Polona Gantar
Inštitut za slovenski jezik Frana Ramovša ZRC SAZU

Moč in nemoč korpusnega pristopa k analizi pomenov

V prispevku prikažemo različne možnosti slovarske metodologije, ki izhaja iz analize virov. Na konkretnih primerih utemeljimo, da je za izdelavo novega slovarja slovenskega jezika pomembno, da nastane na podlagi enotnega korpusnega pristopa s poudarkom na sprotnem posodabljanju referenčnega korpusa, izdelavi specializiranih korpusov in uporabi jezikovnih orodij, prilagojenih za slovenščino.

A corpus approach to the analysis of meaning: powerful and powerless

The article demonstrates different options within the lexicographic methodology that is based on the analysis of primary lexicographic resources. Demonstrating with concrete examples, we argue that it is important for the compilation of the new dictionary of Slovene to be based on a uniform corpus-based approach with the emphasis on the continuous updating of the reference corpus, building of specialized corpora and the use of language technology tools for Slovene.

Ključne besede: slovarski viri, korpusna metodologija, jezikovna orodja za slovenščino

Keywords: lexicographic resources, corpus methodology, language tools for Slovene

Uvod

Za jezikovno skupnost, ki ob koncu 20. stoletja, ko je korpusna leksikografija temeljna metodologija pri izdelavi slovarjev, nima na voljo na korpusu temelječega slovarja, je na nek način pričakovano, da ima do izključne uporabe take metode pomisleke.[1] Slovenci z edinim slovarjem, Slovarjem slovenskega knjižnega jezika (SSKJ), katerega koncept je bil izdelan v 60. letih, leksikalni opis pa zaključen ob koncu 80. let, živimo že vsaj 20 let predolgo. Kljub temu omenjeni slovar v slovenski skupnosti ne živi tako dolgo po naključju. Ob metodologiji, ki je iz računalniške perspektive in z vidika razvoja korpusov, kot smo mu priča danes,

[1] V zvezi s tem gl. prispevek J. Snoj na Strokovnem posvetu o novem slovarju slovenskega jezika in razpravo, ki je sledila (2009: 37−44).

preživeta, velja, da je v številnih svojih opisih uporaben še danes. Seveda pa hkrati tudi velja, da so se svet, dejanskost in z njo človek, ki v njej živi, spremenili tako v globalnem kot v kulturno specifičnem slovenskem prostoru. Zato je nov, na sodobnem jeziku temelječ opis slovenske leksike za slovenskega uporabnika nujno potreben, in to čim prej. Vprašanje, ki se pri tem zastavlja, je, ali lahko zanesljiv sodobni slovar temelji zgolj na referenčnem korpusu, ali pa je zavoljo zanesljivosti pri tem potrebno sistematično pregledovati tudi druge vire, se zanašati na jezikovno intuicijo in t. i. paberkovalno metodo? Preden skušamo odgovoriti na zastavljena vprašanja, si poglejmo nekatere vidike, ki so za nadaljnje razpravljanje ključni.

1 Tri slovarske resnice ali trije slovarski miti

1.1 Vsevedni slovar

Mnenje ali (laična) predstava o tem, da je slovar absolutna avtoriteta v smislu vključenosti vseh podatkov o določenem jeziku, je napačno, kot je napačno tudi prepričanje, da sta pomenska razčlenitev in pomenski opis besed v slovarju absolutna in da podajata nezmotljivo resnico o zunanjem svetu in predmetnosti, ki se odraža v besedah, zvezah in njihovih pomenih. Tak slovar ne obstaja, ne v knjižni, ne v elektronski in ne v spletni obliki. Dejstvo, pogosto znano le leksikografom, je, da ima posamezna beseda v vsakem posameznem kontekstu pravzaprav samosvoj, nekoliko drugačen pomen (Cruse 1986: 53). Povsem enako kot za pomenske opise tudi za večpomenske besede velja, da zanje ni mogoče določiti absolutnega števila pomenov in absolutno veljavnega razmerja med njimi. Pa vendar je legitimna pravica slovarskega uporabnika, da v slovarju ne samo pričakuje, ampak dejansko tudi dobi informacijo, ki jo potrebuje. Zato je smiselno, da slovar pove, kaj lahko uporabnik od njega pričakuje in česa ne, in da posledično na ta vprašanja odgovori že v slovarski zasnovi, ki določa izbor gradiva, način in podrobnost njegove obdelave ter količino in vrsto v slovar vključenih jezikovnih podatkov. Z vidika popolnosti konkretnega slovarskega priročnika je torej pri izdelavi slovarske baze kot na njej nastajajočega slovarja potrebno sprejeti odločitve, ki jih bo mogoče argumentirati in na podlagi argumentov prevzeti odgovornost za vse v slovarju navedene in nenavedene podatke.

1.2 Jezikovna resnica in slovarska verodostojnost

Podobno kot ne obstaja slovarska vsevednost, tudi ne obstaja absolutna jezikovna resnica. Resnica, ko govorimo o jeziku in slovarju, je lahko le jezikoslovna, utemeljena na jezikoslovnih argumentih, pri čemer težimo k temu, da so le-ti empirično podprti, in se hkrati zavedamo, da se jezikoslovne »resnice« spreminjajo tako z jezikovno realnostjo kot z razvojem jezikoslovja in novih teoretičnih in metodoloških pristopov. Podobno velja tudi za slovarje. Pomenski opisi v slovarjih nimajo veliko skupnega s tem, kako se govorci nekega jezika med seboj razumemo. Ljudje v konkretni sporočanjski situaciji ne razmišljamo o pomenih kot o številčno razvrščenih enotah v slovarjih, to počnejo leksikografi. Uslovarjanje jezika kot živega organizma je, če citiram M. Snoja (2012: 103), »nasilje, ki ga upravičuje samo uporabnost«. Če torej razumemo slovar kot umetno, znotraj leksikografskega koncepta dogovorjeno formo in vsebino, so slovarske informacije jezikovno resnične toliko, kolikor so dejansko uporabne in so jezikoslovno zanesljive, v kolikor dosledno sledijo dogovoru – tj. jezikoslovni teoriji in metodi, določeni v slovarski zasnovi.

1.3 Jezikoslovna znanost in teoretična podprtost slovarja

Teoretična podlaga je za slovarski koncept nujna, saj zagotavlja konsistentnost njegove izdelave, hkrati pa omogoča argumentacijo slovarskih rešitev in v slovarju navedenih podatkov. Slovarska teorija je pri tem seveda »dobra« samo v toliko, kolikor je izvedljiva. Tipično lahko tak problem opazujemo pri prenosu slovenske vezljivostne teorije (Žele 2001; 2003) na slovenski vezljivostni slovar (Žele 2008), kjer odsotnost kakršnegakoli praktičnouporabnega vidika kot tudi primerov realne jezikovne rabe povsem blokira prenos sicer teoretično podrobno razdelane vezljivostne teorije do uporabnika.[2] Prav tako kombinacija metodoloških postopkov in združevanje različnih virov za analizo, npr. korpusnega in listkovnega, ne more zagotoviti zanesljive in zadostno dokumentirane slovarske informacije. V potrditev tega je mogoče navesti analizo besedišča za Slovar novejšega besedja slovenskega jezika[3] (SNB; Bizjak Končar et al. 2012), ki je na eni strani

[2] Žal zapleten sistem označevanja stavčnih vzorcev tudi ni uporaben za strojno procesiranje podatkov.

[3] Pri izdelavi slovarja so se sestavljalci oprli na zbirko novejšega besedja, ki je nastala pri projektu Novejša slovenska leksika (v povezavi s spletnimi jezikovnimi viri), kjer je bil kot temeljni gradivni vir uporabljen besedilni korpus Nova beseda (http://bos.zrc-sazu.si/s_beseda.html), za katerega je stroka ugotovila, da ni zasnovan kot referenčni korpus niti ni oblikoslovno označen, zaradi česar ni primeren za vključitev v sodobna orodja za analizo korpusov (Logar Berginc 2009: 157), ter Listkovno kartoteko (po letu 1991), poleg tega so

pokazala odsotnost besed, ki v korpusu Gigafida izkazujejo več kot 1000 pojavitev: *band, biatlonski, bluetooth, beloruski, biatlonka, blaženje, bombon, beleženje, bluz, bovling, biotehničen, brisalnik, prekerni* oz. *prekarni, pogon*, na drugi strani pa vključitev besed, ki v Gigafidi pojavitev ne izkazujejo ali pa so zelo redke: *botrovanec, emocionalec, ožiliti se* idr.

2 Moč in nemoč korpusnega pristopa k analizi pomenov

2.1 V korpusu neizkazani realni pomeni

Korpusna analiza jezikovnega gradiva nujno naleti na problem, kako se odločati v situaciji, ko bodisi leksikografova bodisi uporabnikova jezikovna intuicija sugerira nek realni pomen besede, ki ga v korpusu ni mogoče najti niti s podrobno korpusno analizo. Ker je tako situacijo mogoče povsem legitimno pričakovati, je smiselno že v slovarskem konceptu in še prej pri zasnovi slovarske baze premisliti, kako se z njo spopasti.

Pri izdelavi Leksikalne baze za slovenščino (LBS)[4] smo na tak problem naleteli pri pomenski analizi zelo pogoste leme *šola*. Korpusna analiza, ki je temeljila na pregledu 300 konkordanc in na analizi besedne skice, je za obravnavani samostalnik izločila naslednje pomene in podpomene, ki jih za namene nadaljnjega razglabljanja primerjamo še s stanjem v SSKJ:[5]

LBS	**SSKJ**
šola *samostalnik*	**šola**[1] -e ž
1 o izobraževanju	1. vzgojno-izobraževalna ustanova ...
1.1 ustanova	// vzgojno-izobraževalni proces v tej ustanovi
1.2 stavba	// poslopje te ustanove
1.3 ljudje v stavbi	2. pouk
1.4 izobraževalni proces	3. privrženci, učenci ...
1.5 prostočasna izobraževalna dejavnost	
2 teoretična usmeritev; smer	**šola**[2] -e ž
	1. vneta bulica ...
	2. zadebelina sluznice ... pri konju

uporabili še spletni iskalnik Najdi.si (izjemoma tudi Google) ter priložnostne izpise (Gložančev et al. 2009: 17 in op. 28).

[4] Leksikalna baza za slovenščino je bila izdelana pri projektu Sporazumevanje v slovenskem jeziku, ki so ga financirali Ministrstvo za izobraževanje, znanost in šolstvo ter Evropski socialni skladi. V obliki XML je prosto dostopna na povezavi: http://www.slovenscina.eu/spletni-slovar/prenos.

[5] Ob strani puščamo dejstvo, da je samostalnik *šola* v SSKJ obravnavan kot homonim, v pomenski členitvi pa navajamo le glavne pomene in pomene za dvema poševnicama. Pri LBS navajamo samo pomenske indikatorje.

Splošno jezikovno vedenje in ne nazadnje tudi izpričan pomen v SSKJ (zgoraj podčrtano) pa sta izpostavila vprašanje, kaj narediti s pomenom 'izpuščaj na ustni sluznici; afta', ki ga podrobna korpusna analiza ne izkazuje, hkrati pa ga izkazuje npr. že osnovno iskanje v katerem od spletnih brskalnikov. Možne rešitve so v osnovi vsaj tri:[6]

1. Prva možnost predpostavlja, da je potrebno ob izdelani leksikalni analizi, ki temelji primarno na korpusnem gradivu, vedno (a) »izprašati« tudi lastno jezikovno intuicijo in (b) se o pomenski sliki obravnavane besede posvetovati s slovarji in drugim, npr. problemsko pridobljenim ali časovno zamejenim gradivom. Ta možnost predpostavlja kombinacijo različnih jezikovnih virov, posledično pa tudi različne metodologije gradivne analize, kar je bilo denimo izpeljano pri izdelavi SNB.

2. Druga možnost predvideva enotno metodološko zasnovo slovarja, ki jasno določa (in uporabniku sporoča), da semantični in drugi opisi leksikalnih enot izhajajo izključno iz korpusnega gradiva ter z uporabo točno določenih analitično-sintetičnih postopkov, ki jih omogoča trenutni razvoj korpusne metodologije in orodij za korpusno analizo. Če govorimo o izdelavi splošnega razlagalnega slovarja, je uporabljeni korpus lahko le referenčni, ki zajema besedila s stališča recepcije v jezikovni skupnosti: čim več govorcev in govork dejansko bere določena besedila, tem večji vpliv imajo ta na jezik in toliko bolj so zanimiva za leksikografsko obravnavo. Nasprotno pa specializirana besedila, namenjena in tvorjena s strani strokovnjakov določenega področja, ne morejo biti del referenčnega, pač pa le specializiranega korpusa, posledično pa tudi besedišča in specializiranih pomenskih rab ni mogoče pričakovati v splošnem slovarju. Glede na trenutno stanje korpusnih virov za slovenščino je za detajlno analizo sodobnega splošnega besedišča najbolj primeren korpus Gigafida, skupaj s podkorpusom Kres, zato smo se za tak pristop odločili tudi pri izdelavi LBS.

3. Tretja možnost prav tako predvideva uporabo korpusnega pristopa, pri čemer so glede na namen slovarja, potrebe jezikovne skupnosti in ciljne uporabnike kot gradivni vir poleg referenčnega upoštevani še drugi, zlasti specializirani korpusi, korpus govorjenega jezika in korpusi specializiranih besedil in jezikovnih interakcij (npr. korpus slovenskih tvitov, spletni

[6] Kilgarriff (2013: 77) izpostavlja naslednje prakse: kopiranje iz drugih virov, jezikovno intuicijo in empirično analizo podatkov.

korpus itd.). Tak pristop temelji na enotni korpusni metodologiji, hkrati pa predvideva za posamezne tipe besed (zlasti za zelo pogoste in pomensko razvejane) tudi posebne leksikografske postopke, ki bodo zagotavljali identifikacijo tako jedrnega kot obrobnega besedišča in jezikovnih pojavov.

Z vidika trenutnega stanja razvoja korpusne leksikografije je najbolj optimalen tretji pristop, zato je nujno, da vsako izdelavo slovarja spremljata tudi načrtna nadgradnja referenčnega korpusa in izdelava različnih specializiranih korpusov za področja, ki imajo znotraj jezikovne skupnosti izkazano pomembno vlogo. Čeprav je Gigafida glede na obseg in sestavo drugim evropskim jezikom primerljiv referenčni korpus, je jasno, da temelji prihodnost leksikografije na resnično obsežnih količinah podatkov (Pomikalek et al. 2009; Jakubiček et al. 2013) in orodjih, ki bodo te podatke zmogla analizirati na način, ki bo zadostil slovarskim potrebam različnih uporabnikov. Le želimo si lahko, da bodo šle prihodnje nadgradnje korpusa Gigafida v to smer in da bodo hkrati grajeni različni specializirani korpusi, do takrat pa je smiselno razmisliti, katero od preostalih zgoraj omenjenih izhodišč slovarskega opisa je v danem trenutku in glede na slovensko situacijo smiselno uporabiti pri izdelavi sodobnega slovarja.

Prednosti in slabosti vsebujeta namreč obe sicer povsem legitimni leksikografski odločitvi. Prednosti prve (in tretje) možnosti so v tem, da na tej podlagi izdelan slovar uporabniku zagotovi informacije tudi o jezikovnih pojavih, ki ne predstavljajo del splošnega besedišča oz. so vezane na specializirane govorne položaje ali besedila.[7] Tovrstno besedišče v splošnem jeziku ni zelo pogosto, vendar pa hkrati to še ne pomeni, da dejansko ne obstaja. Nasprotno, z vidika slovarja, namenjenega maternemu govorcu, je upravičeno pričakovati, da najbolj frekventno besedišče in očitni pomeni za splošnega uporabnika ne bodo zanimivi, zanimivi pa bodo tisti, ki so relativno obrobni oz. rabljeni v bolj specializiranih sporazumevalnih situacijah. Predvidevati je torej mogoče, da nas bodo taki pomeni oz. njihova razlaga zanimali prav zato, ker niso splošno znani. Eden od razlogov, da se določeno besedišče in pomenske rabe v referenčnem korpusu ne izkazujejo, je namreč ta, da so vezani na strokovno ali manj razširjeno tematsko področje, za katero v referenčnem korpusu ni besedil.

[7] V zvezi s prednostmi uporabe prve možnosti je glede na to, da analiza korpusov ni sistematična, treba opozoriti, da je evidentirana specifična jezikovna raba lahko povsem naključna in za razliko od tretje možnosti ne zagotavlja nujno evidentiranja vseh za uporabnika zanimivih specifičnih jezikovnih rab.

Na drugi strani ima tak pristop svoje slabosti in pasti. Ena od slabosti je, da je postopek leksikografske analize, ki poleg korpusne metode predvideva za ustrezno dokumentacijo takega pomena veliko dodatnega iskanja po različnih leksikalnih virih, zelo zamuden, to pa podaljša in podraži izdelavo slovarja, še veliko pomembneje pa je, da rezultat ni vedno optimalen, posledično pa tudi leksikografski opis ni konsistenten in v celotnem slovarju enotno izpeljan (prim. Gorjanc 2009: 43). Iz izkušenj pri izdelavi LBS je mogoče povedati, da v zvezi s pomeni, za katere v korpusu nismo našli potrditev, običajno tudi ni bilo mogoče na podlagi drugih virov navesti zanesljivih podatkov o njihovem tipičnem besedilnem okolju, vezljivostnih in drugih pomenskoskladenjskih podatkov, podatkov o slovničnih omejitvah, stilističnih in pragmatičnih posebnostih ter ne nazadnje tudi ustreznih zgledov rabe, kar vse velja tudi za zgoraj izpuščeni pomen samostalnika *šola*. Ker je v takih primerih dejansko na voljo premalo podatkov oz. se v zvezi z navedenimi slovarskimi informacijami na intuicijo[8] ni mogoče zanašati, se postavlja tudi vprašanje, ali je registracija takega pomena z vidika verodostojnosti dejansko upravičena in za uporabnika v resnici nezavajajoča.

Poleg izpostavljene slabosti prinaša omenjeno izhodišče leksikografom še dve zelo znani pasti. Namreč, v trenutku, ko ima leksikograf navodilo, da mora preveriti pomensko zgradbo besede, ki jo analizira, še v drugih virih in slovarjih, se njegova analiza nehote podredi že izdelani pomenski sliki, še zlasti, če je pomene med seboj težko ostro ločevati, kar se v večini primerov dejansko dogaja (Atkins in Rundell 2008: 264–275). Enako je v primerih, v katerih npr. sodobno nasproti starejšemu gradivu izkazuje subtilne, pogostokrat kulturno in družbenopolitično pogojene pomenske spremembe ter spremembe, vezane na tehnološki napredek, družbeno percepcijo ipd. Povedano je mogoče ponazoriti z obravnavo samostalnika *gospodinjec* v SSKJ in korpusu Gigafida.

V SSKJ ima samostalnik *gospodinjec* z oznako »šaljivo« en pomen: »moški, ki opravlja ali vodi domača, hišna dela: biti gospodinjec; dober gospodinjec«. Korpusna analiza ponudi zglede, ki jih glede na pomen lahko razvrstimo v tri skupine:

[8] Hanks (2013: 20) poudarja pomembnost razlikovanja med uporabo jezikovne intuicije za interpretiranje podatkov, ki je za jezikoslovno in leksikografsko delo nujna, in uporabo jezikovne intuicije, ki je namenjena ustvarjanju podatkov. V zvezi z zadnjim pravi: » No reputable scientist (outside linguistics) invents data in order to explain it.« (Noben ugleden znanstvenik (razen v lingvistiki) ne izumlja podatkov z namenom, da bi jih pojasnil. Prevedla P. G.)

1. skupina

- *Verjetno ni gospodinje ali **gospodinjca**, ki se ne bi vsak dan spraševala, kaj dati v lonec.*
- *Tudi tržnice so polne buč, peki in gospodinje ali **gospodinjci** pa pospešeno pečejo bučne pite.*
- *Vsi, ki resno nakupujejo, predvsem pa gospodinje in **gospodinjci** vedo, da je nakupovanje trdo in utrujajoče delo.*

2. skupina

- *Pri zadnjem popisu prebivalstva se je 306 slovenskih moških izreklo za **gospodinjce**.*
- *Barbara Dalton je učiteljica, Bob Dalton pa **gospodinjec**,imata tri otroke in živita v južnem Londonu.*
- *Ženske so še vedno v slabšem položaju na trgu delovne sile, tako da se lahko le malo družin zanese na plačo ženske, moški pa privzame vlogo "**gospodinjca**".*
- *Olga postane uspešna poslovna ženska, Rajko pa še uspešnejši **gospodinjec**.*
- *Opazna izkušnja **gospodinjcev** je izolacija in neizpolnjenost pri gospodinjskem delu, ki ga dojemajo kot monotonega.*
- *John je postal prvi **gospodinjec** v zgodovini popa, skrbel je za sina Seana, ona je skrbela za posle in v petih letih njuno premoženje početverila.*
- *Sicer pa – čeprav je beseda **gospodinjec**, ki sem jo prvi javno uporabljal, zdaj povsem udomačena, je še danes ni v slovarju slovenskega knjižnega jezika. Morda pri urejanju tega slovarja ženske nimajo kaj dosti besede.*
- *Zakaj se vsi (tako M kot Ž) posmehljivo nasmehnemo, če slišimo, da se je kak mož odločil biti **gospodinjec**? Razumem, da se temu rogajo moški, a temu se rogajo tudi ženske!*

3. skupina

- **(a)** *Svoje predloge predmetnika za OŠ so na včerajšnji seji predstavili še učitelji, da bi ugodili zahtevam tehnikov, likovnikov in **gospodinjcev** za povečanje ur njihovim predmetom.*
- **(b)** *Po izračunu bi stroški za eno šolsko leto v POŠ za učitelja in pol delovnega mesta **gospodinjca**, brez materialnih stroškov za delovanje šole, znašali malo več kot 40 tisoč evrov.*

V prvi skupini je mogoče prepoznati rabo samostalnika v pomenu, kot ga predvideva SSKJ, pri čemer se oznaka »šaljivo« v sodobni rabi ne potrjuje. Kar se tiče zaznamovanosti, kot jo izkazuje sodobno gradivo, je mogoče opaziti zapis v narekovajih (2x), kar sugerira distanco do rabe moške oblike v primerjavi z žensko in slabšalno konotacijo v zadnjem zgledu v drugi

skupini, vendar pa gre v vseh treh primerih za pomen, ki ga SSKJ ne registrira: 'moški, ki skrbi za dom in otroke in ne hodi v službo',[9] hkrati pa je tudi očitno, da je ta pomen glede na prvega prevladujoč.

Z vidika slovarske »vsevednosti« in hkrati verodostojnosti leksikografske informacije sta zanimiva primera, ki smo ju uvrstili v tretjo skupino. Z oznakama (a) in (b) pokažemo, da gre za dva različna in hkrati osamljena primera, na podlagi katerih bi težko sklepali na pomen, ki je ustaljen v širši jezikovni skupnosti.

Prav na točki omenjenih primerov je konceptualna zasnova slovarja primorana sprejeti odločitev, ali vključiti posebnosti, ki utegnejo biti za uporabnika zanimive, ali se odločiti za zanesljivo dokumentirano informacijo, ki zadošča vsaj trem temeljnim merilom, po katerih se za ustaljeni del besedišča nekega jezika štejejo besede in pomeni, ki izkazujejo relativno pogosto rabo skozi daljše obdobje in so izpričani v različnih besedilih. V okviru zgoraj navedenih možnosti je seveda optimalna rešitev tista, ki zagotavlja dovolj relevantnih podatkov v korpusu, kar pomeni razširitev v smeri specializiranih besedil, sicer pa je smiselno posebnosti, kot jih zgoraj nakazujeta primera (a) in (b), uvrstiti v slovarsko bazo in s tem omogočiti tudi prehod v slovar v skladu z njegovim namenom in ciljnim uporabnikom.

Druga past, ki izhaja iz pristopa, ki ga omenja prva možnost, je za konsistentnost leksikografskih postopkov in verodostojnost slovarske informacije še bolj problematična. Če se namreč na eni strani zanašamo na vedenje o obstoju nekega v korpusu neizkazanega pomena, smo le korak stran od preverjanja korpusnih realizacij potencialnih jezikovnosistemskih možnosti, ki jih lahko na pomenskoskladenjski ravni izražajo leksikalne enote, za katere v prvi fazi korpusne analize nismo našli potrditev. Govorimo zlasti o manjšalnicah (npr. problem *marelica* v pomenu 'majhna marela'), o obstoju posamostaljenih pridevnikov, ki so jezikovnosistemsko možni tako rekoč za vsak pridevnik, v realni jezikovni rabi pa se realizirajo (ali pa sploh ne) zelo nepredvidljivo. Sem sodi tudi problem deležniških pomenov, ki se v zvezi z nekaterimi pridevniki dejansko ne realizirajo oz. se realizirajo skrajno redko (prim. *noseč* in *obdarjen* glede na stanje v korpusu). Ker gre v vseh primerih vendarle za realne jezikovnosistemske možnosti, obstaja velika verjetnost, da bomo s podrobnim in usmerjenim poizvedovanjem v korpusu ali drugih virih našli tudi konkretne

[9] SNB omenjenih pomenov ne registrira.

pojavitve. V zvezi s tem pa se moramo vprašati, ali vključevanje tovrstnih informacij v slovar dejansko pripomore k njegovi zanesljivosti. Ali ni ne nazadnje prikazovanje obrobne jezikovnosistemske možnosti v slovarju dejansko zavajanje glede njenega realnega statusa v slovenskem besedišču? Spomnimo se denimo samo na argumente za vključitev glagola *prostovoljiti* v SNB.[10]

Pristop, kot ga omenja 2. možnost, zagotavlja konsistentnost slovarskega opisa, saj se osredotoča samo na v zadostni meri v realni jezikovni rabi dokumentirane jezikovne podatke, kar mu v osnovi zagotavlja tudi ponovljivost in sledljivost postopka. Z jasno zasnovo, ki leksikografa seznanja s tem, kateri viri se upoštevajo in na kakšen način se – čim bolj v skladu s trenutnimi možnostmi sodobne leksikografske teorije in prakse – analizirajo, lahko uporabnik pričakuje tako v zvezi s pomenskimi opisi kot tudi v zvezi z izkazanimi ali neizkazanimi pomeni, argumente, utemeljene na dejanski rabi in konceptualnih rešitvah, kot so npr. prag pogostosti za vstop v bazo oz. slovar, ustrezna besedilna razpršenost, zastopanost različnih avtorjev ipd.

2.2 V korpusu izkazani realni pomeni, ki jih korpusna metodologija ne pokaže

Najzgodnejši pristopi korpusne metodologije pri analizi besedišča so temeljili na analizi konkordančnega niza, tj. številnih pojavitev obravnavane besede v njenem neposrednem besedilnem okolju (Church in Hanks 1990). Vendar pa so leksikografi kmalu ugotovili, da tak postopek, ob tem da ne daje končnih odgovorov o vsestranskem obnašanju in rabi preučevane besede in da je ob preučitvi konkordanc za to potrebno še dodatno raziskovalno delo, zahteva tudi veliko dodatnega časa. Eden izmed prvih učinkovitih odgovorov na ta problem so Besedne skice (Kilgarriff et al. 2004). Ker predstavljajo povzetek gramatičnega in kolokacijskega obnašanja obravnavane besede, lahko ponudijo tudi hiter in zanesljiv odgovor o njeni pomenski razvejanosti in skladenjskih lastnostih. Široka uporaba besednih skic v leksikografiji je vplivala tudi na njihov nadaljnji razvoj, kjer je postalo jasno, da sta za optimalno delovanje potrebna (a) zelo velik in dobro oblikoslovno označen korpus ter (b) detajlna slovnica besednih skic, ki mora biti jezikovno specifična, tj. izdelana z upoštevanjem jezikovnosistemskih lastnosti konkretnega jezika. Z analizo besednih skic je leksikografska metoda od stopnje, na kateri je računalniška tehnologija zagotavljala le osnovno podporo v

[10] O tem npr. na spletni strani tednika Družina: http://www.druzina.si/icd/spletnastran.nsf/all/A1C90F85029F7C8AC1257ACC004D8EF2?OpenDocument.

analitičnem leksikografskem procesu, napredovala v identifikacijo leksikalnih zakonitosti (kolokacij, stavčnih vzorcev), ki so usmerile leksikografovo pozornost na ključne elemente besednega obnašanja v sobesedilu. Če je torej za leksikografe večji korpus pomenil težje obvladovanje jezikovnih podatkov, je z jezikovnotehnološkega vidika postalo jasno, da večji kot je korpus, bolj natančne so lahko leksikalne analize. Ali z drugimi besedami: več podatkov kot je na voljo, večja je verjetnost, da bomo v korpusu našli tako frekvenčno izstopajoče kot tudi obrobne jezikovne pojave (Kilgarriff 2013: 84). Prav to izhodišče pa je tudi rešitev za zgoraj omenjeni problem.

V orodju Sketch Engine so bila na podlagi teh ugotovitev in na podlagi vse zahtevnejših potreb leksikografov razvita orodja, ki omogočajo podrobnejšo in bolj usmerjeno leksikalno analizo. Za ponazoritev konkretnega problema se vrnimo k samostalniku *šola*. Ker gre za zelo pogosto lemo, je mogoče pričakovati, da nekateri relativno pogosti pomeni ali skladenjski vzorci pri osnovni analizi ne bodo prišli do izraza. Pri oblikovanju leksikalne baze smo zato v primeru nadpovprečno pogostih lem uporabili dodatne metode. Osnovna analiza besedne skice, ki temelji na prevzetih nastavitvah minimalne frekvence, minimalne statistične izpostavljenosti, največjega števila besed v slovnični relaciji in števila zgledov na kolokator, je za lemo *šola* ponudila seznam 25 kolokatorjev v slovnični relaciji S_kakšen? (tj. za strukturo pridevnik + samostalnik), pri čemer ima kolokator na zadnjem 25. mestu (razvrščenem glede na število pojavitev) v korpusu 2196 pojavitev. To pomeni, da v prvo fazo analize niso bili vključeni vsi tisti kolokatorji, ki predvidevajo frekvenco nižjo od 2.000 pojavitev. S tem pa obstaja tudi velika verjetnost, da bomo spregledali posamezne kolokatorje, ki bodisi sugerirajo samostojne pomene, stalne besedne zveze oz. predvidevajo določene skladenjske vzorce, ki so z vidika slovarske informacije zanimivi. V drugi fazi smo zato analizirali tudi manj pogoste kolokatorje in uporabili še druge funkcije, ki jih predvideva orodje Sketch Engine (primerjalne skice, izdelavo skic za večbesedne enote ipd.). Na ta način smo izločili vzorce: *kaj je dobra šola za koga/kaj*, *kaj je prava šola za koga/kaj* in *kaj je prava šola česa*.[11] Na podlagi analize besedilnega okolja izluščenih vzorcev smo kot relevantno (pomensko in skladenjsko ustaljeno) leksikalno enoto prepoznali frazeološko enoto: *dobra/prava šola za koga/kaj* s pomenom 'poučna izkušnja, priprava na kaj', ki se potrjuje tudi s kolokacijami in korpusnimi zgledi:

[11] Na ta način smo evidentirali tudi številne stalne zveze, npr. *šola za otroke s posebnimi potrebami, devetletna šola, zdrava šola, večerna šola, laična šola, šola za starše, materinska šola, vesela šola* itd., ki v dosedanjih slovarskih priročnikih (niti v najnovejšem SNB) niso evidentirane.

dobra šola za [drugič, naprej, vnaprej, prihodnost]
dobra šola za v [prihodnje, bodoče]
dobra šola za [življenje]

- *Sicer pa ... če je RES kriv, bo pa **dobra šola** za drugič, al pa tud ne?*
- *Srečanje je po njegovem tudi **dobra šola** za člane orkestrov, saj se ob poslušanju drug drugega lahko veliko naučijo.*
- *Je pa to **dobra šola** za bodočnost; namesto da se vsega lotite sami, boste začeli ceniti skupinsko delo.*
- *Na Švedskem sploh ne bi smel nastopiti, saj se mi je vnetje ušesa le poslabšalo. Toda to bo **dobra šola** za v bodoče.*

3 Zaključek

Na podlagi povedanega je mogoče zaključiti, da korpusni pristop v sodobni leksikografiji ni in ne more biti vprašanje (prim. J. Snoj 2009: 43). Z njim sta slovarju zagotovljeni konsistentnost in dokumentirana zanesljivost vseh v slovarju navedenih informacij. Združevanje različnih metodoloških postopkov, ki jih narekujejo različne oblike gradivnih virov, pomeni ne le upočasnitev slovarskega procesa, ampak predstavlja past za predimenzioniranje pomembnosti za jezik obrobne jezikovne informacije kot tudi nezmožnost njenega vsestranskega opisa. Na drugi strani konsistentnost leksikografske metode ne pomirja samo leksikografov, češ izključno korpusna metoda prihrani dodatno delo in mukotrpno stikanje za podatki, pač pa vendarle tudi uporabnike, saj se je pokazalo, da metodološko različni pristopi, kot je kombiniranje procesljivih in neprocesljivih jezikovnih virov, odpovejo tudi pri identifikaciji jedrnega besedišča in z njim povezanih pomenov, o čemer zgovorno priča odsotnost številnih stalnih besednih zvez pri samostalniku *šola* v obstoječih slovarskih priročnikih, zlasti v SNB. Pri problemu zadovoljevanja uporabnikov pa je še najpomembnejše to, da se nam v resnici ni treba odločati med dvema skrajnostma, ali bomo zadovoljili povpraševalce po posebnem, obrobnem ali povpraševalce po jedrnem in osrednjem. Jezikovne pojave, ki se v korpusu bodisi ne kažejo bodisi so relativno slabo zastopani, je vedno mogoče identificirati s podrobnimi korpusnimi analizami. To velja zlasti za zelo pogoste leme, z usmerjenimi korpusnimi poizvedovanji, z uporabo specializiranih orodij za analizo ter seveda tudi vključevanjem analize specializiranih, govornih in drugih podkorpusov. Tako korpuse kot tudi korpusna orodja je namreč vedno mogoče in potrebno

izboljševati, zato je za detajlnost slovarske informacije kot tudi za njeno zanesljivost in verodostojnost smiselno ob vsakem leksikografskem projektu nadgrajevati tudi korpus, slediti razvoju jezikovnih tehnologij ter stremeti k njihovi prilagoditvi za specifike slovenskega jezika.

4 Literatura

- ATKINS, Sue, B. T., RUNDELL, Michael, 2008: *The Oxford Guide to Practical Lexicography.* Oxford: Oxford University Press.
- BIZJAK KONČAR, Aleksandra, SNOJ, Marko (ur.) 2013: *Slovar novejšega besedja slovenskega jezika*. Ljubljana: Založba ZRC, ZRC SAZU.
- CHURCH, Kenneth, HANKS, Patrick, 1990: Word associations norms, mutual information and lexicography. *Proceedings of the 27th Annual Conference of the Association of Computational Linguistics.* Str. 76–82.
- CRUSE, Alan D., 1986: *Lexical Semantics.* Cambridge textbooks in linguistics. Cambridge: Cambridge University Press.
- GLOŽANČEV, Alenka, JAKOPIN, Primož, MIHELIZZA, Mija, URŠIČ, Lučka, in ŽELE, Andreja, 2009: *Novejša slovenska leksika (v povezavi s spletnimi jezikovnimi viri).* Inštitut za slovenski jezik Frana Ramovša ZRC SAZU. Ljubljana: Založba ZRC, ZRC SAZU.
- GORJANC, Vojko, 2009: Jezikovnotehnološka podpora slovarskemu delu. *Strokovni posvet o novem slovarju slovenskega jezika, 23. in 24. oktober 2008.* Ur. Andrej Perdih. Ljubljana: Založba ZRC SAZU, ZRC SAZU. Str. 45–52.
- HANKS, Patrick, 2013: *Lexical Analysis: Norms and Exploitations.* Cambridge: MIT Press.
- JAKUBÍČEK, Miloš, KILGARRIFF, Adam, KOVÁŘ, Vojtěch, RYCHLÝ, Pavel, SUCHOMEL, Vít, 2013: The TenTen Corpus Family. 7th International Corpus Linguistics Conference CL 2013. Lancaster. Str. 125–127.
- KILGARRIFF, Adam, 2013: Using corpora [and the web] as data sources for dictionaries. V: Howard Jackson (ur.) *The Bloomsbury Companion to Lexicography*. Bloomsbury, London. Str. 77–96.
- KILGARRIFF, Adam, RYCHLÝ, Pavel, SMRZ, Pavel, TUGWELL, David, 2004: The Sketch Engine. *Proceedings / XI Euralex International Congress, Lorient*. Université de Bretagne-Sud. Str. 105–116.
- LOGAR BERGINC, Nataša, 2009: O dveh znanstvenomonografskih leksikalnih seznamih. Jezik in slovstvo, 54 (2009), št. 3–4. Str. 153–159.

- POMIKALEK, Jan, RYCHLY, Pavel, in KILGARRIFF, Adam, 2009: Scaling to Billion-plus Word Corpora. *Advances in Computational Linguistics*. Special Issue of Research in Computing Science. Vol 41. Mexico City.
- SNB: *Slovar novejšega besedja slovenskega jezika*. Ur. Aleksandra Bizjak Končar in Marko Snoj. Ljubljana: Založba ZRC, ZRC SAZU.
- SNOJ, Jerica, 2009: Novi slovar: spodbuda za posodobitev slovenskega slovaropisja. *Strokovni posvet o novem slovarju slovenskega jezika, 23. in 24. oktober 2008.* Ur. Andrej Perdih. Ljubljana: Založba ZRC SAZU, ZRC SAZU.
- SNOJ, Marko, 2012: Podgesla v novem slovarju slovenskega jezika. V: Franc Marušič in Rok Žaucer (ur.) *Škrabčevi dnevi 7 – zbornik prispevkov s simpozija 2011*. Nova Gorica: Založba Univerze v Novi Gorici. Str. 96–103.
- SSKJ: *Slovar slovenskega knjižnega jezika na CD-romu z Odzadnjim slovarjem slovenskega jezika in Besediščem slovenskega jezika z oblikoslovnimi podatki*. Ljubljana: SAZU, ZRC SAZU, Inštitut za slovenski jezik, DZS.
- ŽELE, Andreja, 2001: *Vezljivost v slovenskem jeziku (s poudarkom na glagolu).* Ljubljana: Založba ZRC, ZRC SAZU.
- ŽELE, Andreja, 2003: *Glagolska vezljivost: iz teorije v slovar*. Ljubljana: Založba ZRC, ZRC SAZU.
- ŽELE, Andreja, 2008: *Vezljivostni slovar slovenskih glagolov*. Zbirka Slovarji. Ljubljana: Založba ZRC, ZRC SAZU.