Evalvacija avtomatskega luščenja kolokacijskih podatkov iz besednih skic v orodju Sketch Engine

Eva PORI
Filozofska fakulteta, Univerza v Ljubljani

Iztok KOSEM
Filozofska fakulteta, Univerza v Ljubljani; Institut Jožef Stefan

The paper presents the evaluation of automatic extraction of collocations from the reference corpus of Slovene, which was conducted in the research project Collocations as a basis of language description: semantic and temporal perspectives (KOLOS). The main aim was to identify advantages and shortcomings of existing automatic extraction methods using the POS-tagged corpora in the Sketch Engine tool. After conducting a pilot study using the crowdsourcing method to prepare the data annotation task, the qualitative linguistic analyses of collocations in different syntactic structures have provided essential information for the definition of collocation in terms of lexicographic resources for Slovene, and in terms of its distinction to other types of multiword units (compounds, phraseological units etc.). The main issues of the automatic extraction method leading to errors in collocation identification or collocation form were linked to corpus annotation processes (lemmatisation, POS-tagging) or post-processing steps, respectively. An important aspect in which the automatic extraction method can be improved are extended collocations (collocations of collocations), as the analysis revealed that semantically incomplete collocations are quite common, and even very typical for some headwords. On the semantic side, the analysis identified groups of lexicographically less relevant collocates, which are usually very frequent but also very general in use (are used with a large number of other headwords). In sum, the findings of the evaluation

will lead to improvements in automatic extraction of collocations, on the general and structure-specific level, and contribute to more systematic and informed inclusion of collocations in lexicographic resources for Slovene.

Keywords: automatic extraction of data, collocation, semantics, collocationality, Collocations Dictionary of Modern Slovene

1 Uvod

Nedavni trendi v leksikografiji največ pozornosti posvečajo prav avtomatizaciji tistih segmentov jezikovnega opisa, ki so povezani s kolokacijami in zgledi (prim. Kilgarriff in Rychlý 2010; Rundell in Kilgarriff 2011). Dosedanje raziskave prinašajo bistveno ugotovitev, da »avtomatizacija postopkov ne samo skrajša postopek leksikalne analize, ampak tudi izboljša njeno kakovost« (Cook idr. 2013: 50). Ravno na področju leksikografije je v zadnjih letih opazen napredek z vidika identifikacije kolokacij v slovenskem jeziku in izboljšav avtomatskih postopkov. Tu velja izpostaviti nadgrajene postopke za avtomatsko luščenje kolokacij in njihovih zgledov (gl. Gantar idr. 2015, 2016; Kosem idr. 2013), ki predstavljajo temeljni del izdelave Slovarja sodobnega slovenskega jezika (Krek idr. 2013; Gorjanc idr. 2015) in Kolokacijskega slovarja sodobne slovenščine (Kosem idr. 2018a). S pomočjo prilagoditev in izboljšav avtomatskega luščenja leksikalnih podatkov za slovenščino, metodologije API (Kosem idr. 2013), je mogoča učinkovitejša in kakovostnejša obravnava leksikalnih podatkov (izvoz kolokacij in korpusnih zgledov za določen seznam besed).

V slovenskem prostoru najdemo kar nekaj raziskav na temo korpusnega preučevanja leksikalnih enot za slovenščino (npr. Gantar in Krek 2011; Gantar idr. 2009; Kosem idr. 2013), vendar pa obsežnejša in celovita evalvacija različnih slovnično-pomenskih relacij (skladenjskih struktur), ki bi temeljila na avtomatsko izluščenih kombinacijah kolokatorjev, še ni bila opravljena. To dejstvo je pomembno tudi zato, ker je veliko znanega za angleščino, manj pa za morfološko bogate jezike, kot je slovenščina. Primanjkuje študij, ki

bi obravnavale (ne)učinkovitost različnih (korpusnih) metodologij in avtomatsko podprtih postopkov luščenja ob upoštevanju statističnih kriterijev, slovničnih in pomenskih lastnosti leksikalnih enot, študij, ki bi vzpostavile jasno ločnico med t. i. statistično kolokacijo (sopojavitvijo dveh besed oz. lem, ki je statistično pomembna) ter ponudile nastavke za opredelitev semantične kolokacije, kjer sopojavitev vzpostavlja sporočilno oz. jezikoslovno vrednost, tudi v razmerju do slovarske kolokacije, kjer je sopojavitev dovolj relevantna za vključitev v (kolokacijski) slovar.

Pričujoči prispevek zato želi narediti prve korake proti naslavljanju semantično pogojenih problemov, ki jih avtomatsko luščenje lahko reši ali pa tudi ne more rešiti. Na podlagi prikaza procesa evalvacije avtomatskega luščenja večbesednih leksikalnih enot iz korpusa, ki jo je v okviru raziskovalnega projekta *Kolokacije kot temelj jezikovnega opisa: semantični in časovni vidiki* (KOLOS; J6-8255) opravila skupina jezikoslovcev, se osredotoča na identifikacijo prednosti in slabosti uporabe obstoječih metod luščenja z orodjem Sketch Engine. Postopek evalvacije je bil tudi ključna osnova za samo opredelitev kolokacije za namene leksikalnih virov za slovenščino ter z vidika razmerja med kolokacijo in drugimi tipi besednih zvez (gl. Gantar idr. 2020; Kosem idr. 2020). Rezultati so pomembni tudi za izboljševanje metod označevanja in luščenja kolokacij ter hkrati za pohitritev postopka analize kolokacij za slovarske in druge namene.

2 Razvoj orodij za analizo kolokacij

Kolokacije so bile v zadnjih desetletjih deležne vse večje pozornosti različnih disciplin, od korpusnega jezikoslovja in leksikografije do računalniškega jezikoslovja in naravnega procesiranja jezika. To je na eni strani posledica vse večjih korpusov, saj je za temeljito proučevanje kolokacij potrebna dokaj velika količina besedil (gl. npr. prispevek Khokhlova in Benko 2020), na drugi strani pa so k proučevanju, razumevanju in popisovanju kolokacij pripomogla vse boljša orodja za njihovo analizo (za pregled orodij gl. Kilgarriff in Kosem 2012; Kosem 2016). Za jezikoslovce in leksikografe je bil še zlasti pomemben

prihod orodja Sketch Engine (Kilgarriff idr. 2004), ki je od svoje prve predstavitve redno prinašalo nove funkcije za analizo in opis jezika, zaradi česar se je uveljavilo kot vodilno korpusno orodje na svetu.

Najpomembnejša funkcija Sketch Engina za analizo kolokacij je Besedna skica (ang. Word Sketch) (Kilgarriff in Tugwell 2001; Kilgarriff idr. 2004), ki ponudi sliko tipične skladenjske in kolokacijske okolice besede, pri čemer so kolokacije razdeljene v skladenjske strukture.[1] Za izdelavo Besednih skic je potrebna slovnica besednih skic (ang. sketch grammar), ki vsebuje specifikacije oz. definicije slovničnih relacij, značilnih za posamezne strukture. V definiciji vsake slovnične relacije se opredeli, kateri besedni vrsti naj pripadajo kolokatorji v okolici besede oz. iztočnice, ki je predmet analize, koliko besed je lahko med iztočnico in kolokatorjem, katere besedne vrste se med iztočnico in kolokatorjem ne smejo pojavljati ipd.

Z Besednimi skicami so tesno povezani tudi številni mejniki v leksikografski metodologiji. Besedne skice so bile že takoj po vpeljavi uporabljene pri izdelavi angleškega slovarja za nematerne govorce založbe Macmillan (Rundell 2002) in se hitro uveljavile, saj Atkins in Rundell (2008: 111) navajata, da je »tak način leksikalnega profiliranja za mnoge leksikografe postal preferenčno izhodišče pri analizi kompleksnejših iztočnic«. Kot odgovor na intenzivno uporabo Besednih skic in potrebe po pospešitvi leksikografskega dela je nastala funkcija Tick-Box Lexicography (Kilgarriff idr. 2010) oz. slovensko kliksikografija (Gantar 2015), ki je omogočala hitro izbiranje in prenos kolokacij ter z njimi povezanih dobrih zgledov prek funkcije GDEX (ang. Good Dictionary Examples; Kilgarriff idr. 2008) iz Sketch Engina v slovarsko orodje. Kmalu po vpeljavi kliksikografije pa sta Rundell in Kilgarriff (2011) že predlagala naprednejši metodološki pristop k izdelavi slovarjev, ki izkorišča prednosti vseh omenjenih funkcij v Sketch Enginu, in sicer kombinacijo avtomatskega izvoza podatkov (kolokacij, zgledov, oznak ipd.) iz korpusa ter njihove validacije v slovarskem orodju.

Sketch Engine in z njim povezane metode so že od samega začetka uveljavljene tudi v slovenskem prostoru. Najprej je bila

1 Podobno funkcionalnost ponuja tudi DeepDict Lexifier (Bick 2009).

kombinacija analize besedne skice in pregleda naključnega izbora konkordanc (kar omenjata že Atkins in Rundell 2008) uporabljena pri izdelavi Leksikalne baze za slovenščino (Gantar idr. 2012; Gantar 2015; Gantar idr. 2016), na manjšem številu gesel pa se je preizkusilo tudi metodo avtomatskega izvoza podatkov in njihove validacije. Metodologija z uporabo avtomatskih postopkov je postala temeljni del Predloga za izdelavo Sodobnega slovarja slovenskega jezika (Krek idr. 2013) in iz njega izhajajočega koncepta Slovarja sodobnega slovenskega jezika (Gorjanc idr. 2015), v praksi se je uporabila tudi pri izdelavi specializiranih virov, kot je npr. ALEKS (Logar idr. 2019). V tem času so se slovenske različice slovnice besednih skic (Krek 2015) in konfiguracij GDEX (Kosem idr. 2011; Kosem idr. 2013; Kosem 2015) nenehno izboljševale, do mere, ko so bili avtomatski kolokacijski podatki smatrani kot dovolj dobri za neposredno predstavitev uporabnikom, v obliki Kolokacijskega slovarja sodobne slovenščine (KSSS; Kosem idr. 2018).

Priprava vsake nove verzije slovenskih različic funkcionalnosti v orodju Sketch Engine je bila podprta z evalvacijo vzorca podatkov in s povratno informacijo leksikografov. Prepoznava problematičnih delov besednih skic je tako na primer privedla do odločitve, da se določene skladenjske strukture, ki pri večini iztočnic vsebujejo veliko šuma, v KSSS ne vključijo (Kosem idr. 2018b). Pri tem je treba poudariti, da je bila slovnica besednih skic za avtomatsko luščenje kolokacijskih podatkov precej bogatejša v količini definiranih skladenjskih struktur od tiste, namenjene za ročno analizo besednih skic. Se je pa pri izdelavi KSSS pokazala potreba po sistematični evalvaciji metode avtomatskega luščenja podatkov iz korpusov, s katero bi odkrili probleme in rešitve tako na ravni avtomatskega luščenja in postprocesiranja kot na ravni izbire relevantnih kolokacij za vključitev v različne jezikovne vire. V nadaljevanju predstavljamo eksperiment evalvacije jezikoslovcev, v ločenih prispevkih pa smo opisali rezultate študij odnosa uporabnikov do avtomatsko izluščenih podatkov (Pori idr. 2020) in njihove predstavitve v slovarju (Pori idr. 2021).

3 Evalvacija avtomatsko izluščenih kolokacijskih podatkov

Glavni namen evalvacije je bil preveriti zanesljivost avtomatsko izluščenih kolokacijskih podatkov, vendar pa smo hkrati želeli odgovoriti še na druga vprašanja, povezana s postopki avtomatskega luščenja in opredelitve kolokacij:

- kateri so problemi avtomatskega luščenja na ravni prepoznavanja kolokacijskih kandidatov;
- kateri so problemi, povezani s postprocesiranjem izluščenih podatkov;
- katere strukture so kolokacijsko bolj obvestilne oz. slovarsko relevantne;
- kaj je slovarsko relevantna kolokacija oz. katere kolokacije so za leksikalne vire manj relevantne oziroma nerelevantne.

Pri evalvacijskih nalogah smo se poslužili tudi metod, ki jih sicer najdemo predvsem pri postopkih množičenja, pri čemer je glavni poudarek na tem, da je vsaka mikronaloga ločena enota, ki posamezniku ne sme vzeti veliko časa, dokončen pregled vseh rešenih mikronalog pa potem pokaže obseg medsebojnega ujemanja označevalcev ter tudi njihove interne doslednosti pri označevanju podatkov istega tipa, v našem primeru kolokacij določene skladenjske strukture.

Naloge ocenjevanja kolokacijskih kandidatov so se odvijale v odprtokodni platformi za množičenjske naloge Pybossa.[2] Pri vsaki nalogi so imeli označevalci na voljo kolokacijskega kandidata in njegov zgled, izluščen z orodjem GDEX za slovenščino (Kosem idr. 2011; Kosem idr. 2013; Kosem idr. 2015), ki med drugim skuša identificirati zglede, ki kolokacijo prikazujejo v čim bolj tipičnem kontekstu.

Iztočnice in njihove kolokacijske kandidate smo izbirali iz baze Kolokacijskega slovarja sodobne slovenščine (KSSS; Kosem idr. 2018a), ki je bila izdelana s takrat zadnjimi različicami vseh jezikovnih tehnologij (slovnica besednih skic, GDEX) za luščenje

2 https://pybossa.com

kolokacijskih podatkov iz korpusa. Posledično smo že izhodiščno vedeli, da evalvacijski nalogi ne bosta pokrili vseh možnih skladenjskih struktur besednih skic. Pri pripravi podatkov za KSSS so bile namreč na podlagi analize izbrane predvsem kolokacijsko obvestilnejše strukture, manj obvestilne strukture, predvsem strukture z veliko korpusnega šuma, pa so bile izločene, npr. struktura sbz_1 gbz (samostalnik v imenovalniku + glagol)[3]. Z vidika evalvacije je bilo to dejansko smiselno in tudi zaželeno, saj smo se želeli osredotočiti na probleme kolokacijskih podatkov.

3.1 Pilotna naloga

S pilotno nalogo smo želeli predvsem preveriti, ali lahko na podlagi ozkega nabora ponujenih odgovorov 'Da', 'Ne' in 'Ne vem' in osnovnih navodil, s katerimi so označevalci ocenjevali kolokacijske kandidate, pridemo do dovolj sistematičnih analiz zanesljivosti luščenja kolokacijskih podatkov in do jasnih opredelitev, kaj je slovarsko relevantna kolokacija.[4]

Šest označevalcev jezikoslovcev je pri ocenjevanju kolokacijskih kandidatov lahko izbiralo med ponujenimi možnostmi na seznamu oz. so imeli na voljo tri odgovore: 'Da', 'Ne', 'Ne vem'. Označevalcem je bila ponujena tudi podopcija odgovora 'Da', in sicer 'Da (slab zgled)', za katero naj bi se odločali v primerih, ko je bila kolokacija sicer legitimna, zgled pa neustrezen, predvsem zato, ker je bil nejasen oz. jezikovno slab ali pomensko premalo obvestilen. Označevalci so skupaj označili približno 8.800 kolokacijskih kandidatov v 226 različnih skladenjskih strukturah, pri čemer smo za vsakega od kolokacijskih kandidatov zahtevali po 3 odgovore, kar je pomenilo, da vsi

3 Pri navajanju skladenjskih struktur uporabljamo naslednji pristop zapisovanja: za besedne vrste uporabljamo okrajšani zapis, npr. oznake sbz (samostalnik), pbz (pridevnik), gbz (glagol), rbz (prislov) ipd. in podpisane številke, ki podajajo informacijo o sklonu (samostalnika in/ali pridevnika), npr. gbz + sbz_4 (glagol + samostalnik v tožilniku). Pri predložnih strukturah je naveden še predlog, npr. gbz na sbz_5 (glagol + predlog 'na' + samostalnik v mestniku). Okrajšani zapisi so povzeti po Leksikalni bazi za slovenščino in povezani literaturi (Gantar 2012, 2016), kjer so bile strukture prevedene iz formalizma v orodju Sketch Engine.

4 Na tej točki smo bili tudi še odprti za možnost množičenja kolokacij med širšo javnostjo, če bi pilotna raziskava pokazala potencial za to.

označevalci niso označili vseh kandidatov. Ujemanje označevalcev je bilo v razponu 42–76 %, v povprečju 62 % kolokacijskih kandidatov pa sta se v odgovoru strinjala dva označevalca, Cohenova kapa je bila 0,35, kar pomeni srednje ujemanje. Pokazale so se že prve razlike med različnimi strukturami, tj. pri nekaterih strukturah so se označevalci precej bolj strinjali o tem, kaj je oziroma ni slovarsko relevantna kolokacija, kot pa pri drugih.

Po nalogi smo poleg analize podatkov opravili tudi razgovore z označevalci, ki so opozorili na različne pomanjkljivosti pristopa oz. naloge, izpostavljene pa so bile predvsem sledeče:

- premajhen nabor potencialnih odgovorov glede na obliko podatkov. Na odločitve označevalcev o legitimnosti kolokacije je namreč vplivala sama oblika, ki včasih ni ustrezala prevladujoči obliki, podani tudi v zgledu, npr. kolokator ni bil v množini.
- premajhna heterogenost iztočnic na račun širokega nabora skladenjskih struktur. Posledično ni bilo znano, kakšen vpliv imajo na opredeljevanje kolokacije različne lastnosti iztočnic, kot so večpomenskost, povratnost ipd.
- vprašljivost vloge navodil. Označevalci so dobili osnovna navodila, ki so vključevala tudi opredelitev kolokacije, a so komentirali, da bi bilo dejansko bolje označevati brez njih, na podlagi lastnih znanj in predstav o kolokacijah, ter se usklajevati kasneje.
- vsi podatki pomešani v eni nalogi. Označevalci so opozorili, da so morali biti zelo pozorni na preskoke na novo strukturo, ker informacija o strukturi ni bila nikjer eksplicirana.
- Nekatere kolokacije s slabimi zgledi so lahko delovale kot povsem ustrezne, vendar pa zgled ni potrjeval njihove rabe, npr. **zelo ljubiti -> je bil zelo sposoben ljubiti* (prislov določa sledeči pridevnik in ne glagola – *zelo sposoben (ljubiti)*), podobno še: **komentirati nedavno -> je komentiral nedavno sprejeti zakon* (prislov določa sledeči pridevnik in ne glagola – *nedavno sprejeti (zakon)*).[5]

5 V takšnih primerih, ki so bili resda redki, je bilo vedno vprašanje, ali je zgled predstavnik večine rab, pri čemer je šlo potem za nepravilno prepoznano kolokacijo, ali pa je bil zgled zgolj ena redkih nepravilno prepoznanih rab od številnih ustreznih.

Na podlagi povratnih informacij smo pripravili glavno evalvacijsko nalogo.

3.2 Glavna evalvacijska naloga

V izhodišču smo pri glavni evalvacijski nalogi posvetili več pozornosti pripravi nabora iztočnic, in sicer smo za zagotovitev večje reprezentativnosti in heterogenosti pri izbiri iztočnic uporabili različne kriterije (npr. besedna vrsta, večpomenskost, izvor, (ne)številnost, pogostost v korpusu Gigafida ipd.). Končni vzorec je vseboval 333 iztočnic, od tega 154 samostalnikov, 73 glagolov, 81 pridevnikov in 25 prislovov). S heterogenim naborom iztočnic smo želeli identificirati čim več problematičnih mest, saj so večje količine podatkov koristne za različne analize (opredelitev semantične kolokacije, gručenje ipd.), zlasti pa za preizkušanje metod, kot je distribucijska semantika. Iz nabora kolokacij smo izločili tiste, ki smo jih že ocenili v pilotni nalogi, ali pa so bile zabeležene v Leksikalni bazi za slovenščino.

Ocenjevanje kolokacijskih kandidatov se je ponovno odvijalo v platformi Pybossa, vendar tokrat niso bile vse strukture zajete v eni nalogi, pač pa je bila za vsako strukturo pripravljena ločena naloga. Poudarek novega eksperimenta je bil predvsem na tem, da je ocenjevanje kolokacijskih kandidatov temeljilo na lastnem pojmovanju kolokacije in da se kolokativnost (tako temeljno kot slovarsko) opredeli na podlagi analize rezultatov.

Sedem označevalcev jezikoslovcev je še vedno izbiralo med 3 krovnimi odgovori ('Da', 'Ne', 'Ne vem'), a so jim bile ponujene podopcije:

- 'Množina' (podopcija 'Da') za primere, ko je bila kolokacija sicer legitimna, a bi bila ustreznejša množinska oblika kolokatorja; npr. **tihotapljena cigareta -> tihotapljene cigarete*.
- 'Si/Se' (podopcija 'Da') pri glagolskih strukturah, ko je (v kolokacijskem kandidatu manjkajoči) povratni osebni ali svojilni zaimek obvezen, npr. **ogledati prestolnico -> ogledati si prestolnico*.
- 'Največji' (podopcija 'Da') pri pridevnikih in prislovih, ki so v kolokaciji vedno v primerniški ali presežniški obliki, npr. **znatno lahek -> znatno lažji*.

- 'Razširjena kolokacija' (podopcija 'Da'), ki ob sebi predvideva dodaten element; npr. **dnevno brezplačno -> 4-krat dnevno brezplačno*.
- 'Zgled Ne-Kolokacija Morda' za primere, ko zgled ne potrjuje kolokacije, čeprav je sama kolokacija videti povsem legitimna, npr. *doktorski študent -> na doktorski (stopnji) pa 15 študentov*.
- 'Fraze', ko ne gre za kolokacijo, ampak za del fraze, npr. **ne mešati jabolk -> ne mešati jabolk in hrušk.*
- 'Struktura' (podopcija 'Ne'), za primere, kjer je šlo za napako pri oblikoskladenjskem označevanju korpusa (npr. prekrivnost prislova s pridevniško obliko: *medtem ko je grobo mleti sladkor najboljši*).

Skupno je bilo ocenjenih 17.576 kolokacijskih kandidatov v 143 različnih skladenjskih strukturah.

4 Rezultati

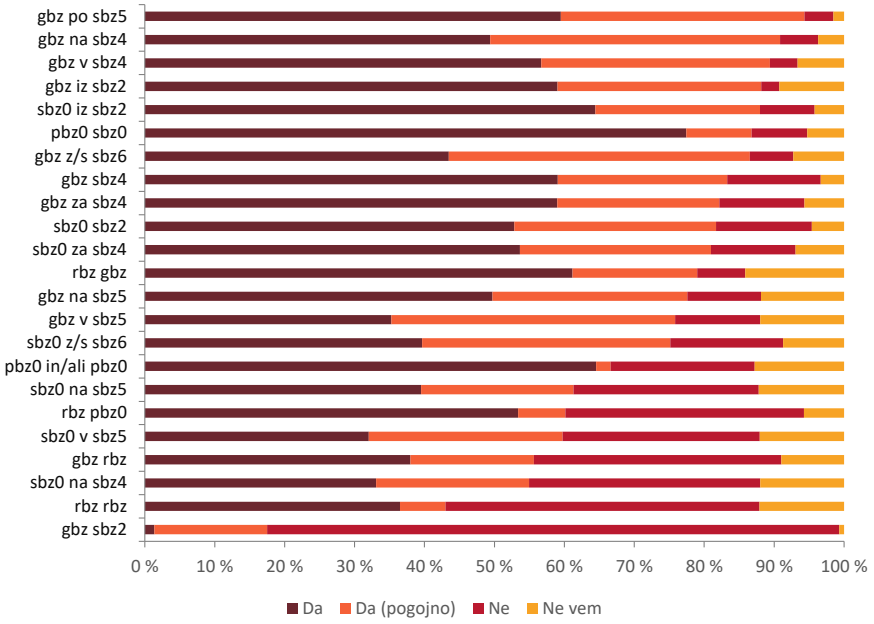
Razporeditev odgovorov označevalcev glede na skladenjsko strukturo prikazuje Slika 1 (prikazanih je 23 struktur z največ kolokacijskimi kandidati).

Strukture z največjim deležem 'Da' (vključno s podopcijami) so bile:

- glagol + po + samostalnik v mestniku (gbz po sbz_5): *poseči po cigareti*;
- glagol + na + samostalnik v tožilniku (gbz na sbz_4): *plezati na jambor*;
- glagol + v + samostalnik v tožilniku (gbz v sbz_4): *prevesti v francoščino*;
- pridevnik + samostalnik (pbz_0 sbz_0): *televizijska cenzura*.

Strukture z največjim deležem 'Ne' in 'Ne vem' pa so bile:

- glagol + samostalnik v rodilniku (gbz sbz_2): *primanjkovati goveda*, *angažirati izvedenca* (tožilnik, ne rodilnik);
- prislov + glagol (rbz gbz): *kako odrezati; zato angažirati;*
- glagol + prislov (gbz rbz): *boleti enako, prebiti tam;*



Slika 1: Prikaz deležev odgovorov označevalcev glede na skladenjsko strukturo kolokacije.

- prislov + prislov (rbz rbz)*: kdaj zboleti*, *treba angažirati*;
- prislov + pridevnik (rbz pbz_0): *bolj debel: vse bolj in bolj debelo plast zraka; bolj ekološki: (je) postal še bolj ekološki in še varčnejši, dnevno sklenjen (promet)*.

Analiza je, kot pri pilotni nalogi, izpostavila različne ravni ujemanja med označevalci glede na skladenjsko strukturo, tj. pri nekaterih strukturah so bila razhajanja precej večja, kar je nakazovalo na njihovo problematičnost z vidika opredeljevanja kolokativnosti. Tabela 1 kaže podatke za deset skladenjskih struktur z največ kolokacijskimi kandidati na ravni strinjanja, deleža kolokacij, pri katerih so se vsi trije označevalci strinjali v odgovoru, ter deleža razhajanj, kjer so upoštevani kolokacijski kandidati, pri katerih sta bila vsaj dva od treh odgovorov označevalcev nasprotujoča ('Da' in 'Ne') ali pa sta bila dva od treh odgovorov 'Ne vem'. Vidimo lahko (Tabela 1), da so deleži razhajanj višji pri strukturah s predlogi, v nekoliko manjši meri pa tudi pri strukturah s prislovi.

Tabela 1: Prikaz ujemanja oz. razhajanj odgovorov označevalcev glede na skladenjsko strukturo (prvih deset struktur po številu kolokacijskih kandidatov).

Struktura	Oznaka strukture	Ujemanje (Cohenova kapa)	Delež kolokacij s popolnim ujemanjem	Delež razhajanj (vsaj en 'Da' in en 'Ne' ali dva 'Ne vem')
pridevnik + samostalnik	pbz_0 sbz_0	0,42	78 %	11,8 %
samostalnik + samostalnik v rodilniku	sbz_0 sbz_2	0,45	63 %	13,3 %
glagol + samostalnik v tožilniku	gbz sbz_4	0,46	73 %	16,1 %
prislov + glagol	rbz gbz	0,37	63 %	19,5 %
prislov + pridevnik	rbz pbz_0	0,46	64 %	15,9 %
samostalnik + predlog 'v' + samostalnik v mestniku	sbz_0 v sbz_5	0,35	50 %	39,0 %
glagol + prislov	gbz rbz	0,47	61 %	19,4 %
glagol + predlog 'v' + samostalnik v mestniku	gbz v sbz_5	0,33	46 %	26,4 %
samostalnik + predlog 'z' + samostalnik v orodniku	sbz_0 z sbz_6	0,42	56 %	23,3 %
glagol + predlog 'z' + samostalnik v orodniku	sbz_0 z sbz_6	0,46	61 %	10,2 %

Strukture s prislovi so zaradi relativno visokih deležev odgovorov 'Ne vem' na eni strani in dokaj visokih deležev razhajanj (a vseeno ne previsokih) v odgovorih označevalcev na drugi predstavljale zelo dobro testno množico za detekcijo (ne)učinkovitosti metod luščenja in obsežnejši poskus opredeljevanja semantične (slovarske) kolokacije, ki sta ga izvedla Pori in Kosem (2018). Po modelu evalviranja kolokacijskih struktur s prislovi pa so bile v nadaljevanju izvedene tudi analize struktur s samostalniki, pridevniki, glagoli, ter ločeno analize večje skupine t. i. predložnih struktur. Zlasti za odgovore 'Struktura'

(podopcije znotraj krovnih opredelitev 'Da', 'Ne', 'Ne vem') velja izpostaviti, da smo jih pri analizah vseh struktur želeli beležiti ločeno od ostalih odgovorov 'Ne', saj gre za napake označevanja, ki so relevantne za nadaljnja prizadevanja izboljševanja oblikoskladenjskih označevalnikov besedil. Podobno velja za odgovore 'Fraze', saj so ti kolokacijski kandidati mogoče relevantni za pripravo postopkov za detekcijo (krajših) frazeoloških enot.

Pregled rezultatov analiz vseh omenjenih kolokacijskih struktur je predvsem razmejil primere, ki prinašajo manj relevantne rezultate, od tistih, ki z vidika opredeljevanja kolokativnosti predstavljajo kompleksnejša ter skladenjsko (ali semantično) bolj problematična mesta, potrebna natančnejše jezikoslovne diskusije. Na podlagi vseh analitičnih prijemov se je pokazalo, katere vrste oz. skupine kolokatorjev so (ne)problematične, nadalje pa predvsem, v kolikšni meri in kdaj so z vidika vključevanja v slovar (ne)problematične posamezne strukture.

4.1 Problemi avtomatskega luščenja

Detektirali smo že znane probleme avtomatskega luščenja, ki so bili popisani že pri preteklih luščenjih bigramov in trigramov (Arhar Holdt 2011; Gorjanc idr. 2015) in se nanašajo na tipično skladenjsko ali besedilno obnašanje leme v korpusu, npr. sopojavljanje z (iz) lastnoimenskimi ali količinskimi poimenovanji, z izrazi v množinski/dvojinski ali zanikani obliki, prevladujoča raba tretjeosebne oblike glagola ali pojavljanje (obveznega/neobveznega) prostega morfema si/se v glagolskih strukturah.

Generična ocena predstavljenih ugotovitev analiz posameznih struktur je izpostavila, da je napak, ki so nastale pri oblikoskladenjskem označevanju korpusa, več vrst in so se pojavljale na različnih ravneh (odvisno od iztočnice, torej samostalniške, pridevniške, glagolske, prislovne): (a) na ravni besede (lematizacija, lastno ime, določena oblika kolokatorja); (b) na ravni skladenjske strukture (napačna besedna vrsta, napačen sklon (im.–tož.), zanikanje) in (c) na ravni celotne kolokacije (kolokacija kot sestavni del ali v celoti del

stalne besedne zveze, skladenjske zveze ali frazeološke enote, razširjena kolokacija).

4.1.1 Problemi, ki izhajajo iz korpusnih podatkov

Avtomatsko luščenje podatkov in s tem povezane funkcije korpusnega orodja, kot je slovnica besednih skic, so odvisni od natančnosti postopkov označevanja korpusnih podatkov, v našem primeru lematizacije in pripisa oblikoskladenjskih oznak. Podatki za KSSS, ki so bili predmet naše analize, so bili izluščeni iz korpusa Gigafida 1.0 (Logar Berginc idr. 2012), ki je bil avtomatsko označen na podlagi smernic JOS, natančnost označevanja na ravni leme je dosegla 97,88 %, na ravni oblikoskladenjskih oznak pa 91,34 % (Grčar idr. 2012). Ob tem je treba poudariti, da gre pri podatkih o natančnosti za povprečji – dejanska slika je taka, da je pri številnih lemah natančnost še precej višja oz. celo 100 %, po drugi strani pa so določene leme oz. skloni še posebej problematični, zlasti tisti, kjer prihaja do prekrivnosti oblike (npr. samostalnika *del* in *delo*; rodilnik in tožilnik živih moških samostalnikov).

Številni kolokacijski kandidati so bili tako pri evalvaciji prepoznani kot napake lematizacije zaradi napačne besedne vrste kolokatorja oz. prekrivnosti enakopisnih oblik samostalnikov, pridevnikov, prislovov ali glagolov z drugimi besednimi vrstami (najpogosteje s pridevniki, tudi s samostalniki; najpogosteje se je namesto pridevnika pojavljal glagol, pri prislovnih strukturah pa smo zaznali tudi prekrivnost zaimkov in prislovov): (rbz pbz_0) **pravo tekmovalen -> pravo tekmovalno vzdušje*; **skrbno aluminijast -> skrbno oblikovanih aluminijastih reber*; (rbz gbz) **premagati zelo -> premagati zlo*. Pri teh napakah se je pokazala visoka raven ujemanja označevalcev, v večini primerov so izbrali možnosti 'Ne' ali 'Ne-Struktura'.

Podobno smo pri evalvaciji prepoznali številne nerelevantne kolokacijske kandidate, ki so izhajali iz neustreznih oblikoskladenjskih oznak (Tabela 2). Med pogostejšimi, tudi zaradi pogostosti strukture, so bili primeri zamenjave imenovalnika s tožilnikom, npr. (gbz sbz_4) **zboleti ovco -> ovca zboli*; **gnezditi lastovke -> lastovke gnezdijo*.

Velika verjetnost napak se je pokazala predvsem pri glagolih stanja ali premikanja, npr. *bivati, dati, odrasti, prestajati, ravnati, smučati*; celoten nabor problematičnih (neprehodnih) glagolov pa je težko pridobiti zgolj z avtomatsko metodo luščenja, saj meja med prehodnimi in neprehodnimi glagoli ni vedno jasna. Posamezni glagoli so lahko problematični le deloma, ker so neprehodni samo v enem od svojih pomenov; kar nekaj prehodnih glagolov *(čutiti ljubezen*, *dojiti otroka*, *parkirati kolo* ipd.*)* pa lahko v določenih primerih prehodnost izgubi oz. je predmetno mesto prazno (*Trenutno parkiram (kolo), te pokličem nazaj!*), kar pa terja ročni pregled kolokacij.

4.1.2 Problemi prepoznave skladenjskih struktur

Na podlagi preteklih študij in evalvacij smo pričakovali tudi določen delež napak zaradi napačno prepoznane strukture (Tabela 2), kljub temu da so bile številne bolj problematične strukture izključene iz luščenja podatkovne množice. Tako smo kot neustrezne, na podlagi največkrat enotne ocene označevalcev ('Ne' oz. 'Struktura' – niso ustrezne zaradi (napak) strukture), identificirali kolokacijske kandidate, ko je šlo za napačno prepoznano strukturno razmerje med posameznimi kolokacijskimi elementi, pri čemer je šlo večinoma za napačno nanašalnost pridevnika na nepravi samostalnik ali pa prislova na nepravi pridevnik: **nagubano oblačilo -> nagubano blago (je nagubano blago oblačil poudarjalo njihovo držo)*; **uspešno doktorski -> uspešno zaključen (v primeru uspešno zaključenega doktorskega študija)*; (podobno še: **dobro kolesarski, *premalo učiteljski*).

Napake oblikoskladenjskega označevanja smo zasledili tudi pri kolokacijskih kandidatih prislovnih struktur, in sicer je šlo za primere, ki izkazujejo tipično povedkovnodoločilno rabo pridevnikov in ne prislovov, oz. pridevniške oblike, ki se prekrivajo z osnovno prislovno obliko (npr. (rbz in/ali rbz) *(spremno besedilo) je karseda *kratko in preprosto*; *ni tako *silovito in strumno*). Pojavljali pa so se tudi primeri, ko prislovi določajo, modificirajo pridevnike/deležnike, pri katerih je treba ločevati navadne prislovne zveze od zloženk dveh

pridevnikov; (rbz in/ali rbz) *mešanica *dolgo in kratko delujočega inzulina.*

Problem na ravni skladenjske strukture so predstavljale tudi zveze besed, kjer je prihajalo do zamenjave pridevnika v vlogi povedkovega določila s pridevnikom v vlogi prilastka, npr. **priložena miška -> miška je priložena*; **kriv hormon -> hormoni so krivi*.

Poseben problem so predstavljale prekrivne strukture oz. strukture, kjer je ena vsebovala tudi podatke druge. Primer so strukture z glagoli, ko kolokacijski kandidati nezanikane strukture temeljijo tudi na zgledih, ki vsebujejo zanikano obliko, npr. **moči brez alkohola -> ne moči brez alkohola.* V takšnih strukturah so sicer kolokacije večinoma videti neproblematične, vendar pa lahko pride do podvajanj in varljivih podatkov o njihovi statistični relevantnosti (frekvenci oz. jakosti).

4.1.3 Problemi postprocesiranja

Precej problematičnih mest, odkritih tudi že med pilotno evalvacijo, je bilo povezanih s postopkom postprocesiranja avtomatsko izluščenih podatkov. Postprocesiranje je bilo potrebno, ker so bile vse kolokacije v Besedni skici izluščene z osnovnimi oblikami tako iztočnice kot kolokatorja. Postopek postprocesiranja je vseboval sledeče korake:

- Vsakemu delu kolokacije je bila pripisana ustrezna oblika glede na zahteve strukture (npr. sklon samostalnika) ali lastnosti iztočnice (npr. spol samostalnika; **velik hiša -> velika hiša*), pri čemer se je za pripisovanje oblik uporabil Slovenski oblikoslovni leksikon Sloleks 1.0 (Dobrovoljc idr. 2013).
- Na podlagi izluščenih zgledov se je pripisala tudi prevladujoča oblika, a samo v primerih male oz. velike začetnice ter vrstnega reda iztočnice in kolokatorja v prirednih strukturah.
- Odstranjene so bile kolokacije, ki so imele vsaj 4 od 5 zgledov povsem identičnih, saj so bile obravnavane kot nerelevantne zaradi zavajajoče statistike. Analiza je pokazala, da je šlo v skoraj vseh primerih za redkejše kolokacije.

Problemi, ki so se zaradi izsledkov pilotne naloge označevali ločeno, kot še vedno legitimne kolokacije z določeno pomanjkljivostjo na ravni oblike, so vključevali:

- kolokacijske kandidate z množinsko obliko besed, pri katerih en del kolokacije predstavlja jedrni kolokator, ki ob sebi predvideva množinsko obliko desnega dopolnila, v množinski obliki pa se lahko pojavljata tudi oba dela kolokacije: **množica emigranta -> množica emigrantov*; **mreža satelita -> mreža satelitov*; **tovarna dežnika -> tovarna dežnikov*; **slika satelita -> slike satelitov;*
- kolokacijske kandidate s pridevniškim ali prislovnim delom v določeni stopnji (primerniku ali presežniku): **kasno nadgraditi -> kasneje nadgraditi*; **natančno povedan -> natančneje povedano*; **blizko želen -> bližje želenemu*; **verjetno odpustiti -> najverjetneje odpustiti*;
- kolokacijske kandidate, ki zahtevajo enega od delov v množini ali dvojini: **zboleti na dihalu -> zboleti na dihalih*; **različna aplikacija -> različne aplikacije*; **zavojček bonbona -> zavojček bonbonov*;
- kolokacijske kandidate v strukturah z glagoli, kjer je obvezen glagolski element morfem (in ne povratni zaimek) *se* ali *si*: **zdeti v redu -> zdeti se v redu.*

Po pričakovanjih je označevanje izpostavilo probleme pri kolokacijah, ki so vsebovale homonimne dele, torej besede, ki so lahko pripadale več kot eni iztočnici v Sloleksu. Najbolj problematični so bili primeri pri iztočnicah z različno sklanjatveno paradigmo, npr. *klòp* – samostalnik moškega spola, *klóp* – samostalnik ženskega spola: **greti klôpa -> greti klóp*; **guliti klôpa -> guliti klóp*; **sedeti v klôpu -> sedeti v klópi*.

Problem, ki je bil tudi pričakovan, so bili kolokatorji (ne pa tudi iztočnice, saj so bile vse v Sloleksu), ki so bili označeni kot neustrezni zaradi pomanjkanja podatkov za postprocesiranje oz. odsotnosti iztočnice v Sloleksu. Problem se je sicer nanašal predvsem na strukture, v katerih je bil kolokator v sklonu, ki ni bil imenovalnik.

Tabela 2: Tipi najpogostejših oblikoskladenjskih napak s primeri po strukturah.

Tip napake	Primeri po strukturah
napačna lematizacija (neustrezna osnovna oblika kolokatorja)	(sbz_0 sbz_2) **plata piva -> plato piva; *palček cimeta -> palčka cimeta; *parti pokra -> partija pokra (sleparji najdejo bogato "tarčo", ki nasede partiji pokra)*
	(rbz pbz_0) **doma ostarel -> dom ostarelih: oskrbovanci bližnjega doma ostarelih; *doma star -> dom starejših: prostore negovalnega dela doma starejših občanov*
	(gbz rbz) **premagati zelo -> premagati zlo*
	(gbz sbz_4) **piliti alkohol -> piti alkohol; *premagati francoz -> premagati Francoza*
napačna besedna vrsta kolokatorja	(rbz pbz_0) **pravo tekmovalen -> pravo tekmovalno vzdušje; *skrbno aluminijast -> skrbno oblikovanih aluminijastih reber*
	(sbz_0 sbz_2) **pivo pite -> pivo piti; *greda stvari -> gredo stvari*
	(pbz_0 iz sbz_2) **težek iz ust -> najtežje iz ust; *težek iz razloga -> težko iz razloga*
zamenjava imenovalnika s tožilnikom	(gbz sbz_4) **zboleti ovco -> ovca zboli; *gnezditi lastovke -> lastovke gnezdijo; *leteti perje -> perje leti; *absorbirati telo -> telo absorbira*
zamenjava rodilnika s tožilnikom	(gbz sbz_4) **primanjkovati surovino -> primanjkovati surovine; *primanjkovati romantika -> primanjkovati romantike*
napačno strukturno razmerje med kolokatorji	(pbz_0 sbz_0) **nagubano oblačilo -> nagubano blago (je nagubano blago oblačil poudarjalo njihovo držo)*
	(rbz pbz_0) **uspešno doktorski -> uspešno zaključen (v primeru uspešno zaključenega doktorskega študija)*
	(sbz_0 sbz_2) **vrtnica plezalka: nakup lilij in vrtnic plezalk*
prekrivnost pridevniške oblike z osnovno prislovno obliko	(rbz in/ali rbz) *(spremno besedilo) je karseda *kratko in preprosto; ni tako *silovito in strumno*
napačna (trdilna) oblika glagolskega kolokatorja namesto nikalne	(gbz sbz_2) **piti piva -> ne piti piva*
	(gbz brez sbz_2) **moči brez alkohola -> ne moči brez alkohola*

Tip napake	Primeri po strukturah
kolokacija v množini (tudi kot del razširjene kolokacije)	(sbz_0 sbz_2) **množica emigranta -> množica emigrantov; *mreža satelita -> mreža satelitov; *tovarna dežnika -> tovarna dežnikov*
	(sbz_0 sbz_2) **zakladnica plašča -> najbogatejše zakladnice mašnih plaščev; *proizvodnja plašča -> proizvodnjo avtomobilskih plaščev in zračnic*
zamenjava pridevnika v vlogi povedkovega določila s pridevnikom v vlogi prilastka	(pbz_0 sbz_0) **priložena miška -> miška je priložena; *kriv hormon -> hormoni so krivi*
homonimi z različno sklanjatveno paradigmo	(gbz sbz_4) **guliti klôpa -> guliti klóp;* (gbz v sbz5) **sedeti v klôpu -> sedeti v klópi*
	(gbz iz sbz_2) **poganjati iz prsta -> poganjati iz prsti;* (gbz v sbz5) **rasti v prstu -> rasti v prsti*
neobvestilnost kolokacije brez manjkajočega elementa	(sbz_0 sbz_2) **informacija značaja -> informacija javnega značaja*
	(sbz_0 z/s sbz_6) **žeja s pivom -> pogasiti žejo s pivom*

4.2 Opredeljevanje slovarsko relevantnih kolokacij

V nadaljevanju se posvečamo kolokacijskim kandidatom, ki so bili prepoznani kot dobri oz. pri evalvaciji označeni z 'Da', kar vključuje tudi oblikovno problematične kolokacijske kandidate iz razdelka 4.1.3. Polnopomenski samostalniški, pridevniški, prislovni in glagolski kolokatorji so bili prepoznani kot semantično smiselni (Kosem idr. 2020) in posledično neproblematični pri vseh obravnavanih strukturah (Tabela 3).

Tabela 3: Prikaz polnopomenskih kolokatorjev glede na posamezne skupine besed (samostalniki, pridevniki, glagoli, prislovi).

samostalniki	– količinski: *ščep, žlica, skodelica (cimeta)* – nekoličinski (del – celota)*: cvet (cvetače); prebivalka (prestolnice)* – izglagolski: *česanje (perja), čiščenje (podstrešja), izdelovanje (venčka)*
pridevniki	– lastnostni:[6] *rdeča (jagoda), rožnate (hlačke); majhna, mala (muca)* – intenzifikator: *droben, hud, izdaten (dež)* – izlastnoimenski (ki so ključni za pomensko členitev): tip *angleška (krona) // švedska (krona)*
glagoli	– dovršniki in nedovršniki: *angažirati (brata), barvati (svilo); poplaviti (njivo), zviti (cigareto)*
prislovi	– lastnostni: *brezplačno (prejeti), natančno (analizirati)* – intenzifikator: *blazno, pošteno, močno, kar (boleti)*

Se je pa že med evalvacijo in tudi na podlagi same analize pokazalo, da obstajajo razlike med označevanjem semantične smiselnosti kolokacije in njene slovarske relevantnosti. Namreč, medtem ko je bila določena kolokacija lahko prepoznana kot statistično, skladenjsko ustrezna in semantično smiselna, so se pojavili dvomi o njeni vključitvi v različne slovarske vire, v našem primeru predvsem v Kolokacijski slovar sodobne slovenščine. Izhodišče za razpravo o slovarski relevantnosti samostalnikov, pridevnikov, glagolov in prislovov kot kolokatorjev so tako predstavljale skupine kolokacijskih kandidatov, pri katerih je bilo identificiranih največ razhajanj v odločitvah označevalcev ('Da', 'Ne', 'Ne vem'):

- kolokacije s pomensko manj obvestilnimi kolokatorji: *posamezna [etaža]*, *podoben [profil]*, *omenjen [sindikat]*, *določena [aplikacija]*; *večinoma [doma]*; *tako [boleti]*; *tukaj [gnezditi]*;
- kolokacije z razširitvenimi elementi, levimi ali desnimi dopolnili samostalnika (tipično: pridevniki pred samostalniki, redkeje tudi zaimki ali členki): *posnemati [računalniško] miško*; *vzeti [sončna] očala*; *barvati [vaš] vsakdanjik*; *ubiti [tega] mačka*; *popravljati [tudi] dežnike*;

6 Tisti lastnostni pridevniki, pri katerih je nabor možnosti znotraj semantičnega tipa barv omejen in ne predstavljajo le ene od vseh možnih barvnih realizacij. Slednje namreč, ki se lahko razvrščajo ob katerokoli kolokacijsko jedro, obravnavamo v razdelku pomensko manj obvestilnih kolokatorjev oz. pridevnikov (4.2.1).

- kolokacije z razširitvenimi zvezami (pridevnika in samostalnika), ki se s kolokatorjem vežejo v prireden zvezo (najdemo jih tudi med razširjenimi): *najdemo zrele plodove in cvetove*;
- kolokacije z lastnoimenskimi kolokatorji in kolokatorji, ki ob sebi predvidevajo odprti naštevalni niz desnih (lastnoimenskih) dopolnil samostalnika ali glagola: *prevod [Aleša, Alje, Kajetana]*; *spremljevalka [Brada]*; *avenija mode ->* (*trgovina) Avenija mode*; tip *okupirati [Evropo, Nemčijo]*, *prevajati [Danteja, Platona]*; tudi tip *selekcija [Slovenije, Avstrije].*[7]

4.2.1 Kolokacijski kandidati s pomensko manj obvestilnimi kolokatorji

Kot pomensko manj obvestilne kolokatorje smo opredelili predvsem tiste, ki nimajo predmetnega pomena oz. je njihov pomen zelo splošen in se kot tak sopojavlja z velikim številom iztočnic.[8] Posledično takšni kolokatorji ne tvorijo slovarsko relevantnih kolokacij. Tako nas pri odločanju, ali je kolokacija slovarsko relevantna, zanima predvsem, kakšen je doprinos kolokatorja k pomenski vrednosti iztočnice, tudi v primerjavi z drugimi podobnimi iztočnicami. Kolokator mora imeti dodano vrednost, ki se izraža predvsem v tipičnosti oz. edinstvenosti. Drugače povedano, bolj omejen je semantični niz kolokatorjev oz. bolj omejen je niz iztočnic, na katere se določeni kolokator veže, bližje je zveza slovarsko relevantni kolokaciji. Če pa je kolokator znotraj semantičnega tipa le eden v nizu mnogih, se poveča verjetnost, da ne gre za slovarsko relevantno kolokacijo. Dober ponazarjalni primer so recimo lastnostni pridevniki, ki označujejo barvo in se navadno lahko razvrščajo ob katerokoli jedro ter predstavljajo le eno od vseh možnih barvnih realizacij, npr. *rdeča [hiša, skodelica, roža]* ali *rdeč [avto, stol, plašč]*. Kolokacija *rdeča* ***hiša*** tako ni slovarsko relevantna, drugače je pri *rdeča* ***jagoda***, kjer se kaže, da je nabor možnosti znotraj semantičnega tipa barv bolj omejen.

7 Te zveze podrobneje in posebej obravnavamo na semantični ravni (glej razdelek 4.2).
8 Glej tudi razdelek o pomensko oslabljenih glagolih v Gantar 2020.

V nadaljevanju navajamo nekaj pri evalvaciji zaznanih tipičnih primerov potencialno pomensko manj obvestilnih kolokatorjev, ki jih zaradi širokega nabora kolokacijskih jeder, ob katera se razvrščajo, v večini kolokacij ne moremo obravnavati kot slovarsko relevantne:

- pridevniki; predvsem splošni (lastnostni, deikti) in deležniški: preostal (*preostal cimet*); nadaljnji, naslednji, sledeči (*nadaljnji dovoz*); različen (*različna embalaža*, *različen hormon*); cel (*cela etaža*); posamezen (*posamezna etaž*a), sam (*sama aplikacija*); omenjen, določen, predstavljen (*omenjen zakon*);
- glagoli; predvsem primarni glagoli (t. i. glagolski primitivi), kot so *biti*, *postati*, *delati*, *narediti* in *imeti*; modalni glagoli: *moči* in *morati* ter fazni glagoli: *začeti*, *končati*, npr. *[hoteti*, *moči] premagati*; *morati potruditi*; *hoteti natančno*; *začeti selekcijo*;
- prislovi; splošni (zlasti deikti: časovni, krajevni, kazalni), stopnjevalni, merni, količinski in številski prislovi, pri čemer pozicija (levo ali desno od samostalniškega, glagolskega jedra ipd.) pomensko manj obvestilnega kolokatorja ne vpliva na pomensko obvestilnost celotne kolokacije, npr. *zakašljati enkrat – enkrat zakašljati*: tako, takole; tu, tukaj, tam; tako, toliko; takoj, danes, dnevno, letno *(tako boleti*, *tukaj komentirati*, *prevajati takole*, *danes pospravljati)*; bolj/najbolj, manj/najmanj, več/največ (*najbolj zdeti*, *prepričati najbolj)*; kako, kaj, kdaj (*kako motivirati*, *komentirati kaj*; *kdaj ljubiti)*; nič, nekaj, več (*nič alkohola*, *nekaj alkohola)*; glede *(glede alkohola)*; nato, potem (*nato križati)*; načeloma *(načeloma morati)*; četrtič; dvakrat *(zmagati četrtič*, *četrtič organizirati*, *zboleti dvakrat)*; prislovi v členkovni, diskurzni rabi: *gotovo (izjemen)*; *očitno (slep)*.

Glavna težava pri iskanju pomensko manj obvestilnih kolokatorjev je v njihovi večpomenskosti in tudi raznolikosti. Tako smo pri poskusih oblikovanja seznamov takšnih kolokatorjev hitro naleteli na izjeme, zaradi katerih določenega kolokatorja ne moremo avtomatično izločiti v vseh kolokacijah (glej tudi Kosem idr. 2021). Tudi pri skupinah kandidatov, kot so npr. izlastnoimenski pridevniki

(npr. *češki, angleški*), tako pri iztočnicah tipa *krona* le-ti predstavljajo glavni razlikovalni element med pomeni (valuta; kraljestvo).

4.2.2 Razširjene kolokacije

Kot dobro izhodišče za debato so se pokazali tudi kolokacijski kandidati s (prevladujočim) odgovorom 'Razširjena kolokacija', kamor so označevalci uvrščali kolokacijske kandidate s potencialno manjkajočim elementom.

Na eni strani smo identificirali tipe besednih zvez, ki so se v svoji avtomatsko izluščeni binarni oz. tridelni predložni strukturi pokazale kot:

(a) semantično smiselne tudi brez razširitvenega elementa, tj. **kolokacije s fakultativnim razširitvenim elementom**, npr. (gbz sbz_4) *izpeljati projekt* -> (gbz + $prid_0$ + sam_0) *izpeljati [zahteven] projekt*); (sbz_0 po sbz_5) *vonj po kuhinji* -> (sbz_0 + po + pbz_0 sbz_0) *vonj po [domači] kuhinji*. V to skupino uvrščamo tudi kolokacije z (variabilnim) razširitvenim elementom, ki deluje kot obvezen, npr. *govoriti jezik -> govoriti [slovenski*, *nemški*, *angleški] jezik*.

(b) semantično nesmiselne zaradi odsotnosti razširitvenega elementa, tj. **kolokacije z nepogrešljivim oz. obveznim razširitvenim elementom**, npr. (gbz sbz_4) **vmešati ščepec* -> (gbz + sbz_0 + sbz_2) *vmešati ščepec [soli]*; (gbz sbz_4) **stopiti žlico* -> (gbz + sbz_0 + sbz_2) *stopiti žlico [masla*, *moke]*[9]. Slednje so predstavljale potencialno dobre kandidate za nadaljnjo obravnavo, saj so v razširjeni obliki slovarsko relevantne kolokacije.

Na drugi strani smo zaznali številne zveze, ki izpolnjujejo osnovne pogoje za kolokacijo (tj. ustrezajo statističnim, skladenjskim in semantičnim kriterijem; gl. Gantar idr. 2020) in so hkrati tudi samostojne (večbesedne) leksikalne enote, ki za razliko od kolokacij potrebujejo svoj pomenski opis in jih posledično ne uvrščamo med

9 Samostalniki, ki izražajo količino v enem od svojih pomenov (npr. *ščepec*, *žlica*), so se v strukturi gbz sbz_4 vedno pojavljali v razširjeni kolokaciji, samostojno pa le v drugem pomenu (npr. *žlica*).

razširjene kolokacije. Ločimo dve skupini: (a) stalne besedne zveze in (b) frazeološke enote. Pri stalnih zvezah gre dejansko za binarne kolokacije dveh leksikalnih enot, enobesedne iztočnice in večbesedne iztočnice, npr. *časopisna kronika -> časopisna črna kronika*; **tanjšati plašč -> tanjšati ozonski plašč*; **organizirati mizo -> organizirati okroglo mizo*; **barvati za noč -> barvati za veliko noč*.[10] Frazeološke enote pa smo lahko zasledili že pri nerazširjenih kolokacijskih kandidatih (npr. *začarani krog*), pri razširjenih pa so se pokazale kot manjkajoči razširitveni element, obvezen za razumevanje pomena celote, npr. **princ na konju -> princ na belem konju*; **obračanje plašča -> obračanje plašča po vetru*; **mešati jabolko -> mešati jabolka in hruške.*

Pri analizi smo zaznali tudi nekatere podskupine skladenjskih zvez,[11] ki so po svoji naravi sicer precej blizu kolokacijam oz. razširjenim kolokacijam, s katerimi jih lahko druži variabilnost enega ali več elementov in relevantnost z vidika vključitve v slovar, npr. **etaža stolpnice -> [tretja] etaža stolpnice*. Gre za zveze z ustaljeno skladenjsko strukturo in omejenim številom predvidljivih kolokabilnih mest in/ali omejenim številom kolokatorjev na predvidenih mestih, ki jih pogosto zasedajo številski ali številski elementi (zapisani s številko ali besedo), npr. *doktorirati leta [x] -> doktorirati leta [1970]*; *[x] [dag] česa -> [50] dag jetrc*; *obiskovati (kaj, koga) [x-krat] na [teden, mesec] -> obiskovati trikrat na teden.* Značilne pa so bile tudi zveze z lastno besedilno funkcijo (povezovalno, organizacijsko, vrednotenjsko), ki opravljajo vlogo diskurznih označevalcev, npr. **rekoč brezplačno -> tako rekoč brezplačno*; **skupaj komentirati -> vse skupaj komentirati*; **imenovana debelost -> tako imenovana trebušna debelost.*

Gledano s strukturnega vidika, razširjene kolokacije presegajo klasično dvodelno, v primeru predložnih pa tridelno strukturo kolokacije. Pri pogostih samostalniških, prislovnih ter predložnih

10 Meje med stalnimi zvezami in kolokacijami pa ni mogoče vedno natančno določiti. Za dodatna merila ločevanja kolokacij in stalnih besednih zvez gl. Gantar 2015 ali prispevek o opredelitvi kolokacije Gantar idr. 2020.

11 Opredelitev skladenjskih zvez v razmerju do kolokacij oz. razširjenih kolokacij gl. v Gantar 2015 ter Gantar idr. 2020.

strukturah smo detektirali strukturne razširitve izhodiščnih skladenjskih struktur predvsem s pridevniki (pred pridevniki ali samostalniki), samostalniki v rodilniku (za samostalniki) in prislovi (pred pridevniki ali glagoli). Spodaj navajamo v naši analizi najpogosteje zaznane možnosti strukturne razširitve dvo- ali tridelnih (predložnih) kolokacij z enim dodatnim elementom po posameznih skladenjskih strukturah.[12] Pri tem je treba poudariti, da gre pri razširjenih kolokacijah dejansko za kombinacijo dveh kolokacij iz dveh (ne nujno različnih) skladenjskih struktur, s tem da lahko obe kolokaciji ali pa samo ena od njiju izkazuje semantično smiselnost.

gbz prid$_4$ sbz$_4$ (gbz sbz$_4$ + pbz$_0$ sbz$_0$): *prevesti slovensko zbirko = prevesti zbirko + slovenska zbirka*

gbz sbz$_4$ sbz$_2$ (gbz sbz$_4$ + sbz$_0$ sbz$_2$): *prevesti zbirko pesmi = prevesti zbirko + zbirka pesmi*

rbz gbz sbz$_4$ (rbz gbz + gbz sbz$_4$): *dobro prevesti zbirko = dobro prevesti + prevesti zbirko*

pbz$_0$ pbz$_0$ sbz$_0$ (pbz$_0$ sbz$_0$ + pbz$_0$ sbz$_0$): *barvni laserski tiskalnik = barvni tiskalnik + laserski tiskalnik*

pbz$_0$ sbz$_0$ sbz$_2$ (pbz$_0$ sbz$_0$ + sbz$_0$ sbz$_2$): *standarden del opreme = standarden del + del opreme*

rbz pbz$_0$ sbz$_0$ (rbz pbz$_0$ + pbz$_0$ sbz$_0$): *nasproti vozeče vozilo = nasproti vozeč + vozeče vozilo*

sbz$_0$ pbz$_2$ sbz$_2$ (sbz$_0$ sbz$_2$ + pbz$_0$ sbz$_0$): *prevod slovenskega avtorja = prevod avtorja + slovenski avtor*

sbz$_0$ sbz$_2$ sbz$_1$ (sbz$_0$ sbz$_0$ + sbz$_0$ sbz$_1$): *plašč znamke Goodyear = plašč znamke + znamka Goodyear*

rbz pbz$_0$ sbz$_0$ (rbz pbz$_0$ + pbz$_0$ sbz$_0$): *lepo vzdrževana trata = lepo vzdrževan + vzdrževana trata*; *dobro založena trgovina = dobro založen + založena trgovina*

gbz rbz rbz (gbz rbz + rbz rbz): *odrezati se zelo slabo = odrezati se slabo + zelo slabo*

12 Možnih kombinacij je še precej več, zlasti pri tridelnih predložnih skladenjskih strukturah.

sbz_0 za pbz_4 sbz_4 (sbz_0 za sbz_4 + pbz_0 sbz_0): *aplikacija za neposredno sporočanje = aplikacija za sporočanje + neposredno sporočanje*

sbz_0 za sbz_4 sbz_2 (sbz_0 za sbz_4 + sbz_0 sbz_2): *aplikacija za vnos podatkov = aplikacija za vnos + vnos podatkov*

gbz iz $prid_2$ sbz_2 (gbz iz sbz_2 + pbz_0 sbz_0): *doktorirati iz ekonomskih ved = doktorirati iz vede + ekonomske vede*

gbz iz sbz_2 sbz_2 (gbz iz sbz_2 + sbz_0 sbz_2): *doktorirati iz področja prava = doktorirati iz področja + področje prava*

gbz v pbz_5 sbz_5 (gbz v sbz_5 + pbz_0 sbz_0): *poslovati v slovenskem jeziku = poslovati v jeziku + slovenski jezik*

Izpostaviti velja še primere razširitev strukture s členki (že, še, le), ki smo jih zabeležili pri analizi struktur s prislovi, npr. (rbz gbz) **vedno skeleti -> [še] vedno skeleti*; (gbz rbz); **uživati naprej -> uživati [še] naprej.* V navedenih primerih členki predstavljajo obvezen razširitveni element, prevladovali pa so večinoma primeri s členki kot neobveznim razširitvenim elementom. Vključitve skladenjskih struktur s členki v izhodiščnem naboru oz. mehanizmu luščenja nismo predvideli, posledično je zaznavanje razširjenih kolokacij tega tipa problematično, saj ne moremo sestavljati podatkov dveh obstoječih struktur. Z vidika razširjenih kolokacij gre torej za problematično skupino besed, ki se sopojavlja z velikim številom komponent in navadno nastopa kot del skladenjskih zvez.

Pri tridelnih predložnih strukturah smo zaznali tudi oblike razširitev z dvema ali več elementi na več pozicijah (levo ali desno od jedra), kar vzpostavlja precej kompleksne pet- ali še večdelne strukture, ki sprožajo vprašanje meja razširjene kolokacije in s tem preseganja kolokacijskosti:

- pbz_0 + sbz_0 + predlog + pbz_4 + sbz_4 (iz **sbz_0 za sbz_4**): *komisija za jezik -> [maturitetna] komisija za [slovenski] jezik*;
- pbz_0 + sbz_0 + predlog + pbz_2 + sbz_2 (iz **sbz_0 do sbz_2**): *toleranca do dejanja -> [nična] toleranca do [kaznivega] dejanja*;
- pbz_0 + sbz_0 + predlog + sbz_4 + sbz_2: (iz **sbz_0 za sbz_4**): *[hidravličen] cilinder za dvig [grebena]*;

- še bolj kompleksni primeri razširitev struktur z več istovrstnimi elementi ali celo ustaljenimi zvezami:
  - sbz_0 + predlog + pbz_2 + pbz_2 + pbz_2 +sbz_2 (iz **sbz_0 iz sbz_2**): *plašč iz snovi -> plašč iz [posebne] [tanke] [ogrevalne] snovi*;
  - pbz_0 + pbz_0 + sbz_0 + predlog + pbz_5 + sbz_5 (iz **sbz_0 v sbz_5**): *aplikacija v kategoriji -> [najboljša] [mobilna] aplikacija v [izbrani] kategoriji*;
  - pbz_0 + sbz_0 + predlog + sbz_2 + predlog + sbz_6 + sbz_2 (iz **sbz_0 sbz_2**): *toleranca do vožnje -> [ničelna] toleranca do vožnje pod vplivom alkohola*.

Na splošno se pri razmislekih o vključevanju razširjenih kolokacij v slovarske vire srečamo predvsem z vprašanjema njihovega beleženja v slovarski bazi na eni strani in predstavitve uporabnikom na drugi. V slovarski bazi je vsako razširjeno kolokacijo smiselno beležiti ločeno, pri čemer je zelo pomembno ohraniti povezavo z izhodiščno binarno kolokacijo, v kolikor je le-ta v samostojni obliki semantično smiselna, kot tudi povezave z vsemi sorodnimi razširjenimi kolokacijami. Pri razširjenih kolokacijah z (obveznim ali pogojno obveznim) variabilnim dodatnim elementom se želimo izogniti pretiranemu naštevanju ter vse variacije iste razširjene kolokacije obravnavati kot povezane oz. gručene. Posledično je smiselno navesti zgolj nekaj najbolj tipičnih predstavnikov variabilnega dodatnega elementa (npr. *prevesti [slovensko, angleško, nemško] zbirko*).

4.2.3 Zveze z lastnoimenskimi kolokatorji

V razpravah o slovarsko relevantni kolokaciji so se lastnoimenska poimenovanja izkazala za precej perečo kategorijo. Med označevalci so sprožala največ dvomov zaradi svoje pomenske specifičnosti oz. nanašalnosti na izključno enega, konkretnega referenta.[13]

13 Na neustreznost (nizko natančnost) in težavnost oblikoskladenjskega označevanja lastnih imen je na več mestih opozorila že obsežna analiza avtomatskega luščenja, predstavljena v monografiji Špele Arhar Holdt, *Luščenje besednih zvez iz besedilnega korpusa z uporabo dvodelnih in tridelnih oblikoskladenjskih vzorcev* (2011).

Kot tipičen, z vidika slovarske relevantnosti problematičen primer so se pokazale iztočnice, ki ob sebi predvidevajo daljši naštevalni niz istovrstnih kolokatorjev. Identificirali smo predvsem kolokacijske kandidate, pri katerih je eden od kolokatorjev: (a) osebno lastno ime: *obleka [Maje], prevod [Aleša]*; *premagati [Avstrijce],* angažirati *[Janeza]*; (b) svojilni pridevnik iz osebnega lastnega imena: *[Uroševa, Tinina] babica*; *[Župančičev, Jesihov] prevod*; (c) stvarno lastno ime: *premagati [Cibono], skupina [Raglje]*; (d) svojilni pridevnik iz stvarnega lastnega imena: *[Saturnov] satelit, [Googlova] aplikacija*; (e) geografsko lastno ime: *okupirati [Evropo], prestolnica [Štajerske], alkohol v [Estoniji]* (vprašanje semantične smiselnosti); (f) vrstni pridevnik iz geografskega lastnega imena, npr. *[štajerska, bavarska, katalonska, dolenjska] prestolnica*; *[slovenski, angleški, nemški] jezik*.

V primerih s stvarnimi in geografskimi lastnimi imeni kot slovarsko relevantne obravnavamo kolokacije z lastnim imenom (*prestolnica [Štajerske, Gorenjske, Dolenjske]*), ki nastopa v vlogi ustreznega semantičnega tipa, npr. *prestolnica [province, države, pokrajine, dežele]* ali *premagati [ekipo, tekmece, nasprotnike, moštvo]*. Analiza je namreč pokazala, da so poimenovanja s konkretnimi lastnimi imeni velikokrat pogostejša in bolj tipična od poimenovanj semantičnih tipov ali pa se poimenovanja za semantične tipe ne pojavljajo kot kolokatorji. V primeru daljšega niza istovrstnih (iz)lastnoimenskih kolokatorjev se sicer zdi smiselno kolokacijo prikazovati kot primer znotraj splošnejšega poimenovanja s semantičnim tipom.

Konceptu slovarske kolokacije ne ustrezajo primeri, pri katerih je celotna zveza lastno ime, torej gre za večbesedno lastnoimensko poimenovanje, ki nastopa kot ena lastnoimenska entiteta. Velja izpostaviti, da je sicer označenost imenskih entitet v korpusu pomembna z vidika reševanja problema avtomatskega luščenja zvez, ki vključujejo del lastnega imena, npr. **mavrični dežnik -> Pod mavričnim dežnikom, *premagati ob Paki -> premagati v Šmartnem ob Paki,* zaradi konkretnega referenta, ki ga zastopajo, pa so slovarsko nerelevantne (vsaj za kolokacijski slovar), npr. geografska lastnoimenska poimenovanja: *Južna Amerika, Podgorska cesta*. Drugače

od teh obravnavamo primere z izlastnoimenskimi pridevniki, ki so pomensko manj obvestilni, npr. poimenujejo smer, lokacijo ipd., v primeru *Podgoriška ulica* ali *podgorska vas* (v pomenu "vas v podgorju"), podobno še: *južna [Italija], severna [Francija]* proti *Južna Amerika, Južna Koreja*. Določene kolokacije z izlastnoimenskimi pridevniki pa pri posameznih iztočnicah vseeno lahko pustimo izpostavljene, če so del krajšega niza ali so celo edine, ki se tipično pojavljajo, npr. *[češka, švedska, norveška] krona // [angleška, francoska, britanska] krona,* kjer gre za niz izlastnoimenskih kolokatorjev, ki so tudi pomensko opredeljevalni in ključni za pomensko členitev.

Vključitev lastnoimenskih imen v slovar poleg podatkovnega obilja prinaša tudi potencialne zaplete z navedbo prepoznavnih osebnih lastnih imen (osebnih podatkov), navajanjem blagovnih znamk in podobnih primerov, ki se nanašajo na stvarna lastnoimenska poimenovanja in so leksikalno nerelevantna: **grajski vitraž -> Grajski vitraž* (prireditev), **škrlaten dež -> Škrlatni dež* (film), **sobotna raglja -> Sobotna raglja* (oddaja). Na drugi strani njihova celovita izpustitev lahko vodi v manko pomembnega dela besedišča, ki v statističnem smislu (tipičnost, pogostost, pojavnost) ustreza kriteriju kolokacijskosti. Kompleksnost vprašanja in možne rešitve se odražajo tudi skozi aspekt mnenj slovarskih uporabnikov in rezultatov uporabniške evalvacije, v okviru katere se je večina uporabnikov do vključenosti lastnoimenskih poimenovanj v slovar sicer opredelila z oceno 'Da' (niso problematična) (gl. Pori idr. 2020: 185–189). Z 'Da' so se opredelili do pomensko relevantnih lastnoimenskih poimenovanj in poudarili, da vsa imena niso enako pomensko (ne)relevantna (*kranjski Janez – Janez Novak*; *delati se Francoza – Francoz*). V določenih primerih se jim je zdela konkretnost, ki jo prispevajo lastna imena, tudi intuitivnejša: kolokacija *klop Reala* ali *klop Liverpoola,* je lahko bolj nazorna in povedna kot *klop prvoligaša*, pri kateri brez konteksta težko razberemo, da gre za nogometni klub. Ravno tako je pomensko obvestilnejša zveza *ljubljanska Olimpija* (ki se nanaša na klub iz Ljubljane) od zveze *pogrešani [Jure, Domen].* Lastna imena se jim zdijo tudi dragocena informacija o tipičnosti ogovornega vzorca, pri čemer pa so poudarili, da konkretnost (osebno ime) ni relevantna

(*dragi + Janez – dragi + [osebno lastno ime]*), pač pa je ključen pri tem podatek o diskurzni funkciji. O primerih daljšega niza istovrstnih kolokatorjev so menili, da so večinoma problematični in moteči, npr. niz izlastnoimenskih pridevnikov: *[češko, belgijsko, angleško, dansko] pivo* ali geografskih lastnih imen (imen mest): *okupirati [Bosno, Ljubljano, Nizozemsko]*, razen v primerih, ko podajajo koristno informacijo o oblikoslovnih kategorijah posameznih besednih vrst, npr. o sklanjatvenem vzorcu in rabi predlogov: *potovati na [Hrvaško, Kitajsko],* vendar *potovati v [Evropo, Azerbajdžan]* (Pori idr. 2020: 186).

5 Diskusija in zaključek

Analize jezikoslovne evalvacije kolokacijsko produktivnih struktur, ki so sledile stopnjam v procesu izdelave celovitega kolokacijskega opisa slovenskih besed, tj. avtomatsko izluščenim kolokacijskim podatkom in pilotni množičenjski nalogi, so se izkazale za zelo učinkovit način opredeljevanja ne samo slovarsko relevantne kolokacije, temveč prek identifikacije nerelevantnih kolokacijskih kandidatov (npr. napak strukture) tudi statistično relevantne kolokacije. Kot se je izkazalo, je za opredeljevanje kolokacije manj bistveno ugotoviti, kaj kolokacija je, precej pomembneje pa opredeliti, kaj kolokacija ni.

Evalvacija označenih avtomatsko izluščenih kolokacij je vsekakor pripomogla k opredelitvi kolokacije, tudi v odnosu do ostalih večbesednih kombinacij besed (za več gl. Gantar idr. 2020). Z izpostavitvijo skupin slovarsko nerelevantnih kolokacij se je pripravila podlaga za testiranje drugih statističnih metod, poleg že uveljavljenih logDice in podobnih mer povezovalnosti, npr. deltaP, za prepoznavo nerelevantnih kolokatorjev oz. bolje rečeno nerelevantnih primerov rabe kolokatorjev (za več gl. Kosem idr. 2021).

Oblikovanih je bilo tudi veliko priporočil za izboljšavo postopkov avtomatskega luščenja in postprocesiranja. Nekatere izboljšave ponujajo že nove funkcije v orodju Sketch Engine, npr. najdaljši skupni niz (ang. longest-commonest match) in ukaz COLLOC v slovnici besednih skic, vendar pa imajo svoje slabosti, npr. najdaljši skupni niz je zaradi statističnih pogojev na voljo le pri določenih kolokacijah.

Poleg tega so bile nekatere izboljšave na podlagi evalvacij že vključene v najnovejše verzije slovnice besednih skic.

Nekatere pomanjkljivosti, povezane z avtomatskim luščenjem na podlagi besednovrstnih oznak in postprocesiranjem, so v okviru tega postopka težko rešljive, zato se postavlja vprašanje, ali je bolje uporabiti skladenjsko razčlenjen korpus, ki naj bi zaradi beleženih povezav med deli kolokacij ponudil zanesljivejše podatke, in hkrati združiti postopek luščenja in postprocesiranja, saj recimo podatek o sklonu, obliki, spolu ipd. najdemo že v oblikoskladenjski oznaki, ki je pripisana vsem pojavnicam. Omenjeni pristop je bil preizkušen v projektu Nova slovnica slovenskega jezika, za več glej Krek idr. (2021).

Ne glede na uporabljeno metodologijo je jasno, da so avtomatski postopki luščenja kolokacijskih podatkov močno odvisni od zanesljivosti korpusnih podatkov, zato je smiselno vlagati v redno izboljševanje postopkov označevanja, kot sta lematizacija in oblikoskladenjsko označevanje. Vendar pa avtomatsko luščenje nikoli ne bo povsem zanesljivo, zato je pomembno izvajati redne evalvacije, kombinirati različne pristope in beležiti rezultate analiz. Ključno je shranjevanje vseh vrst analiziranih kolokacijskih kandidatov, od napačnih, s katerimi lažje prepoznamo napake v prihodnje in se izognemo podvajanju dela, do dobrih (semantično smiselnih in (deloma) slovarsko relevantnih). Slovarsko relevantne kolokacije namreč predstavljajo zgolj podmnožico kolokacij, njihova opredelitev pa se lahko od slovarja do slovarja oz. od jezikovnega vira do jezikovnega vira spreminja. Pri tem nikakor ne gre prezreti mnenj uporabnikov, ki se lahko precej razlikujejo od pojmovanj leksikografov, kaj je npr. slovarsko relevantna oz. uporabna kolokacija (za več gl. Pori idr. 2020).

Zahvala

Projekt *Kolokacije kot temelj jezikovnega opisa: semantični in časovni vidiki* (J6-8255), projekt *Nova slovnica sodobne standardne slovenščine: viri in metode* (J6-8256) in raziskovalni program št. P6-0411 (*Jezikovni viri in tehnologije za slovenski jezik*) je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna.

Reference

Arhar Holdt, Š. (2011): *Luščenje besednih zvez iz besedilnega korpusa z uporabo dvodelnih in tridelnih oblikoskladenjskih vzorcev.* (zb. Trojinski konj). Ljubljana: Trojina, zavod za uporabno slovenistiko.

Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Gantar, P., Gorjanc, V., Klemenc, B., Kosem, I., Krek, S., Laskowski, C. 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 XVIII 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/118/211/2991-1.pdf (30. 6. 2021).

Atkins, B. T. S. in Rundell, M. (2008): *The Oxford Guide to Practical Lexicography.* New York: Oxford University Press.

Bick, E. (2009): DeepDict – A Graphical Corpus-based Dictionary of Word Relations. V *Proceedings of NODALIDA 2009*. *NEALT Proceedings Series*: *Vol. 4*: 268–271. Tartu: Tartu University Library.

Cook, P., Lau, J. H., Rundell, M., McCarthy, D. in Baldwin, T. (2013): A lexicographic appraisal of an automatic approach for detecting new word senses. V I. Kosem idr. (ur.) *Electronic lexicography in the 21st century: thinking outside the paper*: 49–65. Estonia: Proceedings of the eLex conference.

Dobrovoljc, K., Krek, S., Holozan, P., Erjavec, T. in Romih, M. (2013): *Morphological lexicon Sloleks 1.0*. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1033.

Gantar, P., Krek, S., Kosem, I., Šorli, M., Grabnar, K., Pobirk, O., Zaranšek, P. in Drstvenšek, N. (2012): *Leksikalna baza za slovenščino.* Ljubljana: Ministrstvo za izobraževanje, znanost, kulturo in šport. Dostopno prek: http://www.slovenscina.eu/spletni-slovar/leksikalna-baza (30. 6. 2021).

Gantar, P., Kosem, I., Krek, S. in Gorjanc, V. (2015): Collocations dictionary of Slovene: challenge for automatization and crowdsourcing. V G. Corpas Pastor idr. (ur.): *Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives.* Europhras, Malaga.

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

Gantar, P., Kosem, I. in Krek, S. (2016): Discovering Automated Lexicography: The Case of the Slovene Lexical Database. *International Journal of Lexicography*, 29 (2), 200–225. doi: 10.1093/ijl/ecw014

Gantar, P., Krek, S. in Kosem, I. (2021): Opredelitev kolokacij v digitalnih slovarskih virih za slovenščino. V I. Kosem (ur.): *Kolokacije v slovenščini*: 15–41. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani.

Grčar, M., Krek, S., Dobrovoljc, K. (2012): Obeliks: statistical morphosyntactic tagger and lemmatizer for Slovene. V T. Ejavec in J. Žganec Gros (ur.): *Proceedings of the Eighth Language Technologies Conference*: 89–94. Ljubljana: Institut Jožef Stefan.

Gorjanc, V., Gantar, P., Kosem, I. in Krek, S. (ur.) (2015): *Slovar sodobne slovenščine: problemi in rešitve.* Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani. doi: 10.4312/9789612379759

Kilgarriff, A. in Tugwell, D. (2001): WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography. V *Proceedings of the ACL Workshop on Collocations*: 32–38. Toulouse, France.

Kilgarriff, A., Rychly, P., Smrz, P. in Tugwell, D. (2004): The Sketch Engine. V G. Williams in S. Vessier (ur.): *Proceedings of the Eleventh EURALEX International Congress*: 105–116. Lorient: Universite de Bretagne – sud.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. in Rychlý, P. (2008): GDEX: Automatically Finding Good Dictionary Examples in a Corpus. V E. Bernal in J. DeCesaris (ur.): *Proceedings of the 13th EURALEX International Congress*: 425–432. Barcelona: Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra.

Kilgarriff, A. in Rychlý, P. (2010): Semi-automatic Dictionary Drafting. V G.-M. de Schryver (ur.): *A Way with Words: A Festschrift for Patrick Hanks*: 299–312. Kampala: Menha Publishers.

Kilgarriff, A., Kovář, V. in Rychlý, P. (2010): Tickbox Lexicography. V *eLexicography in the 21st century: New challenges, new applications*: 411–418. Presses universitaires de Louvain, Brussels.

Kilgarriff, A. in Kosem, I. (2012): Corpus tools for lexicographers. V S. Granger in M. Paquot (ur.): *Electronic lexicography*. New York: Oxford University Press.

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*: 151–159. Ljubljana: Trojina, Institute for Applied Slovene Studies.

Kosem, I., Gantar, P. in Krek, S. (2013): Automation of lexicographic work: an opportunity for both lexicographers and crowd-sourcing. V I. Kosem idr. (ur.): *Electronic lexicography in the 21st century: thinking outside the paper*: 32–48. Ljubljana: Trojina, Institute for Applied Slovene Studies; Tallinn: Eesti Keele Instituut.

Kosem, I. (2015). Slovarski zgledi. V V. Gorjanc, P. Gantar, I. Kosem in S. Krek (ur.): *Slovar sodobne slovenščine: problemi in rešitve*: 320–339. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani.

Kosem, I. (2016): Interrogating a corpus. V P. Durkin (ur.): *The Oxford handbook of lexicography* [Oxford handbooks in linguistics, 1st ed.]. Oxford: Oxford University Press.

Kosem, I., Krek, S., Gantar, P., Arhar Holdt, Š., Čibej, J. in Laskowski, C. (2018a): *Kolokacijski slovar sodobne slovenščine.* Dostopno prek: https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/download/120/214/3152--1?inline=1 (30. 6. 2021).

Kosem, I., Krek, S., Gantar, P., Arhar Holdt, Š., Čibej, J. in Laskowski, C. (2018b): Collocations dictionary of modern Slovene. V J. Čibej idr. (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 (30. 6. 2021).

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., Logar, N., Dobrovoljc, K. in Ljubešić, N. (2021): Razvrščanje in relevantnost kolokatorjev v slovenščini: novi pristopi. V I. Kosem (ur.): *Kolokacije v slovenščini*: 79–124. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani.

Krek, S., Kosem, I. in Gantar, P. (2013): *Predlog za izdelavo Slovarja sodobnega slovenskega jezika*. Dostopno prek: http://www.sssj.si/datoteke/Predlog_SSSJ_v1.1.pdf (30. 6. 2021).

Krek, S. (2015): Leksikografska orodja za slovenščino: slovnica besednih skic. V V. Gorjanc, P. Gantar, I. Kosem in S. Krek (ur.): *Slovar sodobne slovenščine: problemi in rešitve*: 358–378. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani.

Krek, S., Gantar, P., Kosem, I., Dobrovoljc, K., Arhar Holdt, Š., Čibej, J., Laskowski, C., Klemenc, B. In Krsnik, L. (2021): *Frequency lists of collocations from the Gigafida 2.1 corpus*. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1415.

Logar Berginc, 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.* Ljubljana: Trojina, zavod za uporabno slovenistiko; Fakulteta za družbene vede.

Logar, N., Kosem, I. in Erjavec, T. (2019): *Collocation lexicon of Slovene academic discourse Aleks*. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1245.

Pori, E. in Kosem, I. (2018): V iskanju slovarsko relevantne kolokacije na primeru struktur s prislovi. *Slovenščina 2.0,* 6 (2), 154–185. doi: 10.4312/slo2.0.2018.2.154-185

Pori, E., Kosem, I., Čibej, J. in Arhar Holdt, Š. (2020): The Attitude of Dictionary Users Towards Automatically Extracted Collocation Data: A User Study. *Slovenščina 2.0: Empirical, Applied and Interdisciplinary Research,* 8 (2), 168–201. doi: 10.4312/slo2.0.2020.2.168-201

Pori, E., Kosem, I., Čibej, J. in Arhar Holdt, Š. (2021): Evalvacija uporabniškega vmesnika Kolokacijskega slovarja sodobne slovenščine. V I. Kosem (ur.): *Kolokacije v slovenščini*: 235–268. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani.

Rundell, M. (2002): *Macmillan English Dictionary for Advanced Learners*. Macmillan Education.

Rundell, M. in Kilgarriff, A. (2011): Automating the creation of dictionaries: where will it all end? V F. Meunier (ur.): *A Taste for Corpora. A tribute to Professor Sylviane Granger*: 257–281. Benjamins.