije v slovar, a z vključevanjem terminoloških svetovalcev so skušali leksikografi doseči najboljši možni kompromis med danimi omejitvami glede obsega in specializiranosti ter strokovnostjo razlag:

V terminologiji so najbolj zastopane tehnične vede, manj pa obrti in posamezne stroke, za katere ni bilo mogoče dobiti popolnejšega gradiva (montanistika, radio in televizija, bibliotekarstvo), ali stroke, ki so pri nas šele v razvoju (pomorstvo, kibernetika). (ibid. XVI).

Metodološka izhodišča glede vključevanja terminologije v NSSKJ se v veliki meri naslanjajo na stari SSKJ, izpostavljena pa je potreba po drugačni zastopanosti področij, vključevanju novih področij in uporabi sodobnejših metod terminološkega posvetovanja (Humar 2009). V konceptu NSSKJ (Gliha Komac et al. 2015: 49–51), objavljenem marca 2015, je leksikografska metodologija nadrobneje obrazložena, in sicer avtorji specializirano leksiko delijo na povsem in delno determinologizirano, pri čemer je prva deležna običajne leksikografske obravnave brez dodajanja terminoloških kvalifikatorjev in sodelovanja področnih strokovnjakov, druga skupina pa predvideva oblikovanje posplošenih, a strokovno pravih razlag ob pomoči področnih strokovnjakov ter označevanje specializiranosti s terminološkimi kvalifikatorji. Kriterija za razločevanje med prvo in drugo skupino sta dva, in sicer stopnja determinologizacije ter poznanost izraza splošnemu uporabniku, pri ugotavljanju obeh pa naj bi se leksikograf naslanjal na podatke iz korpusa. Pri tem, po naši oceni ključnem metodološkem vprašanju, je omenjeni osnutek zelo nedorečen.

Če se vrnemo k razlogom za vključevanje terminologije v splošne slovarje, je postopek determinologizacije, ko v splošno rabo prehaja velik del inventarja posameznih strok (informacijske tehnologije) ali pa postajajo strokovna področja predmet širšega zanimanja (zdravje, finance, okolje, šport), intenziven tudi v sodobni slovenščini. Značilnost determinologizacije je, da izrazi močno razširijo svoje prvotno pomensko in ekspresivno polje, ob tem pa ohranijo zgolj del strokovnega pomena. Temu ustrezen mora biti slovarski opis, saj lahko predvidevamo, da bo slovenski uporabnik sodobnega slovarja z njim želel zadostiti tudi informacijskim potrebam po razlagah specializiranega besedišča, vendar pri pomenih, ki so prešli

v splošno rabo, ne bo pričakoval formalne terminološke definicije. Pri ugotavljanju stopnje determinologizacije nam predvsem pomaga korpusna metodologija, ki meri razširjenost izraza v povsem splošnih, poljudnostrokovnih in strokovnih besedilih, pomemben podatek pa je tudi pojavljanje izraza v uporabniško ustvarjenih spletnih besedilih ter v govornem korpusu.

Novi slovar sodobne slovenščine bo imel zagotovo tudi didaktično vlogo, pravzaprav bi močno potrebovali slovar, ki bi se tej vlogi prednostno posvetil. Vprašanje, kako naj slovar najbolje poskrbi za potrebe učech se, najsi bodo rojeni govorci slovenščine ali tujci, presega okvir tega prispevka, se pa uporabniškim potrebam podrobneje posvetimo v naslednjem razdelku.

2.1 Potrebe in pričakovanja uporabnikov

Sodobni splošni slovar je zamišljen primarno v elektronski obliki in nagovarja več tipov uporabnikov, ki jih Arhar Holdt razvršča v tri skupine glede na namen uporabe: a) izobraževanje, kamor sodijo uporabniki slovenščine kot prvega, drugega in tujega jezika, b) z jezikom povezani poklici, kamor se uvrščajo lektorji, novinarji, prevajalci, književniki, akademiki, znanstveniki itd., in c) prosti čas, kjer najdemo jezikovne navdušence, ugankarje in reševalce križank, ljubiteljske jezikovne svetovalce in podobno (Arhar Holdt 2015).

Čeprav se z vprašanjem uporabniških tipov in njihovih slovarskih potreb v digitalni dobi ukvarjajo številni avtorji (Müller-Spitzer 2014; Hult 2014; Lew et al. 2014), je razmeroma malo teh raziskav usmerjenih prav v specializirano leksiko. Ob analizi ugotovitev različnih uporabniških študij in anket pa se izkaže, da so potrebe in pričakovanja uporabnikov v večini slovarskih situacij neločljivo povezani s specializirano leksiko; tako denimo Müller-Spitzer (2014) identificira pomembno, a pogosto spregledano uporabniško skupino jezikovnih entuziastov, ki se ljubiteljsko ukvarjajo z nenavadnimi besedami, rešujejo in sestavljajo križanke ali igrajo Scrabble, pri tem pa v slovarju pogosto iščejo prav specializirane izraze ali tujke. Hult (2014), ki je analizirala komentarje uporabnikov ob spletni različici pedagoškega slovarja švedščine, navaja številne kritike, ki se nanašajo na slabo zastopanost specializiranih izrazov v slovarju.

Posebej podrobno se s pričakovanji uporabnikov glede specializirane leksike ukvarjata Josselin-Leray in Roberts (2005), ki poročata o rezultatih ankete o uporabi eno- in dvojezičnih slovarjev, izvedene med angleško in francosko govorečimi uporabniki. Na vprašanja je odgovorilo skupaj 649 oseb iz treh uporabniških skupin: prevajalci, strokovnjaki in splošna javnost, ki v prispevku žal ni podrobneje opredeljena. Vprašanja so se nanašala na različne vidike uporabe slovarjev, skoraj

vsa – skupaj jih je bilo 19 – pa so bila povezana z deležem specializirane leksike. Če se omejimo le na enega od številnih vidikov, ki jih obravnavata avtorici, in sicer na vprašanje, ali se zdi uporabnikom delež vsebovane terminologije pomemben kriterij pri nakupu splošnega slovarja in ali bi kupili slovar brez terminov, so bili rezultati enoznačno v prid terminologije, in sicer pri vseh skupinah uporabnikov, tudi pri laični, saj je v povprečju kar 60 % vprašanih kategorično zavrnilo možnost nakupa slovarja brez specializirane leksike.

V Sloveniji sorodne raziskave, ki bi se ciljno ukvarjala s pričakovanji in potrebami uporabnikov slovarja glede terminologije, še ni bilo, zato skušamo to vrzel premostiti z analizo strežniških dnevnikov, pridobljenih z nekaterih terminoloških spletišč (Vintar 2015a).

2.2 Delež specializirane leksike

Če predpostavimo, da si torej tudi slovenski uporabnik v slovarju želi specializirano leksiko, je naslednje naravno vprašanje, kolikšen delež besedišča naj bi predstavljale specializirane iztočnice v razmerju do splošnega jezika in do celotnega obsega slovarja.

Pregled tuje literature tudi na to vprašanje ne daje enoznačnih odgovorov. Landau (1974) pri analizi Webstrovega slovarja *Third New International Dictionary* omenja približni delež 40 %, vendar analiza posameznih strani slovarja prikaže vse do 89 % specializiranih gesel. Béjoint se prav tako ukvarja s tem vprašanjem, a se o deležu v odstotkih ne opredeli zaradi zahtevnega določanja meje med termini in netermini (Béjoint 1988: 360). S podobno težavo sta se spopadla tudi Boulanger in L'Homme (1991: 25); v kasnejši študiji pa Boulanger ugotavlja za angleške in francoske enojezične slovarje med 40 in 50 % specializirane leksike (Boulanger 1996: 147). Novejša študija Vrbinc in Vrbinc (2013) raziskuje zastopanost specializirane leksike ter uporabo področnih oznak v slovarjih *Oxford Advanced Learner's Dictionary* tretje, četrte in osme izdaje. Avtorici ugotavljata postopno zviševanje deleža specializiranih leksemov, a ne omenjata števil, ugotavljata pa tudi posodobitve in izboljšave v rabi področnih oznak, ki jih je sicer v *OALD8* devetnajst.

Tu zastopani pristop h gradnji sodobnega slovarja skuša biti pragmatičen v smislu, da mora biti izbrani metodološki okvir glede zajema in obdelave specializirane leksike časovno in finančno izvedljiv, obenem pa učinkovit in trajnosten. Digitalna oblika omogoča popolno svobodo glede količine ponujenih podatkov, računalniške metode luščenja terminologije in definicij iz korpusov pa močno olajšajo pot do leksikografskih podatkov. Ker specializirani izrazi, dokler so pasivno shranjeni

v slovarski bazi in tam čakajo na ustrezno poizvedbo, nikogar ne motijo, bi bili teoretično glede njihovega deleža lahko nadvse radodarni. Še vedno pa je leksikografska obdelava iztočnic zahtevno strokovno delo, ki pri specializiranih enotah zahteva tudi sodelovanje s strokovnjaki, zato so omejitve v luči izvedljivosti slovarskega projekta še kako nujne. Poleg tega ni točno, da velika količina redundantnih specializiranih podatkov v slovarski bazi uporabnika ne moti, saj pri vsaki odprti poizvedbi te iztočnice povzročajo šum in dajejo popačeno sliko tega, kar naj bi slovar splošnega jezika predstavljal.

Ker si je v kontekstu korpusne leksikografije razmeroma nesmiselno vnaprej določiti, kolikšen delež slovarskih iztočnic naj bi bil specializiran, saj bi bilo to podobno, kot če bi pri jemanju vzorca krvi zdravnik vnaprej določil število rdečih in belih krvnih teles, nam na zgoraj zastavljeno vprašanje ostane le zelo krhek odgovor: splošni slovar naj vsebuje ravno pravšnji delež specializiranih enot, da bo z njim zadovoljil kar največ informacijskih potreb, obenem pa ohranil učinkovitost poizvedovanja in karakter splošnega slovarja.

Naslednje vprašanje, ob katerem se bežno ustavimo, je, ali naj bodo posamezna specializirana področja v slovarju zastopana v enakih oziroma uravnoteženih deležih. Leksikografska tradicija kaže, da si tega cilja veliki splošni slovarji navadno niso zastavljali, večinoma so se mu celo izrecno odpovedali iz različnih razlogov (Josselin-Leray 2005: 146). Nemalokrat so na dobro ali slabo zastopanost posameznih strok v slovarjih vplivali tudi povsem idiosinkratični dejavniki: »Eden od urednikov Oxfordovega slovarja angleščine je bil slučajno amaterski mineralog in zato je akademski Oxfordov slovar angleščine še posebej bogat z mineraloško terminologijo« (Béjoint 1988: 361; prevod Š. V.). Če bi želeli zadostiti pogoju uravnoteženosti, bi se namreč takoj soočili z dejstvom, da stroke med seboj niti približno niso primerljive glede količine specializiranih izrazov, ki jih uporabljajo, zato bi bila kakršna koli številčna uravnoteženost nesmiselna. Poleg tega so nekatere stroke za medije in širšo javnost bolj zanimive kot druge, posledično iz takšnih strok več izrazja prehaja v neterminološko rabo (in v splošne korpusse). Ahmad et al. (1995) menijo, naj bi se splošni slovar osredotočal na tiste specializirane izraze, pri katerih se laični in strokovni interes prekrivata. S tega zornega kota bi v splošnem slovarju po vsej verjetnosti pričakovali izraze, kot so *dokapitalizacija*, *hipertenzija* in *prisilna poravnava*, manj pa *magnon* (fizika), *manipul* (zgodovina) ali *izobutil acetat* (kemija).

Privzemimo torej, da splošnemu slovarju ni treba uravnoteženo predstavljati vseh strok tega sveta, namesto tega pa naj bi zajemal izrazje vseh tistih strok, s katerimi se bolj ali manj pogosto srečuje tudi laik. Ob tem privzamemo tudi, da splošni slovar za nobeno od teh strok ne more dosegati enake ravni poglobljenosti in celovitosti kot terminološki slovar (Peters in Fernandez 2013). Če besedišče slovarja

zajemamo iz velikega reprezentativnega korpusa, kakršen je Gigafida, bi lahko domnevali, da je že pojavnost izraza v korpusu dokaz za »srečevanje laičnega in strokovnega interesa« in se tako pri presojanju o vključevanju specializiranih enot v slovar naslanjali kar nanjo. V naslednjem razdelku bomo skušali pojasniti, zakaj tako poenostavljen pristop ne zadošča.

3 GRADIVO ZA ZAJEM SPECIALIZIRANE LEKSIKE

V nadaljevanju v razmislek o metodologiji vključevanja specializirane leksike v splošni slovar uvedemo še najpomembnejši in hkrati najboljčutljivejši parameter, to je korpus.

3.1 (Ne)uravnoteženost korpusa in (ne)uporabnost absolutne pogostosti

Že od začetkov korpusnega jezikoslovja smo priča tehnološkemu razvoju, ki omogoča gradnjo vse večjih korpusov, od razmaha interneta pa so digitalna besedila postala tako vseprisotna, da korpusi z milijardo ali več pojavniciami niso več nikakršna redkost.

Za gradnjo korpusnih splošnih slovarjev uporabljamo referenčne korpuse, ki naj bi – če so tudi dobro uravnoteženi – predstavljali nekakšen skupni imenovalec različnih zvrsti in podzvrsti, ki jih uporabljajo govorci nekega jezika. Za slovenščino imamo tako na voljo Gigafido s prek milijardo pojavnici, ki ni uravnotežena, ter Kres s 100 milijoni pojavnici, ki je vzorčni korpus z izenačenimi deli posameznih žanrov; oba skupaj tvorita tudi vir za ugotavljanje in primerjavo pogostosti leksikalnih enot, kar je za leksikografa pomembna informacija.

Pri zajemu specializirane leksike nam splošni referenčni korpus lahko nudi pomembne dokaze o tem, ali je določeni specializirani izraz prodril tudi v polstrokovno oziroma povsem nespecializirano rabo, medtem ko so podatki o pogostosti uporabni zgolj z zadržkom.

	GF	Kres
paralelogram	107	18
trapez	923	126
deltoid	23	3

Matematični izrazi *paralelogram*, *trapez* in *deltoid* so istoredni termini, ki označujejo like v poglavju srednješolskega učbenika o večkotnikih, a se v korpusu pojavljajo z zelo različnimi frekvencami. Če bi želeli v slovarju pravično zajeti matematično izrazje na srednješolski ravni, bi morali vse tri izraze vključiti, čeprav se pojavljajo s tako različnimi frekvencami. Pri izbranem primeru se hitro pokaže temeljni izziv zajemanja in slovarske obdelave specializiranih iztočnic, *trapez* ima namreč toliko višjo frekvenco zaradi svojih nematematičnih pomenov. V Gigafidi najdemo še naslednje specializirane pomene: polje omejitve (košarka; »podaja iz trapeza«), telovadno orodje (gimnastika; »artistka na trapezu«), pripomoček za dvig z bolniške postelje (medicina; »vstati s pomočjo trapeza«), nastavitev digitalne slike na monitorju (računalništvo; »pogrešam digitalno korekcijo trapeza«), človek za protiutež pri nagibu jadrnice (jadranje; »flokist visi v trapezu«).

Po vsej verjetnosti se bo leksikograf pri obdelavi iztočnice *trapez* moral pri vsakem pomenu posebej odločati, ali ga zapisati v slovar ali ne. Tudi če bi se pri tem lahko opiral na samodejno označene korpusne pogostosti posameznih pomenov, češar jezikovnotehnološka infrastruktura trenutno še ne omogoča, ostaja vprašanje reprezentativnosti in uravnoveženosti korpusa za izbrane specialnoleksikografske namene odprto. V trenutni korpusni situaciji je namreč povsem mogoče, da se računalniški pomen pojavlja nesorazmerno pogosto zgolj zato, ker korpus vsebuje precejšnje število besedil iz revije Monitor, tam pa uredniki redno izvajajo testiranja monitorjev in poročajo o njihovi korekciji trapeza.

Če torej za nespecializirano leksiko lahko pričakujemo, da nam bodo podatki o pogostosti v uravnoveženem referenčnem korpusu predstavljali relevantno leksikografsko osnovo, tega ne moremo z gotovostjo pričakovati za specializirane izraze.

3.2 Prednostna področja in dopolnjevanje gradiva

Opisana situacija nas pripelje do sklepa, da če slovar ne namerava zajemati vseh področij in se po drugi strani pri zajemu specializiranih iztočnic ne more zanašati zgolj na korpusno pogostost, je treba v temeljnem slovarskem konceptu opredeliti tista strokovna področja (v povezavi z žanri in ravniyo specializiranosti), ki jih bo slovar zajemal. Za izbrana področja s povezanimi žanri in registri je treba zagotoviti dovoljšen obseg gradiva, da bo mogoče izvajati avtomatske postopke luščenja in jezikovnotehnološke obdelave z namenom pridobivanja iztočnic ter pripadajočih leksikografskih podatkov.

Pri odločanju, katera področja so prednostna, nas do neke mere vodijo temeljni cilji slovarskega projekta ali z drugimi besedami ciljne skupine uporabnikov, vendar je ob tem treba priznati, da bodo odločitve nujno intuitivne in subjektivne,

saj natančne in merljive opredelitve slovarskih potreb ciljne publike v našem prostoru še nimamo; pomemben prispevek k temu je sklop o uporabnikih v tej monografiji skupaj s prispevkom o analizi dnevnikov iskanja Termanie.

V procesu razmišljanja o prednostnih področjih so se izoblikovala tri načela, ki jim skušamo slediti. Prvo se navezuje na prej omenjeno leksikografsko načelo »stičišča med splošno in (zelo) specializirano rabo«, kar pomeni, da velja namečiti posebno pozornost tistim strokovnim področjem, ki so s svojim izrazjem močneje zastopana v splošnem javnem diskurzu in medijih. O njih nam zmore precej povedati že Gigafida, zato smo izvedli hiter pregled 1.000 najpogostejših samostalniških lem in jim pripisali področja. Nastal je naslednji grobi seznam področij: *politika, šport, pravo, ekonomija, finance, mediji, okolje, uprava, zdravje, kultura, IT, promet, turizem*.

Natančnejšo analizo izstopajočih tem v Gigafidi in primerjavo s korpusom slWaC sta izvedla Logar Berginc in Ljubešić (2013) z metodo tematskega modeliranja. V tej analizi se je osem tem, ki sta jih avtorja ročno poimenovala na podlagi samostalniških lem v vsaki temi, izkazalo kot skupnih obema korpusoma: notranja politika, finance, ekipni šport, vojna in terorizem, publikacije in kultura, lokalna (prostorska) politika, zdravje in pravo. Po drugi strani je nekaj tem opaznejših v enem ali drugem korpusu: v Gigafidi izstopajo naselje in cestni promet (zlasti prometne nesreče), prireditve, televizijski in radijski program, neekipni športi ter zaposlitev; v slWaCu pa film, potovanja in turizem, zunanja politika ter mali oglasi.

Z vidika metodologije moramo ob tem pripomniti dvoje. Lastni pregled tisoč najpogostejših lem iz Gigafide je bil opravljen z mislijo na specializirano leksiko, kar pomeni, da številnih lem nismo uvstili na nobeno tematsko področje, saj so se zdele splošne. Tematsko modeliranje Logar Berginčeve in Ljubešića pa je bilo usmerjeno v prepoznavanje glavnih tematskih poudarkov in razhajanj med obema korpusoma ne glede na raven specializiranosti. Tema seveda ni enako strokovnemu področju, zato je pri morebitnih tovrstnih analizah z namenom dejanskega ugotavljanja zastopanosti posameznih strok v korpusu nujno najprej opredeliti metodologijo ter izbrati členitev področij. Drugi pomembni uvid pa se nanaša na dejstvo, da bo analiza zelo pogostih domnevno specializiranih samostalniških lem vključevala predvsem leksikalne enote, ki so delno ali povsem determinologizirane. V njih sicer prepoznamo izvorno stroko, a so splošno znane, za razumevanje ne terjajo poglobljenega poznavanja področja oziroma so postale pomensko ohlapnejše (*kredit, delnica, zaslon, liga, rak*).

Drugo načelo naslavlja bodoče slovarske uporabnike v izobraževanju in nadaljuje slovarsko tradicijo SSKJ s ciljem, naj novi slovar vključuje terminologijo vseh naravnoslovnih, družboslovnih in humanističnih strok na ravni

srednješolskih učbenikov. V tem segmentu obstoječa Gigafida ni reprezentativna, saj sicer vsebuje 75 osnovno- in srednješolskih učbenikov različnih založb in avtorjev, vendar so nekatera predmetna področja bistveno močnejše zastopana kot druga. Za dobro pokrivanje srednješolske terminologije, ki bi jo v slovar zagotovo morali vključevati selektivno, je torej potrebna premišljena dopolnitev učbeniškega podkorpusa.

Tretje načelo pa izhaja iz spoznanja, da še tako obsežen in reprezentativen korpus ne more zajemati celotnega besedišča nekega jezika. Če pustimo ob strani govoreno slovenščino, ki jo je v vsej bogati zvrstnosti tako rekoč nemogoče reprezentativno predstaviti v korpusu, so tu določene vrste pisne jezikovne rabe, ki so v obstoječih korpusih slabo zastopane. Identificirali smo eno takšnih vrzeli, ki smo jo označili s širokim izrazom življenjski dogodki, vanjo pa uvrščamo izrazje, povezano z različnimi pravno-upravnimi, socialnimi, verskimi in drugimi dogodki, s katerimi se vsaj občasno srečuje veliko število ljudi. Sem sodijo denimo različni bančni, zavarovalniški, poštni in upravni obrazci ter dokumenti, ki lahko vsebujejo pomembne specializirane izraze.

Pričakujemo tudi, da bi se ob podrobnejši korpusni analizi specializiranega besedišča posameznih strok pokazale nadaljnje vrzeli, saj uravnotežena zastopnost strok ter z njimi povezanih žanrov in registrov ni bila temeljna prioriteta korpusa Gigafida, poleg tega sestavo korpusa vselej določajo tudi nepredvidljivi vidiki pri pridobivanju besedil. Gigafida vsebuje 1.377 knjižnih stvarnih enot (4,24 %) in 5.871 enot iz revij (21,51 %), pri čemer niso vse revije specializirane. Sedanja besedilna taksonomija je preširoka, da bi omogočala vpogled v zastopnost posameznih strok, zato je v novi različici korpusa načrtovano podrobnejše tematsko označevanje.

Za tiste stroke, ki bodo za novi slovar prepoznane kot prioritete, je v načrtu podrobni pregled obstoječega gradiva z navzkrižnim luščenjem izrazja ter dopolnjevanje s sodobnimi specializiranimi besedili. Po vsej verjetnosti bo treba vključiti še kakšno strokovno revijo ter poiskati ustrezne in sodobne strokovne publikacije, pri tem pa nimamo namena vključevati znanstvenih revij, monografij in disertacij. Za nekatere stroke bodo v pomoč tudi že obstoječi specializirani korpusi (turizem, odnosi z javnostmi; glej tudi <http://nl.ijs.si/noske>).

V sklopu tretjega načela, ki bi ga lahko imenovali tudi zavedanje o nepopolnosti korpusa, velja omeniti še neologizme. Novi izrazi nastajajo skoraj izključno v povezavi s specializiranimi področji, najmočnejše opazimo poimenovalno potrebo pri novih tehnologijah, napravah in storitvah, ki sprožijo val zanimanja tudi med nespecializirano publiko in v splošnih medijih. Novim izrazom s korpusi težko sledimo, pa vendar je njihovo vključevanje v slovar posebej pomembno, še

posebej v primerih, ko nova poimenovanja ne predstavljajo naključnih avtorskih, uredniških ali prevajalskih domislic, temveč je pri njih prišlo do terminološkega dogovora ali vsaj refleksije.

Dober primer takšnega izraza je eden od razdelkov pričujoče knjige z naslovom *množičenje*. Angleški izraz *crowdsourcing* je bil najprej citatno prevzet in je v Gigafidi izpričan s štirimi pojavitvami, nato pa so se pričeli pojavljati poskusi slovenjenja: *moč množic*, *množicanje*, *množgančkanje*, *množičenje*, *množično zunanje izvajanje*. O najprimernejši ustreznici se je razvnelo kar nekaj (spletnih) razprav, mnenje je podala ekipa Terminologišča na Inštitutu Frana Ramovša za slovenski jezik ZRC SAZU, o izrazu so razpravljali uredniki in komentatorji slovarja informatike Islovar. Zaenkrat sicer ne moremo govoriti o dokončnem terminološkem dogovoru, čeprav sta oba prej omenjena foruma kot zmagovalca izbrala *množično zunanje izvajanje*. Nobena od možnih ustreznic v Gigafidi ni izpričana, pa vendar bi v novem slovarju pričakovali ustrezno leksikografsko obdelavo tega in sorodnega pojma *crowdfunding*.

4 ZAJEM SPECIALIZIRANEGA IZRAZJA

Če je še pred desetletjem ali dvema izdelava slovarskega geslovnika predstavljala najbolj garaško in dolgotrajno leksikografsko delo, nam danes jezikovnotehnološka orodja to delo močno olajšajo in pohitrijo, leksikografovi napori pa so bolj usmerjeni v pregledovanje, urejanje, čiščenje in dopolnjevanje samodejno izluščenihih podatkov.

Za luščenje specializiranihi entot iz besedil so se v preteklosti razvila številna orodja, tudi za slovenščino (Vintar 2010). Orodje LUIZ je bilo razvito za dvojezično luščenje izrazja iz angleško-slovenskih vzporednih in primerljivih besedil, lahko pa se uporablja tudi za enojezično luščenje. Osnovna zamisel luščenihi temelji na primerjavi pogostosti med specializiranim in referenčnim korpusom in s tem sledi predpostavki, da so termini določenega področja pogostejši v besedilih tega področja kot v nespecializiranihi besedilih. Tako ugotovljena »ključnost« za posamezne besede se kombinira z jezikovno odvisnimi oblikoskladenjskimi vzorci ter statistično hevrstiko razvrščanihi ter kot izhodne podatke ponudi seznam eno- in večbesednih terminoloških kandidatov.

Luščilnik LUIZ je bil v preteklosti preskušen za različna področja (Vintar in Erjavec 2008; Vintar in Fišer 2009; Vintar 2010; Logar et al. 2012; Pollak 2014), nikdar doslej pa z namenom pridobivanja izrazja za splošni slovar. Še preden smo se torej odločili luščilnik preskusiti za tukaj opisani namen, smo se dobro zavedali pomanjkljivosti pristopa, saj je zajem izrazja za specializirani slovar povsem

drugačna naloga od luščenja za splošni slovar. Eden od vidikov, ki je zahteval metodološki premislek, je že oblika terminov. Medtem ko v visoko specializiranem besedišču lahko pričakujemo veliko število tro-, štiri- in celo večbesednih terminoloških enot, je pri izrazju za splošni slovar primernejša omejitev na eno- in dvobesedne, izjemoma trobesedne enote. Prav tako je bilo treba temeljito prilagoditi hevrstike glede pogostosti, saj redkih in visokospecializiranih izrazov ne želimo vključevati. O metodologiji polavtomatskega luščenja izrazja podrobneje razpravljamo v naslednjem razdelku.

S pomočjo luščenja ter navzkrižnih primerjav med podkorpusi resda pridobimo sezname potencialnih iztočnic po področjih, ki pa zahtevajo pregled in dopolnjevanje. Dopolnjevanje bo potrebno denimo pri novejših izrazih, ki bodo v korpusu izpričani s premajhno pogostostjo, da bi bili samodejno izluščeni, predvidevamo pa tudi uporabo sodobnih jezikovnotehnoloških metod za luščenje nad-, pod- in protipomenk ter zagotavljanje uravnotežene zastopanosti posameznih pojmovnih polj.

Že v tej fazi bo pri izdelavi področnih geslovníkov potrebna pomoč področnih strokovnjakov, ki bodo pri pregledu izluščenih seznamov prepoznali terminološke variacije, determinologizirane izraze in sopomenke ter s tem leksikografu olajšali razvrščanje izrazja v posamezne kategorije specializiranosti.

5 RAVEN SPECIALIZIRANOSTI IN LEKSIKOGRAFSKI OPIS

Že iz dosedaj opisanih primerov je razvidno, da niso vsi izrazi enako terminološki. Pri luščenju ključnih besed iz podkorpusa glasbenih učbenikov tako med prvimi 100 besedami opazimo mnogim znane glasbene izraze, kot so *ton*, *nota*, *sonata*, *mol*, *tempo*, *sopran*, pa tudi manj znane kompleksnejše izraze, pri katerih bi imeli kot laiki morda že težavo z oblikovanjem strokovno ustrezne razlage: *sinkopa*, *sektakord*, *kromatičen*, *kadenca*, *modulacija*.

Če je izhodišče tu opisanega slovarskega projekta zadovoljiti najmanj tri heterogene skupine uporabnikov, tj. učence vseh stopenj, jezikovne profesionalce in tiste, ki jim je jezik konjiček, se morajo tudi tipi leksikografskega opisa prilagajati tako ciljni skupini kot specializiranosti izraza.

V skladu s tem smo predvideli razvrščanje specializiranih iztočnic v tri skupine, ki jim v nadaljevanju pravimo košarice in ki narekujejo različne pristope k leksikografskemu opisu.

Prva ali splošna košarica je najmanj specializirana in vključuje izraze, pri katerih sicer prepoznamo navezavo na določeno strokovno področje, vendar poznavanje področja ni pogoj za razumevanje in pravilno rabo. Ti izrazi so pogosti tudi v referenčnem korpusu (> 3.000) in se v splošnih besedilih rabijo povsem determinologizirano. Pri leksikografskem opisu ne potrebujejo področne oznake, niti se njihova specializiranost ne izraža v razlagi. Primeri takih iztočnic so *tempo*, *koncert*, *dirigent*, *nota*, *harmonija*.

Druga ali šolska košarica vsebuje izraze, ki jih redkeje srečamo v splošnih besedilih, a poimenujejo temeljne pojme stroke in se kot taki pojavljajo že v učbenikih nižje stopnje. Izkazujejo jasno navezavo na matično področje, a so izpričani tudi v referenčnem korpusu (> 300). Njihova razlaga lahko vsebuje navedbo področja, zlasti takrat, kadar se determinologizirani pomen razlikuje od področnega. Primeri takih izrazov so *sonata*, *akord*, *oboa*, *dur*, *mol*, *trozvok*.

Tretja ali strokovna košarica je najbolj specializirana, vsebuje izraze, ki so splošni publiki manj znani in je za njihovo razumevanje potrebno predznanje. Redko se rabijo v splošnih besedilih in se ne determinologizirajo; za slovar pa so vendarle pomembni z več vidikov: bodisi se pojavljajo v srednješolskih učbenikih in torej kljub specializiranosti sodijo na raven splošne izobrazbe, bodisi sodijo na eno od prioriternih področij in je v referenčnem korpusu izpričana minimalna pojavnost. Pri njihovem leksikografskem opisu se uporabi področna oznaka, ki uporabnika opozarja na terminološkost, razlaga pa je strokovna. Takšni izrazi so denimo *septima*, *aliquotni ton*, *sekstakord*, če se omejimo zgolj na glasbo.

Če izraz ne izpolnjuje nobenega od navedenih kriterijev, se pravi ne izkazuje minimalne pojavnosti v splošnih besedilih, ne sodi na prioritarno področje niti med učbeniško izrazje, prav tako pa ne predstavlja nujnega dela pojmovnega sistema kakega sorodnega – pogostejšega – izraza, lahko sklepamo, da ni relevanten za splošni slovar.

Metodološko še najbolj zahtevni del opisane klasifikacije je učinkovita izraba korpusnih podatkov, saj se pri ugotavljanju specializiranosti določenega izraza nikakor ne moremo opirati zgolj na njegovo pogostost v referenčnem korpusu. Pri izrazih iz prve in druge košarice nam absolutna pogostost ne pove skoraj nič o dejanski pogostosti specializiranega ali iz njega izpeljanega determinologiziranega pomena, saj imajo številni izrazi več pomenov na različnih področjih. Tako je skupna pogostost izraza *sinkopa* v Gigafidi 269, vendar je od tega precejšnji del pojavitve v medicinskem pomenu sinkope kot začasne izgube zavesti ali omedlevice, drugi del se nanaša na za polovico podaljšani ton v glasbi, tretji del pa vsebuje še pojavitve v imenih podjetij, glasbenih skupin in preneseno rabo. Ustrezna leksikografska presoja se torej na pogostost lahko opira le po opravljenem postopku pomenskega razdvoumljanja.

6 ZAKLJUČKI

Dosedanji razmisleki predstavljajo nabor izhodišč, ki jih smiselno dopolnjujejo druga poglavja tega sklopa, obenem pa so izhodišča usklajena s temeljno filozofijo projekta novega slovarja sodobne slovenščine. Ob tem je – kot pri vsaki metodološki zasnovi – že vnaprej jasno, da se bodo sedaj grobo zarisani okvirji obravnave specializiranega besedišča zbrusili in konkretizirali šele skozi leksikografsko prakso, se pravi najprej v iterativnih ciklih samodejnega luščenja podatkov, nato pa v delovno intenzivnih postopkih urejanja, dopolnjevanja in obdelovanja slovarske vsebine.

Čeprav je vsak slovar že v izhodišču kompromis med idealnim in mogočim, pa si za novi sodobni slovar slovenščine želimo, da bi bilo glede specializirane leksike teh čim manj, kajti družba znanja, za katero si prizadevamo, se lahko uresničuje le skozi učinkovito ubesedovanje znanja, k temu pa lahko bistveno pripomorejo jezikovni priročniki.