Opredelitev kolokacij v digitalnih slovarskih virih za slovenščino

Polona GANTAR
Filozofska fakulteta, Univerza v Ljubljani

Simon KREK
Institut Jožef Stefan; Filozofska fakulteta, Univerza v Ljubljani

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

This paper focusses on defining the phenomenon of collocation for the purpose of its use in machine-readable language resources, which will be used in the creation of electronic dictionaries and language applications for Slovene. We first describe three key aspects of collocation that define it as a lexical phenomenon: statistical, syntactic, and semantic. The statistical criterion defines collocation as a statistically significant combination of two or more words, the syntactic criterion expects certain syntactic relations between words, and in order to satisfy the semantic criterion a collocation needs to exhibit a specific communication role. Next, lexicographic relevance is taken as a point of departure for defining collocations within the typology of word combinations (including expanded collocations or collocations of collocations), as well as for distinguishing them from free combinations. In order to distinguish collocations from all multiword lexical units (compounds and phraseological units), we adopt the lexicographic view that multiword lexical units, whose meaning is not a sum of its parts, require a description of their meaning whereas collocations do not. In the final part, we revisit the statistical, syntactic and semantic aspects of collocation and their role in automatic extraction of collocational information from corpora for the purposes of lexicographic analysis. The paper

concludes by summarizing the main points and presenting our ongoing work on collocation identification and extraction and future plans.

Keywords: collocation, typology, word combination, lexicography, lexical resources, definition

1 Uvod

Vključevanje kolokacij v strojno procesljive jezikovne vire, ki služijo za izdelavo elektronskih slovarjev in različnih jezikovnih aplikacij, zahteva njihovo čim bolj natančno, a hkrati dovolj široko opredelitev, ki bo zadostila razvoju jezikovnih tehnologij in uporabi v jezikovnih opisih. Upoštevajoč omenjena izhodišča, ima naša naloga tri cilje, ki so bili opredeljeni tudi v okviru temeljnega raziskovalnega projekta *Kolokacije kot temelj jezikovnega opisa: semantični in časovni vidiki* (ARRS; J6-8255)[1]:

(a) prepoznati lastnosti, ki opredeljujejo kolokacije kot leksikalni jezikovni pojav in posledično kot pomemben del leksike, ki ga je treba vključiti v leksikalne jezikovne vire za slovenščino, v našem primeru v digitalno slovarsko bazo, namenjeno izdelavi jezikovnih priročnikov, jezikovnih aplikacij in nadaljnjemu računalniškemu procesiranju (Klemenc idr. 2017);
(b) opredeliti kolokacije v razmerju do drugih besednih zvez, zlasti na stiku njihovih skladenjskih in pomenskih lastnosti, kar je ključno za obravnavo znotraj slovarske baze kot tudi za določitev načina jezikovnega opisa, namenjenega človeškemu uporabniku;
(c) opredeliti lastnosti kolokacij, ki določajo njihovo slovarsko relevantnost, tj. z vidika njihove pomenske obvestilnosti.

V prispevku najprej opišemo lastnosti, ki kolokacijo opredeljujejo kot leksikalni jezikovni pojav. Različne pristope v kolokacijskih študijah prikažemo s treh ključnih medsebojno povezanih vidikov: (i) statističnega, ki predvideva, da je kolokacija statistično izstopajoča

1 https://www.cjvt.si/kolos/

zveza dveh ali več besed, (ii) skladenjskega, ki predvideva določena skladenjska pravila, ki potekajo med besedami in (iii) pomenskega, ki predvideva, da ima kolokacija določeno leksikalno oz. komunikacijsko vlogo. Prav zaradi zadnjega so kolokacije že od prvih opažanj in opisov (Firth 1957; Altenberg 1991; Sinclair 1991) tudi slovarsko zanimiv leksikalni pojav.

Izhodišče, po katerem je kolokacija vedno zveza vsaj dveh besed, zahteva tako z leksikografskega vidika kot z vidika avtomatskega luščenja iz korpusa opredelitev do vseh drugih besednih zvez, ki obstajajo v jeziku. Pri tem izhajamo iz tipologije večbesednih enot, ki smo jo predhodno zasnovali pri izdelavi Leksikalne baze za slovenščino (Gantar 2015). V nadaljevanju prispevka opredelimo kolokacije tudi z vidika njihove vključitve v slovar, kjer na podlagi statističnih, oblikoskladenjskih in pomenskih kriterijev izpostavljamo tiste lastnosti, ki določajo slovarsko relevantnost kolokacije. Ali z drugimi besedami, opredeliti želimo parametre za avtomatsko luščenje iz korpusa, da bo izplen čim bolj uporaben za slovarske namene. Prispevek zaključimo z evalvacijo izluščenih podatkov in z ugotovitvami, ki jih nameravamo v prihodnje upoštevati pri nadaljnjih iteracijah v tem postopku.

2 Kolokacija kot leksikalni pojav

Obstoj strojno procesljivih jezikovnih virov ter porast zanimanja za procesljive jezikovne podatke, zlasti take, ki imajo semantično naravo, kolokacije vedno znova postavlja v središče leksikalnih analiz in slovarskih praks. Kljub velikemu številu publikacij na področju kolokacij, katerih namen je opisati njihovo naravo (Fontenelle 1994; Herbst 1996), pa pojem kolokacije ostaja izmuzljiv. Raziskave so namreč pokazale, da imajo vse ključne lastnosti prototipičnih kolokacij, kot je izstopajoče sopojavljanje besed in predvidljivost na eni in omejenost izbire na drugi strani, navadno srednje in ne ekstremnih vrednosti (Schmid 2003: 249). V študijah, ki se ukvarjajo s kolokacijami, se pristopi, ki določajo njihove definicijske lastnosti, razlikujejo glede na to, kako na splošno oz. kako specifično opredeliti kolokacijo oz. za kakšen namen jo želijo definirati. Različni pristopi, ki glede

na svoj namen – tip slovarja, učenje jezika, avtomatsko procesiranje jezika ipd. – poudarjajo različne lastnosti kolokacij, definirajo kolokacijo znotraj treh med seboj povezanih kriterijev: statističnega, skladenjskega in pomenskega.

2.1 Statistični vidik

Ena od ključnih lastnosti kolokacij pri prepoznavanju v besedilu je njihova statistična vrednost, ki mora biti večja od naključne sopojavitve besed, ali kot pravita Atkins in Rundell (2008: 302): kolokacija je »ponavljajoča se kombinacija besed, v kateri kaže določen leksikalni element (jedro) očitno tendenco sopojavljanja z drugim leksikalnim elementom (kolokatorjem), s frekvenco, ki je večja od naključne sopojavitve«. Na vprašanje, kdaj je mogoče določeno kombinacijo besed šteti za ponavljajočo, je mogoče najbolj zanesljivo odgovoriti s pomočjo korpusnih analiz. Najpomembnejša pri tem je prav določitev, kako pogosto se mora besedna kombinacija ponoviti, da jo je mogoče prepoznati kot kolokacijo. Pri tem je jasno, da je ustreznost določitve statističnega praga povezana z velikostjo korpusa (Church in Hanks 1990; Clear 1993; Stubbs 1995b; Khokhlova in Benko 2020), pa tudi z drugimi parametri, ki jih določa oblikoskladenjska označenost korpusa in nenazadnje tudi njegova besedilna distribucija (Brezina idr. 2015).

Izstopajoča povezovalnost besed v jeziku je znotraj strojnega procesiranja naravnega jezika vodila v ugotavljanje najbolj zanesljivih in relevantnih statističnih mer, ki omogočajo avtomatsko prepoznavanje pogostih besednih kombinacij v tekočem besedilu. Številne raziskave se osredotočajo na merjenje kolokacijske moči ali t. i. kolokabilnosti (prim. Berry-Rogghe 1973; Church in Hanks 1990; Church idr. 1991; Biber 1993; Manning in Schütze 1999; Evert 2004; Gries 2013). Dober pregled različnih statističnih metod za merjenje besedne povezovalnosti najdemo v Wiechmann (2008), ki primerja 47 različnih asociacijskih mer, in v Pecina (2009), ki primerja več kot 80 različnih statističnih mer za avtomatsko ekstrakcijo kolokacij. Splošne ugotovitve, ki so jih prinesle primerjave, strne Evert (2009),

ki ugotavlja, da različne asociacijske mere kolokatorje razvrščajo popolnoma različno (ibid.: 1218) in da idealna asociacijska mera, ki bi zadostila vsem namenom luščenja, ne obstaja (ibid.: 1236).

Pri avtomatskem luščenju kolokacij za slovarske namene se je izkazalo, da je statistični kriterij po nujnosti narave kolokacij treba upoštevati skupaj z njihovimi pomenskimi in skladenjskimi lastnostmi. Skladenjska zgradba kolokacij je tako za določanje statističnih parametrov ključna, pri čemer poseben izziv predstavljajo kolokacije, katerih sestavni deli v besedilu navadno ne nastopajo skupaj oz. se mednje vrivajo drugi elementi. V naši tipologiji smo jih na podlagi evalvacije avtomatsko izluščenih kolokacij prepoznali kot samostojno podskupino t. i. razširjenih kolokacij (gl. Sliko 1).

2.2 Skladenjski vidik

Drugi temeljni pogoj za obstoj kolokacije je očiten: kolokacijo nujno tvorita vsaj dve besedi. Študije, ki opredeljujejo definicijske lastnosti kolokacij, se zato ne morejo izogniti dejstvu, da kolokacije določa tudi njihova skladenjska zgradba, notranje skladenjsko razmerje in morfološke lastnosti, ki iz tega razmerja izhajajo (Moon 1998; Hausmann 1989; Kilgarriff idr. 2004; Seretan 2010; Baldwin in Kim 2010; Fellbaum 2015) – kar je zlasti pomembno v morfološko bogatih jezikih, kot je slovenščina. Znotraj kolokacijskih študij obstajajo tudi raziskave, ki skušajo natančneje opredeliti status besed v kolokaciji z vidika njihovega medsebojnega razmerja (Sinclair 1966: 415). To razmerje je navadno opredeljeno hierarhično in razlikuje med jedrom kolokacije (ang. node), tj. besedo, ki določa perspektivo, s katere je kolokacija obravnavana, in njenimi kolokatorji. Čeprav gre generalno gledano za tehnični vidik, saj je posamezna beseda lahko v določeni perspektivi jedro, v drugi pa kolokator,[2] Hausmann (1984: 401; 1985: 119) izpostavlja, da je odnos med obema kolokacijskima

2 Upoštevanje omenjenega vidika je pri vključevanju kolokacij v Kolokacijski slovar sodobne slovenščine tesno povezano s statističnimi parametri luščenja in načinom razvrščanja kolokacij v slovarskem vmesniku: tako je denimo kolokacija *dober jezik* pri iztočnici *jezik* navedena, medtem, ko je pri iztočnici *dober* ni oz. ni navedena med najpogostejšimi. Zanimiva bi bila tudi raziskava tovrstne (skladenjske) permutativnosti z vidika +/– spremembe v leksikalni vrednosti kolokacije.

elementoma nujno hierarhičen, v katerem en element, imenovan baza (ang. base), določa drugega, ki je kolokator.

Sintaktični vidik vključuje tudi že omenjeni problem nekontinuiranosti elementov znotraj kolokacijskega niza, saj skladenjska narava besednih zvez v povezavi s pomensko vrednostjo kolokacije lahko zahteva nujno vrivanje elementov (**organizirati mizo -> organizirati okroglo mizo*) kot tudi prilagajanje kontekstu z odpiranjem vezljivostnih mest in zasedanjem pričakovanih stavčnih položajev: *tekmovalni del -> tekmovalni del programa*.

Za avtomatsko luščenje leksikalno relevantnih kolokacij iz korpusa, ki je bilo (prvotno) namenjeno izdelavi Kolokacijskega slovarja (Kosem idr. 2018a; Kosem idr. 2018b; Gantar idr. 2016) je bila zato potrebna premišljena določitev skladenjskih struktur in opredelitev njihovih slovničnih lastnosti, zlasti z vidika specifičnosti slovenščine. Kot je pokazala evalvacija (Pori in Kosem 2021), številni problemi avtomatskega luščenja izhajajo prav iz odločitev pri morfosintaktičnem označevanju korpusa, iz skladenjskega razmerja med elementi kolokacije in iz ustaljenosti posameznih oblik besed v kolokaciji.

2.3 Pomenski vidik

Pri definiranju kolokacij z vidika relevantnosti za vključitev v slovar kot tudi pri razmejevanju kolokacij glede na druge tipe večbesednih enot v jeziku je pomenski vidik najpomembnejši kriterij, hkrati pa ga je tudi najtežje opredeliti, saj so semantične spremembe in omejitve v izbiri, ki jih kaže raba, najbolj očitno povezane z že omenjeno srednjo vrednostjo kolokacij. Če izhajamo iz tipične sopojavitve besed, lahko namreč ugotovimo, da so kolokacije nekje na pol poti med prostimi besednimi zvezami in popolnoma ustaljenimi večbesednimi enotami. Prav omenjena sredinska vrednost kolokacij povzroča številne probleme pri definiranju pojma kolokacije s semantičnega vidika.

Ob splošno sprejetem statističnem in skladenjskem merilu sta se v strokovni literaturi pri definiciji kolokacij s pomenskega vidika oblikovala dva temeljna pristopa, ki upoštevata njihovo leksikalno naravo. Prvi kolokacije prepoznava kot samostojen tip frazeoloških

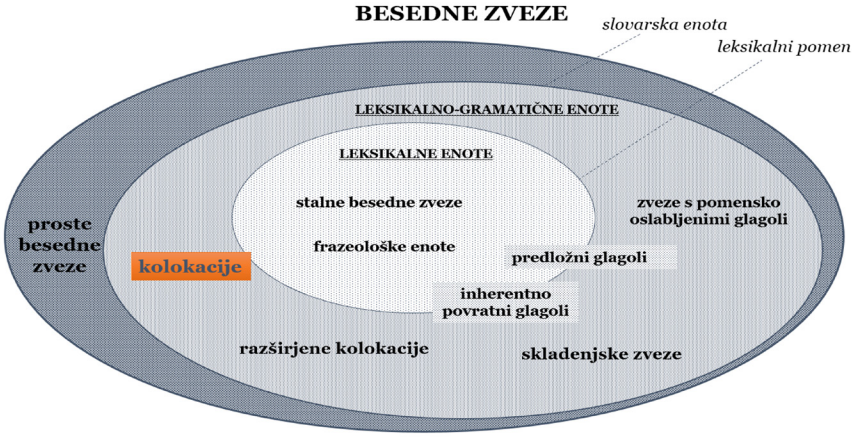
enot, ki so delno ali popolnoma (pomensko in skladenjsko) zamrznjene in so se ustalile skozi ponavljajočo se sobesedilno pogojeno rabo. Ta definicija zajema zlasti t. i. "frazeološke" ali "močne" kolokacije (ang. strong collocations), ki so v leksikalni izbiri svojih sestavnih elementov omejene (Aisenstadt 1981; Cowie 1981), slovarsko pa so zanimive zlasti za učenje tujega jezika. Na drugi strani so pristopi, ki obravnavajo kolokacije širše. Med kolokacije štejejo tudi frekventne besedne zveze, katerih notranja povezovalnost ni ozko zamejena ali celo izključujoča, pač pa je lahko niz sopojavnic tudi razmeroma odprt, npr. (*zeliščni, kamilični, metin, žajbljev, lipov* ...) *čaj*. Atkins in Rundell (2008: 167) jih opredeljujeta kot zveze, ki nimajo lastnega pomena kot celota oz. kot frekventno izstopajoče zveze besed, katerih pomen je razmeroma transparenten. Podobno opredelitev najdemo tudi v Benson idr. (1986), ki kolokacije opredeljujejo kot arbitrarno in ponavljajočo se neidiomatično leksikalno ali gramatično kombinacijo.

Na splošno torej velja, da kolokacije, ki so vključene v splošne slovarje, niso obravnavane kot leksikalne enote, ki potrebujejo pomensko razlago. Njihova vključitev v splošne slovarje je povezana z dejstvom, da tipično razdvoumljajo večpomenske besede (*trnova krona : češka krona : zobna krona*) in so zaradi svoje vsakdanjosti značilne za naravno jezikovno rabo (*trda tema, gosta megla : *trda megla*). Pomenski kriterij, ali natančneje leksikografovo odločitev glede pomenske transparentnosti besedne zveze (katere besedne zveze v slovar vključiti in katere med vključenimi potrebujejo razlago), smo postavili v izhodišče tipologije večbesednih enot, kjer kolokacije obravnavamo kot slovarske enote, vendar zunaj okvira enot z leksikalnim pomenom, s tem pa tudi zunaj frazeologije, kar je pomembno predvsem za način njihove obravnave v digitalni slovarski bazi.

3 Opredelitev kolokacij v razmerju do drugih besednih zvez

Dejstvo, da je kolokacija vedno zveza vsaj dveh besed, v izhodišču zahteva njihovo opredelitev glede na besedne zveze, ki so v jeziku

prav tako pogoste in slovnično ustrezajo določeni skladenjski kombinaciji, a hkrati – za razliko od kolokacij in drugih večbesednih enot – niso predmet slovarskih opisov – t. i. proste besedne zveze. Znotraj slovarskih enot na drugi strani se kolokacije razvrščajo ob bok besednim zvezam v ustaljenih skladenjskih in besedilnih vlogah, ki jih združujemo v skupino t. i. leksikalno-gramatičnih enot, kjer ločimo: razširjene kolokacije, zveze s pomensko oslabljenimi glagoli, različne tipe skladenjskih zvez ter predložne in inherentno povratne glagole, ki jih v nadaljevanju prispevka podrobneje predstavimo. Zadnji dve kategoriji lahko v določenih kontekstih pridobita leksikalno vrednost, s čimer se približujeta večbesednih leksikalnim enotam, ki za razliko od kolokacij in drugih leksikalno-gramatičnih enot v slovarju predvidevajo pomenski opis – tj. frazeološke enote in stalne besedne zveze (Slika 1).



Slika 1: Kolokacije v razmerju do drugih besednih zvez in glede na vključenost v slovarsko bazo.

3.1 Proste besedne zveze

Nekatere besedne zveze, ki vsebujejo slovnične besede, so v jeziku lahko zelo frekventne, vendar nimajo leksikalne ali skladenjske vrednosti in ne razdvoumljajo pomena, zato so pomensko manj informativne in posledično slovarsko nezanimive. Na primer, zvezi kot *nesrečen padec* in *zabeležiti rast* sta tipični besedni kombinaciji

– kolokaciji, ki nam nekaj povesta o besedah, ki jih vsebujeta, tj. da je *padec* lahko *nesrečen* in da *rast* tipično *zabeležimo*. Na drugi strani pogoste zveze kot npr. *je rekla, k meni* in *ta način* te obvestilnosti brez širšega konteksta nimajo. Če upoštevamo zgoraj izpostavljene definicijske vidike, lahko rečemo, da so proste besedne zveze lahko – tako kot kolokacije – v jeziku sicer pogoste besedne kombinacije, vendar pa za razliko od kolokacij nimajo leksikografske vrednosti v smislu, da bi besede pomensko razdvoumljale ali kazale na njihovo tipično in naravno jezikovno rabo.

3.2 Leksikalno-gramatične enote

Kolokacije je treba opredeliti tudi v odnosu do frekventnih večbesednih enot, ki imajo v svoji skladenjski zgradbi besede slovničnih in modifikacijskih besednih vrst, zato so pogosto imenovane tudi gramatične kolokacije (Benson idr. 1986). V naši tipologiji ločimo več podskupin, in sicer t. i. skladenjske zveze, kamor uvrščamo predložne samostalniške in prislovne zveze, zveze s členki, večbesedne veznike ipd.: *za nazaj, po vzoru, na prostem, glede na to da, več kot* ipd. V jeziku opravljajo vlogo stavčnih in besedilnih organizatorjev oz. diskurznih označevalcev (Dobrovoljc 2017; 2018), zato so prav tako kot kolokacije za slovarski opis zanimive, kar jih na drugi strani ločuje od frekventnih prostih besednih zvez. Zanje je tudi značilno, da napovedujejo predvidljiva skladenjska mesta v svoji besedilni okolici, npr. *v prid komu/čemu; v prid koga/česa*. Glede na mednarodno uveljavljene kategorije na preseku med slovarskimi opisi in potrebami računalniškega procesiranja jezika (Gantar idr. 2019b; Bhatia idr. 2017; Candito idr. 2016), ločimo tudi zveze z glagoli z oslabljenim pomenom, npr. *imeti pogum, dati maksimum*, predložne glagole, npr. *gre za, priti do* (česa) in inherentno povratne glagole, npr. *zdeti se, delati se*.

3.2.1 Zveze s pomensko oslabljenimi glagoli

Korpusne analize besedne povezovalnosti posebej izpostavljajo t. i. zveze s pomensko oslabljenimi glagoli (ang. light oz. support verb

constructions; Atkins in Rundell 2008: 175; Baldwin in Kim 2010: 15), ker so na eni strani leksikografsko zanimive in ker predstavljajo izziv za avtomatsko prepoznavanje v korpusih. Tipično gre za zveze pomensko bolj ali manj izpraznjenega glagola in samostalnika ali predložne samostalniške zveze: *sprejeti odločitev, postaviti vprašanje, imeti posledice, biti v dvomih*. Samostalniki navadno pomenijo stanje ali dogodek, glagoli pa nosijo občutno manj pomena kot v številnih drugih zvezah (Atkins in Rundell 2008: 173). Slovarsko so, podobno kot kolokacije, te zveze zanimive zaradi svoje mejne vrednosti, saj so na eni strani kljub svoji pomenski transparentnosti dober pokazatelj jezikovne tipike, hkrati pa je tipičnost povezana tudi z omejenostjo v leksikalni izbiri, npr. *postaviti vprašanje : *položiti vprašanje*. V slovarjih so take zveze vključene na različna mesta slovarske makrostrukture, bodisi kot leksikalne enote, npr. v dvojezičnih slovarjih, ali kot leksikalno-gramatični vzorec pri katerem od pomenov besede v iztočnici (Gantar idr. 2019a).

Za namene vključevanja v slovenske leksikalne vire so bile zveze s pomensko oslabljenimi glagoli definirane okviru projekta PARSEME Shared task na podlagi enotnih smernic za 27 različnih jezikov (Candito idr. 2016; Bhatia idr. 2017) in ročno označene v učnem korpusu ssj500k (Krek idr. 2018). Nabor obsega 130 različnih zvez, ki bodo glede na svoj leksikalno-gramatični status v slovarsko bazo vključene kot leksikalno-gramatične enote, tj. na isti ravni kot kolokacije v povezavi s posameznim pomenom besede.

3.2.2 Razširjene kolokacije

Kot podtip kolokacij obravnavamo tudi t. i. *razširjene kolokacije*, ki so prišle do izraza pri evalvaciji avtomatsko izluščenih kolokacij. Gre za tipične besedne sopojavitve, ki predvidevajo vrivanje leksikalnih elementov, pri čemer so ti elementi bodisi fakultativni: *učiti se (angleški, nemški, francoski ...) jezik,* ali pa obvezni: *organizirati okroglo mizo – *organizirati mizo*. Podrobneje se z opredelitvijo razširjenih kolokacij in obravnavo znotraj pomenskih in skladenjskih lastnosti ukvarja prispevek Pori in Kosem (2021).

3.2.3 Skladenjske zveze

Najbolj tipičen in hkrati heterogen primer leksikalno-gramatičnih enot, ki jih je mogoče prepoznati s korpusno analizo, so t. i. skladenjske zveze. Ker prinašajo v zvezi z lemo, na katero se nanašajo, pomembne leksikalne informacije (Rundell in Atkins 2011: 245), smo jih kot slovarske enote prepoznali že pri izdelavi Leksikalne baze za slovenščino (Gantar 2015: 330). Gre za zveze, ki izkazujejo skladenjsko ustaljenost, hkrati pa – vsaj z vidika naravnih govorcev slovenščine – ne izkazujejo samostojne pomenske vrednosti. V svoji zgradbi predvidevajo nekatere elemente stalnih zvez, tj. določene ustaljene sestavine, katerih izbor je omejen, hkrati pa napovedujejo prosta skladenjska mesta, zamejena z določenimi slovničnimi kategorijami, kot so npr. sklon, število, živost oz. neživost. Najbolj tipične so predložne zveze v različnih prislovnih vlogah, npr. kraja: *na prostem*; časa: *zadnje čase*, *za zdaj*, *ves čas*, *čim prej*, načina: *na nek način*, *v skladu z/s*, *v primerjavi z/s*, *na srečo*, *v celoti*, *v zadregi*, *po naravi*, *s pomočjo*, *pod pogojem*; količine: *kar nekaj*, *več kot*, *kolikor bolj – toliko bolj, do te mere,* vzroka: *od hudega*, *od togote*, *iz maščevanja*, ter zveze, ki vključujejo številske elemente: *[x] [dolarjev*, *tolarjev*, *evrov] žepnine*, *šteti [x] pomladi*, *diplomirati leta [x]*. Glede na vlogo, ki jo skladenjske zveze opravljajo v stavku, lahko prepoznamo zveze v vlogi organizatorjev diskurza (*po besedah*, *v bistvu*, *kar se tiče*) in besedilnih povezovalcev (*glede na*, *medtem ko*, *po eni strani – po drugi strani*). Glede na sorodne lastnosti, ki jih skladenjske zveze delijo s kolokacijami in razširjenimi kolokacijami, jih v slovarski bazi obravnavamo v okviru posameznega pomena besede v iztočnici.

3.2.4 Predložni glagoli

Kot podskupino je znotraj leksikalno-gramatičnih enot mogoče obravnavati tudi t. i. predložne glagole,[3] tj. glagole, ki skupaj s

3 V okviru tipologije, ki je nastala pri projektu PARSEME, smo tovrstne glagole označevali kot *predložnomorfemske glagole* z leksikaliziranim predložnim morfemom (ang. Inherently Adpositional Verbs; Gantar idr. 2019b). V učnem korpusu je ta oznaka pripisana 154 različnim enotam.

predlogom in predvidenim vezljivostnim mestom tvorijo strukturno trdne, lahko pa tudi pomensko samostojne enote. V prvem primeru imamo opraviti s tipičnimi glagolskimi predložnimi zvezami, ki so slovarsko zanimive zaradi svoje strukturne ustaljenosti, zaradi česar tvorijo prepoznavne dele širših glagolskih vzorcev, npr. *veljati za (koga/kaj), sovpadati s (kom/čim), prizadevati si za (koga/kaj), zavzeti se za (koga/kaj).* Ob pomensko razmeroma transparentnih predložnih glagolskih zvezah pa je treba omeniti tudi zveze kot npr. *priti do (česa), obrniti se na (koga/kaj), biti za (koga/kaj), postaviti se za (koga/kaj)* ipd., ki predvidevajo pomenski opis kot samostojna leksikalna enota: *biti za* (kaj) – "strinjati se"; *priti do (česa)* – "zgoditi se". Predložni glagoli so v slovenskih leksikalnih virih kot večbesedne enote označeni v učnem korpusu ssj500k na podlagi smernic, ki so bile določene v projektu PARSEME za različne jezike (Gantar idr. 2019b).

3.2.5 Inherentno povratni glagoli

Kot večbesedne enote je po nekaterih klasifikacijah mogoče obravnavati tudi t. i. inherentno povratne glagole,[4] kjer *se* ali *si* ne nastopata kot povratna zaimka ali kot izrazilo za trpnik, pač pa kot sestavni (morfemski) del glagola, ki brez morfema *se* ne obstaja, npr. *zdeti se, zgoditi se* ipd. Ti glagoli so v slovarjih sicer zapisani kot večbesedne iztočnice, čeprav navadno nimajo statusa večbesedne enote v enakem smislu kot npr. frazeološke enote ali stalne zveze (prim. SSKJ, SSKJ2 in SNB). Problem te kategorije je v prepoznavanju vezanosti določenega pomena na kombinacijo glagola s *se*, pri čemer ni vedno mogoče nedvoumno ločevati pomenske osamosvojitve tipa: *ločiti se* – "prekiniti zakonsko razmerje" od tipičnih trpnih ali povratnih rab: *(koga) se loči od skupine : (kdo) loči (koga) od skupine*, s čimer je povezano tudi leksikografsko vprašanje obravnavanja tovrstnih zvez kot samostojnih leksikalnih enot – iztočnic (*ločiti* in *ločiti se*) ali zgolj ene iztočnice z več pomeni in tipičnimi realizacijami glede na

4 Tudi ta kategorija (ang. inherentlly reflexive verbs) je bila v okviru evropske COST akcije PARSEME definirana na podlagi smernic in označena v učnem korpusu ssj500k. Rezultati analize, ki obsegajo 345 različnih enot, so podrobneje predstavljeni v Gantar idr. (2019b).

morfem oz. zaimek: ločiti: 1. v zvezi s *se*: "prekiniti zakonsko razmerje", 2. "odstraniti iz skupine, celote".

3.3 Leksikalne enote

Ločevanje pomensko transparentnih kolokacij od večbesednih leksikalnih enot, kot so frazeološke enote in stalne besedne zveze, je z leksikografskega vidika pomembno predvsem zato, ker bolj ko se kolokacije približujejo stalnim zvezam in frazeološkim enotam, več leksikografske pozornosti zahtevajo. Kolokacije kot frekventne besedne sopojavitve se v slovarjih približujejo zgledom, saj s svojo tipičnostjo najbolje odražajo realno jezikovno rabo. Stalne zveze in frazeološke enote na drugi strani potrebujejo več slovarskih informacij; v prvi vrsti razlago pomena, lahko pa še opozorila glede pragmatičnih posebnosti ter slovničnih in skladenjskih omejitev.

Pri definiranju kolokacij v razmerju do večbesednih leksikalnih enot, ki sodijo v poimenovalni del jezika, smo zato sledili leksikografskemu merilu, ki ga najbolje opredeljujeta Atkins in Rundell (2008: 167), ki pravita, da so večbesedne leksikalne enote[5] različni tipi zvez, ki imajo določeno stopnjo idiomatičnega pomena oz. se obnašajo idiomatično. Z vidika vključitve v slovar in njihovega slovarskega opisa pa morajo izpolnjevati kriterij, po katerem »je njihov pomen več kot vsota pomenov posameznih sestavin« (ibid.: 167). Ker je tak kriterij seveda relativen in namenjen izključno leksikografski opredelitvi, je pomembno poudariti, da je leksikografova presoja, ali določena besedna zveza zahteva svoj lastni pomenski opis ali ne, nujno odvisna od vrste in namena slovarja.

3.3.1 Stalne besedne zveze

Stalne besedne zveze so besedne zveze, za katere leksikograf – v skladu z določili v slovarskih smernicah – presodi, da zahtevajo v slovarju opis pomena, ker tega ni mogoče v celoti razbrati iz pomena

5 Tu je potrebno omeniti, da Atkins in Rundell (2008: 167) uporabljata izraz *multi-word expressions* za različne tipe večbesednih enot, kamor prištevata tudi kolokacije. V naši tipologiji so nasprotno kolokacije vključene v t. i. leksikalnogramatične enote in ločene od leksikalnih enot prav na podlagi dejstva, da ne potrebujejo pomenskega opisa.

posameznih sestavin, ali z drugimi besedami, je njihov pomen več kot vsota pomenov posameznih sestavin. Bistveno za njihovo razločevanje od frazeoloških enot, kot ga razumemo v naši tipologiji, je, da take zveze nimajo metaforičnega ali ekspresivnega pomena kot celote, npr. *topla greda*: 1. "prostor, v katerem je mogoče gojiti ali prezimovati rastline", 2. "proces otoplitve zemljine atmosfere in površja". Tipično označujejo določeno terminološko ali strokovno vsebino,[6] pojav ali predmet – navadno imajo torej konkretnega referenta. Stopnja terminološkosti je pri tem različna, hkrati pa je včasih težko prepoznati njihovo pomensko samostojnost in jih tako ločevati od kolokacij, npr. *trebušna votlina, jedilna žlica, zeleni čaj, osnovna šola* ipd. Presoja o tem, ali gre za terminološke večbesedne enote ali kolokacije, je v takih primerih izključno leksikografska, pri vključitvi v slovarsko bazo pa take zveze lahko nastopajo kot kolokacije v povezavi s pomenom katerega od svojih sestavnih elementov, npr. *šola* "ustanova": *osnovna šola, višja šola, visoka šola* ..., in hkrati kot besednozvezne enote, ki predvidevajo definicijo: *osnovna šola* "zakonsko določeno obvezno izobraževanje". Poleg tega stalnih zvez navadno ni mogoče neposredno prevesti v tuji jezik, npr. neposredni prevod zveze *dnevna soba* v ang. day room ne ustreza dejanski angleški ustreznici *living room,* ali pa se določena zveza v tujem jeziku ne pojavlja kot večbesedna enota, npr. slovensko *stara mama :* angleško *grandmother.*

Kot stalne zveze obravnavamo tudi tiste zveze, ki so sicer nastale po metaforični poti (prim. *črna luknja* v 1. pomenu), vendar je njihova vloga v prvi vrsti poimenovalna in ne vrednotenjska, npr. *črna luknja* 1. "pojav v vesolju". Take zveze imajo lahko v katerem od svojih pomenov metaforično vrednost, npr. "nepojasnjen vzrok za izginotje česa", kar jih v konkretnem pomenu uvršča med frazeološke enote. Z vidika vključitve v slovarsko bazo je razlikovanje metaforičnosti od nemetaforičnosti pomena zveze kot celote manj pomembno, pomembno pa je táko razlikovanje v načinu pomenskega opisa, ki ga določa tip in namen konkretnega slovarja.

6 Mogoče je govoriti tudi o žargonizmih ali determinologiziranih (poljudno strokovnih) izrazih.

3.3.2 Frazeološke enote

Tudi frazeološke enote so samostojne večbesedne leksikalne enote z lastno pomensko vrednostjo, ki pa imajo, kot rečeno – za razliko od terminoloških enot – metaforični (imenovan tudi preneseni, konotativni ipd. pomen). S komunikacijskega vidika to načeloma pomeni, da želimo z njimi povedati kaj bolj opazno, ekspresivno, drugače, pri čemer imamo v jeziku navadno na voljo tudi nevtralnejše poimenovanje, npr. *delati iz muhe slona* : *pretiravati*. Gre torej za frazeologijo (idiomatiko) v najožjem pomenu, pri čemer je tudi znotraj frazeoloških enot mogoče prepoznati različne strukturno-pomenske tipe, npr. besednozvezne FE: *začarani krog*; stavčne FE oz. besedilno zaključene pregovore in (iz)reke: *čas je denar, počasi se daleč pride*; izraze s pragmatično in vrednotenjsko vlogo: *za vraga, kapo dol* ter izraze v različnih prislovnih, npr. *ena na ena, bolj ali manj* itd. ali sporočanjskih vlogah: *dober tek, vesel božič* ipd.

4 Kolokacija kot slovarska enota

Definicija kolokacije kot leksikalnega fenomena je prvi pogoj za določitev slovarsko relevantnih kolokacij. Ob tem, da je določena kombinacija besed prepoznana kot kolokacija, ta nima nujno tudi enakovredne obvestilne vrednosti za slovarske uporabnike. Nekatere kolokacije so občutno pogostejše kot druge, nekatere vsebujejo zelo splošne in vsebinsko izpraznjene elemente ipd. Odločitve v zvezi z izborom kolokacij, primernih za vključitev v slovar, so v posameznih slovarjih različne in temeljijo na različnih kriterijih. Hkrati je slovarska relevantnost projektno specifična, saj je pri slovarju kolokacij mogoče pričakovati višji prag vključenosti kot pri splošnem slovarju, kjer je fokus na najbolj tipičnih in povednih primerih.

V nadaljevanju prispevka predstavljamo statistična, skladenjska in pomenska izhodišča za luščenje kolokacij, ki so primerne za vključitev v slovar. Naš namen je bil doseči zadostno stopnjo relevantnih in ustreznih avtomatsko izluščenih kolokacij, da jih bo mogoče neposredno ponuditi slovarskim uporabnikom, hkrati pa z analizo

izluščenih podatkov predvideti postopke leksikografske analize pri nadaljnjih posodobitvah slovarja (Kosem idr. 2018b).

4.1 Statistični parametri

Statistične parametre, ki bi zagotavljali najboljši izplen dobrih[7] kolokacij v orodju Sketch Engine, smo prilagajali v več iteracijah (Gantar idr. 2016). Ključni odločitvi, ki smo ju sprejeli na podlagi jezikoslovne evalvacije, sta določitev različnih frekvenčnih skupin za leme znotraj besedne vrste in nastavitev različnih parametrov za posamezne vrednosti znotraj vsake frekvenčne skupine, in sicer: minimalna pogostost kolokatorja, minimalna pogostnost gramatične relacije, minimalna statistična vrednost (tj. logDice) kolokatorja in minimalna statistična vrednost gramatične relacije.

Dodatni parametri, vezani na vrednosti kolokatorja in gramatične relacije, so sicer zmanjšali obseg izluščenih podatkov na relevantnejše primere, hkrati pa so razkrili nove probleme, kot npr. nezadostno število izluščenih kolokacij zlasti za manj frekventne pomene pri visokofrekventnih večpomenskih besedah. V drugi iteraciji smo zato upoštevali izluščene kolokacijske kandidate na podlagi dveh združenih statističnih mer: logDice in absolutne frekvence (več o tem v Gantar idr. 2016).

4.2 Skladenjske strukture

Skladenjske strukture, v katerih se pojavljajo kolokacije, so pri luščenju slovarsko relevantnih kolokacij pomemben parameter, zato je njihov nabor temeljil na predhodni leksikografski analizi pri izdelavi leksikalne baze za slovenščino (Gantar 2015). Vendar pa vse strukture, registrirane pri izdelavi Leksikalne baze, ki je bila v prvi vrsti namenjena izdelavi splošnega slovarja, niso bile relevantne tudi za luščenje kolokacij. Načeloma je mogoče reči, da so se kot kolokacijsko relevantne pokazale strukture, ki so v slovenščini tudi sicer najpogostejše (Tabela 1), čeprav so nekatere, zlasti v smislu

7 »Dobrih« predvsem v smislu izogibanja očitnim napakam oz. nekolokacijam (npr. *stati aranžma*).

vključevanja pomensko splošnih besed, kot so nekateri splošni pridevniki in prislovi ter pomensko izpraznjeni glagoli, npr. *biti* in *postati*, izkazovale tudi kolokacije z manjšo obvestilnostjo.

Tabela 1: Najpogostejše skladenjske strukture v bazi Kolokacijskega slovarja sodobne slovenščine.

	Skladenjska struktura	Število kolokacij v bazi KSSS
1	pridevnik + samostalnik	1.196.130
2	samostalnik + samostalnik v rodilniku	870.956
3	glagol + samostalnik v tožilniku	524.139
4	prislov + glagol	359.459
5	glagol + predlog 'v' + samostalnik v mestniku	351.209

Z vidika relevantnosti za vključitev v slovarske vire so se kot problematične pokazale zlasti strukture, ki so podrobneje določale glagolsko komponento v smislu nedoločnikov in povratnih glagolov, saj bodisi niso zagotavljale ustreznih kolokacij bodisi je bilo ustrezno luščenje kolokacij tega tipa zagotovljeno z drugimi strukturami, npr. samostalnik + glagol v nedoločniku → samostalnik + glagol v tožilniku: *zavračati prezir*. Prav tako smo izločili strukture, ki so predvidevale odvisniške elemente, npr. *zamaknjen, tako da*, in strukture, ki so se pokazale kot relevantne predvsem za luščenje stavčnih vzorcev: kdo/kaj glagol komu kaj. Kot poseben problem pri naboru slovarsko relevantnih kolokacijskih struktur je treba izpostaviti strukture, ki ob relevantnih kolokacijah, npr. samostalnik v imenovalniku + samostalnik v imenovalniku, *raketa nosilka – rakete nosilke, gasilec veteran – gasilca veterana*, z ujemanjem obeh elementov v vseh sklonih paradigme izluščijo tudi številne druge kolokacije, ki so sicer zajete z drugimi strukturami, npr. samostalnik v imenovalniku + samostalnik v rodilniku, npr. *golf igrišče – golf igrišča*.

Ob tem je potrebno izpostaviti, da zahteva skladenjska struktura v slovenščini, ki je morfološko bogat jezik, tudi ustrezne morfološke prilagoditve kolokacijskih elementov v strukturi, kot je npr. ujemanje v spolu in številu (*rdeč -> rdeč**a** jagod**a**; jesenski -> jesensk**o***

*listj**e***) ter ustrezni sklon kolokatorja (*olupiti jabolk**o***$_{\text{TOŽILNIK}}$; *črv v jabolk**u***$_{\text{MESTNIK}}$).[8] Poleg omenjenega se je pokazalo, da je za določene kolokacije ključno tudi preferenčno ali izključno pojavljanje elementov kolokacije v določeni besedni obliki, ki je bodisi ustaljena bodisi izrazito izstopa v številu ali sklonu, npr. *delavnica za otrok ->; delavnic**e** za otrok**e**, oprati jagod**o** -> oprati jagod**e***. Problem smo v prvi fazi luščenja reševali s postprocesiranjem, kjer smo elementom vsake gramatične relacije na podlagi Leksikalne baze za slovenščino (Gantar idr. 2013) avtomatsko pripisali podatek, ali gre za iztočnico, predlog ali kolokator, ter temu ustrezne morfološke podatke na podlagi oblikoslovnega leksikona Sloleks 1.2 (Dobrovoljc idr. 2015; Dobrovoljc idr. 2017), tj. spol, sklon in število kolokacijskega elementa v strukturi.

4.3 Pomenska obvestilnost

Izluščene kolokacije smo z vidika slovarske relevantnosti ovrednotili v jezikoslovni analizi, končni namen pa je bil opredeliti parametre, ki nam bodo pomagali izbrati kolokacije, primerne za vključitev v slovarsko bazo, in opredeliti način njihovega prikaza v slovarskem vmesniku. S tem v mislih smo kolokacije preverjali z vidika njihove pomenske obvestilnosti (močne – šibke kolokacije), z vidika ustreznosti skladenjske zgradbe in z vidika prevladujoče oblike v rabi in korpusnih zgledih.

Evalvacija je jasno potrdila različne stopnje kolokabilnosti med elementi kolokacije, ki v veliki meri odločajo tudi o slovarski relevantnosti kolokacije kot celote. Kot smo izpostavili že pri tipologiji besednih zvez, se kolokacije na eni strani dotikajo besednih zvez, ki vzpostavljajo trdno notranjo povezavo (npr. *trda tema, debela denarnica*), na drugi strani pa obstajajo kolokacije brez »močnih« kolokatorjev, kjer se besede, če citiramo M. Rundella,[9] lahko (in tudi se)

8 Funkcija Besedna skica (ang. Word Sketch), ki smo jo uporabljali v prvi fazi luščenja, namreč prikazuje zgolj seznam kolokatorjev v osnovni obliki ne glede na ustrezno obliko v kolokacijski strukturi in ne glede na prisotnost npr. predložnega elementa v kolokaciji.

9 M. Rundell: Creating and using the Macmillan Collocations Dictionary: https://www.macmillandictionary.com/collocations/features.html.

sopojavljajo tako rekoč s katerokoli besedo, dokler je kombinacija smiselna. Čeprav z našega seznama lem za luščenje kolokacij nismo izločili splošnih besed, kot sta npr. *hiša* in *kupiti*, je velika večina kandidatov, ki po mnenju jezikoslovcev niso pomensko dovolj obvestilni za vključitev v slovar, prav tega tipa (Pori in Kosem 2018). Čeprav se nam te kolokacije za vključitev v kolokacijski slovar niso zdele relevantne, smo jih ohranili v bazi podatkov, ker ustrezajo izbranim statističnim in skladenjskim merilom in jih bo v prihodnjih luščenjih mogoče uporabiti za filtriranje na novo izluščenih kandidatov s šibko kolokacijsko vrednostjo.

Med opaznejšimi lastnostmi izluščenih kolokacij je bilo tudi prekrivanje šibkih kolokacij z obsežnejšim nizom besed, ki smo jih v naši tipologiji prepoznali kot razširjene kolokacije in skladenjske zveze. Zveze kot *zadevati podočnjake*, *formalen smisel*, *zveza z gradnjo* same na sebi ne tvorijo smiselnih celot, saj so del širših zvez z leksikalno-gramatično vrednostjo: *kar zadeva (podočnjake), v (formalnem) smislu, v zvezi z (gradnjo)*. Dodajanje takih primerov na seznam večbesednih slovarskih enot na eni strani omogoča povratno luščenje iz korpusa, na drugi strani pa izogibanje slabim kolokacijskim kandidatom pri nadaljnjih iteracijah.

Samostojen problem pri vrednotenju slovarske relevantnosti izluščenih kolokacijskih kandidatov so tudi lastnoimenske kolokacije, tj. kolokacije, ki so v celoti lastno ime in pogosto odraz kulturne in jezikovne tipike, npr. *Vesele Štajerke*, ter kolokacije, ki vsebujejo lastnoimenske kolokatorje, npr. *prestolnica Lombardije*, *premagati Slovaško*. Ti primeri z vidika slovarske relevantnosti niso povsem enakovredni, kar se je pokazalo tudi pri različnem vrednotenju v jezikoslovni evalvaciji. Medtem ko je večina ocenjevalcev kolokacije tipa *Vesele Štajerke* označila kot nerelevantne za vključitev v slovar, je bila stopnja strinjanja glede izključitve kolokacij tipa *prestolnica Lombardije* manjša, saj so ocenjevalci prepoznavali pomembnost tako kolokacijske trdnosti take zveze kot tudi semantično indikativnost, ki jo tvori zveza *prestolnica* + država/regija. Čeprav ostajajo nekateri dobri argumenti za prikazovanje lastnoimenskih kolokacij tudi v slovarju (prim. Hudeček in Mihaljević 2020), smo se pri oblikovanju

slovarske baze odločili, da bomo te enote obravnavali ločeno in jih v bazi definirali kot večbesedne lastnoimenske enote.

Dejstvo, da je kolokacija v izhodišču statistično izstopajoč pojav, izpostavlja tudi njeno oblikovno ustaljenost, ki odloča o podobi kolokacije, kot bo vidna slovarskim uporabnikom. Evalvacija je pri prepoznavanju oblikovne trdnosti izpostavila vlogo števila, kjer npr. semantične lastnosti elementa kolokacije bodisi zahtevajo **stresti bonbon → stresti bonbone* bodisi preferirajo needninsko obliko: *finančna težava → finančne težave*. Trdnost kolokacije je lahko vezana tudi na obliko pridevnika v določeni stopnji, npr. presežniku: **blizek bife → bližnji bife*. Taki primeri, če so zastopani v osnovni obliki, ne odražajo jezikovne tipike ali pa delujejo napačno, zato je pri nadaljnjem določanju parametrov za luščenje kolokacij iz korpusa smiselno opredeliti kolokacijske elemente tudi na ravni morfoloških oblik. Možnost prepoznavanja tipične kolokacijske oblike je vključena tudi v funkcijo The longest-commonest match (Kilgarriff idr. 2015) v orodju Sketch Engine, ki pa z vidika trenutnih rezultatov za slovenščino potrebuje izboljšave, saj bodisi ne izlušči ustreznih zvez, čeprav te – na podlagi ročnih pregledov – obstajajo, bodisi izlušči zaporedje, ki presega obseg ene kolokacije.

5 Zaključek in prihodnje delo

Kolokacije so tip večbesednih enot, ki so zaradi svojih leksikalnih lastnosti pomembna sestavina slovarjev. Izhajajoč iz pristopov, ki definirajo kolokacije v odnosu do drugih besednih zvez in z vidika njihovega avtomatskega prepoznavanja v korpusu, jih je mogoče opredeliti s treh temeljnih vidikov: statističnega, skladenjskega in pomenskega. Kot prikazujemo v prispevku, so vsi trije vidiki med seboj tesno prepleteni in zahtevajo podrobne odločitve tako pri vzpostavljanju razmerij z drugimi tipi večbesednih enot kot pri načinu vključevanja v strojno procesljive slovarske vire. Pri določanju definicijskih lastnosti kolokacij v razmerju do drugih večbesednih enot postavljamo v izhodišče slovarsko merilo, kjer večbesedne enote ločimo glede na to, ali predvidevajo kakršenkoli slovarski opis, pač

odvisno od slovarskega koncepta in namembnosti, ali ne. Zadnje, imenovane proste besedne zveze, ločujemo od slovarskih enot, ki so v temelju dveh vrst: leksikalne enote predstavljajo večbesedne enote, katerih pomen je več kot vsota pomenov sestavin, zato v slovarju potrebujejo razlago pomena. Znotraj tega izpostavljamo dva tipa: stalne besedne zveze in frazeološke enote. Od večbesednih leksikalnih enot ločujemo heterogeno množico leksikalno-gramatičnih enot, ki ne zahtevajo pomenske razlage, so pa relevantni deli slovarja v smislu tipične zgradbe, skladenjske vloge ali vloge diskurznega označevalca. Učinkovito prepoznavanje različnih tipov večbesednih enot nam omogoča boljšo organizacijo in povezljivost podatkov v slovarski bazi.

Kot je pokazala evalvacija avtomatsko izluščenih kolokacij iz korpusa, prinaša praktična aplikacija teoretičnih izhodišč za slovarske namene nove izzive tako v smislu izboljševanja parametrov za avtomatsko luščenje kot tudi pri prepoznavanju slovarsko relevantnih kolokacij in zadovoljevanju uporabniških pričakovanj. Naša prizadevanja bodo še naprej usmerjena v čim boljše rezultate avtomatskega luščenja, kar v prvi vrsti pomeni v čim večji meri znebiti se slabih kolokacij, ki izhajajo iz napak (ali odločitev) pri morfosintaktičnem označevanju korpusa, in zagotoviti njihovo ustrezno in hkrati najbolj prepoznavno obliko, v kateri nastopajo v slovarju. Tudi vprašanje slovarske relevantnosti pri nadaljnjih izboljšavah ni zgolj vprašanje statistične relevantnosti, ampak predvsem vprašanje semantične obvestilnosti, ki jo določajo slovarski uporabniki.

Trenutno preizkušamo luščenje kolokacij na podlagi na novo definiranih skladenjskih struktur, ki jih določa število, zaporedje in tip kolokacijskih elementov v njej. Nov način luščenja za razliko od gramatičnih relacij, definiranih v Besednih skicah, vključuje tudi odvisnostna razmerja med elementi kolokacije, vsak element kolokacije pa je v zapisu definiran tudi na morfološki ravni in na ravni reprezentacije, ki določa zapis oblike kolokacije v bazi oz. slovarju. Začetni rezultati luščenja na tej podlagi so obetavni in rešujejo nekatere izpostavljene probleme, kot so zmanjšanje števila neustreznih kolokacij ter prevladujoča oblika kolokacije v smislu sklona, števila

in pridevniške oblike, npr. *zadnja leta, različne oblike, dnevni red, širša javnost* namesto manj ustreznih: *zadnje leto, različna oblika, dneven red* in *široka javnost*.

Ključen za naša nadaljnja prizadevanja ostaja cilj izdelave celostne digitalne slovarske baze, v kateri bodo kolokacije obravnavane kot samostojni tip večbesednih enot in jim bodo pripisane informacije, ki jih pričakujejo slovarski uporabniki, raziskovalna ter računalniška skupnost. Pri tem ni odveč znova poudariti, da bo baza pod odprto kodo na voljo celotni raziskovalni skupnosti in jo bo mogoče uporabiti pri nadaljnjih izdelavah in izboljšavah jezikovnih virov za slovenščino.

Zahvala

V prispevku so opisani rezultati, ki so nastali v okviru projekta *Kolokacije kot temelj jezikovnega opisa: semantični in časovni vidiki* (J6-8255), projekta *Nova slovnica sodobne standardne slovenščine: viri in metode* (J6-8256) 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

Aisenstadt, E. (1981): Restricted Collocations in English Lexicology and Lexicography. *ITL - International Journal of Applied Linguistics*, 53 (1), 53–61.

Altenberg, B. (1991): Amplifier Collocations in Spoken English. V S. Johansson in A. B. Stenström (ur.): *English Computer Corpora. Selected Papers and Research Guide:* 127–147. Berlin/New York: Mouton de Gruyter.

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

Baldwin, T. in Kim, S. N. (2010): *Multiword expressions. V Handbook of Natural Language Processing* (2nd ed.). CRC Press, Taylor and Francis Group.

Benson, M., Benson, E. in Ilson, R. (1986): *The BBI Dictionary of English Word Combinations.* John Benjamins, Amsterdam.

Berry-Rogghe, G. L. (1973): The computation of collocations and their relevance in lexical studies. In *The computer and literal studies*: 103–112. Edinburgh/New York: University Press.

Bhatia, A., Bonial, C., Candito, M., Cap, F., Cordeiro, S., Foufi, V., Gantar, P., ..., Walsh, A. (2017): PARSEME shared task 1.1 annotation guidelines (last updated on November 30, 2017). Dostopno prek: http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.1/ (27. 4. 2021).

Biber, D. (1993): Representativeness in Corpus Design. *Literary and Linguistic Computing*, 8 (4): 243–257.

Brezina, V., McEnery, T. in Wattam, S. (2015): Collocations in context: a new perspective on collocation networks. *International Journal of Corpus Linguistics,* 20 (2): 139–173.

Candito, M., Cap, F., Cordeiro, S., Foufi, V., Gantar, P., Giouli, V., ..., Vincze, V. (2016): PARSEME shared task 1.0 annotation guidelines - version 1.6b (last updated on November 26, 2016). Dostopno prek: https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.0/ (27. 4. 2021).

Church, K. W., Gale, W., Hanks, P. in Hindle, D. (1991): Using statistics in lexical analysis. V U. Zernik (ur.): *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon*: 116–164. Erlbaum, Hillsdale, NJ.

Church, K. in Hanks, P. (1990): Word association norms, mutual information and lexicography. *Computational Linguistics*, 6 (1): 22–29.

Clear, J. (1993): From Firth principles. Computational Tools for the Study of Collocation. V M. Baker, G. Francis in E. Tognini-Bonelli (ur.): *Text and Technology. In honour of John Sinclair*: 271–292. Philadelphia, Amsterdam: John Benjamins,

Cowie, A. P. (1981): The treatment of collocations and idioms in learners' dic-tionaries. In A. P. Cowie (ur.): *Lexicography and its Pedagogical Applica-tions* (Thematic issue). *Applied Linguistics,* 2 (3): 223–235.

Dobrovoljc, K. (2017): Multi-word discourse markers and their corpus-driven identification: the case of MWDM extraction from the reference corpus of spoken Slovene. *International journal of corpus linguistics*, 22 (4): 551–582.

Dobrovoljc, K. (2018): Raba tipično govorjenih diskurznih označevalcev na spletu. *Slavistična revija*, 66 (4): 497–513.

Dobrovoljc, K., Krek, S. in Erjavec, T. (2017): Leksikon besednih oblik Sloleks in smernice njegovega razvoja. V V. Gorjanc, P. Gantar, I. Kosem in S. Krek (ur.): *Slovar sodobne slovenščine: problemi in rešitve:* 80–105. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani.

Dobrovoljc, K., Krek, S., Holozan, P., Erjavec, T. in Romih, M. (2015), *Morphological lexicon Sloleks 1.2*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1039.

Evert, S. (2004): The statistics of word cooccurrences: Word pairs and collocations. PhD Thesis, University of Stuttgart.

Evert, S. (2009): Corpora and collocations. V A. Lüdeling in M. Kytö (ur.): *Corpus Linguistics: An International Handbook*: *Vol. 2*: 1212–1248. Berlin/New York: Mouton de Gruyter.

Fellbaum, C. (2015): Syntax and grammar of idioms and collocations. V T. Kiss in A. Alexiadou (ur.): *Syntax: Theory and analysis: Vol. 2*: 776–802. Berlin/New York: Mouton de Gruyter.

Fontenelle, T. (1994): What on earth are collocations. *English today*, 10 (4): 42–48.

Gantar, P. (2015): *Leksikografski opis slovenščine v digitalnem okolju*. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani. Dostopno prek: http://www.ff.uni-lj.si/sites/default/files/Dokumenti/Knjige/e-books/leksikografski.pdf (27. 4. 2021).

Gantar, P., Arhar Holdt, Š., Čibej, J. in Kuzman, T. (2019a): Structural and semantic classification of verbal multi-word expressions in Slovene. *Prispevki za novejšo zgodovino*, 59 (1): 99–119.

Gantar, P., Colman, L., Parra Escartín, C. in Marínez Alonso, H. (2019b): Multiword Expressions: Between Lexicography and NLP. *International Journal of Lexicography*, 32 (2): 138–162.

Gantar, P., Kosem, I. in Krek, S. (2016): Discovering automated lexicography: the case of Slovene lexical database. *International journal of lexicography*, 29 (2): 200–225.

Gorjanc, V., Gantar, P., Kosem, I. in Krek, S. (ur.). (2017): *Dictionary of Modern Slovene: Problems and Solutions*. Ljubljana: Ljubljana University Press, Faculty of Arts.

Grčar, M., Krek, S. in Dobrovoljc, K. (2012): Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. V T. Erjavec in J. Žganec Gros (ur.): *Zbornik Osme konference Jezikovne tehnologije*: 89–94. Ljubljana: Institut Jožef Stefan.

Gries, S. (2013): 50-something years of work on collocations. *International Journal of Corpus Linguistics*, 18 (1): 137–165.

Halliday, M. A. K. (1966): Lexis as a Linguistic Level. *Journal of Linguistics*, 2 (1): 57–67.

Hausmann, F. J. (1984): Wortschatzlernen ist Kollokationsletnen. Zum Lehren und Lernen französischer Wortverbindungen. *Praxis des neusprachlichen Unterrichts*, 31: 395–406.

Hausmann, F. J. (1989): Le dictionnaire de collocations. V F. J. Hausmann idr. (ur.): *Wörterbücher: ein internationales Handbuch zur Lexikographie*: 1010–1019. Berlin/New York: De Gruyter.

Herbst, T. (1996): What are Collocations: Sandy Beaches or False Teeth. *English Studies*, 4: 379–393.

Hudeček, L. in Mihaljević, M. (2020): Collocations in Croatian Web Dictionary – Mrežnik. *Slovenščina 2.0: Empirical, Applied and Interdisciplinary Research*, 8 (2): 78–111.

Khokhlova, M. in Benko, V. (2020): Size of Corpora and Collocations: the Case of Russian. *Slovenščina 2.0: Empirical, Applied and Interdisciplinary Research*, 8 (2): 58–77.

Kilgarriff, A., Baisa, V., Rychlý, P. in Jakubíček, M. (2015): Longest–commonest Match. V I. Kosem, M. Jakubíček, J. Kallas in S. Krek (ur.): *Electronic Lexicography in the 21st Century: Linking Lexical Data in the Digital Age. Proceedings of the eLex 2015 Conference*: 397–404. Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd.

Kilgarriff, A., Rychly, P., Smrz, P. in Tugwell, D. (2004): The Sketch Engine. V G. Williams in S. Vessier (ur.): *Proceedings of the 11th EURALEX In--ternational Congress*: 105–116. Lorient: France.

Klemenc, B., Robnik Šikonja, M., Fürst, L., Bohak, C. in Krek, S. (2017): Tech-nological design of a state-of-the-art digital dictionary. V V. Gorjanc, P. Gantar, I. Kosem in S. Krek (ur.): *Dictionary of Modern Slovene: Problems and Solutions*: 10–22. Ljubljana: Ljubljana University Press, Faculty of Arts.

Kosem, I., Husák, 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 the eLex 2011 Conference*: 151–159. Ljubljana: Trojina, Institute for Applied Slovene Studies.

Kosem, I., Gantar, P., Krek, S., Arhar Holdt, Š., Čibej, J., Laskowski, C., Pori, E., Klemenc, B., Dobrovoljc, K., Gorjanc, V. in Ljubešić, N. (2018a): *Kolokacije 1.0: Kolokacijski slovar sodobnega slovenskega jezika*. Dostopno prek: https://viri.cjvt.si/kolokacije/slv/# (27. 4. 2021).

Kosem, I., Krek, S., Gantar, P., Arhar Holdt, Š., Čibej, J. in Laskowski, C. (2018b): Kolokacijski slovar sodobne slovenščine. V D. Fišer in A. Pančur (ur.): *Zbornik konference Jezikovne tehnologije in digitalna humanistika*: 133–139. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani. Dostopno prek: http://nl.ijs.si/jtdh18/JTDH-2018--Proceedings.pdf (27. 4. 2021).

Krek, S., Gantar, P., Kosem, I., Gorjanc, V. in Laskowski, C. (2016): Baza kolokacijskega slovarja slovenskega jezika. V T. Erjavec in D. Fišer (ur.): *Proceedings of the Conference on Language Technologies and Digital Humanities*: 101–105. Ljubljana: Academic Publishing Division of the Faculty of Arts.

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

Manning, C. D. in Schütze, H. (1999): *Foundations of statistical natural lan--guage processing. Cambridge, Massachusetts: The MIT Press*, Chap. 5. Collocations.

Moon, R. (1998): *Fixed Expressions and Idioms, a Corpus-Based Approach*. Oxford: Oxford University Press.

Pecina, P. (2009): Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44 (1–2): 137–158.

Pori, E. in Kosem, I. (2018): In the Search of Lexicographically Relevant Collocation: The Example of Grammatical Relations Containing Adverbs. *Slovenščina 2.0: Empirical, Applied and Interdisciplinary Research*, 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.

Pori, E. in Kosem, I. (2021): Evalvacija avtomatskega luščenja kolokacijskih podatkov iz besednih skic v orodju Sketch Engine. V I. Kosem (ur.): *Kolokacije v slovenščini*: 43–77. Ljubljana: Znanstvena založba Filozofske fakultete.

Schmid, H. J. (2003): Collocation: hard to pin down, but bloody useful. *ZAA*, 51 (3): 235–258.

Seretan, V. (2010): *Syntax-Based Collocation Extraction* (1st ed.). Berlin, Heidelberg: Springer-Verlag.

Sinclair, J. (1966): Beginning the Study of lexis. V Bazell idr. (ur.): *In Memory of J.R. Firth*: 410–430. London: Longman.
Sinclair, J. (1991): *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
Stubbs, M. (1995): Collocations and Semantic Profiles. On the cause of the Trouble with Quantitative Studies. *Functions of Language*, 2: 23–55.
Wiechmann, D. (2008): On the computation of collostruction strength. *Corpus Linguistics and Linguistic Theory*, 42: 253–290.