Kolokacije v Slovarju sopomenk sodobne slovenščine: evalvacija podatkov in predlog za izboljšavo

Špela ARHAR HOLDT
Filozofska fakulteta; Fakulteta za informatiko in računalništvo,
Univerza v Ljubljani

The Thesaurus of Modern Slovene includes collocational information, which the users can use to compare the contextual behaviour of two synonyms. Collocations have been automatically extracted from the reference corpus Gigafida, and are grouped by syntactic structures. In this paper, we present a qualitative linguistic analysis of collocates for 48 synonym pairs for 24 headwords: eight nouns, eight adjectives, and eight verbs. In the analysis, we separate incorrectly identified, incomplete and semantically irrelevant data from syntactically complete and for synonym comparison relevant collocations. We describe the issues that can be addressed with adapting the data extraction method, and define aspects for further improvement. Then, using the corpus Gigafida 2.0 and the Sketch Diff functionality in the Sketch Engine tool, we analyse each pair of synonyms and their syntactic structures, identifying the most relevant structure for synonym comparison in the Thesaurus. We conclude the paper by outlining the future developments of the Thesaurus such as the improvement of the data extraction method, and the selection and presentation of collocations, including the improved list of syntactic structures.

Keywords: Thesaurus of Modern Slovene, synonymy, collocations, automatic data extraction, linguistic evaluation

1 Uvod

Ustrezno strukturiran in formaliziran opis jezika v zadnjih desetletjih postaja vedno pomembnejši ne le za človeškega uporabnika, ampak tudi za potrebe strojne obdelave naravnega jezika in razvoja jezikovnih tehnologij. V kontekstu priprave temeljne digitalne jezikovne infrastrukture so kolokacijski podatki ključni tako za obravnavo v samostojnih specializiranih jezikovnih virih kot tudi za podporo pri opisu drugih jezikovnih pojavov, od katerih nas v pričujočem prispevku zanima sopomenskost.

Slovar sopomenk sodobne slovenščine je odprto dostopna zbirka slovenskih sopomenk, ki v različici 1.0 prinaša 105.473 iztočnic in 368.117 sopomenk.[1] Slovar je bil po modelu odzivnega slovarja pripravljen s strojnimi postopki, njegov nadaljnji razvoj pa poteka odprto in v sodelovanju s širšo jezikovno skupnostjo (Arhar Holdt idr. 2018). Sopomensko gradivo, ki ga prinaša slovar, je bilo pridobljeno iz razpoložljivih jezikovnih virov, med katerimi sta glavna Veliki angleško-slovenski slovar Oxford–DZS (Šorli idr. 2006) in referenčni korpus pisne slovenščine Gigafida (Logar idr. 2012). Gradivo je avtomatsko urejeno glede na moč pomenske povezanosti sopomenke z iztočnico ter glede na pomensko podobnost sopomenk (Krek idr. 2017). Slovaropisno urejanje gradiva, vključno s pomenskim členjenjem oz. opisom ter kvalificiranjem, je prepuščeno kasnejšim korakom slovarske gradnje, medtem ko je že v prvi različici slovarja omogočen vpogled v kontekst jezikovne rabe, in sicer s povezavo sopomenskega gradiva s kolokacijskimi podatki in zgledi iz referenčnega korpusa Gigafida.[2]

Uporabniške evalvacije Slovarja sopomenk so pokazale, da je vključitev kolokacijskih podatkov v splošnem všeč 68 % sodelujočih (N=671), ostalim pa je bilo za vključitev vseeno ali se do nje niso znali

1 Slovar je dostopen na spletni strani https://viri.cjvt.si/sopomenke/slv/, kot podatkovna baza pa je skupnosti na voljo na repozitoriju CLARIN.SI (Krek idr. 2018: http://hdl.handle.net/11356/1166).

2 Upoštevati je treba, da je Slovar sopomenk sodobne slovenščine prvi slovenski jezikovni vir, ki vključuje avtomatsko pridobljene kolokacije. Kolokacijski slovar sodobne slovenščine je izšel šele kasneje, prinaša pa številne nove premisleke na ravni vmesniških elementov, izbire in postavitve kolokacijskega gradiva (Kosem idr. 2018).

opredeliti. Pokazalo se je tudi, da kolokacijski podatki od vseh novosti, ki jih slovar uvaja, uporabnike najbolj prepričajo: sodelujoči, ki so slovar že uporabljali, so se do vključitve opredeljevali signifikantno bolj pozitivno kot tisti, ki slovarja niso dobro poznali (Arhar Holdt 2020). Pričujoči prispevek želi splošni uporabniški oceni dodati natančnejšo jezikoslovno evalvacijo uporabljenih rešitev, predvsem na ravni izbire kolokacijskih besednozveznih vzorcev in drugih parametrov luščenja ter predstavitve kolokatorjev v slovarskem vmesniku. S kvalitativno analizo kolokatorjev za 48 sopomenskih parov želimo preveriti predpostavko, da lahko pri predstavitvi sopomenskosti kolokacijski podatki do določene mere nadomestijo pomenski in stilni opis, ter podati predlog za podatkovno izboljšavo oz. nadgradnjo.

2 Kolokacije v Slovarju sopomenk sodobne slovenščine

Slovar sopomenk sodobne slovenščine uvrščamo med odzivne slovarje. To poimenovanje izvira iz dejstva, da se slovar neprestano razvija in dinamično odziva tako na spremembe v jeziku kot na mnenja in potrebe uporabniške skupnosti. Glavne značilnosti odzivnega slovarja so, da je predviden za uporabo v digitalni obliki in zanjo tudi ciljno razvit; slovarska baza nastaja z uporabo naprednih računalniških metod; rezultati so skupnosti odprto na voljo takoj, ko so relevantni za nadaljnji razvoj, četudi so še ročno neprečiščeni; slovarski podatki se razvijajo, kar zahteva transparentno sledenje spremembam in dostop do arhiviranih različic baze; in pri razvoju slovarja lahko sodeluje širša jezikovna skupnost (Arhar Holdt idr. 2018).

V slovarju so dostopni tudi kolokacijski podatki, ki so vedno primerjalne narave: na enotnem ekranu so urejeni kolokatorji dveh sopomenskih besed, kar naj bi uporabniku omogočilo, da vidi podobnosti in razlike v njuni rabi. Kolokatorji so razvrščeni v tri razdelke glede na to, ali se pojavljajo (a) z obema primerjanima besedama, (b) samo s prvo ali (c) samo z drugo od besed.[3] Pri primerjavi pridev-

3 V prispevku nekoliko posplošeno podatke pod točko (a) imenujemo 'prekrivni' in pod (b) ter (c) 'samostojni'.

nikov *ekološki* in *zelen* so v prvem razdelku denimo navedeni primeri *zelena / ekološka [luč, barva, solata],* v drugem *ekološka [kmetija, sanacija, katastrofa]* (ne pa tudi **zelena [kmetija, sanacija, katastrofa]*) in v tretjem *zelen list, zelena zima, zelene oči* (ne pa tudi **ekološki list, *ekološka zima, *ekološke oči*).[4] Velik del kolokacij je na klik povezan z zgledi iz referenčnega pisnega korpusa Gigafida, ki so bili izvoženi s pomočjo parametrov GDEX za slovenščino (Kosem idr. 2011). Kadar izvoženi zgledi niso na voljo, slovar ponuja povezavo do konkordančnega niza neposredno v korpusnem vmesniku.

Kolokacijski podatki so urejeni v stolpce glede na to, v katerem besednozveznem vzorcu se pojavljajo, pri čemer so pri različnih besednih vrstah uporabljeni različni vzorci, kot prikazuje Tabela 1. Nekateri vzorci obsegajo po dva stolpca. V vsakem stolpcu je prostor za 15 kolokatorjev, skupno torej prikaz lahko obsega (do) 60 kolokatorjev. V Tabeli 1 za vsak vzorec navajamo po en ponazoritveni primer.

Kolokacijski podatki so iz referenčnega pisnega korpusa pridobljeni avtomatsko in ustrezajo statističnim in skladenjskim kriterijem kolokacijskosti (Kosem idr. 2020), niso pa (še) urejeni na ravni pomena. Statistične in skladenjske kriterije določajo parametri slovaropisnega orodja Sketch Engine (Kilgarriff idr. 2014), v katerem je na voljo funkcionalnost Primerjalne skice, ki primerja kolokacijsko in koligacijsko okolico dveh (uporabniško opredeljenih) besed enake besedne vrste. V primerjalni postavitvi so statistično razvrščeni kolokatorji prikazani glede na pogostost sopojavljanja z eno ali drugo od izbranih besed, pri čemer so podatki urejeni po besednozveznih oz. relacijskih vzorcih, ki jih za slovenščino opredeljujejo Besedne skice (Krek idr. 2016). V poenostavljeni obliki je primerjalni pogled ohranjen tudi v Slovarju sopomenk, za katerega pa so uporabljeni le izbrani, nekoliko posplošeni vzorci. Če glede na parametre luščenja kolokacijski podatki v korpusu niso na voljo, se v slovarju v naboru kolokatorjev pojavljajo prazna mesta.

4 V slovarju so izpisani samo kolokatorji, običajno v lematizirani obliki (načinu izpisa pri posameznem kolokacijskem besednozveznem vzorcu se posvečamo v nadaljevanju). V besedilu prispevka za lažje branje kolokacije navajamo v besednozvezni in kategorialno ustrezno prilagojeni obliki.

Tabela 1: Besednozvezni vzorci, zajeti v primerjalni prikaz kolokacij v Slovarju sopomenk.

Bes. vrsta	Stolpec 1	Stolpec 2	Stolpec 3	Stolpec 4
Samostalnik (*jezik*)	pridevnik kot ujemalni levi prilastek (*slovenski jezik*)	predložna zveza s samostalnikom v različnih sklonih kot desni prilastek (*jezik na Slovenskem*)	glagol s predlogom, ki zahteva različne sklone samostalnika (*prevesti v jezik*)	glagol, ki mu sledi samostalnik v tožilniku (*uporabljati jezik*)
Pridevnik (*odrezan*)	pridevnik kot ujemalni levi prilastek sledečega samostalnika (*odrezano steblo*)		predložna zveza s samostalnikom v različnih sklonih kot desni prilastek (*odrezan od sveta*)	zveza z določujočim prislovom (*popolnoma odrezan*)
Glagol (*boleti*)	predložna zveza s samostalnikom v različnih sklonih (*boleti v trebuhu*)		glagol, ki mu sledi samostalnik v tožilniku (*boleti človeka*)	zveza z določujočim prislovom (*pošastno boleti*)
Prislov (*lepo*)	zveza s pomensko določanim glagolom (*lepo pozdravljati*)	zveza s pomensko določanim pridevnikom (*lepo ohranjen*)	predložna zveza s samostalnikom v različnih sklonih kot desni prilastek (*lepo na dopustu*)	

Upoštevati je treba, da je Slovar sopomenk sodobne slovenščine trenutno še v prvi fazi, ki ni ročno pregledana in urejena. Vedno ko v tem prispevku govorimo o sopomenkah, bi bilo torej ustrezneje govoriti o besedah, ki so kandidatke za sopomenke. Enako opozorilo velja za kolokacije. V nadaljevanju prispevka skušamo opredeliti, kako bi bilo mogoče nabor in prikaz kolokacij izboljšati, da bi bolje ustrezale namenu, ki ga imajo znotraj obravnavanega slovarja, pri čemer rešitve iščemo v okviru predhodno uporabljene metodologije luščenja s pomočjo Primerjalnih skic orodja Sketch Engine. V okviru nacionalnih projektov Kolokacije kot temelj jezikovnega opisa: semantični in časovni vidiki[5] in Nova slovnica sodobne standardne slovenščine: viri

5 Nacionalni projekt J6-8255, spletna stran projekta: https://www.cjvt.si/kolos/.

in metode[6] so predvidene tudi številne druge izboljšave identifikacije, luščenja in strojnega razvrščanja kolokacij za slovenščino, ki bodo naslovile in odpravile nekatere naštete težave.

3 Gradivo in metoda

3.1 Gradivo

V analizo smo vključili 48 sopomenskih parov za 24 slovarskih iztočnic, in sicer 8 samostalniških, 8 pridevniških in 8 glagolskih.[7] Iztočnice so bile izbrane s seznama 333-ih lem iz Leksikalne baze za slovenščino (Gantar 2015), ki je bil pripravljen in uporabljen za predhodne analize kolokacijskosti v slovenščini (Pori in Kosem 2018). Kot pišeta avtorja (ibid.: 160), so lemam pripisane za jezikoslovne analize relevantne značilnosti, npr. samostalnikom spol, številnost, pogostost v korpusu Gigafida ipd. Pri izbiri 24-ih iztočnic za pričujočo analizo so bile naštete značilnosti upoštevane, poleg njih pa tudi nekatere dodatne značilnosti, kot je število jedrnih in bližnjih sopomenk[8] v slovarju. V vzorec smo tako vključili karseda raznoliko, vendar z vidika značilnosti uravnoteženo besedišče, kot je prikazano v Tabeli 2.

Tabela 2: Izbor 24-ih iztočnic za analizo in njihove razlikovalne lastnosti.

Iztočnica	Upoštevane značilnosti	Frekvenca v Gigafidi	Število sopomenk (jedrne; bližnje)
jezik	samostalnik, večpomensko, m. spol, neživo, število	100.000–499.999	14; 11
stanovanje	samostalnik, večpomensko, s. spol, neživo, število, besedotvorno iz glagola	100.000–499.999	14; 7
konj	samostalnik, večpomensko, m. spol, živo in nečloveško, število	50.000–99.999	2; 2

6 Nacionalni projekt J6-8256, spletna stran projekta: http://slovnica.ijs.si/.

7 Ker je bila nedavno opravljena natančna analiza kolokacij v prislovnih besednozveznih vzorcih (Pori in Kosem 2018), prislovov ne vključujemo v obravnavo, so pa izsledki v veliki meri prenosljivi tudi nanje.

8 Razlika med jedrnimi in bližnjimi sopomenkami je njihova strojno ocenjena pomenska bližina z iztočnico (Krek idr. 2017).

Iztočnica	Upoštevane značilnosti	Frekvenca v Gigafidi	Število sopomenk (jedrne; bližnje)
kos[9]	samostalnik, enopomensko, m. spol, neživo, števno, uporablja se za izražanje količine	50.000–99.999	32; 28
babica	samostalnik, večpomensko, ž. spol, živo in človeško[10], števno	10.000–49.999	10; 7
cigareta	samostalnik, enopomensko, ž. spol, neživo, števno, v kolokacijah se pogosto pojavlja v množini	10.000–49.999	2; 2
kiks	samostalnik, enopomensko, m. spol, neživo, števno, stilno zaznamovano (pogovorno)	1.000–4.999	6; 4
vznesenost	samostalnik, enopomensko, ž. spol, neživo, neštevno (pojmovna neštevnost), besedotvorno iz pridevnika	1.000–4.999	8; 6
športen	pridevnik, vrstni, večpomensko, besedotvorno iz samostalnika	100.000–499.999	8; 5
izjemen	pridevnik, lastnostni, večpomensko, intenzifikator, besedotvorno iz samostalnika	50.000–99.999	53; 45
ekološki	pridevnik, vrstni, enopomensko, besedotvorno iz samostalnika	50.000–99.999	6; 4
slep	pridevnik, lastnostni, večpomensko, konverznost	10.000–49.999	7; 6
oljen	pridevnik, snovni, večpomensko, besedotvorno iz samostalnika	5.000–9.999	5; 3
odrezan	pridevnik, lastnostni, večpomensko, besedotvorno iz glagola	5.000–9.999	21; 21
nagajiv	pridevnik, lastnostni, enopomensko, besedotvorno iz glagola	1.000–4.999	24; 22
pikčast	pridevnik, lastnostni, enopomensko, besedotvorno iz samostalnika	1.000–4.999	8; 5

9 Samostalnik *kos* je enakopisnica (*kos* – ptica v primerjavi s *kos* – del celote), vendar v Slovarju sopomenk najdemo sopomensko gradivo samo za drugi primer.

10 Za *babica* – riba v Slovarju sopomenk ni jedrnih in bližnjih sopomenk, zato pomena s podspolom nečloveškega tukaj ne upoštevamo.

Iztočnica	Upoštevane značilnosti	Frekvenca v Gigafidi	Število sopomenk (jedrne; bližnje)
doživeti	glagol, dovršni, enopomensko, prehodnost	100.000–499.999	12; 8
dovoliti	glagol, dovršni, večpomensko, prehodnost	50.000–99.999	18; 13
obljubiti	glagol, dovršni, enopomensko, prehodnost	50.000–99.999	8; 7
boleti	glagol, nedovršni, večpomensko, neprehodnost, brezosebnost	10.000–49.999	13; 9
blesteti	glagol, nedovršni, večpomensko, neprehodnost	10.000–49.999	5; 3
cveteti	glagol, nedovršni, večpomensko, neprehodnost, brezosebnost	10.000–49.999	7; 4
angažirati	glagol, dvovidski, večpomensko, prehodnost, povratnost	5.000–9.999	4; 4
izroditi	glagol, dovršni, večpomensko, neprehodnost, povratnost	1.000–4.999	7; 7

Za vsako od 24-ih iztočnic smo nato iz Slovarja sopomenk sodobne slovenščine izbrali po dve sopomenki. V analizo so bile zajete enobesedne in preferenčno jedrne sopomenke.[11] Kot pri izbiri iztočnic, smo tudi pri izbiri sopomenk skušali zajeti raznovrstno besedišče: splošne (npr. *obljubiti – priseči*) in strokovne (npr. *izroditi se – degenerirati*) ter pogovorne (npr. *cigareta – čik*); pogoste (npr. *angažirati – naročiti*) in redke (npr. *blesteti – briljirati*)*;* enopomenske (npr. *oljen – oljnat*) in večpomenske (npr. *blesteti – sijati*); sopomenske v dobesednem (npr. *stanovanje – nastanitev*) in prenesenem (npr. *konj – koza*) pomenu; lematizacijsko enostavne (npr. *vzvesenost – zmagoslavje*) in zahtevnejše (npr. *kos – del*) besede. Rezultate prikazuje Tabela 3, v kateri navajamo izbrane sopomenke in njihovo pogostost v korpusu Gigafida 2.0.

11 Gradivo, ki so ga v slovar v času od izida dodali uporabniki, ni vključeno v analizo, ker za uporabniške sopomenke kolokacijski podatki (še) niso na voljo.

Tabela 3: Nabor iztočnic in sopomenk za analizo, skupaj s frekvenco v korpusu Gigafida 2.0.

Iztočnica	Sopomenka 1	Sopomenka 2
jezik: 227.962	govorica: 51.383	slog: 84.736
stanovanje: 216.857	bivališče: 15.216	nastanitev: 7928
konj: 84.676	žrebec: 4440	koza: 11.676
kos: 66.042	del: 960.422	odlomek: 12.769
babica: 34.241	babi: 1240	porodničarka: 127
cigareta: 32.864	čik: 1337	tobak: 10.293
kiks: 1032	spodrsljaj: 13.958	napaka: 215.154
vznesenost: 1205	radost: 13.453	zmagoslavje: 11.261
športen: 304.627	fer: 475	neformalen: 15.609
izjemen: 123.104	izreden: 97.316	ubijalski: 2790
ekološki: 55.790	zelen: 144.324	okoljevarstven: 12.468
slep: 31.409	nekritičen: 1847	prazen: 84.738
oljen: 7339	naften: 35.716	oljnat: 509
odrezan: 6226	odsekan: 1539	prirezan: 523
nagajiv: 2962	navihan: 2645	poreden: 2475
pikčast: 1068	pikast: 600	pegast: 1286
doživeti: 141.224	izkusiti: 8924	utrpeti: 19.619
dovoliti: 92.506	tolerirati: 3238	pristati: 72.271
obljubiti: 65.102	obvezati: 4966	priseči: 9088
boleti: 27.554	mučiti: 23.217	gristi: 5520
blesteti: 23.670	briljirati: 616	sijati: 9241
cveteti: 16.728	uspevati: 24.980	razvijati se: 77.485
angažirati: 8093	najeti: 33.349	naročiti: 60.793
izroditi: 1353	degenerirati: 141	sprevrniti: 183

3.2 Postopek analize

Analiza je potekala v dveh korakih. Najprej smo za vsak sopomenski par iz Tabele 3 v Slovarju sopomenk sodobne slovenščine pregledali vključene kolokacije. Zanimala so nas problematična mesta: napačno označeni ali izluščeni, pomanjkljivi ter pomensko nerelevantni podatki. Na drugi strani smo pozitivno ocenili vzorce, ki prinašajo skladenjsko celovite in za primerjavo sopomenk uporabne podatke. Uporabnost v evalvaciji ocenjujemo subjektivno, glede na vtis o

informacijski vrednosti kolokacijskih podatkov za razumevanje razlik in podobnosti v pomenu in rabi obravnavanih sopomenskih besed. Pri tem je seveda treba upoštevati, da je jezikoslovna ocena uporabnosti lahko precej drugačna od rezultatov, ki bi jih pokazala primerljivo natančna uporabniška evalvacija.

V drugem koraku analize smo uporabili Primerjalne skice orodja Sketch Engine, in sicer na korpusu Gigafida 2.0 (Krek idr. 2020), ki je aktualna verzija referenčnega pisnega korpusa sodobne (standardne) slovenščine. Za vsak par sopomenk smo izdelali primerjalno skico, pri čemer smo določili besedno vrsto in ohranili frekvenčno mejo za prikaz podatkov na 3 pojavitve v korpusu. Ker v Primerjalnih skicah ni mogoče primerjati večbesednih enot (npr. *obvezati se*), smo vse primerjave opravili na posameznih besedah (npr. *obvezati*). Za analizo so nas zanimali podatki o frekvenci oz. relativni frekvenci vzorcev iz Besednih skic (Krek idr. 2016), v katerih se primerjani besedi pojavljata. Ideja pristopa je, da razlike v zastopanosti vzorcev lahko razkrijejo značilnosti jezikovne rabe, ki jih je mogoče pri luščenju in prikazu kolokacijskih podatkov upoštevati. Vzorci, ki so se s tega vidika izkazali za potencialno relevantne, so bili nadalje analizirani, kot predstavljamo v nadaljevanju. Rezultati analize so urejeni po besedni vrsti obravnavanih slovarskih iztočnic in znotraj tega glede na besednozvezne kolokacijske vzorce (Tabela 1).

4 Analiza in diskusija

4.1 Samostalnik

4.1.1 Zveze s pridevnikom kot levim ujemalnim prilastkom

V tem delu kolokacijskih podatkov se pojavi 34 praznih mest (14 %), kar je v primerjavi z drugimi vzorci malo.[12] Vrzeli so na predvidljivih mestih, in sicer pri samostalnikih, ki so v korpusu redki,[13] tako pri pogovornih (*babi, čik* in *kiks*) kot nezaznamovanih (*porodničarka, vznesenost*). Napak lematizacije v tem delu podatkov ni veliko,

12 Deleže podatkovne (ne)pokritosti računamo od 240 kolokatorjev: 16 slovarskih ekranov s primerjavo kolokatorjev, v kateri posamezen stolpec zajema do 15 kolokatorjev.

13 Kot kaže Tabela 3, imajo vse te besede v korpusu frekvenco pojavitev nižjo od 1500.

pojavljata se le primera *polna radost* namesto *poln radosti* in *rotacijski kos* namesto *rotacijska kosa*. Tudi težave z nezaključenimi oz. delnimi kolokacijami niso izstopajoče, čeprav jih je mestoma opaziti, npr. *prodajana cigareta* namesto *najbolje prodajana cigareta*. Omeniti gre še svojilne pridevnike, npr. *donov žrebec* (zapis kolokatorja z malo, v korpusnih zgledih z veliko) ali *Uroševa babica,* ki sicer niso napačni, vendar za primerjavo pomena sopomenk niso posebej obvestilni.

Več možnosti za izboljšave je pri izpisu podatkov v besednozvezno ustrezni obliki, pri čemer je prva težava pri pridevnikih moškega spola, ki so v lematizirani nedoločni obliki težje berljivi, npr. *povoden konj* namesto *povodni konj*, ali dvoumni, npr. *učen jezik* namesto *učni jezik*.[14] Kot drugo, prikaz besed v lematizirani ednini je problematičen pri primerih, kjer je raba v določenem slovničnem številu vezana na pomen iztočnice, ki ni sopomenski z drugo od primerjanih besed. Samostalnika *jezik* in *govorica* sta denimo sopomenska v pomenu, ki ga izražajo primeri *nemški jezik – nemška govorica, tuj jezik – tuja govorica,* ne pa v pomenu, ki ga izpričujejo primeri *[nepreverjene, neresnične, lažne] govorice*. Kadar je tendenca k množini vezana na vse pomene besede, npr. *[ponarejene, drage, lahke] cigarete,* je problem manjši, še najbolj moteči so morda primeri tipa *številna napaka,* kjer bi bilo vezavo z množino mogoče predvideti na ravni kolokatorja in temu prilagoditi tudi prikaz v slovarju.

4.1.2 Predložna zveza s samostalnikom kot desnim prilastkom

Za razliko od prejšnjega vzorca je delnost kolokacij tu večji problem. Ker postopek luščenja dovoljuje prosta mesta med predlogom in sledečim samostalnikom, v rezultatih najdemo številne primere tipa *jezik na ravni, babica za dan, del za industrijo* in podobno. Težave so tudi z nezaključenimi zvezami, npr. *stanovanje v izmeri, cigareta v vrednosti.* Na drugi strani je najti tudi višje število praznih mest v podatkih, in sicer 60 (25 %).

14 Po hitri oceni bi izpis ustreznih oblik izboljšal prikaz pri 39 od 88 oblik (44 %) za pridevnike moškega spola.

Nebeležena množina je najbolj problematična, kadar slovnično število vpliva na sopomenskost, kot je opredeljeno v 4.1.1, npr. *govorica o [prevzemu, poroki, odhodu]* namesto *govorice o [prevzemu, poroki, odhodu]*. Prikaz množine bi bil smiseln tudi pri desnem samostalniku, npr. v primerih tipa *jezik za zobom; bivališče na kolu*. Tudi tu pa so redke težave z lematizacijo, ki so na predvidljivih mestih pri napačnem razdvoumljanju samostalnika *delo* v npr. *del na cesti,* glagola *kosati* v npr. *kos s konkurenco* in samostalnika *slog* v npr. *Sloga na Primskovem.*

4.1.3 Zveze z glagolom in predlogom, ki mu sledi samostalnik v sklonu

Podobno kot pri 4.1.2 je problematična delnost prikazanih kolokacij, ki v kombinaciji s širokim oknom za luščenje, ki relacije išče na obeh straneh glagola, prinaša slabo razumljive in neustrezne rezultate, npr. *potrebovati za kos, povedati na jezik* (*za 12 kosov potrebujemo*, *brez dlake na jeziku povem*). Kot drugo, ker se nekateri predlogi lahko pojavljajo z različnimi skloni, samostalnika pa v izpisu kolokacije ne dodajamo, mora uporabnik slovarja ustrezni sklon samostalnika predvideti sam. Pri tem so seveda v pomoč korpusni zgledi, ki kolokacijo razdvoumijo, vendar se izkazuje, da bi bilo preglednost slovarskih informacij mogoče pomembno izboljšati, če bi na primerjalnem ekranu izpisovali celotne kolokacije, ne le kolokatorjev. Razumljivost rezultatov bi izboljšala tudi vključitev povratnega zaimka, npr. *[požvižgati se na, odzivati se na, zmeniti se za] govorice* namesto trenutnega *[požvižgati na, odzivati na, zmeniti za] govorice*.

Praznih mest v naboru kolokatorjev za ta vzorec je 57 (24 %). Težav z lematizacijo ni prav dosti, najti je npr. *obcutiti na kozi,* kjer gre za problem manjkajočih šumnikov, ali *potrditi o bivališču* namesto *potrdilo o bivališču.* Kolokacije tega vzorca smo posebej pregledali tudi z vidika morebitne pojavnosti zvez s pomensko izpraznjenimi glagoli, npr. *iti za [odlomek, kos]*. Izkaže se, da so tovrstni primeri vsaj v analiziranem gradivu redki in jih imamo lahko za neproblematične.

Če bi se izkazalo, da jih je preveč in da motijo pomensko primerjavo, bi jih lahko iz prikaza izločili.

4.1.4 Glagol s samostalnikom v tožilniku

Obravnavani vzorec ima med vsemi drugo najvišje število praznih mest, tj. 76 (37 %). Tako visoka podatkovna nepokritost je signal, da je treba preveriti metodologijo luščenja in razmisliti o morebitnih prilagoditvah. Vzorec prinaša glagolske kolokatorje, ki se pojavljajo s samostalnikom v tožilniku. V skladenjskem smislu gre torej večinoma za stavčne predmete, pri čemer se v rezultatih mestoma pojavljajo napake, ki zaradi dvoumnosti tožilniških oblik z imenovalniškimi vključujejo stavčne osebke, npr. *pripetiti spodrsljaj, zgoditi kiks.* Napak lematizacije sicer (ponovno) ni veliko, npr. *nabrusiti kos* namesto *koso* ali *peti kos / odlomek* namesto *pojesti kos, peti odlomek.*

Za razliko od 4.1.2 in 4.1.3 je delnost kolokacij pri teh podatkih manj moteča. Široko luščenje je problematično predvsem v primerih, kadar identificira skupne točke v rabi primerjanih samostalnikov, za katere se izkaže, da nastopajo v pomensko precej različnih zvezah, npr. 'prekrivna' raba *najti jezik / slog,* kjer gre v resnici za zveze *najti [skupni, primeren] jezik* in na drugi strani *najti [svoj, osebni, lasten] slog.* Analiza je identificirala tudi vprašanje ločenega prikaza dovršnih in nedovršnih glagolov. Na načelni ravni je ločevanje seveda metodološko utemeljeno, v praksi pa lahko v izpisu petih kolokatorjev vidski pari zavzamejo precej prostora in s tem nižajo informativno vrednost podatkov, kot npr. pri *[občutiti, doživeti, čutiti, začutiti, doživljati] vznesenost / radost.*

4.1.5 Uporabnost podatkov za primerjavo rabe in pomena samostalnikov

Od obravnavanih štirih vzorcev se zdijo zveze z ujemalnim levim prilastkom (4.1.1) najbolj uporabne za primerjavo rabe in pomena sopomenk. Dobri rezultati se kažejo tako pri opredeljevanju prekrivnosti, npr. *[pokajen, poceni, uvožen] tobak* in *[pokajena, poceni, uvožena] cigareta,* kot tudi pri izpostavljanju razlik, npr. *[lahka,*

ponarejena, ponujena] cigareta v primerjavi z *[okrasni, rezani, pridelani] tobak*. Ali denimo *[zasebno, luksuzno] stanovanje / nastanitev,* vendar *[novo, dvosobno]* stanovanje v primerjavi z *[nekajdnevna, kratkotrajna] nastanitev*. Kot drugo, uporabne rezultate dajejo tudi kolokacije glagola in samostalnika v tožilniku, katerih pridobivanje in prikaz bi se sicer dalo še izboljšati. Kolokacije dobro odražajo predvsem razlike v pomenu sopomenk, vendar je zaradi problemov, navedenih v 4.1.4, trenutno relevanten samo del podatkov. Tipična primera koristne primerjave sta denimo *[kupovati, prodajati, oddajati] stanovanje* v primerjavi z *[izdolbsti, obdati] bivališče;* ali *[ugasniti, vzeti, odnesti] cigareto* v primerjavi z *[gojiti, pridelovati, njuhati] tobak.*

Slabše rezultate dajejo zveze s predložnim samostalnikom (4.1.2). Pogosto opredeljuje krajevne okoliščine npr. *jezik [na Slovenskem, v Italiji]*, *stanovanje [v Celju, v bloku]*, kar razen v določenih izjemah za primerjavo rabe in pomena nima visoke vrednosti. Tudi sicer so uporabni primeri pri tem vzorcu redkejši, morda se med petimi kolokatorji pojavita dva ali eden, ki ga je mogoče izpostaviti, denimo *konj za preskakovanje* v primerjavi z *žrebec za pripust* ali *vznesenost [ob revoluciji, nad občutjem]* v primerjavi z *radost [do življenja, v srcu, na snegu]*. Kot relevantni se pogosto izkažejo kolokatorji, v katerih nastopa en sam predlog, npr. *za,* ki v sledečem primeru nakazuje namembnost samostalnika: *tobak [za kajenje, za pipo, za žvečenje, za njuhanje]*, medtem ko se pri primerjani besedi *cigareta* ta predlog ne pojavi. Kot zadnje, tudi zveze z glagolom in predlogom (4.1.3) imajo omejeno uporabno vrednost. Kadar so na voljo kolokatorji za obe besedi, so podatki sicer lahko koristni, npr. *[preseliti v, vstopiti v, vdreti v] stanovanje* / *bivališče*, vendar *zagoreti v stanovanju, zapreti v stanovanje* ter *izbrisati z bivališča, odjaviti z bivališča.* Vzorec se nekoliko bolje odreže pri samostalnikih, ki so v nesopomenskem delu rabe dovolj različni, npr. *narezati na kose* v primerjavi z *izpisati iz odlomka* ali *prevesti v jezik* v primerjavi z *opremiti v slogu.* Najpogosteje pa uporabnost podatkov zavira abstrahirani in delni prikaz besedne zveze, skupaj s podatkovnimi vrzelmi, kot je omenjeno v 4.1.3.

4.1.6 Analiza ostalih vzorcev v orodju Sketch Engine

Analiza pokaže, da se pri večini obravnavanih samostalnikov v samem vrhu[15] frekvenčnega seznama pojavlja eden, ki v slovar ni vključen, in sicer vzorec, v katerem samostalnik nastopa kot desni neujemalni prilastek v rodilniku, npr. *najemnik stanovanja*. Podatki kažejo, da bi vključitev tovrstnih primerov ponudila uporabnejše rezultate, kot velja za nekatere trenutno vključene vzorce. Dober primer je denimo *[hrbet, hlev, predstavitev, sedlo] konja / žrebca* v primerjavi z *[moč, jahanje, pasma, žegnanje] konja* in *[linija, licenciranje, odbira, seme] žrebca.* Analiza pokaže tudi, da se (samo) pri samostalnikih, ki opredeljujejo količino, v vrh frekvenčnega seznama prebije vzorec, kjer obravnavani samostalnik nastopa kot jedro zveze z desnim neujemalnim samostalnikom v rodilniku, npr. *kos / odlomek [posode, keramike, kosti, besedila]* v primerjavi s *kos [pohištva, kruha, mesa, orožja]* ter *odlomek [pesmi, pisma, knjige, romana].* To značilnost je mogoče izkoristiti in tovrstne primere med iztočnicami slovarja najprej strojno identificirati, nato pa jih ponuditi z ustrezno prilagojenim naborom kolokacijskih vzorcev.

4.2 Pridevnik

4.2.1 Zveze s samostalnikom kot jedrom ujemalne zveze

Te zveze v slovarju trenutno obsegajo dva stolpca. V prvem je najti samo 18 praznih mest v podatkih, pri drugem 36 (7,5 % ter 15 %). Podobno kot pri 4.1.1, ki je v skladenjskem smislu soroden vzorec, je tudi tukaj glavni razlog za prazna mesta redkost obravnavanega besedišča, npr. pri pridevnikih *fer, oljnat, prirezan, pikast.*[16] Ob tem se prazna mesta tokrat pojavijo tudi na drugih mestih, in sicer pri pogostih in pomensko zelo podobnih pridevnikih *izjemen / izreden*, kjer slovar navaja kolokatorje za oba pridevnika (npr. *dosežek, pomen, uspeh*) in na drugi strani primere tipa *izredna [seja, odpoved, skupščina]*, ne pa tudi kolokatorjev samo za pridevnik *izjemen.* Takšna distribucija sugerira, da sta pridevnika sopomenska v vseh pomenih

15 Pogosto na 2. mestu, za vzorcem z ujemalnim levim pridevniškim prilastkom.

16 Kot je razvidno iz Tabele 3, se vsi ti primeri v korpusu pojavljajo z manj kot 1000 pojavitvami.

besede *izjemen,* ne pa tudi v vseh pomenih besede *izreden,* kar se zdi na prvi pogled ustrezno. Problematično pa je, da tovrstnih primerov v slovarju trenutno ni mogoče enostavno vizualno ločevati od primerov, kjer podatki manjkajo zaradi redkosti.

Analiza pokaže, da bi bilo smiselno pri izbiri kolokatorjev za prikaz filtrirati (oz. na potisniti na dno seznama kandidatov za luščenje) lastna imena, predvsem osebna, morda pa tudi kratična poimenovanja. Problem z lastnimi imeni in kraticami pride zlasti na površje pri redkih (in v korpusu posledično lahko napačno označenih) besedah, kot npr. pri primeru *fer [play, plej, Il, Tampere, boa, Leipajas, fajt, Hans, FBK, Kaunas]*. Razen tega večjih lematizacijskih ali označevalnih težav – tako pri tem vzorcu kot obeh ostalih, ki sta v rabi za pridevnike – ni opaziti.

4.2.2 Predložna zveza s samostalnikom kot desnim prilastkom

Od vseh obravnavanih vzorcev je ta najbolj problematičen z vidika praznih mest: pri obravnavanih primerih je najti kar 175 praznih mest, kar pomeni 73 % podatkovno nepokritost. Prazna mesta, ki se pojavljajo pri večini obravnavanega gradiva, so lahko posledica redkosti te zveze v rabi, morda pa nakazujejo napake v postopku luščenja, kar bi bilo treba preveriti. Podobno kot pri 4.1.2, je vzorec problematičen tudi zaradi delnosti in prikaza rezultatov ter napak, ki so posledica širokega okna za iskanje skladenjskih enot, npr. *športen [od brata, od tekmeca]* ali *izreden [ob točki, na študij]* – kar je denimo posledica zvez tipa *prehod z izrednega na redni študij*.

4.2.3 Zveze pridevnika s pomensko določujočim prislovom

Pri teh zvezah je najti 59 praznih mest (25 %). Za razliko od vseh do sedaj navedenih primerov je glavni problem distribucijski: v rezultatih so našteti prislovi, ki so pomensko precej splošni in se v rabi pojavljajo z veliko pridevniki, npr. *[tako, res, lahko] izjemen / ubijalski* ali *[bolj, tako, vedno, nekaj, veliko] ekološki / zelen.* Tovrstni podatki za primerjavo pomena pridevnikov nimajo prave vrednosti. Podobno kot prej pa je problematična tudi napačna interpretacija skladenjskih

odnosov, ki se pojavi zaradi širokega okna za luščenje, npr. *približno ekološki* (*približno 1.000 ekoloških kmetij*), *domov prazen* (*vrnili domov praznih rok*), *gensko oljen* (*gensko spremenjena oljna repica*), *rusko naften* (*rusko naftno družbo*) ipd. Tudi pri tem vzorcu bi torej omejitev okna za luščenje kolokacij lahko pomembno izboljšala kvaliteto rezultatov.

4.2.4 Uporabnost podatkov za primerjavo rabe in pomena pridevnikov

Ujemalne zveze pridevnika in samostalnika dajejo uporabne rezultate, kot je omenjeno že v 4.1.5. Primer, ki dobro predstavi potencial teh kolokacij za primerjave sopomenskosti, je npr. *slepa / nekritična [ljubezen, vera, poslušnost]* oz. *slepo / nekritično [posnemanje, navdušenje]* v primerjavi s *slepa [ulica, pega, oseba, miš]* ali *nekritično [jemanje, povzdigovanje, pretiravanje, objavljanje]*. Primerljivo uporabne so kolokacije za pridevnike, ki so bolj terminološke narave, npr. *oljna / naftna [ploščad, industrija, črpalka]* v primerjavi z *oljna [ogrščica, repica, slika, barva]* ter *naftni [trg, plin, velikan, kartel]*. Omejena uporabnost pa se po pričakovanjih izkaže, kjer je prekrivnost pomena vezana na zelo ozko področje rabe, nesopomenski pomen pridevnikov pa je v rabi pogost. Posledično se v kolokacijah sopomenska prekrivnost ne pokaže, npr. v primeru *športna / neformalna [zveza, vzgoja, pot],* ne pa tudi *oblačilo* ali kaj podobnega.

Slabo uporabne so zveze s predložnim samostalnikom na desni (4.2.2). Podatki so pomanjkljivi in redko zares relevantni, v najboljšem primeru se med kolokatorji najde eden ali dva, ki ju je mogoče uporabiti za primerjavo. Izjema so (redki) primeri, kjer pridevnik močneje kolocira z določenim predlogom, še zlasti, če je ta vezljivost s sopomensko primerjavo povezana. Tak primer je denimo *nekritičen do [dela, stanja, odplake]* kot sopomenka pridevniku *slep.* Zdelo bi se smiselno, da bi se v prihodnjih korakih tovrstna pridevniška vezljivost s predlogi identificirala na vsem gradivu in kadar je to utemeljeno, tudi vključila v izpis sopomenk v slovarju, npr. *izoliran – odrezan od*

namesto *izoliran – odrezan.* Obravnavani kolokacijski vzorec pa bi bilo najbolje nadomestiti s kakim drugim.

Tudi kolokacije s prislovi so glede uporabnosti omejene (kar ugotavljata že Pori in Kosem (2018), ki se ukvarjata s prislovnimi zvezami v Kolokacijskem slovarju sodobne slovenščine). Ob vseh težavah (4.2.3) so kolokacije, ki dobro nakazujejo podobnosti in razlike v rabi ter pomenu pridevnikov, pri tem vzorcu redke. Še najboljši primer je morda *[gospodarsko, zdravstveno, ekonomsko, razvojno] ekološki* v primerjavi s *[temno, svetlo, olivno, živo] zelen.* Če bi prislove, ki se tipično pojavljajo z zelo veliko pridevniki, zlasti deiktične in merne (npr. *takoj, tako, jutri*), in tiste, ki so v slovnični vlogi (npr. *bolj, najbolj, lahko*), potisnili na dno seznama kandidatov za luščenje, bi se v slovar predvidoma uvrstili bolj pomenonosni podatki, npr. *[elegantno, rekreativno, poudarjeno] športen* namesto *[nekaj, več, tako] športen.*

4.2.5 Analiza ostalih vzorcev v orodju Sketch Engine

Analiza frekvenčnih seznamov pridevniških vzorcev v Sketch Engine pokaže, da sta v vrhu seznamov, v slovarju pa (še) nezajeta, vzorca dveh vrst: (a) vzorec, ki prinaša z vezajem povezane priredne zloženke in (b) priredne zveze, ki jih povezuje veznik *in*. Oba vzorca sta kandidata za vključitev v slovar, vendar nista brez težav. Vzorci z vezajem se slabo obnesejo za prikaz prekrivnosti, razlike pa so dokaj pomenonosne, npr. *ribolovno-ekološki, razvojno-ekološki, turistično-ekološki* v primerjavi z *belo-zelen, rdeče-zelen, socialdemokratsko-zelen*. Na drugi strani so zveze z *in* koristne tudi za primerjavo prekrivnosti, vendar bi njihova vključitev uporabnike slovarja lahko zmedla, saj podatki zaradi enake besedne vrste na videz spominjajo na sopomensko gradivo, ki ga slovarju prinaša na drugih mestih, v resnici pa gre za nabor besedišča, ki vključuje različna pomenska razmerja. Primer kolokacij je npr. *športen / neformalen in [sproščen, prijateljski, oseben]* ter na drugi strani *športen in [kulturen, družaben, zabaven]* v primerjavi z *neformalen in [formalen, priložnosten, odprt]*. V primeru vključitve tega vzorca bi bilo torej v izogib

zmedi treba poskrbeti, da so kolokacije prikazane čimbolj celovito in nedvoumno.

4.3 Glagol

4.3.1 Zveze glagola s predlogom, ki mu sledi samostalnik

Te zveze v slovarju trenutno obsegajo dva stolpca, pri čemer je praznih mest v podatkih 44 v prvem ter 57 v drugem stolpcu (18 % in 24 %). Opozoriti gre, da se pri izpisu v stolpcih pojavljajo težave, npr. izpis v drugem stolpcu namesto v prvem (npr. pri *dovoliti / tolerirati*) ali v tretjem namesto v drugem (npr. pri *izroditi se / degenerirati*), zato štetje ni povsem natančno. Pri analizi kolokatorjev je opaziti podobne težave kot pri podobnih zvezah (4.1.2 in 4.2.2). Izstopajo zlasti napake v interpretaciji skladnje, ki so posledica širokega okna luščenja, in pa primeri nezaključenih zvez, npr. *doživeti v krogu* (*v četrtem krogu doživela še en poraz*), *priseči na pismo* (*sveto pismo*). Redkejši in manj moteč problem je izpis lematiziranih oblik na mestu, kjer se v rabi pojavlja množina, npr. *obljubiti pred volitvijo* namesto *obljubiti pred volitvami* ali *gristi z zobom* namesto *gristi z zobmi.* Med kolokatorji se pojavljajo tudi lastna imena, kar pa večinoma ni problematično, npr. *doživeti v Ljubljani, priseči na Apolona.*

4.3.2 Glagol s samostalnikom v neimenovalniškem sklonu

V podatkih se pojavlja 46 praznih mest (19 %). Kolokatorji pri tem vzorcu se trenutno izpisujejo na dva različna načina: v primerjalni postavitvi ostajajo v osnovni obliki, v samostojni postavitvi pa so preoblikovani v ustrezen sklon in včasih tudi ustrezno število. Primer so npr. kolokacije *doživeti / utrpeti [spremememba, nesreča, poškodba]* v primerjavi z *utrpeti [rane, odrgnino, deformacijo]*. Razlika ustvarja nekaj zmede, ki bi se ji dalo izogniti z ločenim izpisom celotnih kolokacij tudi v primerjalni postavitvi. Kot drugo, trenutno na izpis oblike vpliva zanikanost glagola, vendar nikalnica v kolokacijo ni vključena. Tako dobimo rezultate tipa *dovoliti vstopa* namesto *dovoliti vstop* ali

ne dovoliti vstopa. Pri nadgradnji prikaza podatkov bi bilo treba to težavo nasloviti, dodati pa je treba tudi preverjanje ponovljenih kolokatorjev: pri primerjavi *boleti / mučiti* se kolokacija *boleti vrat* pojavi dvakrat, predvidoma zato, ker se isti samostalnik lahko pojavlja z oznakami za različne sklone, kar se pri luščenju interpretira kot različne kolokacije.

Dokaj pogosto se razkriva tudi napačno označevanje samostalnika v imenovalniku, npr. *mučiti težavo, blesteti vratarja, pristati helikopterja* (*nato pa sta v bližini pristala helikopterja*). Problematične so tudi zveze, ki sugerirajo stavčni predmet, v resnici pa gre za prislovno določilo, v katerem manjka del zveze, npr. *blesteti teden* (*blesteti naslednji teden*). Tudi te težave bi bilo mogoče nasloviti z natančnejšo opredelitvijo parametrov za luščenje. Analize potrdijo, da so pri neprehodnih glagolih težave še večje, saj se v podatkih redko sploh pojavi uporabna kolokacija, npr. *blesteti [Milivoja, Penelope, soigralki, libero, znamenje]* ali *cveteti [poletje, maja, spomlad, junija, julija]*.

4.3.3 Zveze glagola s pomensko določujočim prislovom

V teh podatkih se pojavlja samo 33 praznih mest (14 %). Kot pri 4.2.3 se pokaže, da je velik del vključenih prislovov distribucijsko zelo razpršenih, torej se v rabi pojavljajo z veliko glagoli, zaradi česar predvsem primerjalni pogled ne daje veliko informacij. Tipična primerjava je npr. *[lepo, dobro, naprej, hitro, tako] cveteti / razvijati se.* Težave s prekinjenostjo ali nepopolnostjo zvez tukaj niso tako moteče kot pri nekaterih drugih vzorcih, pojavi pa se nekaj težav z označevanjem, npr. *primexu tolerirati, rano obvezati se, ku boleti* in mestoma primeri napačne skladenjske interpretacije zvez, npr. *svobodno dovoliti* (*dovolila svobodneje zadihati*). Kakovost rezultatov bi lahko izboljšali z omejitvijo luščenja na prislove, ki se pojavljajo levo (ne pa tudi desno) od glagola, kot je že bilo predlagano predhodno (4.2.4), pa tudi s potiskom distribucijsko zelo razpršenih prislovov na dno seznama kandidatov za luščenje.

4.3.4 Uporabnost podatkov za primerjavo rabe in pomena glagolov

Kolokacije s samostalnikom v neimenovalniškem sklonu (4.3.2) se zdijo posebej koristne za primerjavo rabe in pomena prehodnih glagolov. Tipična primera sta denimo *angažirati / najeti [odvetnika, strokovnjaka, detektiva]* v primerjavi z *angažirati [gledalca, Francoza, reprezentantko]* ter *najeti [posojilo, stanovanje, sobo]* ali pa *doživeti / utrpeti [spremembo, poraz, poškodbo]* v primerjavi z *doživeti [premiero, orgazem, razcvet]* ter *utrpeti [rane, odrgnino, zvin]*. Za neprehodne glagole je vzorec bistveno manj uporaben in bi ga bilo bolje nadomestiti z imenovalniškim vzorcem (samostalnik kot stavčni osebek).

Tudi zveze s prislovom se zdijo uporabne za pomensko primerjavo, vendar bi bilo treba njihovo luščenje nadgraditi (4.3.3). Že sedaj je v analiziranem gradivu najti uporabne kolokacije, npr. *[nikoli, spet, končno] dovoliti / pristati* v primerjavi s *[preveč, javno, zakonsko] dovoliti ter [zasilno, mehko, uspešno] pristati;* ali pa *[naravnost, najbolj, znova] blesteti / sijati* in *[strelsko, intelektualno, svetovno] blesteti* ter *[toplo, prijetno, prijazno] sijati.* Primer, v katerem smo za vtis o vrednosti predlaganih metodoloških izboljšav distribucijsko razširjene prislove odstranili ročno, je npr. *doživeti / izkusiti.* Trenutno stanje v slovarju je: *[lahko, nekaj, veliko, kar, več] doživeti / izkusiti* in na drugi strani *[lani, zelo, precej, spet, najbolj] doživeti* ter *[dodobra, šestič] izkusiti.* Izboljšani primer je *[osebno, neposredno, zares, skupaj, živo] doživeti / izkusiti* v primerjavi s *[končno, zagotovo, težko, nedvomno, nepričakovano] doživeti* ter *[dodobra, boleče, grenko, bridko] izkusiti*.

Zveze s predlogom in samostalnikom (4.3.1) se zdijo manj uporabne, vendar bi se tudi ti rezultati lahko izboljšali, če bi postavili strožje skladenjske pogoje in vnaprej izbran nabor kolokatorjev potisnili na dno seznama za luščenje. Slednje velja denimo za deiktične opredelitve časa in kraja, npr. *doživeti v [soboto, nedeljo, sredo, petek]*. Glede na omejeno uporabnost podatkov se zdi prikaz v dveh stolpcih odveč, smiselno je obdržati en stolpec. Uporabnost vzorca se potrjuje predvsem pri primerih, kjer se razlika v pomenu

kaže skozi opredelitev okoliščin dejanja, npr. *doživeti [na festivalu, v Ljubljani, na odru]* v primerjavi z *utrpeti [v nezgodi, pri trčenju, v ujmi]*; ali pa *obljubiti / priseči [pred bogom, v cerkvi, na slovesnosti]* v primerjavi z *obljubiti [pred volitvami, na sestanku, na konferenci]* ter *priseči [pred predsednikom, pred sodnikom, na ustavo]*. Pogosto se v teh podatkih kaže tudi frazeologija, npr. *doživeti / izkusiti na koži, boleti pri srcu, gristi v jezik*.

4.3.5 Analiza ostalih vzorcev v orodju Sketch Engine

Primerjava frekvenčnih seznamov vzorcev za sopomenska glagola lahko relativno enostavno razkrije primere, kjer se določen glagol pogosteje pojavlja z določenim predlogom (in zahtevanim sklonom). Pozornost zahtevajo primeri, kjer je pri enem od glagolov relativna frekvenca takega vzorca 0, pri drugem pa 0,02 ali več. Te razlike nakazujejo primere, ki jih je treba v nadaljevanju ročno pregledati, saj bi bilo v prikaz glagolskih sopomenk mestoma možno in koristno vključiti tudi predlog in njegovo vezljivost, npr. pri paru *dovoliti – pristati na* ali *pričakovati – računati na*. Upoštevanje vezljivosti bi lahko v nadaljevanju pomagalo avtomatsko ločevati podatke, ki so sedaj prikazani skupaj. Primer je denimo par *obljubiti – obvezati se,* za katerega s predlaganim postopkom najdemo pomensko različni skupini kolokacij *obvezati se s [povojem, trakom, gazo]* ter *obvezati se k [sodelovanju, plačilu, zmanjšanju]*; sopomenskost z *obljubiti* je vezana samo na drugo skupino primerov, kar je pri prikazu kolokacij v slovarju mogoče upoštevati.

Frekvenčni seznami vzorcev so lahko v pomoč tudi pri odločitvi, ali v kombinaciji z določenim glagolskim parom kazati samostalnike v imenovalniku (kot stavčne osebke) ali v neimenovalniškem sklonu (primarno kot stavčne predmete). Za primera *blesteti / sijati* in *cveteti / uspevati* bi prehod na kombinacije z imenovalnikom bistveno izboljšal kvaliteto podatkov. Pri tovrstnih glagolih se v frekvenčnih seznamih imenovalniški vzorci dejansko pojavljajo v samem vrhu in jih je mogoče na tej osnovi zelo enostavno identificirati. V primeru, da bi v slovarju v prihodnosti kolokacije izpisovali v besednozvezni

obliki, je glagol mogoče izpisati v ednini tretje osebe. Primer ročno pripravljenih podatkov je npr. *[zvezda, igralka, trener] blesti / sije* v primerjavi z *[vratar, igralec, napadalec] blesti* ter *[sonce, luna, svetloba] sije*; oziroma drugi primer: *[rastlina, posel, roža] cveti / uspeva* v primerjavi s *[trgovina, rožica, turizem] cveti* ter *[zelenjava, koruza, gozd] uspeva.* Kot je razvidno iz primera, bi lahko tovrstni podatki bistveni prispevali k primerjavi pomena in rabe sopomenskih neprehodnih glagolov.

5 Diskusija in zaključek

Evalvacija je razkrila šibka in močna mesta kolokacijskih podatkov v prvi različici Slovarju sopomenk sodobne slovenščine. Potrdila so se predvidevanja, da je tudi znotraj obstoječega metodološkega okvira luščenje in prikaz kolokacij v slovarju mogoče pomembno izboljšati. Kot je že bilo omenjeno, pa je v pripravi tudi metodološka nadgradnja luščenja in urejanja slovenskih kolokacij, ki bo določene identificirane težave samodejno odpravila.

Predpostavka, da lahko primerjava izbranega nabora kolokacij **nakaže podobnosti in razlike v pomenu** sopomenk iste besedne vrste, se je potrdila. Pri vsaki od obravnavanih besednih vrst (samostalnik, pridevnik, glagol) se pojavlja vsaj en vzorec, ki že v trenutni, prvi različici slovarja prinaša koristne informacije, vsaj polovica vzorcev pa ima dober potencial ob predvidenih metodoloških nadgradnjah. Na drugi strani kolokacije v analiziranem vzorcu morebitne **stilne ali časovne zaznamovanosti** besed ne razkrivajo. Trenutno so slovarski podatki pridobljeni iz referenčnega pisnega korpusa, ki je v zadnji ediciji postal tudi korpus standardnega (sodobnega) jezika. Če se zaznamovano besedišče pojavlja, je v korpusu redko in redki (ali neobstoječi) so tudi kolokacijski podatki. Če bi želeli v primerjavo rabe sopomenk zajeti tudi nestandardno ali nesodobno gradivo, bi bilo treba v metodologijo luščenja vključiti druge korpusne vire in razviti način za skupen primerjalni prikaz.

Problematične so **podatkovne vrzeli oz. prazna mesta v naboru kolokatorjev,** kjer je mogoče izpostaviti tri težave. Prva je odsotnost

kolokatorjev zaradi splošne redkosti primerjanih besed. Druga je vezana na redkost izbranega skladenjskega vzorca, pri čemer izrazito izstopajo zveze pridevnika in predložnega samostalnika, kjer manjka kar 73 % podatkov. Večina vzorcev sicer izkazuje do 25 % nepokritost. Omenjene vrzeli bi se dalo v prvi vrsti nasloviti z menjavo (pre)redkih vzorcev, pogojno pa tudi z nižanjem frekvenčnega praga za luščenje kolokacij. Tretja težava je, da iz trenutne vizualizacije podatkov v slovarskem vmesniku uporabnik ne more ugotoviti, ali vrzeli v podatkih odslikavajo dejansko redkost opazovanih pojavov ali so zgolj posledica redkosti besedišča. V tem smislu bi bilo koristno v slovar vključiti tudi informacijo o pogostosti primerjanih besed, morda tudi izluščenih kolokatorjev.

K uporabnosti podatkov bi prispevalo **skladenjsko strožje luščenje kolokacij.** Trenutna metodologija je usmerjena v priklic podatkov: vključuje široko okno za iskanje kolokatorjev levo in desno od iztočnice, kar je nujno za pripravo jezikovnih virov, ki ciljajo na karseda širok nabor kolokacijskega gradiva. Za primerjavo rabe sopomenk pa je bistveno pomembnejša natančnost rezultatov. Velik del podatkov je trenutno neuporaben zato, ker vsebujejo skladenjsko napačno interpretirane, nezaključene ali pomensko pomanjkljive besedne zveze. Široko luščenje obenem viša možnosti za napačno združevanje rezultatov pri prikazu kolokacijsko prekrivne rabe obeh besed. Nenazadnje bi omejitev iskalnega okna na mesta tik ob iztočnici olajšala povezovanje slovarja s korpusom Gigafida. Povezave na konkordančni niz, ki skušajo reproducirati široka iskalna okna, trenutno javljajo napake in so za uporabnike brez prave vrednosti.

Kot možna rešitev pri izbiri vzorcev za posamezni par besed se ponuja **prilagojeni prikaz,** ki bi upošteval pomembnejše značilnosti iztočnic, npr. (ne)prehodnost glagolov, kjer se lahko odločimo za prikazovanje zvez bodisi s skladenjskim osebkom ali predmetom. Predpostavka, da razlike v frekvenčni zastopanosti posameznih vzorcev lahko razkrijejo značilnosti, ki jih je mogoče upoštevati pri nadgradnji luščenja kolokacijskih podatkov, se je potrdila. Frekvenčni seznami vzorcev, v katerih se pojavlja posamezna beseda, se zdijo zelo dobro izhodišče za opredelitev npr. neprehodnih glagolov, samostalnikov,

ki izražajo količino, ter glagolov in pridevnikov, pri katerih bi bilo v obravnavno sopomenskosti smiselno vključiti predložni morfem.

Za boljšo uporabniško izkušnjo (hitrejši in intuitivnejši vpogled v podatke) bi bilo treba zagotoviti **popolnejši izpis besedne zveze.** Glede na rezultate analize se zdi bolje izpisovati celotne kolokacije, ne le posameznih kolokatorjev, pri tem pa zagotoviti besednozvezno ustrezno prilagoditev, ki lahko upošteva kategorialne značilnosti, od katerih se kažejo kot najpomembnejše sklon, spol in določnost, mestoma tudi množina. Lematizirani prikaz v kombinaciji z ostalimi trenutnimi težavami namreč pogosto prinaša zmedeno kolokatorsko sliko, ki sama na sebi nima prave informativne vrednosti – torej zahteva zamudno klikanje in razdvoumljanje s korpusnimi zgledi in povezavami.

Uporabnost podatkov za primerjavo pomena sopomenk bi bilo mogoče izboljšati tako, da se pri izbiri kolokatorjev nekatere vrste podatkov **potisnejo na dno seznama kandidatov za luščenje.** V prvi vrsti se kaže takšna rešitev koristna za vse obravnavane zveze s prislovi, koristila pa bi tudi pri primerih, kjer se pojavljajo lastna imena (bodisi kot samostalniki ali kot svojilni pridevniki), kratice in deikti različnih besednih vrst.

Napačne lematizacije in šuma, ki nastane zaradi težav označevanja, v opazovanih podatkih ni veliko. Težave so dokaj predvidljive, npr. vezane na dvoumne oblike pri besedah *peti, delo, kos*, na drugi strani pa na lastna in kratična imena.

Za vse obravnavano gradivo velja, da so rezultati slabši pri parih, kjer je primerjani **pomen prenesen in redek** v primerjavi z osnovnim, npr. *konj / koza*, ki sta sopomenki v pomenu športnega rekvizita. Ker kolokatorji tu odražajo skoraj izključno osnovni pomen (žival), primerjava za namene opredeljevanja sopomenskosti nima pravega smisla. Podobno so le delno uporabni podatki pri parih, kjer je prekrivni sopomenski del v rabi redek pri eni od besed, npr. *kos / odlomek.* Tu so prekrivni kolokatorji zelo splošni, npr. *[kratek, izbran, lep] odlomek / kos,* razlike v pomenu pa se kažejo precej specifično, npr. *[majhen, drag, oblačilni] kos* v primerjavi s *[svetopisemski, prozni, prevedeni] odlomek.*

Pri vsem skupaj je treba upoštevati, da v prispevku ostajamo v domeni strojnega pridobivanja in urejanja podatkov. Pravi preskok v kvaliteti je seveda pogojen **z ročnim pregledom** najprej sopomenskega, nato kolokacijskega gradiva in zgledov. Nadaljnji koraki vključujejo odstranitev nerelevantnih sopomenskih kandidatov, pomensko členjenje besedišča in opredelitev sopomenskosti na ravni specifičnih pomenov ali načinov rabe besede. V strojnem smislu je naloga za prihodnje delo tudi adaptacija metodologije za obravnavo **večbesednih sopomenk,** ki so v slovar sicer zajete (npr. *dovoliti – izdati dovoljenje*; *doživeti – biti priča*), vendar kolokacije in zgledi zanje (še) niso na voljo. Nenazadnje pa je treba nadaljevati z zbiranjem podatkov o **slovarski rabi.** V uvodu prispevka omenjena uporabniška anketa (Arhar Holdt 2020) je pokazala, da sodelujoči v raziskavi kolokacije v Slovarju sopomenk sodobne slovenščine vidijo kot pomemben doprinos. Ker odzivni slovarji beležijo uporabniške aktivnosti v vmesniku, tudi klikanje na posamezne razdelke in povezave (Arhar Holdt idr. 2018: 407–408), bi bilo v nadaljevanju smiselno analizirati, v kolikšni meri in pri katerih iztočnicah so kolokacijski podatki najpogosteje v rabi.

Zahvala

V prispevku so opisani rezultati, ki so nastali v okviru projekta *Kolokacije kot temelj jezikovnega opisa: semantični in časovni vidiki* (J6-8255) ter programskih skupin P6-0411 – *Jezikovni viri in tehnologije za slovenski jezik* in P6-0215 – *Slovenski jezik – bazične, kontrastivne in aplikativne raziskave*, ki jih financira Javna agencija za raziskovalno dejavnost Republike Slovenije.

Reference

Arhar Holdt, Š. (2020): How Users Responded to a Responsive Dictionary: The Case of the Thesaurus of Modern Slovene. *Rasprave Instituta za hrvatski jezik i jezikoslovlje,* 46 (2): 465–482. doi: 10.31724/rihjj.46.2.1.

Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Gantar, P., Gorjanc, V., Klemenc, B., Kosem, I., Krek, S., Laskowski, C. A. in Robnik Šikonja, M. (2018):

Thesaurus of modern Slovene: by the community for the community. V J. Čibej, V. Gorjanc, I. Kosem in S. Krek (ur.): *Proceedings of the 18th EURALEX International Congress: Lexicography in global contexts*: 401–410. Ljubljana: Ljubljana University Press, Faculty of Arts. Dostopno prek: https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/3000-1 (12. 2. 2021).

Gantar, P. (2015): *Leksikografski opis slovenščine v digitalnem okolju*. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani. doi: 10.4312/9789612377922.

Gorjanc, V., Gantar, P., Kosem, I. in Krek, S. (ur.) (2017): *Slovar sodobne slovenščine: problemi in rešitve*. (1. elektronska izdaja). Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani. doi: 10.4312/9789612379759.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. in Suchomel, V. (2014): The Sketch Engine: ten years on. *Lexicography*, 1 (1): 7–36.

Kosem, I., Krek, S. in Gantar, P. (2020): Defining Collocation for Slovenian Lexical Resources. V I. Kosem in P. Gantar (ur.): Kolokacije v leksikografiji: trenutne rešitve in izzivi za prihodnost [tematska številka]. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave*: 8 (2): 1–27. doi: 10.4312/slo2.0.2020.2.1-27.

Kosem, I., Krek, S., Gantar, P., Arhar Holdt, Š., Čibej, J. in Laskowski, C. A. (2018): Collocations dictionary of modern Slovene. V J. Čibej, V. Gorjanc, I. Kosem in S. Krek (ur.): *Proceedings of the 18th EURALEX International Congress: Lexicography in global contexts*: 989–997. Ljubljana: Ljubljana University Press, Faculty of Arts. Dostopno prek: https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/3000-1 (12. 2. 2021).

Kosem, I., Husak, M. in McCarthy, D. (2011): GDEX for Slovene. V I. Kosem in K. Kosem (ur.): *Electronic lexicography in the 21st century: New applications for new users: Proceedings of eLex 2011*: 150–159. Ljubljana: Trojina, Institute for Applied Slovene Studies. Dostopno prek: http://www.trojina.si/elex2011/elex2011_proceedings.pdf (12. 2. 2021).

Kosem, I., Gantar, P., Krek, S., Arhar Holdt, Š., Čibej, J., Laskowski, C., Pori, E., Klemenc, B., Dobrovoljc, K., Gorjanc, V. in Ljubešić, N. (2019): *Collocations Dictionary of Modern Slovene KSSS 1.0*. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1250.

Krek, S., Arhar Holdt, Š., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešić, N., Kosem, I. in Dobrovoljc, K. (2020): Gigafida 2.0: the reference corpus of written standard Slovene. V N. Calzolari (ur.): *Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*: 3340–3345. Paris: ELRA – European Language Resources Association. Dostopno prek: http://www.lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf (12. 2. 2021).

Krek, S., Laskowski, C. A. in Robnik Šikonja, M. (2017): From translation equivalents to synonyms: creation of a Slovene thesaurus using word co-occurrence network analysis. V I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek in V. Baisa (ur.): *Electronic lexicography in the 21st century: proceedings of eLex 2017 Conference*: 93–109. Brno: Lexical Computing. Dostopno prek: https://elex.link/elex2017/proceedings/eLex_2017_Proceedings.pdf (12. 2. 2021).

Krek, S. in Kilgarriff, A. (2006): Slovene word sketches. V T. Erjavec in J. Žganec Gros (ur.): *Language technologies: proceedings of the 9th International Multiconference Information Society IS 2006*: 62–67. Ljubljana: Institut Jožef Stefan.

Krek, S., Laskowski, C., Robnik Šikonja, M., Kosem, I., Arhar Holdt, Š., Gantar, P., Čibej, J., Gorjanc, V., Klemenc, B. in Dobrovoljc, K. (2018): *Thesaurus of Modern Slovene 1.0*. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1166.

Logar, N., Grčar, M., Brakus, M., Erjavec, T., Arhar Holdt, Š. in Krek, S. (2012): *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba* (1. izd.). Ljubljana: Trojina, zavod za uporabno slovenistiko: Fakulteta za družbene vede.

Pori, E. in Kosem, I. (2018): V iskanju slovarsko relevantne kolokacije na primeru struktur s prislovi. V Š. Arhar Holdt, P. Gantar, V. Gorjanc in R. Grošelj (ur.): Slovnične raziskave za jezikovni opis [tematska številka]. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave*: 6 (2): 154–185. doi: 10.4312/slo2.0.2018.2.154-185.

Šorli, M., Grabnar, K., Krek, S. in Košir, T. (2006): Oxford-DZS comprehensive English-Slovenian dictionary. V E. Corino, C. Marello, C. Onesti in M. Alvar Ezquerra (ur.): *Proceedings of the XII EURALEX International Congress*: 631–637. Torino: Edizioni dell`Orso: Universita di Torino: Academia della Crusca.