Urednik:
IZTOK KOSEM

KOLOKACIJE V SLOVENŠČINI

Univerza *v Ljubljani*

Kataložni zapis o publikaciji (CIP) pripravili v
Narodni in univerzitetni knjižnici v Ljubljani

COBISS.SI-ID= 78637571
ISBN ISBN 978-961-06-0537-9 (PDF)

www.slovenščina.eu
sporazumevanje

Kolokacije v slovenščini

Urednik: Iztok Kosem

Univerza v Ljubljani
FILOZOFSKA FAKULTETA

Kolokacije v slovenščini

Zbirka: Sporazumevanje (e-ISSN 2738-4527)
Urednika zbirke: Špela Arhar Holdt, Vojko Gorjanc

Urednik: Iztok Kosem
Recenzenta: Tamara Mikolič Južnič, Darinka Verdonik
Tehnično urejanje: Jure Preglau
Prelom: Aleš Cimprič
Oblikovanje naslovnice: Kofein dizajn

Založila: Znanstvena založba Filozofske fakultete Univerze v Ljubljani
Izdal: Center za jezikovne vire in tehnologije Univerze v Ljubljani
Za založbo: Mojca Schlamberger Brezar, dekanja Filozofske fakultete

Ljubljana, 2021
Prva izdaja, e-izdaja
Publikacija je brezplačna.
Publikacija je dostopna na: https://e-knjige.ff.uni-lj.si

Projekt Kolokacije kot temelj jezikovnega opisa: semantični in časovni vidiki (šifra ARRS: J6-8255) je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna.

Raziskovalni program Jezikovni viri in tehnologije za slovenski jezik (šifra ARRS: P6-0411) je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna.

Raziskovalni program Slovenski jezik - bazične, kontrastivne in aplikativne raziskave (šifra ARRS: P6-0215) je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna.

Kazalo vsebine

Uvod

Kolokacija kot tipična sopojavitev vsaj dveh leksikalnih enot je jezikovni pojav, ki je že dolgo predmet raziskav na različnih področjih, od korpusnega jezikoslovja in leksikografije do učenja in poučevanja jezikov. Pa vendar izzivov za jezikoslovce ne zmanjka – korpusi postajajo vse večji, kolokacijskih podatkov je posledično vse več in iščejo se metode za čim bolj učinkovito prepoznavanje, spremljanje rabe in selekcijo kolokacij za različne namene. Ob naraščajočem vključevanju kolokacij v slovarje in druge leksikalne vire pa se zastavljajo tudi vprašanja o odnosu uporabnikov do tovrstnih podatkov.

Pričujoča monografija, ki je nastala kot rezultat temeljnega raziskovalnega projekta KOLOS (*Kolokacije kot temelj jezikovnega opisa: semantični in časovni vidiki*), prinaša v znanstveni prostor pomembna nova teoretska in metodološka spoznanja o različnih vidikih kolokacij v slovenskem jeziku. Hkrati so teoretska razmišljanja, rezultati opravljenih analiz in evalvacij ter novo nastali viri in orodja relevantni za vpenjanje slovenske jezikoslovne skupnosti v mednarodni prostor.

V prvem prispevku Polona Gantar, Simon Krek in Iztok Kosem predstavijo opredelitev kolokacije za namene izdelave digitalnih slovarskih virov za slovenščino. Pri tem opredelijo tri ključne kriterije, ki določajo kolokacijo, nato pa prikažejo mesto kolokacije v odnosu do drugih tipov večbesednih enot. Prispevek izpostavi koncept slovarske relevantnosti kolokacij in merila za njihovo vključevanje v slovarske vire.

Eva Pori in Iztok Kosem v svojem prispevku opišeta rezultate in ugotovitve obsežne evalvacije avtomatskega luščenja kolokacijskih podatkov z uporabo orodja Sketch Engine z jezikoslovnega in metodološkega vidika. Temeljni namen prispevka je predstaviti glavne prednosti in slabosti postopka prepoznave kolokacij, ki se uporablja

pri izdelavi številnih slovarskih in drugih jezikovnih virov, ter predlagati izboljšave.

S statistično problematiko kolokacij se ukvarjajo Iztok Kosem, Nataša Logar, Kaja Dobrovoljc in Nikola Ljubešić, ki predstavijo rezultate treh ločenih preizkusov, od katerih se dva ukvarjata z razvrščanjem kolokacij (besedne vložitve in deltaP), eden pa s statistično prepoznavo slovarsko nerelevantnih kolokacij (razpršenost). Prispevek je inovativen v tem, da gre za prvo sistematično evalvacijo treh statističnih metod na slovenskih podatkih.

Prispevek Špele Arhar Holdt zaokrožuje analize statističnega vidika kolokacij s predstavitvijo rezultatov anketne raziskave med 457 uporabniki, ki so podali mnenje o najbolj optimalnem razvrščanju kolokacij in tako imenovanem jagodnem izboru, tj. kolokacijskih podatkih, ki so v kolokacijskem slovarju uporabnikom predstavljeni najprej oz. na najbolj vidnem mestu.

Semantični vidik kolokacij je naslovljen v prispevku Iztoka Kosma in Eve Pori, ki predstavita ontologijo semantičnih tipov za samostalnike SLONEST-sam na podlagi slovenskega gradiva. Ontologija bo služila kot osnova pri opredeljevanju pomenskih konceptov v digitalni slovarski bazi, pa tudi pri (avtomatskem) gručenju kolokacij in prepoznavanju posameznih pomenov. Avtorja podrobno opišeta vse krovne kategorije in njihove podkategorije ter izpostavita podobnosti oz. razlike glede na obstoječe ontologije, ki se uporabljajo v leksikografske namene, v mednarodnem in slovenskem prostoru.

Prispevek Iztoka Kosma in Jake Čibeja prinaša v slovenski prostor pomembne osvetlitve o diahroni rabi oz. časovnih trendih kolokacij v slovenščini. Avtorja prepoznavata različne vzorce časovnih trendov z vidika relevantnosti za jezikovni opis, pa tudi z vidika (nadaljnjih) analiz obnašanja kolokacij glede na družbene trende, besedilne žanre, socio-politično situacijo itd. V prispevku prikažeta metode zaznavanja naraščajoče oz. padajoče rabe kolokacij in z njimi povezanih pomenov. Metoda zaznavanja trendov pa omogoča tudi prepoznavanje korpusnega šuma in slovarsko nerelevantnih kolokacij, kar je metodološko uporabno tudi pri izdelavi in posodabljanju jezikovnih virov.

Eva Pori, Iztok Kosem, Jaka Čibej in Špela Arhar Holdt predstavijo rezultate kvalitativne jezikoslovne analize uporabniške evalvacije vmesnika Kolokacijskega slovarja sodobne slovenščine, pri kateri je bila uporabljena metoda glasnega razmišljanja. Prispevek prinaša številna nova spoznanja glede organizacije podatkov in njihove predstavitve v spletnem vmesniku, pa tudi o odnosu uporabnikov do avtomatsko pridobljenih podatkov. Prispevek ponudi tudi konkretna priporočila za izboljšavo slovarskega vmesnika pri nadaljnjih posodobitvah vira.

Podrobne analize uporabniških spoznanj se loteva tudi prispevek Špele Arhar Holdt, ki predstavlja rezultate uporabniške evalvacije kolokacijskih podatkov v Slovarju sopomenk sodobne slovenščine. Avtorica v prispevku uporabniško oceno podrobno analizira z jezikoslovnega vidika, sinergijo obeh vidikov pa združi v konkretna priporočila za izboljšavo metodologije luščenja in načina vključevanja kolokacijskih podatkov v Slovar sopomenk sodobne slovenščine.

Rdeča nit prispevkov v monografiji je njihova izhodiščna dvojna naravnanost: po eni strani se ukvarjajo z jezikoslovno analizo kolokacij z več vidikov ter oceno njihove jezikoslovne ali slovarske relevantnosti, po drugi strani pa ponujajo konkretna metodološka priporočila ali rešitve za izboljšavo postopkov pri pripravi slovarskih virov. Poleg temeljnega namena, ki smo ga zasledovali pri pripravi monografije, namreč prikazati rezultate projekta, nas je pri izboru prispevkov vodila tudi želja po predstavitvi problematike, ki odpira in spodbuja nadaljnje raziskave tega v jeziku tako pomembnega pojava. Vsako novo spoznanje na tem področju ima namreč vrednost tudi za nadaljnjo izgradnjo jezikovne infrastrukture, ki se neposredno kaže v možnosti izboljšanja obstoječih in nastanku novih leksikalnih virov za slovenščino.

Ob izidu monografije bi se rad zahvalil vsem recenzentom, ki so pripomogli k izboljšanju prispevkov, tehničnim sodelavcem, brez katerih priprava podatkov ne bi potekala tako hitro oz. sploh ne bi bila mogoča, in številnim sodelujočim v uporabniških raziskavah.

Še posebna zahvala gre sodelavki Evi Pori, ki mi je priskočila na pomoč v zaključnih korakih, pa tudi dajala prepotrebno spodbudo pri spopadanju z uredniškimi izzivi.

Rad bi se zahvalil tudi Znanstveni založbi Filozofske fakultete za vso potrpežljivost in pomoč pri procesu dokončne priprave monografije.

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.

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

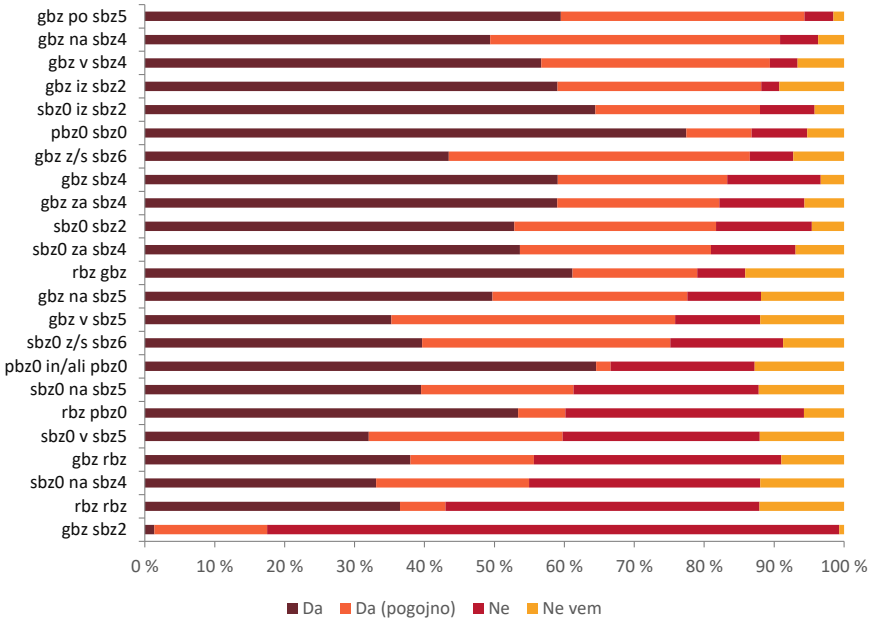
Razporeditev odgovorov označevalcev glede na skladenjsko strukturo prikazuje Slika 1 (prikazanih je 23 struktur z največ kolokacijskimi kandidati).

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.

Razvrščanje in relevantnost kolokatorjev v slovenščini: novi pristopi

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

Nataša LOGAR
Fakulteta za družbene vede, Univerza v Ljubljani

Kaja DOBROVOLJC
Institut Jožef Stefan; Filozofska fakulteta, Univerza v Ljubljani

Nikola LJUBEŠIĆ
Institut Jožef Stefan

Automatic identification of collocations has been significantly improved over the years, however due to larger corpora and thus larger amounts of collocations, the ordering of collocations and the identification of their (lexicographic) relevance have become increasingly important for both lexicographers and researchers. In this paper, we present three separate, but related, experiments that focussed on testing three approaches to collocation extraction: word embeddings, deltaP, and collocate distribution (co-occurrence with a number of different headwords). Word embeddings and deltaP, which have not yet been systematically tested on the Slovene data, have been compared to the association measure logDice used in the Slovene lexicographic projects. The findings showed that, quantitatively, word embeddings perform better than logDice, however the qualitative analysis conducted by the lexicographers revealed that word embeddings perform worse than logDice, especially on the specialised corpus of academic texts. Similarly, logDice performed somewhat better than deltaP, but deltaP showed some potential for identifying (terminological) compounds. It should be pointed out that considerable differences in the performance

of different measures were observed on a headword level. Finally, the distribution information proved to be much more informative, as many collocates that were deemed lexicographically less relevant (i.e. not informative for the users) displayed a high level of distribution. The paper concludes by summarizing the main findings and providing guidelines for their implementation into lexicographic practice.

Keywords: word embeddings, deltaP, distribution, collocation, corpus, Slovene

1 Uvod

Analiza kolokacij je v 21. stoletju doživela velik razmah predvsem zaradi vse večjih korpusov, ki so omogočali prepoznavanje tipičnih besednih kombinacij (gl. npr. Khokhlova in Benko 2020). Velika količina podatkov – še pred tridesetimi leti tako težko dosegljiv cilj – pa je leksikografom in drugim uporabnikom korpusov prinesla novo težavo: ločevanje med pomembnim, tipičnim in splošnim na eni strani ter nepomembnim, posebnim in obrobnim na drugi. Pri kolokacijah, ki so z nastankom korpusov kot jezikovni pojav, ki nedvomno sodi v jezikovni opis, pridobile največ, to težavo rešujejo različne mere kolokacijske jakosti.

Obstaja veliko raziskav o merjenju jakosti kolokacij oz. kolokativnosti (npr. Berry Rogghe 1973; Church in Hanks 1990; Church idr. 1991; Biber 1993; Manning in Schütze 1999; Evert 2004; Gries 2013). Jezikoslovci po vsem svetu skupaj z jezikovnimi tehnologi uporabljajo in razvijajo zlasti različne statistične metode, zato so te tudi redna tema primerjalnih študij. Dober pregled mer povezovalnosti trenutno ponujata dve raziskavi: Wiechmannova (2008), v kateri je avtor primerjal 47 različnih mer, in Pecinova (2009), v kateri je avtor opravil primerjavo več kot 80 mer. Splošne ugotovitve tovrstnih raziskav je smiselno povzel Evert (2009: 1218, 1236): »Različne mere pripeljejo do povsem različnih razvrščanj kolokatorjev.« In še: »Idealna mera ujemanja, ki bi zadostila vsem različnim potrebam, ne obstaja.«

Ena od pogosto uporabljenih mer povezovalnosti, zlasti v leksikografiji, je logDice (Rychlý 2008). Metoda logDice je bila prvič predstavljena v orodju Sketch Engine (Kilgarriff idr. 2004), vodilnem kolokacijskem orodju v leksikografiji in korpusnem jezikoslovju. Posledično je bil logDice uporabljen in vrednoten v mnogih leksikografskih ter terminografskih projektih po vsem svetu, tudi v Sloveniji (prim. Gantar idr. 2012; Gantar 2015; Kosem 2018; Logar in Kosem 2013; Logar idr. 2019). Kljub temu da se je mera logDice v splošnem izkazala za zelo koristno, pa smo tudi pri izdelavi jezikovnih virov za slovenščino zaznali njene pomanjkljivosti. Tako so bile npr. pri pripravi Kolokacijskega slovarja sodobne slovenščine (Kosem idr. 2018) slovarsko relevantne kolokacije v nekaterih primerih razvrščene v spodnji del seznama in pod prag parametrov, določenih za avtomatski izvoz (gl. Gantar idr. 2016), kar za redakcijsko delo ni ugodno. Ker so bile takšne kolokacije večinoma zelo pogoste, smo v tem primeru izvoz kolokacij, razvrščenih po logDice, kombinirali z izvozom kolokacij, razvrščenih po pogostosti, obenem pa smo se začeli ozirati tudi po izboljšavah in novih metodah pridobivanja kolokacij. V zadnjih letih so se namreč pojavile nekatere nove metode, ki naj bi imele prednosti bodisi pri prepoznavanju kolokacij bodisi pri razvrščanju kolokatorjev, nekatere tudi pri obojem. Med njimi sta tudi metoda besednih vložitev (ang. word embedding; Levy in Goldberg 2014) in metoda deltaP (Gries 2013).

Ne glede na metodo luščenja pa ostajajo problematične kolokacije s t. i. pomensko manj obvestilnimi oz. splošno rabljenimi kolokatorji, za katere je značilno, da se lahko sopojavljajo s tako rekoč katerokoli besedo – njihova raba je razpršena. Analize slovenskega gradiva (npr. Pori in Kosem 2018; Pori in Kosem 2021) so namreč pokazale, da so kolokacije s takšnimi kolokatorji velikokrat visoko na seznamih kolokacijskih kandidatov, čeprav v vir (praviloma slovar) na koncu niso vključene.

V prispevku bomo predstavili rezultate treh preizkusov uporabe različnih metod, pri čemer smo se osredotočili na:

a) prepoznavanje kolokacij in razvrščanje kolokatorjev ter
b) razpršenost kolokatorjev in slovarsko relevantnost kolokacij.

Ker se bo v prispevku večkrat pojavil izraz »slovarsko nerelevantna kolokacija«, naj že tu pojasnimo, da smo kolokacijo razumeli kot slovarsko nerelevantno v primeru, ko sicer izpolnjuje nekatere skladenjske, pomenske ali statistične pogoje za kolokacijo (gl. Kosem idr. 2020; Gantar idr. 2021), vendar pa za slovarske uporabnike nima obvestilne vrednosti. Kriteriji za vključitev kolokacij v slovar so seveda odvisni od zahtev posameznega projekta; v našem primeru smo imeli v mislih pripravo kolokacijskega slovarja, v katerem so kolokacije najbolj temeljito obdelane.

Glavni raziskovalni vprašanji, po katerih smo usmerjali raziskavo, sta bili:

1. *Ali lahko z metodo besednih vložitev in metodo deltaP izboljšamo avtomatsko luščenje kolokacij in razvrščanje kolokatorjev v primerjavi z rezultati metode logDice?*
2. *Ali lahko z analizo razpršenosti prepoznamo slovarsko nerelevantne kolokatorje, ki bi jih lahko uporabili za pohitritev redakcijskega dela?*

Metoda dela je natančneje pojasnjena v nadaljevanju pri vsakem preizkusu posebej. Preizkusi so bili, kot rečeno, trije: v prvem smo ocenjevali rezultate metode besednih vložitev (2.1), v drugem rezultate metode deltaP (2.2), v tretjem pa rezultate kolokacijske razpršenosti (2.3). Rezultate smo sproti že tudi komentirali, kratek povzetek vseh analiz in odgovor na obe raziskovalni vprašanji pa nato sledita v razdelku 2.4. V sklepu smo zaključni misli dodali še nekaj leksikografskih priporočil, ki so jih kot smiselna pokazali tukajšnji in siceršnji projektni pregledi številnih ter raznovrstnih kolokacijskih podatkov.

2 Raziskava

2.1 Besedne vložitve

Metoda besednih vložitev (npr. Levy in Goldberg 2014) je postala v zadnjem desetletju pomemben pristop v procesiranju naravnih jezikov. Temelji na vektorski razdalji med besedami, pri čemer upošteva

besedno semantiko. Vektorji pokažejo, v katerih sopojavitvah z drugimi besedami se določena beseda pojavlja ter kolikšna je razdalja med njo in njimi. Lahko bi rekli tudi, da gre pri besednih vložitvah za izračun semantične podobnosti (oz. različnosti) besed na podlagi metrike, torej za matematični podatek o tem, kako zelo so besede pomensko blizu (oz. oddaljene) druga od druge.

Metoda besednih vložitev je bila na več tujih jezikih preizkušena že tudi za potrebe razpoznavanja kolokacij, pri čemer je bil njen uspeh razmeroma ali celo zelo dober (za pregled gl. Ljubešić idr. 2021), v okviru projekta KOLOS pa smo jo ovrednotili na dveh zbirkah kolokacij: zbirki KOLOS oz. Kolokacijskem slovarju sodobne slovenščine 1.0 (Kosem idr. 2018), pridobljeni iz korpusa Gigafida (Logar idr. 2012), in zbirki KAS (Logar idr. 2019), pridobljeni iz istoimenskega korpusa akademske slovenščine (Erjavec idr. 2016).[1]

Preizkus je potekal primerjalno. Zanimalo nas je, kako dobre rezultate daje metoda besednih vložitev, ki – kot rečeno – vključuje semantiko besed in je bila v preizkusu, opisanem v Ljubešić idr. (2021), izvedena s strojnim učenjem na ročno označenih podatkih, v primerjavi z zelo uveljavljeno metodo logDice, ki izhaja iz kombinacije: (a) pogostosti iztočnice in kolokatorja ter (b) pogostosti celotne kolokacije. Predmet preizkusa so bili seznami kolokatorjev (gl. primer v Tabeli 1) – vprašali smo se torej, po kateri metodi dobimo leksikografsko bolj informativen seznam kolokatorjev (zlasti v njegovem vrhu): po metodi logDice ali po metodi besednih vložitev.

Odgovor na vprašanje smo dobili na dva načina:

- **Kvantitativno** po metriki AUC ROC (Davis in Goadrich 2006), ki je trenutno najbolj uporabljana metrika v primerih: (1.) ko imamo v učni množici tako pozitivne kot negativne kandidate (v našem primeru je šlo za zbirko avtomatsko izluščenih kolokacij, ki so bile nato ročno označene kot slovarsko (a) relevantne/pozitivne in (b) nerelevantne/negativne) in (2.) ko sta ti dve skupini po obsegu precej različni (podatkovna zbirka KOLOS je npr. imela skoraj 14.000 pozitivno označenih kolokacij, a le nekaj manj

1 Nadaljevanje tega razdelka pretežno povzema raziskavo Ljubešić idr. (2021); podrobnejše podatke o pripravi, izvedbi in rezultatih zato gl. tam.

Tabela 1: Kolokatorji za iztočnico *dvojček*, razvrščeni po metodi logDice in metodi besednih vložitev; struktura: pridevnik + samostalnik (p0-s0; podatki iz zbirke KOLOS).

Mesto	LogDice	Besedne vložitve
1.	dvojajčen	ameriški
2.	newyorški	newyorški
3.	tušev	novorojen
4.	zloben	nerojen
5.	zraščen	leten
6.	identičen	enodružinski
7.	enodružinski	zloben
8.	sejemski	dveleten
9.	samski	zaporeden
10.	parazitski	sedemleten
11.	atrijski	atrijski
12.	dveleten	rojen
13.	porušen	pravi
14.	beneški	slaven
15.	nerojen	parazitski
16.	novorojen	samski
17.	zaklet	porušen
18.	sedemleten	siamski
19.	soroden	enojajčen
20.	stanovanjski	star
21.	znamenit	dvojajčen
22.	zaporeden	stanovanjski
23.	rojen	soroden
24.	slaven	zraščen
25.	leten	majhen
26.	star	znamenit
27.	enojajčen	podoben
28.	siamski	identičen
29.	podoben	sejemski
30.	pravi	zaklet
31.	ameriški	beneški
32.	majhen	tušev

kot 4.000 negativnih). AUC ROC številsko ovrednoti rezultate razvrščanja, pri čemer dobi najslabše možno razvrščanje oceno 0,0 (vsi negativni kandidati so tu razvrščeni višje kot vsi pozitivni kandidati), najboljše možno razvrščanje dobi oceno 1,0 (vsi pozitivni kandidati so tu razvrščeni višje kot vsi negativni kandidati), medtem ko vmesna ocena 0,5 (ali njena bližina) pomeni, da je bilo razvrščanje (povsem) naključno.

- **Kvalitativno** s pomočjo jezikoslovcev, ki so pregledali različno dolge sezname kolokatorjev v obliki dvojnih stolpcev za skupno 143 iztočnic v 14 različnih slovničnih relacijah (primer takega seznama prikazuje Tabela 1).

Kvantitativni del ocene je pokazal, da dobimo bolj informativen seznam kolokatorjev s pomočjo nove metode, in to na obojih podatkih, iz zbirke KOLOS in iz zbirke KAS. Na drugi strani – pri leksikografih (v raziskavi jih je sodelovalo 9, niso pa vsi ocenjevali vseh podatkov) – pa je bilo večinsko mnenje drugačno: izmed možnih odgovorov 'Stolpec 1 je bolj informativen', 'Stolpec 2 je bolj informativen' in 'Oba stolpca sta približno enako (ne)informativna' so se leksikografi večinoma odločali za zadnji odgovor, tj. odgovor 'niti-niti'.[2] Primerjava med zbirkama KOLOS in KAS je pokazala še to, da se je jezikoslovka, ki je ocenjevala sezname iz KAS-a,[3] le v 4 % primerov odločila, da je razvrščanje kolokatorjev, pridobljeno z metodo besednih vložitev, boljše kot razvrščanje, pridobljeno z metodo logDice, medtem ko so leksikografi pri podatkih iz KOLOS-a dali prednost razvrstitvi po novi metodi pri 23 % iztočnic.

Dokončne primerjalne prednosti nove metode za leksikografske potrebe torej nismo potrdili, lahko pa jo kot način razpoznavanja in razvrščanja kolokatorjev vsekakor postavimo ob bok že uveljavljeni(m). Na tem mestu zato rezultate analize zaključujemo odprto: s še tremi problemsko izbranimi primeri (Tabela 2, 3 in 4), ki kažejo – po našem mnenju – prav ambivalentni 'niti-niti', ko gre za

2 Tudi primer *dvojčka* v Tabeli 1 je tak: 4 (67 %) od 6 jezikoslovcev, ki so ga ocenjevali, se je odločilo, da sta stolpca kolokatorjev približno enakovredna.

3 Ker je šlo le za eno ocenjevalko, je njena ocena zgolj informativna in je ni mogoče posplošiti.

odločitev, kateri seznam ima leksikografsko prednost. Bralci si lahko torej mnenje o rezultatih metode besednih vložitev v primerjavi z rezultati metode logDice ustvarijo še sami.

Tabela 2: Kolokatorji za iztočnico *alkohol*, razvrščeni po metodi logDice in metodi besednih vložitev; struktura: glagol + samostalnik v tožilniku (gg-s4; podatki iz zbirke KOLOS).

Mesto	LogDice	Besedne vložitve
1.	točiti	poskusiti
2.	piti	ponujati
3.	konzumirati	dodati
4.	presnavljati	streči
5.	zavohati	kupiti
6.	razgrajevati	odstraniti
7.	uživati	povzročiti
8.	zaužiti	prodajati
9.	piliti	kupovati
10.	zlorabljati	prenašati
11.	botrovati	vsebovati
12.	streči	zavohati
13.	opustiti	točiti
14.	vsebovati	opustiti
15.	popiti	piliti
16.	prepovedati	uživati
17.	poskusiti	popiti
18.	kupovati	botrovati
19.	prodajati	zaužiti
20.	zadevati	zadevati
21.	odstraniti	zlorabljati
22.	kupiti	prepovedati
23.	prenašati	konzumirati
24.	dodati	presnavljati
25.	povzročiti	piti
26.	ponujati	razgrajevati

Tabela 3: Kolokatorji za iztočnico *empiričen*, razvrščeni po metodi logDice in metodi besednih vložitev; struktura: pridevnik + samostalnik (p0-s0; podatki iz zbirke KAS).

Mesto	LogDice	Besedne vložitve
1.	raziskava	preverba
2.	del	pristop
3.	raziskovanje	dejstvo
4.	študija	podatek
5.	analiza	rezultat
6.	dokaz	študija
7.	preverba	ugotovitev
8.	preverjanje	metoda
9.	ugotovitev	vidik
10.	podatek	spoznanje
11.	spoznanje	raziskava
12.	delo	literatura
13.	proučevanje	delo
14.	preučevanje	dokaz
15.	konstanta	gradivo
16.	enačba	raziskovanje
17.	rezultat	analiza
18.	testiranje	testiranje
19.	dejstvo	preverjanje
20.	metoda	preučevanje
21.	model	model
22.	gradivo	proučevanje
23.	literatura	konstanta
24.	vidik	enačba
25.	pristop	del

Tabela 4: Kolokatorji za iztočnico *tematika*, razvrščeni po metodi logDice in metodi besednih vložitev; struktura: samostalnik + samostalnik v rodilniku (s0-s2; podatki iz zbirke KAS).

Mesto	LogDice	Besedne vložitve
1.	hotenje	razvoj
2.	samomor	raziskava
3.	pogovor	izobraževanje
4.	naloga	delo

Mesto	LogDice	Besedne vložitve
5.	del	besedilo
6.	vojna	naloga
7.	besedilo	odnos
8.	nasilje	vojna
9.	zaposlovanje	vprašanje
10.	vprašanje	zaposlovanje
11.	delo	samomor
12.	odnos	hotenje
13.	izobraževanje	pogovor
14.	raziskava	nasilje
15.	razvoj	del

2.2 DeltaP

Danes velja za leksikografsko dejstvo, da lahko količino korpusnega šuma in ostalih neustreznih – v našem primeru kolokacijskih – kandidatov zmanjšamo z izboljšavo postopkov korpusnega označevanja na eni strani in postopkov luščenja podatkov, ki nas zanimajo, na drugi. Kljub temu pa tudi po tovrstnih izboljšavah na strojno pridobljenih seznamih pogosto ostaja veliko enot (kolokacij), med katerimi mora leksikograf izbrati slovarsko relevantne. Pri kolokacijskih kandidatih tu igra ključno vlogo statistični kriterij, po katerem za kolokacijsko velja vsaka zveza, v kateri se besedi pojavljata skupaj s pogostostjo, ki je višja od naključne (Manning in Schütze 1999). Gre za enega od treh »gradnikov« kolokacije kot jezikoslovnega dejstva (Gantar idr. 2021), ki posledično določa, katere kolokacije bodo leksikografu predstavljene najprej (vrh seznama) oz. – če so že pred tem določeni minimalni številski parametri za izvoz – katere mu bodo sploh na voljo.

Eno od pomembnih vprašanj pri analizi kolokacij, zlasti v leksikografskem kontekstu, je, ali je kolokacija, ki jo identificiramo pri določeni iztočnici, relevantna tudi za kolokator, ko ta postane iztočnica. Z drugimi besedami: koliko je razmerje kolokator – iztočnica simetrično. Če ponazorimo s primerom: pri iztočnici *obetaven* je

kolokacija *obetaven nogometaš* dobra za ponazoritev pomena *obetaven*, pri iztočnici *nogometaš* pa je le ena izmed mnogih semantično podobnih kolokacij (*nadarjen nogometaš*, *najboljši nogometaš*, *odličen nogometaš*, *vrhunski nogometaš* ipd.), poleg tega pa je za sam pomen iztočnice *nogometaš* tudi manj relevantna. Postavlja se torej vprašanje, kako zaznati to potencialno (ne)relevantnost različnih delov kolokacije oz. njeno usmerjenost (ang. directionality). Večina statističnih mer (kolokacijskih mer oz. mer povezovalnosti), ki se uporabljajo v korpusnih orodjih, usmerjenosti kolokacije ne zaznava (Gries 2013: 141), kar pri zgornjem primeru pomeni, da ima kolokacija *obetaven nogometaš* pri iztočnici *obetaven* in pri iztočnici *nogometaš* isto vrednost.

Ena izmed metod, ki upošteva omenjeno usmerjenost, je mera deltaP[4] (Gries 2013).[5] DeltaP izhaja iz kognitivnega jezikoslovja in psiholingvistike, konkretneje iz teorije asociativnega učenja, pri kateri raziskovalci merijo asociacijsko napovedovalnost besed v kombinacijah. Drugače povedano, deltaP ponudi podatek, kako verjetna je asociacija oz. izbira druge besede (npr. besede 2), ko nam je kot iztočnica (namig) dana izhodiščna beseda (beseda 1).

Pri prenosu te teorije na kolokacije je Gries (2013) uporabil dve enačbi:

- **deltaP_12** = (a / (a + c)) – (b / (b + d))
- **deltaP_21** = (a / (a + b)) – (c / (c + d))

Pri čemer so:

- a = število skupnih pojavitev besede1 in besede2
- b = število pojavitev besede1 brez besede2
- c = število pojavitev besede2 brez besede1
- d = število pojavitev niza brez besede1 in besede2

4 Poleg avtorjevega izhodiščnega poimenovanja Δ P se v literaturi uporablja tudi razvezana različica (delta P oz. deltaP), čemur sledimo tudi v tem prispevku.

5 Gries (2013) omenja še pristope Michelbacherja idr. (2007a; 2007b), vendar ti po njegovem mnenju zahtevajo veliko procesorske moči in časa, ne ponujajo pa dosti boljših rezultatov od obstoječih statističnih mer.

Prva enačba ponudi podatek o tem, kako verjetna je izbira prve besede, če je prisotna druga, druga enačba pa podatek o tem, kako verjetna je izbira druge besede, če je prisotna prva. Pri izračunih vrednosti deltaP Gries (2013) uporablja n-grame, torej nize besed v celotnem korpusu. Pri n-gramih, pa tudi kolokacijah o opredelitvi prve in druge besede odloča njun vrstni red.

Že takoj je v zvezi s kolokacijami, ki nas tu zanimajo, mogoče reči, da Griesova enačba v premajhni meri upošteva skladenjski vidik. Če npr. raziskujemo, v katerih primerih je pridevnik v strukturi pridevnik + samostalnik bolj obvestilen, nas zanima konkretna struktura, ne pa tudi vse ostale možnosti, npr. pridevnik + predlog + samostalnik. Ker bi bilo to za preizkus uporabnosti metode na kolokacijah prevelika omejitev, smo enačbi za naše namene prilagodili na naslednji način:

- **deltaP_12** = (k / (k + k2)) – (k1 / (k1 + k0))
- **deltaP_21** = (k / (k + k1)) – (k2 / (k2 + k0))

Pri čemer so:

- k = število pojavitev besede1 in besede2 v določeni strukturi (pogostost kolokacije)
- k1 = število pojavitev besede1 brez besede2 v določeni strukturi (beseda1 v drugih kolokacijah z isto strukturo)
- k2 = število pojavitev besede2 brez besede1 v določeni strukturi (beseda2 v drugih kolokacijah z isto strukturo)
- k0 = število pojavitev kolokacij v določeni strukturi brez besede1 in besede2

V naši, po Griesu prirejeni enačbi tako vrednost *a* oz. v prilagojeni enačbi *k* predstavlja število obeh besed skupaj v določeni skladenjski strukturi (torej pogostost celotne kolokacije, npr. *granatno jabolko*), vrednosti *b* oz. *k1* in *c* oz. *k2* pomenita število pojavitev vsake izmed besed v celotnem korpusu izven dane zveze, vendar pa še vedno v isti skladenjski strukturi (torej samo *granatno* ali samo *jabolko* v strukturi pridevnik + samostalnik (p0-s0), a brez pojavitev, ko sta skupaj), medtem ko je vrednost *d* oz. *k0* število vseh drugih kolokacij

v tej skladenjski strukturi, torej brez *granaten* in brez *jabolko*.

Tabela 5 prikazuje podatke, ki jih zahteva enačba, za primer *granatno jabolko*. Podatke smo pridobili iz korpusa Gigafida 2.0 (Krek idr. 2019, 2020; tudi nadaljnji podatki v tem razdelku so iz tega korpusa).

Tabela 5: Pogostost različnih kombinacij pridevnika *granaten* in samostalnika *jabolko*; struktura: pridevnik + samostalnik (p0-s0).

	jabolko$_{DA}$	**jabolko$_{NE}$**	**SKUPAJ v strukturi p0-s0**
granaten$_{DA}$	989 (k)	163 (k1)	1152
granaten$_{NE}$	9531 (k2)	98.037.812 (k0)	98.047.343
SKUPAJ v strukturi p0-s0	10.520	98.037.975	98.048.495

Upoštevajoč podatke iz Tabele 5, je izračun naslednji:

- **deltaP_12** = (989 / 10.520) – (163 / 98.037.975) = 0,094
- **deltaP_21** = (989 / 1.152) – (9.531 / 98.047.343) = 0,858

Kolokacija *granatno jabolko* ima torej v korpusu Gigafida 2.0 vrednost deltaP_12 = 0,094 (verjetnost pojavljanja besede *granatno* ob besedi *jabolko*) in vrednost deltaP_21 = 0,858 (verjetnost pojavljanja besede *jabolko* ob besedi *granatno*). Razlika med obema vrednostma (deltaP_21 – deltaP_12) je tako veliko večja od 0,5,[6] kar izrazito nakazuje asimetričnost kolokacije, in sicer v smeri večje napovedovalnosti od *granaten* proti *jabolko*. Drugače povedano, *granaten* je veliko boljši napovedovalec besede *jabolko* kot obratno.

2.2.1 DeltaP in razvrščanje kolokatorjev v seznam

Kot ugotavlja Gries (2013: 152), bi bila mera deltaP v leksikografiji lahko koristna pri umeščanju večbesednih kombinacij, kot so kolokacije, v gesla njihovih sestavnih delov. Zato smo želeli preveriti, ali nam lahko deltaP pomaga pri prepoznavanju (slovarsko)

6 Vrednost 0,5 je sicer arbitrarna, Gries (prav tam) je interpretiral razliko med deltaP_12 in deltaP_21, ki je ≥ 0,5 ali ≤ –0,5, kot kazalnik asimetričnosti in v tem smo mu sledili.

nerelevantnih kolokatorjev za dano iztočnico. Ravno ti so namreč pri analizi kolokacij velikokrat problematični, saj zaradi svoje pogostosti (in posledično visokih vrednosti pri mnogih statističnih merah) na seznamih zasedajo visoka mesta ter leksikografom jemljejo čas in pozornost.

Za preizkus smo najprej izbrali po tri naključne pogostejše iztočnice za tri strukture: pridevnik + samostalnik (p0-s0; samostalniki *bife*, *drama*, *priloga*), glagol + samostalnik v tožilniku (gg-s4; glagoli *prevajati*, *okrasiti*, *prihraniti*) in prislov + glagol (r-gg; glagoli *investirati*, *prevajati*, *prisluškovati*). Za vsako od iztočnic smo izluščili kolokacije z izračunanimi vrednostmi deltaP_12 in deltaP_21. Minimalna pogostost kolokacij je bila 5.

Prvi cilj je bil preveriti, ali razvrščanje po meri deltaP ponudi koristne informacije o kolokacijah, upoštevajoč predpostavko, da bodo kolokatorji z višjo vrednostjo deltaP, torej bolj napovedovalni kolokatorji, na seznam uvrščeni višje.

Pri tem je treba poudariti, da je bila izbira deltaP_12 ali deltaP_21 glede na položaj iztočnice pomembna, saj smo preverjali prav napovednost pojavitve iztočnice (!) ob kolokatorju in ne obratno.[7] Analizirane sezname prikazujejo Tabele 6, 7 in 8, pri čemer so s krepkim tiskom označene kolokacije, ki so po naši oceni slovarsko nerelevantne. Pri vrhnjih 10 kolokacijah gre torej za besede, s katerimi najhitreje asociiramo dane iztočnice, pri spodnjih 10 pa za besede, ob katerih je asociacija najšibkejša.

7 Pred sabo smo sicer imeli sezname z obema deltaP, zato smo hitro ugotovili, da je druga deltaP, torej tista, ki kaže napovednost kolokatorja ob iztočnici, skoraj vedno enaka razvrstitvi po pogostosti.

Tabela 6: Vrhnjih 10 in spodnjih 10 kolokacij po metodi deltaP s samostalniškimi iztočnicami *bife, drama* in *priloga*; struktura: samostalnik + samostalnik (p0-s0; krepki tisk = slovarsko nerelevantna kolokacija).

Iztočnica	DeltaP_21: vrhnjih 10 kolokacij	DeltaP_21: spodnjih 10 kolokacij
bife	**zajtrkovalni bife** /naslov oddaje/	prijeten bife
	solatni bife	**tamkajšnji bife**
	samopostrežni bife	letni bife
	zanikrn bife	majhen bife
	zakoten bife	odprt bife
	hladni bife	mali bife
	improviziran bife	**ljubljanski bifeji**
	bližnji bife	**nekdanji bife**
	potujoči bife	**številni bifeji**
	priročni bife	**nov bife**
drama	talska drama	**posebna drama**
	Hauptmannova drama	**Posamezne drame**
	Ibsenova drama	močna drama
	krimi drama	**dodatna drama**
	Albeejeva drama	finančna drama
	Calderonova drama	**letošnja drama**
	konverzacijska drama	**Mlada drama**
	Biografska drama	**različne drame**
	Shakespearjeva drama	javna drama
	Strindbergova drama	svetovna drama
priloga	Delova priloga	strokovna priloga
	Sobotna priloga	nogometna priloga
	tarifna priloga	poslovna priloga
	Večerova priloga	**lanska priloga**
	Dnevnikova priloga	**posamezne priloge**
	zelenjavna priloga	skupna priloga
	škrobne priloge	gospodarska priloga
	kartografske priloge	**dobra priloga**
	revijalna priloga	slovenska priloga
	tematska priloga	**Velika priloga**

Tabela 7: Vrhnjih 10 in spodnjih 10 kolokacij po metodi deltaP z glagolskimi iztočnicami *prevajati*, *okrasiti* in *prihraniti*; struktura: glagol + samostalnik v tožilniku (gg-s4; krepki tisk = slovarsko nerelevantna kolokacija).

Iztočnica	DeltaP_12: vrhnjih 10 kolokacij	DeltaP_12: spodnjih 10 kolokacij
prevajati	prevajati Danteja	prevajati filme
	prevajati leposlovje	prevajati imena
	prevajati Biblijo	prevajati sporočila
	prevajati toploto	prevajati izjave
	prevajati pesnike	prevajati naslove
	prevajati poezijo	prevajati zgodbe
	prevajati dražljaje	prevajati odgovore
	prevajati književnost	prevajati programe
	prevajati tok	prevajati občutke
	prevajati prozo	**prevajati del**
okrasiti	okrasiti jelko	okrasiti trg
	okrasiti drevesce	**okrasiti vrh**
	okrasiti smrečico	**okrasiti stran**
	okrasiti balkon	okrasiti izdelek
	okrasiti torto	okrasiti šolo
	okrasiti avlo	okrasiti avtomobile
	okrasiti pogrinjek	okrasiti model
	okrasiti obod	**okrasiti del**
	okrasiti smreko	okrasiti mesto
	okrasiti izložbe	okrasiti vrata
prihraniti	prihraniti sitnosti	prihraniti delo
	prihraniti cent	prihraniti besede
	prihraniti tolar	prihraniti korak
	prihraniti zelenje	**prihraniti večino**
	prihraniti marinado	prihraniti življenje
	prihraniti nevšečnost	prihraniti delež
	prihraniti ponižanje	**prihraniti stvari**
	prihraniti muke	prihraniti dan
	prihraniti sramoto	prihraniti zgodbo
	prihraniti brskanje	prihraniti mesto

Tabela 8: Vrhnjih 10 in spodnjih 10 kolokatorjev po metodi deltaP z glagolskimi iztočnicami *investirati*, *prevajati* in *prisluškovati*; struktura: prislov + glagol (r-gg; krepki tisk = slovarsko nerelevantna kolokacija).

Iztočnica	DeltaP_21: vrhnjih 10 kolokacij	DeltaP_21: spodnjih 10 kolokacij
investirati	veliko investirati	**spet investirati**
	več investirati	**prej investirati**
	lani investirati	**prvič investirati**
	letos investirati	**zdaj investirati**
	največ investirati	**takrat investirati**
	raje investirati	**nato investirati**
	toliko investirati	**vedno investirati**
	ogromno investirati	dobro investirati
	doslej investirati	**danes investirati**
	letno investirati	**tako investirati**
prevajati	sproti prevajati	**tam prevajati**
	veliko prevajati	**znova prevajati**
	dobro prevajati	**nikoli prevajati**
	simultano prevajati	**spet prevajati**
	dobesedno prevajati	**bolj prevajati**
	slabo prevajati	**najprej prevajati**
	neposredno prevajati	**nato prevajati**
	trenutno prevajati	**skupaj prevajati**
	večinoma prevajati	**danes prevajati**
	pogosto prevajati	**tako prevajati**
prisluško-vati	nezakonito prisluškovati	**naprej prisluškovati**
	napeto prisluškovati	**tam prisluškovati**
	pozorno prisluškovati	**vedno prisluškovati**
	skrivaj prisluškovati	**potem prisluškovati**
	dolgo prisluškovati	**takrat prisluškovati**
	očitno prisluškovati	**rad prisluškovati**
	ponoči prisluškovati	**skupaj prisluškovati**
	verjetno prisluškovati	**tako prisluškovati**
	tajno prisluškovati	**nato prisluškovati**
	zunaj prisluškovati	**zdaj prisluškovati**

Iz tabel je hitro razvidno, da je razvrščanje po deltaP koristno, saj se je na dnu seznama znašlo precej slovarsko nerevantnih kolokacij, se je pa že zgolj pri osmih naključno izbranih iztočnicah v treh skladenjskih strukturah pokazalo še nekaj: da so razlike v uspešnosti razvrščanja kolokatorjev tudi na ravni struktur precejšnje. Zato smo v naslednjem koraku opravili še obsežnejšo tovrstno analizo.

2.2.2 DeltaP in razvrščanje kolokatorjev v seznam po strukturah

V analizo deltaP po strukturah smo vključili kolokacije 63 iztočnic (36 samostalnikov, 14 glagolov, 9 pridevnikov in 4 prislovov) v 25 različnih skladenjskih strukturah (Priloga 1).[8] Zaradi manjše pogostosti določenih struktur v primerjavi s strukturami v prvotnem preizkusu smo tu mejo pogostosti znižali na 4. Pri vsaki iztočnici smo največ pozornosti zopet posvetili vrhnjim in spodnjim 10 kolokacijam, v primerih, ko je bilo kolokacij več kot 100, pa vrhnjim in spodnjim 15 kolokacijam. Na podlagi teh 10 ali 15 vrhnjih in spodnjih kolokacij smo vsaki strukturi glede na delež slovarsko nerelevantnih kolokacij pripisali oceno:

1. ocena 'zelo dobro' je pomenila, da na vrhu seznama skoraj ni bilo slovarsko nerelevantnih kolokacij (oz. na dnu relevantnih kolokacij);
2. ocena 'dobro do zelo dobro' je pomenila, da je bilo nekaj iztočnic z zelo dobro razvrstitvijo kolokatorjev, nekaj pa z dobro;
3. ocena 'dobro' je pomenila, da je bilo na vrhu nekaj nerelevantnih kolokacij, a ne pri vseh iztočnicah;
4. ocena 'niti dobro niti slabo' je pomenila, da je bilo na vrhu veliko nerelevantnih kolokacij, in to skoraj pri vseh iztočnicah;
5. ocena 'slabo' pa je pomenila, da so na vrhu prevladovale slovarsko nerelevantne kolokacije oz. na dnu slovarsko relevantne.

Iztočnic pri posameznih strukturah, ki so že vnaprej izkazovale bodisi same slovarsko relevantne bodisi same slovarsko nerelevantne kolokacije (zadnje so bile posledica napak v luščenju), pri katerih zato razvrstitve kolokatorjev ni bilo smiselno ocenjevati za tukajšnje potrebe, nismo vključili v ocenjevanje.

Sezname so ocenjevali trije jezikoslovci, dokončna ocena je bila v primeru nestrinjanja usklajena.

Kot kaže Priloga 1, izkazuje deltaP pri razvrščanju kolokacij pri večini struktur zelo dobre rezultate. Če najprej pogledamo spodnji del seznamov: za razvrščanje slovarsko nerelevantnih kolokacij na

8 Prvotno smo sicer izluščili podatke za 75 struktur, a smo nato analizirali le strukture, v katerih so bile vsaj 3 iztočnice z vsaj 10 kolokacijami.

dno seznama je oceno 'zelo dobro' dobilo 7 struktur od 25, oceno 'dobro' 11 struktur, pri 4 strukturah pa smo izbrali oceno 'dobro do zelo dobro'. Tudi pri vrhu seznamov, torej pri razvrščanju slovarsko relevantnih kolokacij na vrh, so bili rezultati podobni, a je bilo precej več struktur z oceno 'dobro' (17), oceno 'zelo dobro' so dobile 3 strukture, 'dobro do zelo dobro' pa 2 strukturi.

Slabše rezultate smo zaznali samo pri treh strukturah: glagol + povratni osebni zaimek + samostalnik v rodilniku (gg-zp-s2), glagol + povratni osebni zaimek + glagol v tožilniku (gg-zp-s4) in nikalnica + glagol + samostalnik v rodilniku (l-gg-s2), ki pa so problematične že zaradi zahtevnosti luščenja (npr. prepoznave prave povezave s *si* in *se*) in manjšega deleža iztočnic, pri katerih dobimo pomensko smiselne kolokacije. Pri vseh ostalih strukturah deltaP na dno uspešno potiska slovarsko nerelevantne kolokacije, na vrh pa relevantne, pri čemer je, v celoti gledano, ta metoda nekoliko uspešnejša pri prepoznavanju slovarsko nerelevantnih kolokacij kot slovarsko relevantnih. Pomemben je tudi podatek, da so rezultati dobri pri strukturah, ki so v slovenščini najpogostejše, kot so pridevnik + samostalnik (p0-s0), samostalnik + samostalnik v rodilniku (s0-s2) in glagol + samostalnik v tožilniku (gg-s4).

2.2.3 DeltaP proti *logDice: razvrščanje kolokatorjev*

Enako kot pri besednih vložitvah smo rezultate metode deltaP na koncu kvalitativno primerjali še z rezultati metode logDice. Ostali smo pri istih podatkih kot zgoraj (63 iztočnic, 25 struktur), zopet smo se osredotočili na vrhnjih in spodnjih 10 ali 15 kolokacij, tokrat seveda razvrščenih po obeh metodah (primer seznama prikazuje Priloga 2).

Tudi te sezname so ocenjevali trije jezikoslovci, dokončna ocena je bila v primeru nestrinjanja usklajena.

Ugotovitve smo povzeli v Tabeli 9. Ta razkriva, da obstajajo med rezultati obeh metod precejšnje podobnosti, vendar pa daje pri 7 strukturah logDice boljše rezultate, medtem ko je metoda deltaP nekoliko boljša le pri 2 strukturah. V primerih, ko nobena metoda ni

bistveno boljša, se razlike kažejo le pri posameznih iztočnicah. Podrobnejši vpogled še pokaže, da so glavne razlike skoraj vedno omejene na vrh seznama kolokacij, medtem ko sta pri potiskanju slovarsko manj relevantnih ali nerelevantnih kolokacij na dno seznama metodi skoraj enako uspešni. Metoda deltaP večkrat na vrh ali v bližino vrha razvršča redke in terminološke kolokacije, medtem ko pri metodi logDice v vrhu najdemo pogostejše in splošnejše kolokacije, npr.:

- glagol + samostalnik v tožilniku (gg-s4): *kitara*
  - deltaP: *brenkati kitaro* (14 pojavitev), *uglaševati kitaro* (12), *nažigati kitaro* (11), *špilati kitaro* (4), *uglasiti kitaro* (16)
  - logDice: *igrati kitaro* (1641), *poučevati kitaro* (58), *uglasiti kitaro* (16), *prijeti kitaro* (74), *brenkati kitaro* (14)
- pridevnik + samostalnik (p0-s0): *dopust*
  - deltaP: *rodniški dopust* (4), *porodniški dopust* (3515), *sabatni dopust* (21), *posvojiteljski dopust* (61), *očetovski dopust* (902)
  - logDice: *porodniški dopust* (3515), *bolniški dopust* (1999), *letni dopust* (4787), *očetovski dopust* (902), *starševski dopust* (715)
- pridevnik + samostalnik (p0-s0): *akuten*
  - deltaP: *akutni bronhiolitis* (18), *akutni enterokolitis* (5), *akutna timpanija* (4), *akutni laringitis* (9), *akutni pankreatitis* (17)
  - logDice: *akutno vnetje* (364), *akutna levkemija* (192), *akutna okužba* (449), *akutna zastrupitev* (166), *akutni sindrom* (272)

Tabela 9: Ocena relevantnosti razvrstitve kolokacij po metodi logDice in metodi deltaP v 25 strukturah.

Struktura	Relevantnejša razvrstitev kolokacij: deltaP ali logDice	Komentar ocene
glagol + predlog + samostalnik v rodilniku gg-d-s2	logDice	razvrstitev je enaka ali zelo podobna, pri nekaterih iztočnicah je na vrhu logDice boljši
glagol + predlog + samostalnik v tožilniku gg-d-s4	oba	razlike so predvsem na vrhu, včasih je pri razvrščanju boljši logDice, drugič deltaP

Struktura	Relevantnejša razvrstitev kolokacij: deltaP ali logDice	Komentar ocene
glagol + predlog + samostalnik v mestniku gg-d-s5	logDice	velika podobnost v razvrstitvi, vendar pa je na vrhu zaradi splošnejših kolokacij boljši logDice
glagol + predlog + samostalnik v orodniku gg-d-s6	oba	razlike so predvsem na vrhu, v nekaj primerih je boljši logDice, v drugih deltaP
glagol + samostalnik v dajalniku gg-s3	oba	velika podobnost v razvrstitvi, razlike so pri istih kolokacij
glagol + samostalnik v tožilniku gg-s4	oba	pri nekaterih iztočnicah so razlike predvsem na vrhu, v 4 primerih je boljši logDice, v 4 deltaP
glagol + povratni osebni zaimek + predlog + samostalnik v rodilniku gg-zp-d-s2	oba	velika podobnost v razvrstitvi
glagol + povratni osebni zaimek + predlog + samostalnik v tožilniku gg-zp-d-s4	oba	velika podobnost v razvrstitvi, v enem primeru je deltaP boljši
glagol + povratni osebni zaimek + predlog + samostalnik v mestniku gg-zp-d-s5	deltaP	pri iztočnicah, kjer so razlike, je deltaP večinoma boljši
glagol + povratni osebni zaimek + predlog + samostalnik v orodniku gg-zp-d-s6	logDice	razlike so zgolj na vrhu, kjer je logDice boljši
glagol + povratni osebni zaimek + samostalnik v rodilniku gg-zp-s2	oba	velika podobnost v razvrstitvi; ko so razlike, je enkrat boljši logDice, drugič deltaP
glagol + povratni osebni zaimek + samostalnik v tožilniku gg-zp-s4	oba	velika podobnost v razvrstitvi; ko so razlike, so na vrhu, enkrat je boljši logDice, drugič deltaP
nikalnica + glagol + samostalnik v rodilniku l-gg-s2	oba	na vrhu je boljši logDice, ki pa ima na dnu tudi slovarsko relevantne kolokacije

Struktura	Relevantnejša razvrstitev kolokacij: deltaP ali logDice	Komentar ocene
pridevnik + predlog + samostalnik v tožilniku p0-d-s4	oba	velika podobnost v razvrstitvi
pridevnik + predlog + samostalnik v mestniku p0-d-s5	oba	velika podobnost v razvrstitvi; ko so razlike, so na vrhu, enkrat je boljši logDice, drugič deltaP
pridevnik + predlog + samostalnik v orodniku p0-d-s6	oba	velika podobnost v razvrstitvi; razlik je malo, takrat je nekoliko boljši deltaP
pridevnik + samostalnik p0-s0	logDice	razlike so zgolj na vrhu, kjer je velikokrat boljši logDice
prislov + glagol r-gg	oba	na vrhu so velike razlike, a je relevantnost kolokacij pri obeh metodah podobna
prislov + povratni osebni zaimek + glagol r-zp-gg	oba	velika podobnost v razvrstitvi; ko so razlike na vrhu, je enkrat boljši logDice, drugič deltaP
samostalnik + predlog + samostalnik v rodilniku s0-d-s2	logDice	razlike so predvsem na vrhu, kjer je logDice večinoma boljši
samostalnik + predlog + samostalnik v tožilniku s0-d-s4	logDice	razlike so predvsem na vrhu, kjer je logDice večinoma boljši
samostalnik + predlog + samostalnik v mestniku s0-d-s5	logDice	boljši je logDice, pri nekaterih iztočnicah tudi pri dnu
samostalnik + predlog + samostalnik v orodniku s0-d-s6	deltaP	na vrhu je pri nekaterih iztočnicah boljši deltaP
samostalnik + samostalnik v rodilniku s0-s2	oba	pri nekaterih iztočnicah je boljši logDice, pri drugih deltaP
samostalnik v imenovalniku + pomožni glagol + pridevnik v imenovalniku s1-gp-p1	oba	razlike so opazne, a je relevantnost razvrstitve v celoti podobna
SKUPAJ	relevantnejša deltaP: 2 relevantnejši logDice: 7 oba enako: 16	

Na podlagi opravljene analize lahko v zvezi z razvrščanjem kolokatorjev po metodi deltaP sklepno ugotovimo, da deltaP tu ne more nadomestiti metode logDice, lahko pa služi kot dodatno potrjevanje rezultatov te uveljavljene metode. Kaže pa se potencial metode deltaP pri odkrivanju (redkejših) terminoloških kolokacij oz. kandidatov za stalne zveze.

2.3 Razpršenost kolokatorjev kot kazalnik slovarske nerelevantnosti

Glede na to, da se pri obstoječih merah povezovalnosti kot eden najbolj perečih problemov kaže dejstvo, da se na seznamih izluščenih kolokacij pogosto pojavljajo tudi slovarsko nerelevantne kolokacije s pomensko manj obvestilnimi, splošno rabljenimi kolokatorji (za slovenščino *posamezen*, *velik* itd.; gl. tudi Pori in Kosem 2018), smo v zadnjem preizkusu z merjenjem razpršenosti oz. distribucije kolokatorjev skušali izdelati še sezname potencialnih slovarsko nerelevantnih kolokatorjev v različnih strukturah. Pri tem smo se opirali na Rundellovo tezo (2020), da se takšni kolokatorji vežejo s praktično vsako besedo, samo da je zveza smiselna.

Preizkus smo izvedli tako, da smo izračunali razpršenost kolokatorjev, ki pove, ob koliko različnih iztočnicah se dani kolokator pojavlja v isti strukturi. Izhajali smo iz predpostavke, da se pomensko manj relevantni kolokatorji pogosteje vežejo na širok nabor iztočnic, se pravi, da so uporabljeni zelo razpršeno; njihova razpršenost je torej visoka. Izračun razpršenosti smo naredili tako, da smo iz korpusa Gigafida 2.0 izluščili čisto vse kolokacije – tudi tiste z enkratno pojavitvijo – za strukturi pridevnik + samostalnik (p0-s0; 7.559.093 kolokacij) in glagol + samostalnik v tožilniku (gg-s4; 2.839.984 kolokacij), potem pa prešteli število različnih lem, ob katerih se je vsak od elementov kolokacije pojavljal.[9] Npr.: v kolokaciji *današnji razpis* se *današnji* v strukturi p0-s0 pojavlja ob 9.148 različnih samostalnikih (*razpis* je torej le eden od njih), *razpis* pa s 1.078 različnimi pridevniki (*današnji* je le eden od njih).

9 Ta izračun je bil opravljen avtomatsko že med samim luščenjem kolokacijskih podatkov.

Najprej so nas zanimali pridevniki, saj je bilo v analizah ugotovljeno, da se prav med njimi pojavlja največ nerelevantnih kolokatorjev (Pori in Kosem 2021). Tabela 10 prikazuje prvih 50 pridevnikov z najvišjo razpršenostjo v strukturi p0-s0. Na seznamu najdemo zelo pogoste oz. ene najpogostejših lastnostnih pridevnikov (*nov*, *velik*, *star* ipd.), vrstne pridevnike iz zemljepisnih lastnih imen (*slovenski*, *ameriški*, *nemški*, *evropski* ipd.), nanašalne pridevnike (*omenjen*, *tovrsten*, *določen*), pridevnike, ki ob sebi zahtevajo samostalnik v množini (*številen*, *različen*, *razen* ipd.), in pridevnike iz časovnih prislovov (*dosedanji*, *letošnji* ipd.). Lahko vidimo, da na seznamu prevladujejo pridevniki, ki so bili tudi pri evalvaciji avtomatsko izluščenih kolokacijskih podatkov najpogosteje izpostavljeni kot slovarsko nerelevantni kolokatorji (gl. Pori in Kosem 2021), mnoge pa smo srečali tudi na dnu seznamov kolokacij izbranih iztočnic pri analizi metode deltaP.

Tabela 10: 50 pridevnikov z najvišjo razpršenostjo; struktura: pridevnik + samostalnik (p0-s0).

Mesto	Pridevnik	Število iztočnic, ob katerih se pojavlja
1.	nov	30.239
2.	velik	27.871
3.	star	21.821
4.	slovenski	21.020
5.	sam	21.002
6.	dober	19.870
7.	mlad	18.730
8.	pravi	17.567
9.	omenjen	15.289
10.	mali	14.998
11.	domač	14.489
12.	majhen	14.193
13.	znan	14.059
14.	ameriški	12.689
15.	nekdanji	12.419
16.	zadnji	12.195

Mesto	Pridevnik	Število iztočnic, ob katerih se pojavlja
17.	številen	11.622
18.	različen	11.530
19.	nemški	11.068
20.	poseben	10.784
21.	visok	10.752
22.	podoben	10.476
23.	močen	9836
24.	lep	9759
25.	italijanski	9678
26.	lasten	9612
27.	evropski	9605
28.	edin	9489
29.	odličen	9243
30.	današnji	9148
31.	političen	9100
32.	sodoben	9041
33.	klasičen	8956
34.	znamenit	8954
35.	francoski	8678
36.	pomemben	8511
37.	običajen	8423
38.	tovrsten	8390
39.	razen	8214
40.	navaden	8145
41.	priljubljen	7823
42.	popoln	7790
43.	morebiten	7642
44.	glaven	7623
45.	dodaten	7619
46.	uspešen	7608
47.	ljubljanski	7548
48.	sedanji	7498
49.	legendaren	7491
50.	letošnji	7481

Če torej izhajamo iz Rundellove teze, bi lahko rekli, da je tak seznam dobro izhodišče za filtriranje slovarsko nerelevantnih kolokacij.

Uporabnost za potencialno filtriranje za ta preizkus pripravljenega seznama smo dalje preverili tako, da smo vzeli prvih 100 kolokacij, razvrščenih po logDice, za vsakega od prvih 20 pridevnikov z največjo razpršenostjo in jih razvrstili na slovarsko relevantne (tako za pridevniško kot samostalniško iztočnico) in slovarsko nerelevantne (oz. potencialno relevantne samo za pridevniško iztočnico). Potencialne stalne zveze oz. frazeološke enote ali pa njihove dele smo označili s posebnimi oznakami in jih izločili iz nadaljnje analize. Po pričakovanjih so povsem slovarsko nerelevantni nanašalni pridevniki; pridevniki, ki ob sebi zahtevajo samostalnik v množini; in pridevniki, ki so nastali iz časovnih prislovov. Tudi vrstni izlastnoimenski pridevniki so povečini slovarsko nerelevantni, vendar pa najdemo tudi nemalo kolokacij, ki bi jih zaradi geografske pogojenosti koncepta ali izrazite tipičnosti (in velikokrat tudi pogostosti) uvrstili tudi v gesla samostalniških iztočnic (npr. *nemška kanclerka*, *ameriško veleposlaništvo*, *ameriški igralec*, *slovenski jezik*, *slovenska manjšina*). Podobno lahko rečemo tudi za lastnostne pridevnike, pri katerih se sicer kaže še večja medsebojna raznolikost. Tako med kolokacijami z *znan* dejansko ni slovarsko relevantnih kandidatov, pri kolokacijah z *nov* pa jih je kar nekaj (npr. *nova tehnologija*, *nov izziv*, *novi prostori*, *nova podoba*, *najnovejše raziskave*, *nova pridobitev*).

V tretjem koraku smo za vsakega od 20 pridevnikov pregledali še približno 100 kolokacijskih kandidatov, razvrščenih po logDice, z dna in 100 s sredine seznama. Rezultati te analize precej bolj podpirajo argument izdelave filtrirnih seznamov slovarsko nerelevantnih kolokatorjev, saj so bili deleži slovarsko relevantnih kandidatov od sredine proti dnu seznamov, če odštejemo stalnozvezne, dejansko majhni.

Seznama z najvišjo razpršenostjo (prvih 50) smo na koncu izdelali še za samostalnike (struktura p0-s0; Tabela 11) in glagole (struktura gg-s4; Tabela 12).

Tabela 11: 50 samostalnikov z najvišjo razpršenostjo; struktura: pridevnik + samostalnik (p0-s0).

Mesto	Samostalnik	Število iztočnic, ob katerih se pojavlja
1.	delo	14.086
2.	beseda	13.565
3.	del	10.520
4.	sistem	10.313
5.	skupina	9349
6.	mesto	8786
7.	hiša	8773
8.	zgodba	8337
9.	svet	8217
10.	program	8123
11.	pot	7989
12.	družina	7535
13.	način	7518
14.	življenje	7503
15.	mnenje	7497
16.	stran	7357
17.	oblika	7294
18.	družba	7131
19.	čas	6881
20.	prostor	6774
21.	dan	6463
22.	model	6439
23.	projekt	6368
24.	igra	6356
25.	podjetje	6320
26.	knjiga	6251
27.	pogled	6246
28.	človek	6187
29.	leto	6183
30.	podoba	6153
31.	ekipa	6110
32.	ime	5949
33.	film	5932

Mesto	Samostalnik	Število iztočnic, ob katerih se pojavlja
34.	izdelek	5907
35.	center	5809
36.	šola	5735
37.	vloga	5696
38.	dejavnost	5674
39.	izjava	5617
40.	politika	5564
41.	država	5443
42.	vrsta	5379
43.	odnos	5370
44.	prijatelj	5367
45.	slika	5351
46.	načrt	5344
47.	roka	5324
48.	otrok	5271
49.	žena	5250
50.	glas	5235

Tabela 12: 50 glagolov z najvišjo razpršenostjo; struktura: glagol + samostalnik v tožilniku (gg-s4).

Mesto	Glagol	Število iztočnic, ob katerih se pojavlja
1.	imeti	24.886
2.	dobiti	13.218
3.	najti	12.739
4.	videti	12.475
5.	postaviti	9539
6.	zamenjati	9410
7.	poznati	9280
8.	izbrati	9219
9.	pomeniti	9012
10.	narediti	8902
11.	vzeti	8637
12.	predstavljati	8620

Mesto	Glagol	Število iztočnic, ob katerih se pojavlja
13.	uporabljati	8590
14.	potrebovati	8494
15.	predstaviti	8263
16.	premagati	8214
17.	omeniti	7948
18.	dodati	7674
19.	iskati	7608
20.	dati	7588
21.	uporabiti	7567
22.	spremljati	7500
23.	opaziti	7177
24.	prinesti	7135
25.	pripraviti	7074
26.	imenovati	7048
27.	poslati	6910
28.	pustiti	6881
29.	pripeljati	6798
30.	sprejeti	6758
31.	spoznati	6721
32.	pričakovati	6604
33.	zahtevati	6424
34.	odkriti	6322
35.	ponujati	6304
36.	pokazati	6287
37.	omogočati	6144
38.	doseči	6028
39.	kupiti	5899
40.	ponuditi	5891
41.	vključevati	5879
42.	gledati	5836
43.	spraviti	5748
44.	obiskati	5733
45.	vsebovati	5666
46.	opazovati	5494
47.	pogledati	5473

Mesto	Glagol	Število iztočnic, ob katerih se pojavlja
48.	podpirati	5471
49.	ustvariti	5427
50.	zadevati	5348

Iz Tabel 11 in 12 je razvidno, da je v primerjavi s pridevnikom kot kolokatorjem pri samostalnikih in glagolih kot kolokatorjih precej manj lem, ki se v celoti ali pretežno pojavljajo v slovarsko nerelevantnih kolokacijskih zvezah, npr. pri samostalniku *način* in *del*,[10] pri glagolu pa *imeti* in *dati*. Pri glagolih se kaže tudi strukturna specifičnost seznama – zelo očitna je odsotnost glagola *biti*, ki je v mnogih drugih strukturah, ki niso omejene zgolj na polnopomenske glagole, na vrhu seznama razpršenosti.

2.4 Povzetek analiz in ključne ugotovitve

V predstavljeni raziskavi smo se osredotočili na dvoje: na (a) prepoznavanje kolokacij in razvrščanje kolokatorjev ter na (b) razpršenost kolokatorjev in slovarsko relevantnost kolokacij. Izvedli smo tri preizkuse, vse z veliko količino podatkov. V preizkusih smo testirali dve na slovenščini še nepreverjeni metodi prepoznavanja besedne povezovalnosti, in sicer metodo besednih vložitev in metodo deltaP. Rezultate obeh metod – sezname kolokacijskih kandidatov – smo primerjali z rezultati že vrsto let uveljavljene mere povezovalnosti logDice. Dodatno smo relevantnost kolokatorjev preverjali še s številom iztočnic, s katerimi se statistično značilno povezujejo, in v rezultatih prepoznavali pomensko prazne leme, ki bi lahko sodile v filter tipa 'odstrani/umakni'.

Opravljene analize sta vodili dve leksikografsko konkretni raziskovalni vprašanji:

10 Dejansko bi moral biti pri samostalniškem seznamu na vrhu samostalnik *del*, saj gre pri *delo* za napako v lematizaciji (veliko pojavitev samostalnika *delo* je dejansko pojavitev samostalnika *del*).

1. *Ali lahko z metodo besednih vložitev in metodo deltaP izboljšamo avtomatsko luščenje kolokacij in razvrščanje kolokatorjev v primerjavi z rezultati metode logDice?*
2. *Ali lahko z analizo razpršenosti prepoznamo slovarsko nerelevantne kolokatorje, ki bi jih lahko uporabili za pohitritev redakcijskega dela?*

Kratek odgovor na prvo vprašanje se glasi 'ne', kratek odgovor na drugo vprašanje pa 'da'. Ali če smo nekoliko natančnejši: preizkusi novih pristopov k razvrščanju in prepoznavanju slovarsko relevantnih kolokatorjev v slovenščini so pokazali naslednje:

- Razvrščanje kolokatorjev po metodi besednih vložitev je bilo v primerjavi z razvrščanjem po metodi logDice slabše; še zlasti je to veljalo za kolokatorje, izluščene iz specializiranega korpusa KAS.
- Tudi metoda deltaP ni ponudila bistvenih izboljšav v primerjavi z logDice; dejansko se je logDice pri več strukturah izkazal kot uporabnejši za nadaljnjo leksikografsko urejanje, s tem da je razvrstitev kolokatorjev po deltaP v nekaj primerih pokazala večji potencial za prepoznavanje (terminoloških) stalnih zvez.
- Ker se razlike med merami kažejo pravzaprav na mikroravni, torej na ravni konkretnih iztočnic, je pri avtomatskih luščenjih kolokacijskih kandidatov iskanje optimalne mere za njihovo razvrščanje pri besednovrstno različnih iztočnicah in v različnih strukturah dejansko kontraproduktivno.
- Obetavni rezultati so se pokazali šele pri tretji preverbi. Ker so predhodne evalvacije kolokacijskih kandidatov, izluščenih z metodo logDice (Pori in Kosem 2021), pokazale prevladujočo oz. popolno slovarsko nerelevantnost določenih kolokatorjev, smo tu z analizo razpršenosti kolokatorjev znotraj posameznih struktur natančneje ocenili še možnost, da bi take kolokatorje izločili že ob samem izvozu (ali da bi jih npr. opremili z opozorilom). Pokazalo se je, da gre pri mnogih lemah, ki se statistično značilno povezujejo z zelo velikim številom različnih iztočnic, za slovarsko nerelevantne kolokatorje, ki jih je smiselno pred izvozom kolokacijskih podatkov dati na poseben filtrirni seznam. Tak seznam

lahko vsekakor pomaga pri optimiziranju nadaljnjega ročnega leksikografskega dela.

3 Priporočila za nadaljnjo leksikografsko prakso in zaključek

Eden od pomembnih ciljev tukajšnjih preizkusov, pa tudi projekta KOLOS nasploh je bila priprava priporočil za nadaljnjo leksikografsko prakso. Na podlagi analize rezultatov metode besednih vložitev in metode deltaP (tudi primerjalno z logDice) na slovenskih podatkih predlagamo naslednje:

- Pri izvažanju kolokacij je med predstavljenimi metodami še naprej priporočljivo računanje po metodi logDice. V nabor drugih potencialno uporabnih klasičnih mer povezovalnosti je kot dopolnilno smiselno vključiti še katero od mer simetričnosti, npr. deltaP. Pri tem priporočamo izračun deltaP po strukturno prilagojeni formuli, predstavljeni v tem prispevku.
- Za prepoznavanje določenega dela slovarsko nerelevantnih kolokacij priporočamo predhodno oblikovanje seznama kolokatorjev z največjo razpršenostjo rabe. Ker pa zanesljivost tovrstnih seznamov ni stoodstotna, je smiselno, da v razvidu, ki ga bo dobil leksikograf, potencialno slovarsko nerelevantne kolokacije vendarle ohranimo, a ločeno od ostalih (bodisi z grafično rešitvijo, kakršna je klicaj, ali s premikom v drug dokument). Na ta način bodo tudi pomensko prazne kolokacije ostale vidne, lahko pa jih hitro odstranimo, če bo pogled potrdil, da niso relevantne.
- Avtomatsko luščenje je časovno zamudno, poleg tega je veliko lažje obdelovati podatke in iskati skupne vzorce šele po izvozu kot pa dodajati nove delne izvoze vsake toliko časa. Poleg običajnega izvoza kolokacij za vnaprej znan geslovnik je zaradi računalniške učinkovitosti, metodološkega preverjanja, spremljanja razvoja jezika, omogočanja raznovrstnih raziskav različnih segmentov jezika itd. koristno hkrati narediti tudi izvoz vseh kolokacij v korpusu nasploh ter omogočiti dostop do njih čim širši zainteresirani skupnosti.

- Postopek izvažanja kolokacijskih podatkov je treba predvsem v luči novih metod nenehno vrednotiti in izboljševati. V postopke vrednotenja je nujno zajeti tudi rezultate ročnih leksikografskih analiz, na podlagi katerih nastajajo učne množice dobrih in slabih kolokacijskih kandidatov, pri čemer velja upoštevati posebnosti konkretnih slovarskih virov.

Kolokacije so vsekakor kompleksen jezikovni pojav. Čeprav za njihovo prepoznavanje v korpusih in razvrščanje v smislu večje (vrhnje) ali manjše slovarske relevantnosti obstaja več statističnih metod, je človeški pregled rezultatov teh metod še vedno nujen. V primerjavi s časom pred dobrim desetletjem (in še prej), ko je bilo treba kolokacije, ki bodo zapisane v slovar, najti z ročnim pregledom stotin in stotin konkordančnih vrstic, gre tako rekoč za čas leksikografskega »blagra«, v katerem je redakcija hitrejša, analitični napor manjši, prikaz jezika pa realnejši. A ta ni zato – priznajmo si – nič manj frustrirajoč; rezultati statističnih izračunov jezikovnih dejstev namreč še vedno niso brezhibni, še tako skrben in podroben človeški ogled pa prav tako ne. Pa vendar se izboljšanju obojega ne bomo odrekli niti jezikoslovci niti računalničarji, statistiki ali matematiki ter vsi drugi, ki si skupaj prizadevamo za (skoraj) popolne metode ujetja in prikaza kolokacij ali kateregakoli drugega jezikovnega pojava. Poti do tja je več. Ena se kaže tudi v usklajenosti raziskovalne skupnosti, ki bi si enotno prizadevala za eno (v našem primeru slovensko) kolokacijsko bazo, oblikovano na eni strani z možnostjo nenehnih dopolnitev in – v delu, ki bi hranil tudi »slabe« podatke – razvoja nadaljnjih strojnih postopkov; na drugi pa z možnostjo selektivnega izvoza za potrebe priprav številnih in raznoterih jezikovnih virov.

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

Berry-Rogghe, G. L. (1973): The Computation of Collocations and their Relevance in Lexical Studies. V A. J. Aitken idr. (ur.): *The Computer and Literal Studies*: 103–112. Edinburgh/New York: Edinburgh University Press.

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

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. Hillsdale: Lawrence Erlbaum Associates.

Church, K. in Hanks, P. (1990): Word Association Norms, Mutual Information and Lexicography. *Computational Linguistics*, 6 (1): 22–29.

Davis, J. in Mark, G. (2006): The Relationship between Precision-Recall and ROC Curves. *ICML '06: Proceedings of the 23rd International Conference on Machine Learning*: 233–240. New York: Association for Computing Machinery.

Erjavec, T., Fišer, D., Ljubešić, N., Logar, N. in Ojsteršek, M. (2016): Slovenska akademska besedila: prototipni korpus in načrt analiz. V T. Erjavec in D. Fišer (ur.): *Zbornik konference Jezikovne tehnologije in digitalna humanistika*: 58–64. Ljubljana: Znanstvena založba Filozofske fakultete.

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.

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

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.

Gantar, P., Krek, S. in Kosem, I. (2021): Opredelitev kolokacij v digitalnih slovarskih virih za slovenščino. V I. Kosem (ur.): *Kolokacije v slovenščini*: 13–39. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani.

Gantar, P., Krek, S., Kosem, I., Šorli, M., Grabnar, K., Pobirk, O., Zaranšek, P. in Drstvenšek, N. (2012): *Slovene Lexical Database*. Slovenian Language Resource Repository CLARIN.SI, http://hdl.handle.net/11356/1030.

Gries, S. (2013): 50-something Years of Work on Collocations. *International Journal of Corpus Linguistics*, 18 (1): 137–165.

Khokhlova, M. in Benko, V. (2020): Size of Corpora and Collocations: The Case of Russian. *Slovenščina 2.0*, 8 (1): 58–77.

Kilgarriff, A., Rychlý, P., Smrz, P. in Tugwell, D. (2004): The Sketch Engine. V G. Williams in S. Vessier (ur.): *Proceedings of the 11th EURALEX International Congress*: 105–116. Lorient: Université de Bretagne–Sud, Faculté des lettres et des sciences humaines.

Kosem, I., Krek, S. in Gantar, P. (2020): Defining Collocation for Slovenian Lexical Resources. *Slovenščina 2.0*, 8 (2): 1–27.

Kosem, I. idr., ur. (2018): *Kolokacije 1.0: Kolokacijski slovar sodobne slovenščine*. Dostopno prek: https://viri.cjvt.si/kolokacije/ (23. 12. 2020).

Krek, S. idr., ur. (2019): *Gigafida 2.0: Korpus pisne standardne slovenščine*. Dostopno prek: https://viri.cjvt.si/gigafida/ (23. 12. 2020).

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.

Levy, O. in Goldberg, Y. (2014): Neural Word Embedding as Implicit Matrix Factorization. V Z. Ghahramani idr. (ur.): *Proceedings of the 27th International Conference on Neural Information Processing Systems, Volume 2 (NIPS 2014)*: 2177–2185. Cambridge: MIT Press.

Ljubešić, N., Logar, N. in Kosem, I. (2021): Collocation Ranking: Frequency vs Semantics. *Slovenščina 2.0*, v tisku.

Logar, N., Grčar, M., Brakus, M., Erjavec, T., Arhar Holdt, Š. in Krek, S. (2012): *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. 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.

Manning, C. D. in Schütze, H. (1999): *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press.

Michelbacher, L., Evert, S. in Schütze, H. (2007a): Asymmetric Association Measures. V G. Angelova idr. (ur.): *Proceedings of the 6th International Conference on Recent Advances in Natural Language Processing (RANLP)*: 367–372. Borovets: Linguistic Modelling Department, Institute

for Parallel Processing, Bulgarian Academy of Sciences; Association for Computational Linguistics.

Michelbacher, L., Evert, S. in Schütze, H. (2007b): Asymmetry in Corpus-derived and Human Word Associations. *Corpus Linguistics and Linguistic Theory*, 7 (2): 245–276.

Pecina, P. (2009): Lexical Association Measures and Collocation Extraction. *Language Resources and Evaluation*, 44 (1–2): 137–158.

Pori, E. in Kosem, I. (2021): Evalvacija avtomatskega luščenja podatkov. V I. Kosem (ur.): *Kolokacije v slovenščini*: 43–77. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani.

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

Rundell, M. (2020): Creating and Using the Macmillan Collocations Dictionary. Dostopno prek: https://www.macmillandictionary.com/collocations/features.html (11. 6. 2021).

Rychlý, P. (2008): A Lexicographer-friendly Association Score. V P. Sojka in A. Horák (ur.): *Proceedings of Second Workshop on Recent Advances in Slavonic Natural Languages Processing (RASLAN 2008)*: 6–9. Brno: Masaryk University.

Wiechmann, D. (2008): On the Computation of Collostruction Strength: Testing Measures of Association as Expressions of Lexical Bias. *Corpus Linguistics and Linguistic Theory*, 4 (2): 253–290.

Prilogi

Priloga 1: DeltaP: ocena razvrstitve kolokacij za 63 iztočnic v 25 skladenjskih strukturah s primeri z vrha in dna seznama.

Pomen ocen v predzadnjem in zadnjem stolpcu:

1. 'zelo dobro': na vrhu seznama skoraj ni bilo slovarsko nerelevantnih kolokacij (oz. na dnu ne relevantnih kolokacij)
2. 'dobro do zelo dobro': na seznamu je nekaj iztočnic z zelo dobro razvrstitvijo kolokatorjev, nekaj pa z dobro
3. 'dobro': na vrhu seznama je nekaj nerelevantnih kolokacij, a ne pri vseh iztočnicah
4. 'niti dobro niti slabo': na vrhu seznama je veliko nerelevantnih kolokacij, in to skoraj pri vseh iztočnicah
5. 'slabo': na vrhu seznama prevladujejo slovarsko nerelevantne kolokacije, na dnu pa slovarsko relevantne

Struktura[11]	Analizirani vzorec	Primeri kolokacij z vrha seznama	Primeri kolokacij z dna seznama	Ocena relevantnosti kolokacij na vrhu seznama	Ocena relevantnosti kolokacij na dnu seznama
gg-d-s2	6 glagolov, 14 samostalnikov	brati s telepromterja brati iz Tore brati z ustnic prisluškovati brez naloga prisluškovati brez odredbe	brati s koncev brati od konca brati konec leta prisluškovati od avgusta prisluškovati od leta	ZELO DOBRO	ZELO DOBRO
gg-d-s4	6 glagolov, 10 samostalnikov	investirati v obveznice investirati v bitcoin investirati v kriptovalute brenkati na kitaro zabrenkati na kitaro preigravati na kitaro	investirati v svet investirati na način investirati v način iti za kitaro biti za kitaro iti po kitaro	DOBRO	DOBRO

11 Za razvezavo okrajšanih poimenovanj struktur gl. Tabelo 9.

Struktura	Analizirani vzorec	Primeri kolokacij z vrha seznama	Primeri kolokacij z dna seznama	Ocena relevantnosti kolokacij na vrhu seznama	Ocena relevantnosti kolokacij na dnu seznama
gg-d-s5	5 glagolov, 7 samostalnikov	okrasiti po želji okrasiti po domišljiji okrasiti na krožniku sporazumevati v jeziku žlobudrati v jeziku programirati v jeziku	okrasiti v letih okrasiti v primeru okrasiti ob strani biti po jeziku biti o jeziku biti ob jeziku	DOBRO	DOBRO
gg-d-s6	6 glagolov, 5 samostalnikov	okrasiti z bunkicami okrasiti s krebuljico okrasiti z meliso preobleči z blagom kupčevati z blagom podložiti z blagom	okrasiti pod vodstvom okrasiti z glavami okrasiti pred leti biti z blagom imeti z blagom biti pod blagom	DOBRO	ZELO DOBRO
gg-s3	3 samostalniki	poveljevati armadi ukazati armadi zadati armadi zamakniti bolnišnicam donirati bolnišnici darovati bolnišnici	pomagati armadi dati armadi slediti armadi pripasti bolnišnici pripadati bolnišnici ustrezati bolnišnicam	DOBRO	DOBRO
gg-s4	3 glagoli, 10 samostalnikov	izplaziti jezik jezikati jezike govoriti jezik cariniti blago pozvanjati blago ocariniti blago pomanjšati aplikacije nameščati aplikacije lansirati aplikacijo reciklirati embalažo sortirati embalažo odlagati embalažo	dajati jezik pripraviti jezik omogočati jezik postaviti blago omogočati blago pripraviti blago dati aplikacijo zahtevati aplikacijo sprejeti aplikacijo imeti embalažo videti embalažo postaviti embalažo	DOBRO	DOBRO do ZELO DOBRO
gg-zp-d-s2	4 samostalniki	prevesti se iz jezika prevajate se iz jezika zameriti se zaradi jezika odriniti se od obale oddaljiti se od obale potopiti se od obale	znajti se zaradi jezika razviti se iz jezika razlikovati se do jezika umakniti se od obale vrniti se od obale posloviti se od Obale	DOBRO	DOBRO

Struktura	Analizirani vzorec	Primeri kolokacij z vrha seznama	Primeri kolokacij z dna seznama	Ocena relevantnosti kolokacij na vrhu seznama	Ocena relevantnosti kolokacij na dnu seznama
gg-zp-d-s4	7 samostalnikov	odpravljati se na dopust oditi se na dopust odpraviti se na dopust vlagati se v panoge usmerjati se v panoge razviti se v panogo	vrniti se na dopust nanašati se na dopuste odpraviti se za dopust preseliti se v panogo uvrstiti se med panoge vrniti se v panogo	DOBRO	DOBRO
gg-zp-d-s5	2 glagola, 6 samostalnikov	mučiti se na kontroli mučiti se v fitnesu mučiti se v peklu zdraviti se v bolnišnici zbuditi se v bolnišnici okužiti se v bolnišnici odklopiti se na dopustu spočiti si/se na dopustu odpočiti se/si na dopustu	mučiti se v času mučiti se na koncu mučiti se v letih prikazati se v bolnišnici dogajati se po bolnišnicah zgoditi se pri bolnišnici pogovarjati se na dopustu pojaviti se na dopustu začeti se na dopustu	DOBRO do ZELO DOBRO	ZELO DOBRO
gg-zp-d-s6	1 glagol, 5 samostalnikov	utaboriti se pred bolnišnico zbirati se pred bolnišnico zbrati se pred bolnišnico nacejati se z alkoholom opijati se z alkoholom omamljati se z alkoholom osliniti si z jezikom oblizniti si/se z jezikom ovlažiti si z jezikom	zgoditi se z bolnišnico pogovarjati se z bolnišnico ukvarjati se z bolnišnicami spopadati se z alkoholom začeti se z alkoholom ukvarjati se z alkoholom začeti se z jezikom znajti se med jeziki pogovarjati se z jezikom	ZELO DOBRO	ZELO DOBRO
gg-zp-s2	1 glagol, 3 samostalniki	pregrizniti si jezika odgrizniti si jezika učiti se jezika točiti se alkohola popiti se alkohola pritakniti se alkohola	spomniti se jezika zavedati se jezika lotiti se jezika rešiti se alkohola privoščiti si alkohola znebiti se alkohola	NITI DOBRO NITI SLABO (veliko napak strukture)	NITI DOBRO NITI SLABO (veliko napak strukture)

Struktura	Analizirani vzorec	Primeri kolokacij z vrha seznama	Primeri kolokacij z dna seznama	Ocena relevantnosti kolokacij na vrhu seznama	Ocena relevantnosti kolokacij na dnu seznama
gg-zp-s4	1 glagol, 3 samostalniki	preživeti si/se dopust odobriti se/si dopust odšteti se dopuste pregrizniti si jezik razvezati se/si jezik odgrizniti si jezik	prislužiti si dopust želeti si dopust zagotoviti si dopust omisliti si jezik privoščiti si jezik ogledati si jezik	NITI DOBRO NITI SLABO	SLABO
l-gg-s2	1 glagol, 4 samostalniki	ne otresati jezika ne šparati jezika ne stegniti jezika ne užiti alkohola ne pokusiti alkohola ne streči alkohola	ne poznati jezika ne potrebovati jezika ne imeti jezika ne potrebovati alkohola ne dobiti alkohola ne imeti alkohola	NITI DOBRO NITI SLABO	NITI DOBRO NITI SLABO
p0-d-s4	4 samostalniki	polnjen v embalažo pakiran v embalažo zapakiran v embalažo prepeljan v bolnišnico odpeljan v bolnišnico pripeljan v bolnišnico	namenjen za embalažo odgovoren za embalažo primeren za embalažo namenjen v bolnišnico primeren za bolnišnico odgovoren za bolnišnico	DOBRO	ZELO DOBRO
p0-d-s5	3 samostalniki	zdraviti v bolnišnici hospitaliziran v bolnišnici zdrav v bolnišnici govorjen v jeziku odpet v jeziku podnaslovljen v jeziku	visok v bolnišnicah rojen v bolnišnici velik v bolnišnicah dober v jeziku znan v jeziku zaposlen v jeziku	DOBRO	DOBRO do ZELO DOBRO
p0-d-s6	4 samostalniki	nalit z alkoholom natopljen z alkoholom zasvojen z alkoholom preoblečen z blagom prodajalen z blagom oblazinjen z blagom	povezan z alkoholom napolnjen z alkoholom pogojen z alkoholom okrašen z blagom povezan z blagom zadovoljen z blagom	ZELO DOBRO	DOBRO

Struktura	Analizirani vzorec	Primeri kolokacij z vrha seznama	Primeri kolokacij z dna seznama	Ocena relevantnosti kolokacij na vrhu seznama	Ocena relevantnosti kolokacij na dnu seznama
p0-s0	9 pridevnikov, 13 samostalnikov	alkoholna dehidrogenaza alkoholni bledež alkoholno vrenje Bajni denarci Bajni Matevž bajni zaslužki brik embalaža vračljiva embalaža nekomunalna embalaža konfetni aranžmaji Nicejski aranžma pihalski aranžmaji PARENTERALNA APLIKACIJA nativne aplikacije prednameščene aplikacije	alkoholni program alkoholna skupina alkoholni del bajni načrt Bajno življenje bajno mesto evropski embalaž visoka embalaža slaba embalaža Letošnji aranžma Slovenski aranžmaji star aranžma Slaba aplikacija slovenska aplikacija Zadnja aplikacija	DOBRO	DOBRO do ZELO DOBRO
r-gg	5 glagolov, 4 prislovi	zverinsko mučiti grozovito mučiti sadistično mučiti interpretativno brati površno brati mrmraje brati simultano prevajati sinhrono prevajati sproti prevajati molče trobentati molče stopalo molče obsedeti ironično pripomniti ironično pripominjati ironično ošvrkniti	letos mučiti takoj mučiti dobro mučiti uspešno brati zelo brati nekoliko brati danes prevajati večkrat prevajati letos prevajati molče povedati molče postati molče priti ironično postati ironično iti ironično imeti	DOBRO	ZELO DOBRO
r-zp-gg	5 glagolov, 3 prislovi	tekoče se brati obetavno se brati gladko se brati strašansko se mučiti dolgo se/si mučiti zakaj se/si mučiti glasno se zakrohotati glasno se odhrkati glasno se usekniti molče se spogledati molče se ukloniti molče se zastrmeti	letos se brati rad se/si brati nato se brati pozno se mučiti pogosto se mučiti rad se mučiti glasno se pojaviti glasno se začeti glasno se odločiti molče se vrniti molče si/se ogledati molče se lotiti	DOBRO	DOBRO do ZELO DOBRO

Struktura	Analizirani vzorec	Primeri kolokacij z vrha seznama	Primeri kolokacij z dna seznama	Ocena relevantnosti kolokacij na vrhu seznama	Ocena relevantnosti kolokacij na dnu seznama
s0-d-s2	8 samostalnikov	trakoma iz blaga serviete iz blaga prtič iz blaga podgesla iz jezikov sposojenka iz jezika prevajalnik iz jezika navodilo z embalaže koda z embalaže trgovina brez embalaže	podjetje od blaga odnos do blaga pot do blaga ljudje do jezika dostop do jezika pot do jezika Izdelki iz embalaže izdelki brez embalaže odnos do embalaže	DOBRO	DOBRO
s0-d-s4	8 samostalnikov	kotlovnica na biomaso kogeneracije na biomaso kurilnice na biomaso Menuet za kitaro brenkanje na kitaro ojačevalec za kitaro tolmačica za jezik slovensko v jezik albanščina za jezik	projekt na biomaso prehod na biomaso center za biomaso zanimanje za kitaro oddelek za kitaro denar za kitaro čas za jezike zavod za jezike Kandidat za jezik	DOBRO	DOBRO
s0-d-s5	5 samostalnikov	nektarji v embalaži žganje v embalaži jogurt v embalaži solaža na kitari kitara v roki akordi na kitari dlake na jeziku Vezljivost v jeziku papile na jeziku	potreba po embalaži podatki o embalaži zakon o embalaži pesem ob kitari poudarek na kitari koncert na kitari hiša v jeziku delo o jeziku delo pri jeziku	DOBRO	DOBRO
s0-d-s6	5 samostalnikov	Dialog med civilizacijami Prelomnica med civilizacijami most med civilizacijami preklapljanje med aplikacijami upravljavec z aplikacijami rokav z aplikacijo	primerjave s civilizacijami zveza s civilizacijo sklad s civilizacijo zveza z aplikacijami povezava med aplikacijo sodelovanje med aplikacijami	DOBRO	DOBRO

Struktura	Analizirani vzorec	Primeri kolokacij z vrha seznama	Primeri kolokacij z dna seznama	Ocena relevantnosti kolokacij na vrhu seznama	Ocena relevantnosti kolokacij na dnu seznama
s0-s2	5 samostalnikov	reklamacija aranžmaja rezervacija aranžmaja odpovedovanje aranžmajev neizrabe dopusta koriščenje dopusta	primer aranžmajev odstotki aranžmajev svet aranžmajev odstotek dopusta cena dopusta program dopusta	DOBRO do ZELO DOBRO	DOBRO
s1-gp-p1	2 pridevnika, 6 samostalnikov	proizvodnja je avtomatizirana skladišče je avtomatizirano Ogrevanje je avtomatizirano drame so uprizarjane drama je uprizorjena drama je nominirana alkohol ni topen alkohol je prepovedan alkohol je zaznaven	delo je avtomatizirano večina je avtomatizirana del je avtomatiziran drama je pomembna drama je dobra drama je potrebna alkohol je visok alkohol je dober alkohol je pomemben	DOBRO	ZELO DOBRO

Priloga 2: DeltaP_21: razvrstitev kolokacij z iztočnico *kitara*; struktura: glagol + samostalnik v tožilniku (gg-s4), skupaj z vrednostjo logDice in pogostostjo v korpusu Gigafida 2.0.

	Kolokacija	DeltaP_21	LogDice	Pogostost
1.	brenkati kitaro	0,29771	6,85098	14
2.	uglaševati kitaro	0,10795	6,60554	12
3.	nažigati kitaro	0,09549	6,47858	11
4.	špilati kitaro	0,05179	5,03277	4
5.	uglasiti kitaro	0,04195	6,92753	16
6.	poprijeti kitaro	0,02639	5,56502	6
7.	igrati kitaro	0,01852	9,18973	1641
8.	drgniti kitaro	0,0107	5,95439	9
9.	žgati kitaro	0,00858	4,90164	4

	Kolokacija	DeltaP_21	LogDice	Pogostost
10.	poučevati kitaro	0,00858	7,49098	58
11.	priklopiti kitare	0,00477	5,05105	5
12.	prijeti kitaro	0,00445	6,92239	74
13.	vihteti kitaro	0,00443	5,64661	9
14.	vaditi kitaro	0,00431	6,31667	21
15.	brusiti kitare	0,00386	4,73468	4
16.	učiti kitaro	0,00363	6,34	28
17.	preigravati kitaro	0,00312	5,09436	6
18.	nasloniti kitaro	0,0031	4,90645	5
19.	študirati kitaro	0,00279	6,2005	37
20.	odložiti kitaro	0,00278	6,22924	41
21.	vzeti kitaro	0,00272	6,49429	270
22.	zagrabiti kitaro	0,00261	5,2655	8
23.	privleči kitaro	0,00254	4,99789	6
24.	zažgati kitaro	0,00223	5,33555	10
25.	podariti kitaro	0,00219	6,03144	52
26.	zaigrati kitaro	0,00214	5,59087	16
27.	obvladati kitaro	0,00211	5,86181	32
28.	razbiti kitaro	0,00172	5,5411	23
29.	pokloniti kitaro	0,00161	4,60019	5
30.	razbijati kitare	0,00158	4,73595	6
31.	priključiti kitaro	0,00134	4,73673	7
32.	kupiti kitaro	0,00125	5,45967	113
33.	mučiti kitaro	0,0012	4,43052	5
34.	zgrabiti kitaro	0,00111	4,50331	6
35.	pograbiti kitaro	0,00104	4,46278	6
36.	posoditi kitaro	0,00095	4,48641	7
37.	izvleči kitaro	0,00083	4,49949	9
38.	držati kitaro	0,00073	4,71829	32
39.	prodati kitaro	0,00068	4,68813	55
40.	obesiti kitaro	0,00066	4,03133	5
41.	slišati kitaro	0,00054	4,39168	29
42.	zamenjati kitaro	0,00053	4,4089	46
43.	odigrati kitaro	0,00052	4,38164	42
44.	prinesti kitaro	0,0004	4,12828	55

	Kolokacija	DeltaP_21	LogDice	Pogostost
45.	odnesti kitaro	0,0004	3,97501	14
46.	odvreči kitaro	0,00039	3,54673	4
47.	peti kitaro	0,00038	3,52515	4
48.	posneti kitaro	0,00038	3,99112	20
49.	ukrasti kitaro	0,00035	3,83984	13
50.	izdelovati kitare	0,00031	3,76198	14
51.	oboževati kitare	0,00026	3,2878	4
52.	zaslišati kitaro	0,00024	3,36829	6
53.	snemati kitare	0,00024	3,4151	7
54.	pospraviti kitaro	0,00022	3,18422	4
55.	vleči kitaro	0,00021	3,14221	4
56.	vreči kitaro	0,00019	3,27634	8
57.	vrniti kitaro	0,00013	3,10849	11
58.	hraniti kitare	0,00013	2,89475	4
59.	uporabljati kitaro	0,00013	3,17128	34
60.	naročiti kitaro	0,00013	2,93741	5
61.	obvladovati kitaro	0,00012	2,83431	4
62.	pobrati kitaro	0,00011	2,8383	5
63.	predati kitaro	0,00011	2,86917	5
64.	dodati kitaro	0,0001	3,00671	24
65.	prirediti kitaro	0,00009	2,69109	4
66.	spraviti kitaro	0,00008	2,76912	7
67.	poslušati kitaro	0,00007	2,75289	8
68.	nositi kitaro	0,00004	2,61444	16
69.	položiti kitaro	0,00004	2,55107	6
70.	izdelati kitaro	0,00003	2,52215	8
71.	prodajati kitaro	0,00002	2,39505	7
72.	povezovati kitaro	0,00001	2,22033	4
73.	obdržati kitaro	–0,00001	2,05058	4
74.	prispevati kitaro	–0,00002	1,98328	4
75.	ponuditi kitaro	–0,00002	2,07682	13
76.	zbirati kitare	–0,00003	1,88525	5
77.	pustiti kitaro	–0,00003	2,01736	8
78.	dobiti kitaro	–0,00003	2,06548	69
79.	dati kitaro	–0,00004	1,89653	20

	Kolokacija	DeltaP_21	LogDice	Pogostost
80.	prejeti kitaro	–0,00006	1,70569	13
81.	podati kitaro	–0,00006	1,61748	4
82.	imeti kitaro	–0,00007	1,61326	180
83.	odkriti kitaro	–0,00007	1,43521	5
84.	postaviti kitaro	–0,00007	1,54867	13
85.	delati kitare	–0,00008	1,29787	5
86.	videti kitaro	–0,00009	1,16465	10
87.	izbrati kitaro	–0,00009	1,04211	6
88.	uporabiti kitaro	–0,00009	1,09905	6
89.	narediti kitaro	–0,0001	0,87809	10
90.	potrebovati kitaro	–0,0001	0,89652	9
91.	ustvariti kitaro	–0,0001	0,83221	4
92.	prevzeti kitaro	–0,00011	0,63154	6
93.	predstaviti kitaro	–0,00011	0,65689	9
94.	najti kitaro	–0,00011	0,54584	11
95.	poslati kitaro	–0,00012	0,10499	4

Razvrstitev kolokacij v slovarskem vmesniku: uporabniške prioritete

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

The Collocations Dictionary of Modern Slovene contains a large amount of collocations, which were automatically extracted from the reference corpus of written Slovene Gigafida. In the dictionary interface, the collocations can be listed according to logDice (default setting) and other parameters. In this paper we focus on the so-called collocational cream of the crop, the information that is shown as the main, most easily accessible and noticeable part of an entry. This information summary is a sort of a preview of the entry and the point of departure for further exploration in the dictionary. The criteria of the users for this cream of the crop have been identified by using a survey, which was completed by 457 participants, comprising of proofreaders, translators, teachers and other users. In a questionnaire, the participants first made a selection of collocations of the headwords *belina* and *izolirati*, and then described the criteria and the preferences used for ordering. Nearly all the participants named frequency as the main criterion when making the cream of the crop selection. Also often used criteria were understandability of the collocate, the presence in the general language (not rare or terminological), and the typicality in a certain syntactic structure. The analysis of actual orderings of the participants also confirms that their decisions are more similar to the ordering by frequency than by the logDice association score. Based on these and similar other findings we have formed research hypotheses that can be tested on a larger sample of dictionary material.

Keywords: Collocations Dictionary of Modern Slovene, sorting collocations, user preferences

1 Uvod

Z razvojem digitalnega medija in strojno podprte metodologije je (tudi) na področju uporabnega jezikoslovja na voljo vedno obširnejša količina jezikovnih virov in podatkov. Podatkovno obilje, s katerim se srečujemo prvič v zgodovini, s sabo prinaša nova raziskovalna in razvojna vprašanja. Na eni strani se pojavljajo izzivi in rešitve na področju pridobivanja podatkov iz jezikovnih virov ter njihovega strukturiranja, razvrščanja, označevanja in opisovanja v strojno berljivih slovarskih bazah, na drugi strani vprašanja, kako količinsko bogate in raznorodne jezikovne informacije predstaviti jezikovni oz. uporabniški skupnosti. V tej točki se digitalno slovaropisje, podprto s spoznanji vmesniškega oblikovanja ter uporabniških raziskav, trudi zagotoviti prijetno uporabniško izkušnjo z digitalnimi slovarskimi viri, ki naj bi omogočali hiter dostop do jezikovnih informacij, njihovo preglednost, razumljivost, povezanost, napredne možnosti iskanja, filtriranja, razvrščanja in izvažanja.

V tem kontekstu se prispevek posveča kolokacijskim podatkom za slovenščino, ki so na voljo v Kolokacijskem slovarju sodobne slovenščine (Kosem idr. 2018a). Slovar je uporabnikom prosto na voljo prek spletnega slovarskega vmesnika,[1] kot odprto dostopna slovarska baza pa na repozitoriju CLARIN.SI (Kosem idr. 2019). Osrednje vprašanje prispevka je razvrstitev kolokacij v slovarskem vmesniku, pri čemer izhajamo iz predpostavke, da slovarski uporabniki pri svojem delu potrebujejo hiter vpogled v najrelevantnejše kolokacijske informacije o določeni besedi, ne pa zgolj poljubno urejenega seznama vseh možnih, čeprav legitimnih, kolokatorjev. Kolokacijski "jagodni izbor" nam pomeni podatke, ki jih kažemo na prvem oz. najhitreje dostopnem, najbolj vidnem mestu kolokacijskega slovarja in so izbrani tako, da uporabniku podajo učinkovito izhodišče za nadaljnje delo. Kaj so kriteriji za jagodni izbor po mnenju potencialnih uporabnikov slovarja, ugotavljamo s pomočjo ankete, ki jo je izpolnilo 457 sodelujočih.

Raziskava je nastala pod okriljem projekta KOLOS – Kolokacije kot temelj jezikovnega opisa: semantični in časovni vidiki (ARRS

1 https:/viri.cjvt.si/kolokacije/slv/

J6-8255), ki se je posvečal tudi merjenju kolokativnosti, mdr. evalvaciji oz. določitvi optimalnih statističnih metod za namene različnih ciljnih uporabnikov.[2] Vsebinsko se prispevek povezuje z uporabniško evalvacijo vmesnika Kolokacijskega slovarja sodobne slovenščine, ki je bila pod okriljem istega projekta opravljena med predstavniki različnih uporabniških skupin (Pori idr. 2020; 2021): sodelujoči so ocenjevali preglednost, intuitivnost, funkcionalnost vmesnika, torej možnosti, ki jih ima glede razvrščanja kolokacij v vmesniku na voljo uporabniška skupnost. Raziskavo, ki jo predstavljamo v tem poglavju monografije, zanima korak prej: kako naj kolokacijske podatke uredijo razvijalci slovarja, da bo izhodiščna postavitev v slovarskem geslu skladna z uporabniškimi željami in pričakovanji.

V prispevku najprej opredelimo raziskovalno področje, v katerega se študija umešča, nadaljujemo s predstavitvijo anketnega vprašalnika, metapodatkov sodelujočih jezikovnih uporabnic in uporabnikov ter izbranih rezultatov ankete. Ker je podatkov preveč za celovito navedbo v prispevku, so v pregledni tabelarični obliki odprto dostopni prek spletne strani projekta Kolos, za preverljivost in ponovljivost raziskave pa je na voljo tudi celotni anketni vprašalnik.[3]

2 Raziskovalna izhodišča

2.1 Uporabniške raziskave na področju digitalnega slovaropisja

Slovaropisne uporabniške raziskave imajo v evropskem prostoru relativno dolgo tradicijo, čeprav so se v primerjavi z drugimi vrstami slovarskih raziskav pojavile pozno. Po nekaj desetletjih utemeljevanja je področje zacvetelo z vstopom slovarjev v digitalni svet, ki je znatno razširil nabor metodoloških možnosti za identifikacijo uporabniških navad, potreb in želja, kar opisujejo npr. Tarp (2009), Welker (2003a; 2003b), Lew in De Schryver (2014). Predvsem pristop z

2 Spletna stran projekta: https://www.cjvt.si/kolos/.

3 Vprašalnik: https://www.cjvt.si/kolos/wp-content/uploads/sites/9/2021/03/Anketni-vprasalnik-Razvrstitev-podatkov-v-kolokacijskem-slovarju.pdf, podatki: https://www.cjvt.si/kolos/wp-content/uploads/sites/9/2021/03/Anketni-rezultati-Razvrstitev-podatkov-v-kolokacijskem-slovarju.xlsx.

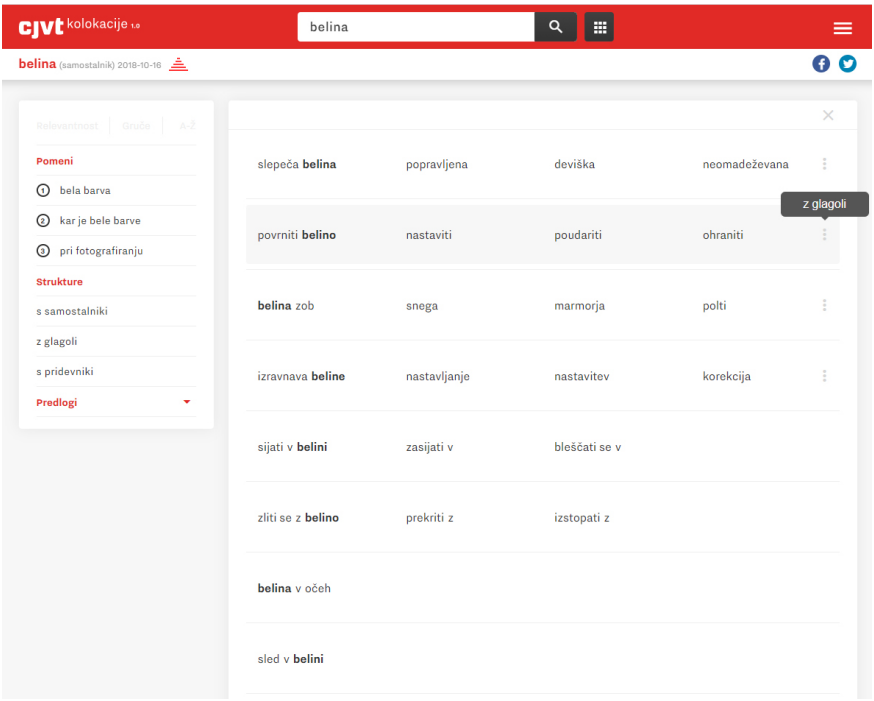
anketiranjem, ki v sodobnem času postaja skoraj izključno spletno, je bil v literaturi tudi že kritično ovrednoten (Bogaards 2003; Tarp 2009; Bergenholtz 2011), čemur so sledile jasneje zasnovane in metodološko premišljene anketne študije (npr. Lorentzen in Theilgaard 2012; Müller-Spitzer 2014; Kosem idr. 2018b), ki jim sledimo tudi v pričujočem prispevku. Idejno izhajamo iz funkcijske teorije slovaropisja, po kateri »so slovaropisni izdelki po temeljnih značilnostih enaki kateremu koli drugemu človeškemu orodju, zasnovani so namreč – oz. bi morali biti – za zadovoljevanje določene vrste človeških potreb« (Fuertes-Olivera in Tarp 2014: 45, prevod Š. A. H.).

Po opozorilih, da akutno primanjkuje empiričnih podatkov o slovenskih slovarskih uporabnikih (npr. Stabej 2009; Logar 2009), in načrtih, kako to vrzel odpraviti (Gorjanc idr. 2017), so v zadnjih letih uporabniške raziskave v porastu tudi pri nas, še zlasti v povezavi z odzivnimi jezikovnimi viri, ki uporabnike vključujejo ne le kot naslovnike, ampak tudi kot sodelujoče pri razvoju digitalne jezikovne infrastrukture za slovenščino. K že omenjenim uporabniškim raziskavam, ki se dotikajo Kolokacijskega slovarja sodobne slovenščine (Pori idr. 2020; 2021), je mogoče dodati anketno raziskavo Arhar Holdt (2020), ki identificira odnos potencialnih slovarskih uporabnikov (N = 671) do novosti odzivnega koncepta, kot so stalno posodabljanje, digitalni format, povezave na korpus, uporabniško vključevanje, vsebnost napak v strojno pridobljenih podatkih in podobno. Anketo, ki jo predstavljamo v pričujočem prispevku, je mogoče razumeti kot naslednji korak v pridobivanju uporabniške povratne informacije s pomočjo spletnega anketiranja, pri čemer je interes to pot usmerjen v izboljšavo specifičnega jezikovnega vira. Za ta namen so bile nujne določene prilagoditve metodologije, čemur se posvečamo v razdelku 3.

2.2 Razvrstitev kolokacij v Kolokacijskem slovarju sodobne slovenščine

Pred nadaljevanjem razprave je treba nekaj besed nameniti trenutnemu prikazu kolokacijskih podatkov v Kolokacijskem slovarju

sodobne slovenščine. Vstopni zaslon posameznega gesla trenutno (različica 1.0) prikaže besednozvezne strukture, v katerih se iztočnica pojavlja, za vsako od struktur pa so prikazani tudi (do) štirje kolokatorji (Slika 1). Na levem robu ekrana je na voljo filter, s katerim je strukture mogoče filtrirati glede na besedno vrsto. Gesla na višji stopnji slovaropisne pregledanosti v filtru ponujajo tudi indikatorje pomenov, ki jih iztočnica (lahko) izkazuje. Na vstopnem zaslonu si uporabnik izbere strukturo, ki ga zanima, in tako nadaljuje z raziskovanjem kolokacijskih podatkov v slovarju (na Sliki 1 je denimo izbrana druga struktura v seznamu; klik na ikono treh pik vodi do več kolokacijskih podatkov).[4]



Slika 1: Vstopni zaslon pri iztočnici *belina*.

4 Vstopne zaslone za iztočnici *belina* in *izolirati*, ki ju obravnavamo v tem prispevku, si je mogoče ogledati na povezavah: https://viri.cjvt.si/kolokacije/slv/headword/48924# in https://viri.cjvt.si/kolokacije/slv/headword/1684#. Pri prvi iztočnici so na voljo tudi pomenski indikatorji, ki so pridobljeni iz Leksikalne baze za slovenščino (Gantar 2015).

Kot opisujejo Kosem idr. (2018a), so kolokacije avtomatsko pridobljene iz referenčnega korpusa pisne slovenščine Gigafida (Logar idr. 2012). Pri izvozu kolokacij iz korpusa se v podatkovni bazi zabeleži statistika logDice (Rychlý 2008), kot tudi pogostnost kolokacije v korpusu. Pri razvrstitvi podatkov na vstopnih zaslonih se trenutno upošteva kolokacijska moč: pri posamezni strukturi so izpisani tisti štirje kolokatorji, ki imajo najvišjo vrednost logDice. Načeloma je izpis padajoč glede na vrednost, vendar se vrstni red lahko spremeni tako, da na prvo mesto pride tista od štirih besed, ki je najkrajša. Prva kolokacija je namreč v tem delu vmesnika izpisana v celoti in pri uporabi kratkih kolokatorjev lahko ostane izpisana v eni vrstici, kar je z vidika vmesniškega oblikovanja preferenčno. Na drugi strani je vrstni red samih struktur v trenutni različici slovarja univerzalen in sledi vnaprej pripravljenemu seznamu.

Ker je interpretacija anketnih rezultatov v nadaljevanju tega prispevka v veliki meri povezana s statističnimi podatki o kolokatorjih besed *belina* in *izolirati*, v Tabelah 1 in 2 predstavljamo podatke, ki so vključeni v Kolokacijski slovar sodobne slovenščine. V prvem stolpcu je navedena struktura, kot je opredeljena v slovarju. Kadar opis strukture v slovarju manjka, smo dodali zvezdico * in opisali strukturo po analogiji z obstoječimi opisi. Sledi število kolokatorjev (ki predstavlja kolokacijsko bogatost strukture); mesto oz. zaporedna številka strukture v slovarskem geslu; nato pa so kolokatorji našteti padajoče glede na logDice, pri čemer sta v oklepaju navedeni tako statistična vrednost logDice kot korpusna pogostnost. Tisti kolokatorji, ki so del trenutnega vstopnega zaslona pri posameznem geslu, so v tabeli prikazani s krepkim tiskom.

Tabela 1: Kolokatorji za samostalnik *belina* v Kolokacijskem slovarju sodobne slovenščine.

Struktura	Število kolokat.	Mesto v slovarju	Kolokator (logDice / pogostnost v korpusu Gigafida)
s samostalniki v rodilniku / s samostalniki v vseh sklonih[5]	32	3. in 4.	**izravnava** (8,38 / 67); **nastavljanje** (7,17 / 49); **nastavitev** (7,03 / 201); **korekcija** (5,94 / 22); uravnavanje (5,49 / 17); ravnovesje (5,43 / 36); določanje (5,39 / 46); prilagajanje (5,25 / 41); zajemanje (4,99 / 8); popravljanje (4,09 / 5); kanček (3,28 / 7); odtenek (2,27 / 8); določitev (2,14 / 7); košček (1,51 / 5); temperatura (1,46 / 15); stopnja (1,06 / 26); videz (1,05 / 8); simbol (0,96 / 5); nivo (0,96 / 5); **zob** (5,1 / 70); **snega** (4,65 / 85); **marmorja** (4,6 / 8); **polti** (4,2 / 9); perila (3,62 / 17); platna (3,61 / 19); papirja (2,98 / 49); stene (2,56 / 25); zidu (2,03 / 12); kamna (1,97 / 12); obleke (1,38 / 11); kože (0,9 / 12); obraza (0,67 / 10)
s pridevniki v vseh sklonih	21	1.	**popravljena** (8,22 / 79); **slepeča** (8,03 / 19); **deviška** (7,6 / 38); **neomadeževana** (7,59 / 17); snežna (7,55 / 197); bleščeča (6,75 / 63); brezmadežna (6,53 / 14); nenatančna (6,41 / 12); nedolžna (6,18 / 64); neskončna (6,02 / 36); nedotaknjena (5,72 / 12); opojna (5,54 / 8); prosojna 5,23 / 9); brezhibna (5,1 / 11); sijoča (5,06 / 10); nežna (4,29 / 26); popolna (3,84 / 62); čista (3,83 / 61); mlečna (3,79 / 13); sveža (2,67 / 20); naravna (2,44 / 34)
z glagoli + predlog 'v' *+ tožilnik	7	9.	**odeti** (6,62 / 22); **oviti** (4,18 / 7); **zazreti se** (4,01 / 8); **potopiti se** (2,6 / 5); obleči (1,79 / 6); zaviti (1,13 / 4); spremeniti (–1,71 / 6)
s pridevniki v vseh sklonih + predlog 'v'	4	14.	**odet** (7,58 / 35); **ovit** (5,27 / 5); **potopljen** (5,16 / 6); **oblečen** (1,77 / 4)

5 V slovarju so skupaj prikazane kolokacije tipa *izravnava beline* (19 kolokacij) in *belina zob* (13 kolokacij), kar je bilo pri dosedanjih evalvacijah že identificirano kot problematično in bo v prihodnji različici odpravljeno. Za večjo preglednost kolokatorje v tabeli ločujemo v dve skupini. Napačno je tudi poimenovanje strukture »s samostalniki v vseh sklonih«, ker gre pri obeh skupinah za samostalnike z rodilniško obliko.

Struktura	Število kolokat.	Mesto v slovarju	Kolokator (logDice / pogostnost v korpusu Gigafida)
z glagoli	4	2.	**nastaviti** (4,43 / 27); **povrniti** (2,68 / 13); **poudariti** (1,41 / 13); **ohraniti** (0,75 / 11)
z glagoli + predlog 'v' *+ mestnik	3	5.	**zasijati** (5,23 / 9); **bleščati se** (4,1 / 5); **sijati** (3,78 / 8)
*z glagoli + predlog 's/z'	3	6.	**zliti se** (2,9 / 5); **prekriti** (2,05 / 4); **izstopati** (1,21 / 4)
*s pridevniki v vseh sklonih + predlog 's/z'	2	12.	**prekrit** (4,12 / 5); **obseden** (3,62 / 4)
*s predlogom 'v' + samostalnik	1	7.	**očeh** (–0,79 / 5)
*s samostalniki + predlog 'v'	1	8.	**sled** (1,94 / 9)
*s samostalniki + predlog 'na'	1	11.	**pogled** (–0,72 / 8)
*z glagoli + predlog 'po'	1	10.	**drseti** (None / None)[6]
*z glagoli + predlog 'od'	1	13.	**bleščati** (4,11 / 5)

6 Pri urejanju slovarskega gesla je bila ena od kolokacij oz. struktur dodana ročno na osnovi pregleda posodobljenega referenčnega korpusa Gigafida 2.0 (Krek idr. 2020). Celovita posodobitev podatkov z novim korpusom (tudi statistik logDice in korpusne pogostnosti) se načrtuje za slovarsko različico 2.0.

Tabela 2: Kolokatorji za glagol *izolirati* v Kolokacijskem slovarju sodobne slovenščine.

Struktura	Število kolokat.	Mesto v slovarju	Kolokator (logDice / pogostnost v korpusu Gigafida)
s samostalniki v tožilniku	54	1.	**podstrešje** (8,66 / 18); **protein** (8,03 / 12); **bakterijo** (8 / 19); **spojino** (7,92 / 11); steno (7,6 / 37); radij (7,52 / 6); gen (7,42 / 18); alkaloid (7,29 / 5); strop (7,22 / 10); virus (7,02 / 13); glivo (6,97 / 4); vlaknino (6,96 / 4); učinkovino (6,88 / 7); Hrvaško (6,76 / 7); celico (6,68 / 24); inzulin (6,65 / 4); molekulo (6,51 / 5); povzročitelja (6,48 / 4); substanco (6,37 / 5); snov (6,37 / 28); beljakovino (6,33 / 4); kabino (6,3 / 5); antibiotik (6,22 / 5); zid (6,05 / 11); fasado (5,97 / 5); ostrešje (5,96 / 4); skladišče (5,83 / 5); streho (5,81 / 14); notranjost (5,64 / 8); hrup (5,56 / 5); tla (5,48 / 10); zvok (5,37 / 11); Nemčijo (5,3 / 4); sestavino (5,27 / 7); cev (4,98 / 4); bolnika (4,8 / 5); stavbo (4,79 / 6); posameznika (4,77 / 4); hišo (4,12 / 21); zapis (4,06 / 4); element (3,96 / 6); telo (3,55 / 9); objekt (3,34 / 5); stran (3,25 / 10); ploščo (3,08 / 4); prostor (2,56 / 13); dom (2,42 / 4); državo (2,35 / 6); skupino (2,35 / 5); Slovenijo (2,24 / 4); del (1,51 / 7); sistem (1,36 / 4); vrsto (1,35 / 5); človeka (1,01 / 4)
s prislovi	28	2.	**toplotno** (11,36 / 131); **zvočno** (10,3 / 64); **socialno** (8,49 / 22); **mednarodno** (7,03 / 7); politično (5,78 / 12); popolnoma (5,15 / 70); dodatno (5,08 / 51); povsem (4,96 / 76); rekoč (4,08 / 6); primerno (3,92 / 9); ustrezno (3,49 / 9); pravilno (3,39 / 11); temeljito (3,26 / 9); prvič (3,15 / 35); dobro (2,77 / 101); bolj (2,44 / 43); treba (2,08 / 58); odlično (2,03 / 6); preveč (1,99 / 8); težko (1,55 / 20); uspešno (1,18 / 7); najprej (1,16 / 16); takoj (1,1 / 10); pogosto (1,07 / 11); hkrati (1,02 / 6); torej (0,5 / 6); nato (0,32 / 8); lahko (0,28 / 79)

Struktura	Število kolokat.	Mesto v slovarju	Kolokator (logDice / pogostnost v korpusu Gigafida)
predlog 'od' + s samostalniki v rodilniku	12	10.	**okolice** (10,55 / 24); **sveta** (9,75 / 64); **okolja** (9,63 / 13); **dogajanja** (8,38 / 6); skupnosti (8,07 / 6); vrstnika (7,77 / 4); izjave (7,72 / 5); Evrope (7,67 / 6); družbe (7,49 / 12); družine (7,4 / 6); medija (7,4 / 4); človeka (6,49 / 10)
predlog 'od' + s samostalniki v rodilniku	11	8.	**tkiva** (9,49 / 9); **droge** (9,17 / 5); **rude** (9,1 / 5); **bakterije** (8,96 / 4); krvi (8,93 / 10); rastline (8,82 / 12); zarodka (8,63 / 4); možganov (8,34 / 4); celice (7,45 / 6); življenja (5,11 / 4); vira (4,36 / 4)
*z glagoli v nedoločniku[7]	9	9.	**uspeti** (3,72 / 62); **poskušati** (2,23 / 19); **nameravati** (2 / 8); **želeti** (1,59 / 1); dati (1,56 / 6); skušati (1,52 / 11); pomagati (1,46 / 5); hoteti (1,18 / 7); znati (1,11 / 9)
predlog 'v' + s samostalniki v mestniku	5	3.	**laboratoriju** (5,51 / 8); **politiki** (3,82 / 4); **sobi** (3,33 / 6); **obliki** (2,95 /); letu (–1,21 / 5)
z zanikanimi glagoli v nedoločniku[8]	4	12.	**želeti** (2,34 / 10); **uspeti** (2,14 / 6); **smeti** (1,3 / 12); **moči** (1,29 / 41)
predlog 'z' + s samostalniki v orodniku	4	4.	**volno** (9,94 / 7); **fasado** (9,32 / 5); **materialom** (5,89 / 4); **oljem** (4,75 / 5)
*predlog 'pred' + s samostalniki v orodniku	2	11.	**svetom** (8,25 // 13); **novinarjem** (7,71 // 8)
*predlog 'na' + s samostalniki v mestniku	1	5.	**otoku** (3,56 // 5)
*predlog 'za' + s samostalniki v tožilniku	1	6.	**čas** (2,75 // 5)
*predlog 'na' + s samostalniki v tožilniku	1	7.	**način** (1,1 // 12)

3 Zasnova ankete

Prednost spletnih anket je, da so enostavne in poceni za pripravo ter diseminacijo, zato je z njimi mogoče doseči veliko število potencialnih

7 V slovarju je struktura napačno poimenovana »z glagoli + predlog Inf-GBZ«.
8 V slovarju je struktura napačno poimenovana »z zanikanimi glagoli + predlog Inf-GBZ«.

sodelujočih.[9] Pristopi, ki sodelujoče sprašujejo po mnenju, so po drugi strani pomanjkljivi, ker se vprašani (lahko) opredeljujejo le do poznanega stanja (Müller-Spitzer 2014: 169) in ker sodelujoči pri anketah lahko vprašanja narobe interpretirajo, nanje nimajo odgovora, v ponujenih odgovorih podzavestno iščejo tiste, ki so po njihovi presoji bolj všečni in podobno (Groves idr. 2004: 209–226). Anketo smo zato zasnovali tako, da so za izbrani besedi – samostalnik *belina* in glagol *izolirati* – sodelujoči morali razvrščanje kolokacijskih struktur in kolokatorjev najprej opraviti sami, šele nato smo jih vprašali po prioritetah in mnenju. Na tak način so anketiranci pri odgovorih lahko izhajali iz lastne izkušnje s problemom in ne iz predvidevanj, kako bi različne rešitve v slovarju (lahko) funkcionirale. Strukturo anketnega vprašanja, ki je v celoti na voljo na spletni strani projekta, prikazuje Tabela 3.

Tabela 3: Struktura in vsebina anketnega vprašalnika.

Razdelek vprašalnika	Vsebina
Uvod	• Uvodni nagovor in poljudna definicija pojma »kolokacija« s primeri. • V1: Samoopredelitev razumevanja: »Ali se vam zdi, da dovolj dobro razumete pojem kolokacija, da lahko nadaljujete z reševanjem ankete?«
Priklic kolokacij (po spominu)	• V2: »Vpišite tri kolokacije z besedo *belina*, ki vam najprej padejo na pamet.« • V3: »Vpišite tri kolokacije z besedo *izolirati*, ki vam najprej padejo na pamet.«
Razvrščanje struktur	• V4: Na ekranu je 6 kolokacijskih struktur s samostalnikom *belina*. Sodelujoči/-a izbere tri skupine, ki bi jih v slovarju rad/-a videl/-a na prvem mestu in jih nato razvrsti glede na relevantnost (po lastni presoji). • V5: Na ekranu so 4 kolokacijske strukture z glagolom *izolirati*. Sodelujoči/-a izbere tri skupine, ki bi jih v slovarju rad/-a videl/-a na prvem mestu in jih nato razvrsti glede na relevantnost (po lastni presoji).

9 Pri delu smo uporabili zmogljiv spletni servis za anketiranje 1KA (https://www.1ka.si/), ki zbrane podatke tudi uredi in vizualizira, kar preprečuje napake obdelave.

Razdelek vprašalnika	Vsebina
Razvrščanje kolokacij	• Napoved nalog za razvrščanje kolokacij in navodila. • V6-V12: Na ekranu je nabor (okrog 20) kolokacij z besedo *belina* ali *izolirati*. Sodelujoči/-a izbere 5 kolokacij, ki bi jih v slovarju rad/-a videl/-a na prvem mestu in jih nato razvrsti glede na relevantnost (po lastni presoji).
Opredelitev kriterijev	• V13: Sodelujoči/-a opredeli, katere kriterije za razvrščanje kolokacij je upošteval/-a in kateri kriteriji pri odločanju niso igrali vloge. • V14 in V15: Sodelujoči/-a opredeli, ali je opazil/-a, da nastopata besedi *belina* in *izolirati* v različnih pomenih, in ali je to igralo vlogo pri razvrščanju.
Dodatni komentarji	• V16: Odprto vprašanje, kjer je mogoče vnesti morebitne dodatne komentarje.[10]
Metapodatki	• V17: Starost. • V18: Status oz. zaposlitev. • V19: Opredelitev uporabniške skupine. • V20: Dosežena izobrazba.

Pri diseminaciji ankete smo ciljali na izbrane skupine potencialnih slovarskih uporabnikov (Arhar Holdt idr. 2016): lektorje, prevajalce ter učitelje slovenščine kot prvega (na različnih stopnjah šolanja) in kot drugega ali tujega jezika. Pripadnost uporabniškim skupinam smo v vprašalniku identificirali (V19 v Tabeli 3), nismo pa sodelovanja v anketi nikakor omejevali. Zastopanost ciljnih skupin v vzorcu je zato višja, ne pa izključna in uravnotežena, zato rezultati niso uporabni za posploševanje na celotno populacijo slovarskih uporabnikov. Dodaten zadržek pri posploševanju je tudi, da spletne ankete, ki temeljijo na prostovoljnem sodelovanju, običajno pritegnejo sodelujoče, ki jih tematika bolj zanima in so zato lahko do vsebine atipično nekritični ali kritični. V nadaljevanju prispevka rezultate interpretiramo naštetim omejitvam primerno.

Anketa je bila aktivna med 22. 5. in 19. 9. 2018. V tem času jo je delno izpolnilo 820 sodelujočih, končalo pa 457 sodelujočih. Kompleksnost ankete je povzročila precejšen osip sodelovanja: najvišji upad je bil na začetku ankete, med samo izvedbo nalog razvrščanja

10 Zaradi omejenega prostora v prispevku teh komentarjev ne analiziramo, po večini gre za pojasnila sodelujočih glede odločitev v anketi ali spodbudne besede glede Kolokacijskega slovarja sodobne slovenščine.

pa je obupalo cca. 100 sodelujočih. V preverbi razumevanja pojma kolokacija (V1 v Tabeli 3) je 28 sodelujočih (3 %) odgovorilo, da teme ne razumejo dovolj, da bi z anketo nadaljevali.[11] Vseeno je treba omeniti, da po avtomatski oceni programa za spletno anketiranje 1ka anketa še vedno spada med enostavne in srednje dolge; tako ocenjeni kot dejanski čas reševanja je bil cca. 10 minut.

Ker gre za prvo predstavitev raziskave, rezultate v nadaljevanju prikazujemo opisno, izpostavimo pa tudi možnosti za nadaljnje analize in raziskovalne hipoteze, ki bi jih bilo smiselno preveriti na večji količini gradiva. Pri tem je treba upoštevati, da so bili metapodatki v vprašalnik vključeni nazadnje (V17–20 v Tabeli 3), zato bo za primerjalne statistične analize mogoče uporabiti le do konca izpolnjene ankete, pri katerih sodelujoči teh vprašanj obenem niso preskočili (okrog 415 anket, gl. razdelek 4.4).

4 Rezultati z diskusijo

4.1 Priklic kolokacij po spominu

Naštevanje kolokacij po spominu (V2 in V3 v Tabeli 3) smo v anketo vključili za dodatno preverbo, ali sodelujoči koncept kolokacije razumejo, in oceno, v kolikšni meri se tak kolokacijski priklic prekriva s kolokacijskim naborom, ki ga ponuja Kolokacijski slovar sodobne slovenščine. Predvidevali smo, da se bodo nekatere kolokacije v odgovorih sodelujočih pojavljale pogosteje, in zanimalo nas je, kolikšne so v teh primerih statistične mere in korpusna pogostnost.

Tabeli 4 in 5 predstavljata rezultate. V prispevek vključujemo tiste kolokacije, ki se v podatkih pojavijo s pogostnostjo 10 ali več.[12] Takšnih je pri obeh iztočnicah primerljivo število: 29 za *belina* in 30

11 Ta filter smo vključili, da lahko pri interpretaciji odgovorov računamo s tem, da so sodelujoči osnovne pojme (vsaj po lastni oceni) ustrezno razumeli. Razumevanje posredno preverjata tudi vprašanji V2 in V3.

12 Ker so sodelujoči v odgovore pisali tako kolokacije kot posamezne kolokatorje, smo rezultate uredili tako, da smo posamezne kolokatorje razširili v kolokacije (npr. *neskončna – neskončna belina*). Pri tem smo ohranili morebitne razlike v slovničnem številu (npr. *izolirati človeka – izolirati ljudi*), kot kaže Tabela 5. Če bi seštevali vse primere v edninski obliki, bi se v tabelo prebila še kakšna dodatna kolokacija, vendar so za analizo in razmisleke o prikazu kolokacij v slovarju zanimive tudi tovrstne razlike v kategorialnih lastnostih.

za *izolirati*. V prvem stolpcu tabel navajamo kolokacijo, sledi podatek, kolikokrat se je pojavila v anketnih odgovorih, ustrezajoča struktura in statistike (oboje povzemamo iz Tabele 1 in 2). Dodali smo tudi mesto, ki ga iz korpusa pridobljena kolokacija zaseda na seznamu, če je slednji urejen padajoče glede na logDice ali glede na korpusno pogostnost. Kolokacije, ki se že zdaj pojavljajo na vstopnem zaslonu slovarskega gesla (gl. razdelek 2.2), so v tabeli prikazane s krepkim tiskom.

Tabela 4: Najpogostejše uporabniško naštete kolokacije s samostalnikom *belina*.[13]

Uporabniško naštete kolokacije z *belina*	Št. v anketi	Ustrezajoča struktura	logDice	pogostnost	Mesto glede na logDice	Mesto glede na pogostnost
snežna belina	344	s prid. v vseh sklonih	7,55	197	5 od 21	1 od 21
čista belina	236	s prid. v vseh sklonih	3,83	61	18 od 21	6 od 21
popolna belina	71	s prid. v vseh sklonih	3,84	62	17 od 21	5 od 21
belina snega	61	s samost. v rodilniku	4,65	85	2 od 13	1 od 13
bleščeča belina	60	s prid. v vseh sklonih	6,75	63	6 od 21	4 od 21
neskončna belina	52	s prid. v vseh sklonih	6,02	36	10 od 21	8 od 21
belina zob	46	s samost. v rodilniku	5,1	70	1 od 13	2 od 13
belina perila	45	s samost. v rodilniku	3,62	17	5 od 13	6 od 13
sveža belina	37	s prid. v vseh sklonih	2,67	20	20 od 21	11 od 21
slepeča belina	37	s prid. v vseh sklonih	8,03	19	2 od 21	12 od 21
očesna belina	27	s prid. v vseh sklonih				ni v slovarju
belina papirja	24	s samost. v rodilniku	2,98	49	7 od 13	3 od 13
zobna belina	23	s prid. v vseh sklonih				ni v slovarju
zaslepljujoča belina	23	s prid. v vseh sklonih				ni v slovarju
Janez Belina[14]	23	s samost. v vseh sklonih				ni v slovarju
sijoča belina	19	s prid. v vseh sklonih	5,06	10	15 od 21	19 od 21
belina stene	17	s samost. v rodilniku	2,56	25	8 od 13	4 od 13
zimska belina	17	s prid. v vseh sklonih				ni v slovarju
belina neba	16	s samost. v rodilniku				ni v slovarju
belina obraza	15	s samost. v rodilniku	0,67	10	13 od 13	11 od 13

13 Prvo kolokacijo je vneslo 642 sodelujočih, drugo 623 in tretjo 595 sodelujočih.
14 Gre za osebno lastno ime znanega televizijskega lika.

Uporabniško naštete kolokacije z *belina*	Št. v anketi	Ustrezajoča struktura	logDice	pogostnost	Mesto glede na logDice	Mesto glede na pogostnost
nebeška belina	15	s prid. v vseh sklonih				ni v slovarju
bela belina	14	s prid. v vseh sklonih				ni v slovarju
brezhibna belina	13	s prid. v vseh sklonih	5,1	11	14 od 21	18 od 21
prostrana belina	12	s prid. v vseh sklonih				ni v slovarju
neomadeževana belina	11	s prid. v vseh sklonih	7,59	17	4 od 21	13 od 21
belina dneva	11	s samost. v rodilniku				ni v slovarju
nedolžna belina	11	s prid. v vseh sklonih	6,18	64	9 od 21	3 od 21
velika belina	10	s prid. v vseh sklonih				ni v slovarju
brezmadežna belina	10	s prid. v vseh sklonih	6,53	14	7 od 21	14 od 21

V Tabeli 5 se pojavljajo primeri, kjer so sodelujoči pri vpisovanju kolokatorjev oz. kolokacij upoštevali slovnično število, in sicer pri primerih *izolirati* [*bolnika / bolnike; steno / stene; človeka / ljudi*]. Pri teh primerih smo v drugi stolpec dodali klicaj, ki opozarja, da je celostno teh primerov več.

Tabela 5: Najpogostejše uporabniško naštete kolokacije z glagolom *izolirati.*[15]

Uporabniško naštete kolokacije z *izolirat*i	Št. v anketi	Ustrezajoča struktura	logDice	pogostnost	Mesto glede na logDice	Mesto glede na pogostnost
izolirati hišo	216	s samost. v tožilniku	4,12	21	39 od 54	4 od 54
izolirati bolnika	83 (!)	s samost. v tožilniku	4,8	5	36 od 54	31 od 54
dobro izolirati	72	s prislovi	2,77	101	15 od 28	2 od 28
toplotno izolirati	55	s prislovi	11,36	131	1 od 28	1 od 28
zvočno izolirati	52	s prislovi	10,3	64	2 od 28	6 od 28
izolirati streho	44	s samost. v tožilniku	5,81	14	28 od 54	8 od 54
izolirati osebo	39	s samost. v tožilniku				ni v slovarju
izolirati stene	35 (!)	*s samost. v tožilniku – množina	v slovarju samo ednina		(5 od 54)	(1 od 54)

15 Prvo kolokacijo je vneslo 576 sodelujočih, drugo 570 in tretjo 556 sodelujočih.

Uporabniško naštete kolokacije z *izolirat*i	Št. v anketi	Ustrezajoča struktura	logDice	pogostnost	Mesto glede na logDice	Mesto glede na pogostnost
popolnoma izolirati	32	s prislovi	5,15	70	6 od 28	5 od 28
izolirati človeka	28 (!)	s samost. v tožilniku	1,01	4	54 od 54	40 od 54
izolirati žico	28	s samost. v tožilniku				ni v slovarju
izolirati sobo	27	s samost. v tožilniku				ni v slovarju
izolirati prostor	27	s samost. v tožilniku	2,56	13	46 od 54	9 od 54
izolirati se[16]	26	/				ni v slovarju
izolirati steno	26 (!)	s samost. v tožilniku	7,6	37	5 od 54	1 od 54
izolirati stanovanje	24	s samost. v tožilniku				ni v slovarju
izolirati kabel	21	s samost. v tožilniku				ni v slovarju
izolirati fasado	20	s samost. v tožilniku	5,97	5	25 od 54	34 od 54
izolirati stavbo	19	s samost. v tožilniku	4,79	6	37 od 54	26 od 54
izolirati virus	17	s samost. v tožilniku	7,02	13	10 od 54	10 od 54
izolirati ljudi	15 (!)	*s samost. v tožilniku – množina	v slovarju samo ednina		(54 od 54)	(40 od 54)
izolirati okna	15	*s samost. v tožilniku – množina				ni v slovarju
izolirati sebe	15	/				ni v slovarju
izolirni trak[17]	14	/				/
izolirati elektriko	13	s samost. v tožilniku				ni v slovarju
izolirati bolnike	12 (!)	*s samost. v tožilniku – množina	v slovarju samo ednina		(36 od 54)	(31 od 54)
izolirati pacienta	12	s samost. v tožilniku				ni v slovarju
povsem izolirati	12	s prislovi	4,96	76	8 od 28	4 od 28
izolirati cev	10	s samost. v tožilniku	4,98	4	35 od 54	47 od 54
izolirati posameznika	10	s samost. v tožilniku	4,77	4	38 od 54	46 od 54

Med najpogostejšimi anketnimi vnosi je kot problematične v smislu razumevanja koncepta kolokacije najti samo *izolirni trak, izolirati* [*sebe / se*] ter *Janez Belina*, vendar tudi zadnja dva primera

16 *Izolirati se* in *izolirati sebe* v slovar nista vključeni kot kolokaciji; *izolirati se* bi bila potencialno iztočnica.

17 V ta odgovor so šteti tako primeri, ko je sodelujoči napisal zvezo *izolirni trak*, kot primeri s samo besedo *trak*. Pri teh vnosih gre za asociacijo, ne pa dejansko kolokacijo z glagolom *izolirati*.

ne nasprotujeta poenostavljeni definiciji, ki je bila podana v uvodu ankete.[18] V obeh tabelah je opaziti, da najpogostejši oz. najpogostejša odgovora glede zastopanosti močno izstopata, nato pogostnost postopoma pada. Hipoteza, da obstajajo primeri, ki jih po spominu navaja večina sodelujočih, se za dani dve iztočnici torej potrdi in bi jo bilo zanimivo preveriti tudi na večji količini gradiva. Podobno velja za strukture: v tabelah je najti samo po dve različni strukturi (izstopa tudi sicer atipičen primer *Janez Belina*): pri *belina* 20 kolokacij sodi med strukture z ujemalnimi pridevniki (npr. *nežna belina*) in 8 z določujočim desnim samostalnikom v rodilniku (npr. *belina snega*); pri *izolirati* se 22 primerov pojavlja s samostalnikom v tožilniku (npr. *izolirati hišo*) in 5 z določujočim prislovom (npr. *toplotno izolirati*).

Prekrivnost po spominu naštetih kolokacij s korpusno pridobljenimi je nižja od pričakovanega. Od 29 najpogosteje naštetih kolokacij s samostalnikom *belina* jih je v slovarju 18 (62 %), od 30 z glagolom *izolirati* pa 19 (63 %), pri čemer je treba upoštevati, da se nekateri od teh primerov v korpusu pojavljajo, niso pa dosegli frekvenčnih pragov za umestitev v slovar. Tipičen tak primer je *zimska belina*, ki se v korpusu Gigafida pojavi 20-krat, v Gigafida 2.0 pa 21-krat, ali *izolirati hišo,* ki se v Gigafida 2.0 pojavi 15-krat. Nadaljnje analize na večji količini gradiva bodo pomagale opredeliti, ali in v katerih primerih bi sprememba parametrov lahko pozitivno vplivala na reprezentiranost podatkov v slovarju. Pri tem gre kot izziv izpostaviti zlasti kolokacije, ki so izvorno terminološke in v referenčnem korpusu redke, vendar očitno prisotne v pojmovnem svetu sodelujočih, kot npr. *izolirati* [*žico, kabel, elektriko*].

Kot protiprimer je omeniti kolokacije, ki se pojavljajo v slovarju, ni pa jih med najpogostejšimi po spominu naštetimi primeri. Pregled širšega nabora uporabniških vnosov pokaže, da se večina primerov najde med redkeje naštetimi: pri samostalniku *belina* najdemo *mlečna* (9 odgovorov), *deviška* (8), *nežna* (8), *prosojna* (7), *naravna*

18 Ubeseditev: »Kolokacije so besede, ki v jeziku tipično nastopajo skupaj, npr. *hitra vožnja*, *varna vožnja*, *spretna vožnja* itd. Kolokacije so po zgradbi različne, npr. glagol + samostalnik (*nadaljevati vožnjo*, *učiti se vožnje*), pridevnik + samostalnik (*hitra vožnja*, *varna vožnja*), samostalnik + samostalnik (*inštruktor vožnje*, *vožnja avtomobila*) itd.«

(2), *nedotaknjen*a (1) in *opojna* (1 odgovor). Ne pojavita pa se med uporabniškimi vnosi dva terminološka primera s področja fotografije, *popravljena belina* in *nenatančna belina*, čeprav glede na logDice in korpusno pogostnost kotirata precej visoko (Tabela 1).

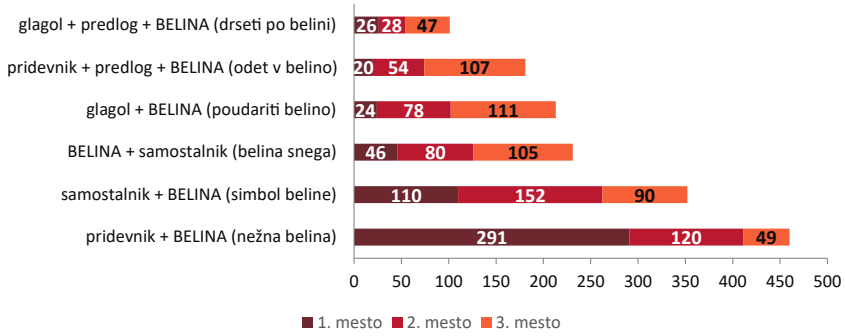
Pregled statističnih mer pokaže, da je v podatkih nekoliko večji delež takih kolokacij, ki so v seznamu korpusne pogostnosti više kot v logDice. Pri *belina* je takih primerov 11, pri *izolirati* 10 – v primerjavi s 7 in 4 primeri, kjer je kolokacija višje glede na logDice. Zlasti kolokacije, ki jih je naštevalo največ sodelujočih (*snežna belina, čista belina, izolirati hišo*), so na seznamu pogostnosti bistveno više, kot so na seznamu, urejenem glede na logDice.

4.2 Izbira in razvrščanje besednozveznih struktur

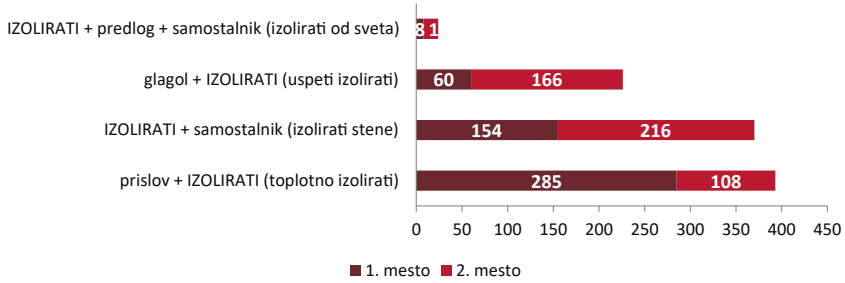
Naslednji korak ankete je vključeval selekcijo in razvrščanje kolokacijskih struktur (V4 in V5 v Tabeli 3). Za vsako od obeh besed je bilo naštetih nekaj struktur[19] (6 za *belina* in 4 za *izolirati*), ki so se v vprašalniku prikazale v naključnem vrstnem redu. Sodelujoči so izbrali polovico struktur, ki bi jih glede na lastne preference želeli videti na prvem mestu slovarja, v drugem koraku pa so izbrane strukture tudi razvrstili. Rezultate prikazujeta Slika 2 in Slika 3, kjer barve nakazujejo, kolikokrat je bila posamezna struktura izbrana za prvo, drugo ali tretje mesto slovarja.

Rezultati kažejo, da bi na prvo mesto slovarja sodelujoči po večini umestili zveze samostalnika z levim pridevniškim prilastkom (npr. *nežna belina*), sledita obe strukturi z samostalnikom v rodilniku (*simbol beline, belina snega*). Pri zvezah z glagolom *izolirati* so prepričljivo na prvem mestu zveze s prislovom (*toplotno izolirati*) in samostalnikom v sklonu (*izolirati stene*). Vsaka od struktur je bila (tudi za prvo mesto) izbrana vsaj nekajkrat, vendar se kažejo trendi, ki potrjujejo hipotezo, da je mogoče identificirati večinske preference uporabniške skupnosti glede razvrstitve kolokacijskih struktur.

19 Natančno členjene strukture, kot so navedene v Kolokacijskem slovarju sodobne slovenščine, smo za vprašalnik združili v robustnejše skupine. V vprašalniku so bili za boljšo predstavo na voljo tudi konkretni primeri različnih kolokacij, ki ustrezajo posamezni strukturi.



Slika 2: Razvrstitev kolokacijskih struktur za samostalnik *belina*.



Slika 3: Razvrstitev kolokacijskih struktur za glagol *izolirati*.

Za vtis, kako se razvrstitev struktur primerja z ostalimi razpoložljivimi podatki, navajamo Tabelo 6. Strukturam iz ankete je pripisano, na katero mesto jih razvrščajo sodelujoči (Slika 2 in 3) in na katero mesto se uvrščajo glede na: (a) pojavljanje med kolokacijami, ki so najpogosteje priklicane po spominu (Tabeli 4 in 5); (b) kolokacijsko bogatost, tj. število kolokacij v posamezni strukturi (Tabela 1 in 2); (c) skupno frekvenco teh kolokacij v referenčnem korpusu (Tabela 1 in 2).

Primerjava pokaže, da je uporabniška razvrstitev dokaj skladna s priklicem po spominu, z izjemo dejstva, da se med najpogosteje naštetimi po spominu strukture tipa *simbol beline* ne pojavljajo, pri razvrščanju pa so sodelujoči ta tip zveze umestili na drugo mesto v slovarskem prikazu. Druga zanimiva ugotovitev, ki bi jo bilo smiselno

preveriti na večji količini gradiva, je, da v primerjavi s korpusnimi statistikami uporabniške preference nekoliko slabše vrednotijo predložne zveze. Ta podatek nakazuje težnjo sodelujočih k prioretiziranju krajših, dvodelnih zvez s polnopomenskimi elementi. V splošnem podatki pokažejo, da je za dana dva primera skupna frekvenca kolokacij v korpusu skladnejša s preferencami sodelujočih v raziskavi kot število kolokacij v strukturi, kar je mogoče nadalje testirati in uporabiti za dinamično razvrščanje struktur znotraj slovarskih iztočnic.

Tabela 6: Primerjava razvrstitve kolokacijskih struktur.

	Uporabniška razvrstitev	Uporabniški priklic po spominu	Število kolokacij v obravnavani strukturi	Skupna frekvenca kolokacij v korpusu
BELINA				
nežna belina	1. mesto	1. mesto	1. mesto (21)	1. mesto (805)
simbol beline	2. mesto	-	2. mesto (19)	2. mesto (578)
belina snega	3. mesto	2. mesto	4. mesto (13)	3. mesto (339)
poudariti belino	4. mesto	-	6. mesto (4)	5. mesto (64)
odet v belino	5. mesto	-	5. mesto (6)	6. mesto (59)
drseti po belini	6. mesto	-	3. mesto (15)	4. mesto (98)
IZOLIRATI				
toplotno izolirati	1. mesto	2. mesto	3. mesto (28)	1. mesto (891)
izolirati stene	2. mesto	1. mesto	1. mesto (54)	2. mesto (469)
uspeti izolirati	3. mesto	-	4. mesto (13)	4. mesto (237)
izolirati od sveta	4. mesto	-	2. mesto (37)	3. mesto (322)

4.3 Izbira in razvrščanje posameznih kolokacij

V nadaljevanju ankete je sledilo sedem nalog, pri katerih so sodelujoči s seznama kolokacij izbrali tiste, ki po njihovem mnenju najbolj sodijo na prvo mesto v slovarju (V6–V12 v Tabeli 3). Sodelujoči so za vsako od danih struktur izbrali po pet kolokacij in jih razvrstili. Že pred reševanjem nalog so bili tudi obveščeni, da bodo po razvrščanju kolokacij opredelili kriterije za svoje odločitve.

Rezultati razvrščanja so v celoti na voljo na projektni spletni strani. V prispevku navajamo samo najbolj kolokacijsko bogate strukture in tiste kolokacije, ki so jih sodelujoči izbrali za prva mesta v slovarju, torej kolokacijski jagodni izbor po mnenju sodelujočih.[20] V Tabeli 7 in Tabeli 8 je pri vsaki posamezni kolokaciji navedeno, kolikokrat so jo sodelujoči izbrali za prvo, drugo, tretje, četrto ali peto mesto v slovarskem prikazu. Izbor primerjamo s kolokacijami, ki so jih sodelujoči najpogosteje navajali po spominu ter z jagodnima izboroma, ki ju ponujata razvrstitev glede na logDice in korpusno pogostnost.

Tabela 7: Jagodni izbor kolokacij s samostalnikom *belina* po mnenju sodelujočih v anketi.

Kolokacije	1	2	3	4	5	Skupaj	Najpogostejše po spominu	Top 5 glede na logDice	Top 5 glede na pogostnost
pridevnik + BELINA									
snežna belina	261	97	46	34	24	462	da	da	da
bleščeča belina	107	115	66	48	39	375	da	ne	da
brezhibna belina	45	50	46	44	45	230	da	ne	ne
naravna belina	21	43	66	55	38	223	ne	ne	ne
sijoča belina	7	30	58	42	63	200	da	ne	ne
BELINA + samostalnik									
belina zob	75	119	89	63	43	389	da	da	da
belina snega	193	73	49	31	33	379	da	da	da
belina perila	40	43	53	55	50	241	da	da	ne
belina papirja	25	34	37	47	46	189	da	ne	da
belina kože	39	32	40	35	35	181	ne	ne	ne
samostalnik + BELINA									
odtenek beline	105	69	50	37	28	289	ne	ne	ne
kanček beline	98	42	53	41	29	263	ne	ne	ne
videz beline	44	52	48	43	39	226	ne	ne	ne
določanje beline	49	47	38	35	33	202	ne	ne	da
nastavitev beline	48	46	34	26	25	179	ne	da	da
glagol + predlog + BELINA									
odeti se v belino	87	77	51	33	17	265	ne	da	da

20 Navajamo po 5 kolokacij, ki so jih sodelujoči najpogosteje izbrali za prioritetni prikaz v slovarju. V primeru, ko za peto mesto tekmujeta dve zvezi, smo v tabelo vključili obe.

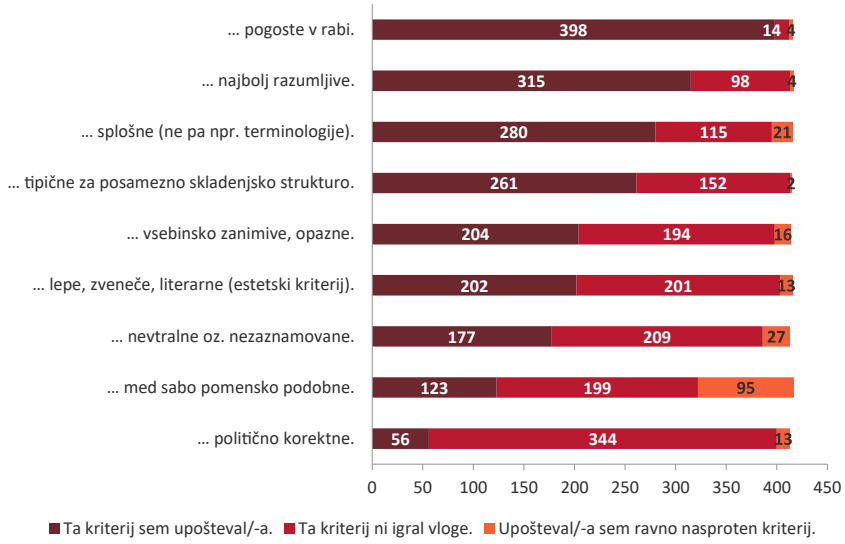
Kolokacije	1	2	3	4	5	Skupaj	Najpogostejše po spominu	Top 5 glede na logDice	Top 5 glede na pogostnost
zliti se z belino	18	33	39	58	70	218	ne	ne	ne
bleščati se od beline	110	25	21	20	23	199	ne	da	ne
zazreti se v belino	23	30	52	58	34	197	ne	ne	da
izginjati v belini	16	32	28	30	40	146	ne	ne	ne
potopiti se v belino	19	31	31	29	36	146	ne	ne	ne
Skupaj 'Da'							**8 od 21**	**7 od 21**	**9 od 21**

Tabela 8: Jagodni izbor kolokacij z glagolom *izolirati* po mnenju sodelujočih v anketi.

Kolokacije	1	2	3	4	5	Skupaj	Najpogostejše po spominu	Top 5 glede na logDice	Top 5 glede na pogostnost
prislov + IZOLIRATI									
zvočno izolirati	65	114	73	72	50	374	da	da	da
toplotno izolirati	122	91	72	59	28	372	da	da	da
socialno izolirati	17	36	67	57	70	247	ne	da	ne
dobro izolirati	114	29	31	33	24	231	da	ne	da
dodatno izolirati	35	60	36	33	41	205	ne	ne	ne
IZOLIRATI + samostalnik									
izolirati bolnika	110	59	49	35	36	289	da	ne	ne
izolirati hišo	88	62	47	46	23	266	da	ne	da
izolirati fasado	39	45	35	19	20	158	da	ne	ne
izolirati človeka	43	31	27	18	25	144	da	ne	ne
izolirati stene	14	29	21	33	27	124	da	da	da
IZOLIRATI + predlog + samostalnik									
izolirati od družbe	67	66	39	25	24	221	ne	ne	da
izolirati pred vlago	60	40	30	40	46	216	ne	ne	ne
izolirati od sveta	36	47	41	52	39	215	ne	da	da
izolirati od ljudi	29	48	49	47	26	199	ne	ne	ne
izolirati pred hrupom	46	47	39	22	20	174	ne	ne	ne
Skupaj 'Da'							**8 od 15**	**5 od 15**	**7 od 15**

Kot velja za predhodne rezultate, se tudi pri jagodnem izboru kolokacij korpusna pogostnost izkaže kot mera, ki je (nekoliko)

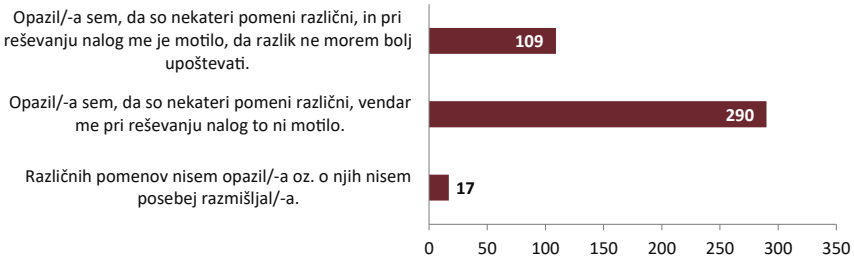
skladnejša s preferencami sodelujočih v anketi, kar je mogoče v nadaljevanju testirati na večji količini gradiva in uporabiti pri razvrščanju podatkov v slovarju. Da je pogostnost v jezikovni rabi – oziroma vtis sodelujočih glede pogostnosti – dejansko igrala pomembno vlogo pri razvrščanju, pa povsem neposredno pokažejo odgovori na vprašanje v anketi, kjer so sodelujoči opredelili kriterije, po katerih so kolokacije izbirali. Kot kaže Slika 4, so skoraj vsi sodelujoči opredelili, da so izbirali kolokacije, ki so pogoste v rabi. Sledijo kriteriji razumljivosti, splošnosti in tipičnosti.



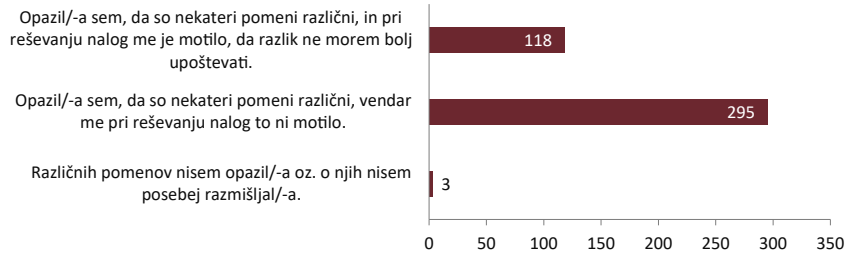
Slika 4: »Pri razvrščanju sem izbiral/-a kolokacije, ki so ...« (N = 417).

Ker v anketi kolokacijski podatki niso bili urejeni glede na različne pomene besed *belina* in *izolirati* – takšno je namreč tudi trenutno stanje pri večini gesel v Kolokacijskem slovarju sodobne slovenščine – smo na koncu ankete sodelujoče vprašali, koliko jih je pri razvrščanju motilo, da pomenov ne morejo upoštevati. Rezultati, ki jih kažeta Sliki 5 in 6, so primerljivi za obe obravnavani besedi. Večina sodelujočih je razlike v pomenu opazila, moteče so bile za približno tretjino in nemoteče za približno dve tretjini sodelujočih. Rezultate

je mogoče uporabiti za nadaljnje analize, zlasti v luči temeljne značilnosti odzivnih slovarjev, da v prvi, avtomatsko pripravljeni fazi informacij, ki zahtevajo ročni leksikografski pregled, po večini ne prinašajo.



Slika 5: Ne/upoštevanje pomenov pri samostalniku *belina* (N = 416).



Slika 6: Ne/upoštevanje pomenov pri glagolu *izolirati* (N = 416).

4.4 Anketni vzorec

Na koncu ankete smo zbrali še metapodatke o sodelujočih v raziskavi. Kot omenjeno (razdelek 3), je kompleksnost ankete povzročila precejšen osip sodelovanja. Za primerjalne statistične analize bo zato mogoče uporabiti približno 415 anket, ki jih predstavljamo v nadaljevanju.

Kot kažejo Tabele 9, 10 in 11, je večina sodelujočih starih med 26 in 55 let, z dokončano univerzitetno izobrazbo, zaposlenih v javnem sektorju, samozaposlenih ali študentov.

Tabela 9: »V katero starostno kategorijo spadate?« (N = 417)

Odgovori	Frekvenca
1 (do 15 let)	0
2 (od 16 do 25 let)	62
3 (od 26 do 35 let)	102
4 (od 36 do 45 let)	119
5 (od 46 do 55 let)	82
6 (od 56 do 65 let)	42
7 (66 let ali več)	10

Tabela 10: »Kakšen je vaš status oziroma glavno področje angažiranja (dela, zaposlitve)?« (N = 416)

Odgovori	Frekvenca
1 (dijak/-inja)	1
2 (študent/-ka)	69
3 (samozaposlen/-a)	63
4 (zaposlen/-a v javnem sektorju (uprava, šolstvo, zdravstvo, sociala, kultura ...))	194
5 (zaposlen/-a v neprofitnem sektorju (društva, združenja ...))	8
6 (zaposlen/-a v podjetju)	43
7 (vzdrževanje gospodinjstva)	2
8 (upokojen/-a)	18
9 (v iskanju zaposlitve)	13
10 (na porodniškem ali starševskem dopustu)	2
11 (Drugo:)	3

Tabela 11: »Kakšna je vaša najvišja dosežena izobrazba?« (N = 415)

Odgovori	Frekvenca
1 (Nedokončana osnovna šola)	0
2 (Osnovna šola)	2
3 (Poklicna šola)	1
4 (Štiriletna srednja šola)	50
5 (Višja šola)	9
6 (Visokošolski strokovni študij)	1

Odgovori	Frekvenca
7 (Visoka šola)	9
8 (Univerzitetni študij)	240
9 (Magisterij)	49
10 (Doktorat)	50
11 (Specializacija)	4

V zadnjem vprašanju so sodelujoči opredelili, v kakšnem kontekstu se srečujejo z jeziko(slo)vnimi vprašanji, izbrali so lahko več odgovorov. Prednjačijo posameznice in posamezniki, ki se ukvarjajo z lektoriranjem ali/in prevajanjem, velik delež je tudi aktivnih v izobraževanju in piscev različnih vrst besedil. Podatki kažejo, da je zajem želenih ciljnih skupin (gl. razdelek 3) ustrezen in da so podatki primerni za nadaljnje primerjalne statistične analize.

Tabela 12: »Če igra, je igral ali bo igral pri vašem delu jezik posebno vlogo, označite vse ustrezajoče kategorije.« (N = 405)

Odgovor (možnih je bilo več izbir)	Frekvence	% – Veljavni (N = 405)
Lektoriranje	223	55 %
Prevajanje	193	48 %
Poučevanje slovenščine kot 1. jezika – osnovna šola	69	17 %
Poučevanje slovenščine kot 1. jezika – srednja šola	46	11 %
Poučevanje slovenščine kot 2. ali tujega jezika	55	14 %
Predavanje jezikoslovnih predmetov na višji / univerzitetni ravni	38	9 %
Novinarstvo	38	9 %
Marketing	37	9 %
Pravo	10	2 %
Bibliotekarstvo	29	7 %
Javna uprava	40	10 %
Administracija	37	9 %
Menedžment	20	5 %
Beletristika	29	7 %
Kreativno pisanje, blogerstvo	84	21 %

Odgovor (možnih je bilo več izbir)	Frekvence	% – Veljavni (N = 405)
Strokovno in znanstveno pisanje	105	26 %
Jezikoslovne raziskave	68	17 %
Leksikografija	29	7 %
Ljubiteljsko raziskovanje jezika	99	24 %
Drugo:	20	5 %

5 Zaključek in prihodnje delo

Raziskava potrjuje, da se je pri razvoju Kolokacijskega slovarja sodobne slovenščine smiselno posvetiti razvrščanju kolokacij glede na relevantnost za uporabniško skupnost in opredelitvi parametrov za prikaz kolokacijskega »jagodnega izbora«: kolokacijskih struktur in kolokatorjev, ki omogočijo hiter vpogled v besedišče, ki je po izbranih kriterijih najbolj relevantno, in ponujajo prvi vtis o (sicer količinsko zelo obsežnem) kolokacijskem gradivu. Rezultati kažejo, da preference oz. preferenčne tendence med sodelujočimi v anketi obstajajo in so metodološko opredeljive. Delo s specifičnimi kolokacijami je sodelujočim v raziskavi omogočilo vpogled v konkretne podatke in urejevalno izkušnjo z njimi, kar je bil predpogoj za zanesljivejše opredeljevanje kriterijev in želja za razvrščanje kolokacij v slovarju. Identificirane preference (Slika 4) so torej posledica konkretne izkušnje z gradivom, kar je močna točka raziskave, na drugi strani pa tako natančna in kvalitativna poglobitev v gradivo (lahko) zajame izredno omejen nabor iztočnic. Izsledke raziskave je zato mogoče uporabiti zlasti za identifikacijo perečih raziskovalnih vprašanj in oblikovanje raziskovalnih hipotez, kot sledi v nadaljevanju.

Izkazuje se, da so – vsaj za obravnavane podatke – preference sodelujočih prekrivnejše s korpusno pogostnostjo kot kolokacijsko jakostjo. V nadaljevanju je mogoče to hipotezo preveriti na večji količini gradiva in posledično podatke urediti po pogostnosti, kot je bilo denimo preizkušeno pri pripravi baze za Estonski kolokacijski slovar (Kallas idr. 2015). Pogostnost je mogoče vključiti kot del kombiniranega pristopa, ki upošteva mesto kolokacije v seznamu, urejenem

tako glede na statistično jakost kot tudi pogostnost, kot je bilo pri luščenju kolokacij za slovenščino preizkušeno v Gantar idr. (2016: 215). Kombinirani pristopi na ravni luščenja so posledica spoznanja, da zanašanje na eno samo statistično mero povzroči izpad določenih relevantnih kolokacij. Vprašanje, ki se ga dotikamo v pričujočem prispevku pa je, kako mere kombinirati tudi pri urejanju (karseda celovito izluščenih) podatkov, da na prva mesta v slovarju pridejo kolokacije, ki jih uporabniki tam želijo in pričakujejo.

Rezultati kažejo obstoj kolokacij, ki jih po spominu navaja večji delež sodelujočih: za dani iztočnici v uporabniških odgovorih močno izstopajo kolokacije *snežna belina, čista belina* in *izolirati hišo*, nato pogostnost odgovorov strmo pada. Primerljive podatke, vendar za veliko večji nabor iztočnic, bo mogoče pridobiti s pomočjo aplikacije Igra besed. Aplikacija podpira model odzivnega slovarja z elementom igrifikacije, ki je namenjena čiščenju in urejanju avtomatsko pridobljenega kolokacijskega ter sopomenskega gradiva (Arhar Holdt idr. 2020). V enem od igralnih modulov igralke in igralci po spominu vpisujejo kolokacije za izbrane iztočnice, kar bo omogočilo nadaljnje analize v prispevku zastavljenega vprašanja. Hipoteze, ki jih postavlja pričujoča raziskava, so: v jeziku obstajajo kolokacije, ki jih po spominu navaja velik delež jezikovne skupnosti; korpusna frekvenca je boljši napovedovalec teh primerov kot statistična jakost; tovrstno besedišče je skoraj vedno splošno, ne terminološko; v vrhu po pogostnosti je manjši nabor primerov, tj. eden ali dva.

Primerjava po spominu naštetih kolokacij s korpusno pridobljenimi podatki posredno evalvira izbrane frekvenčne prage in mere, ki so kriterij za umestitev kolokacije v slovar. Primeri, kot je po spominu pogosto navajana *zimska belina,* se pojavljajo v korpusu, ne pa v slovarju, zato so lahko indikator za spust frekvenčnega praga za vključitev gradiva. Na drugi strani podatki jasno osvetlijo problematičnost terminologije v referenčnem korpusu: na eni strani so v korpusu redke, vendar po spominu pogosto navajane zveze, kot npr. *izolirati* [*žico, kabel, elektriko*], na drugi strani se v slovarju na najbolj vidnih mestih pojavljajo terminološke kolokacije, kot npr. [*popravljena*, *nenatančna*] *belina*, ki jih sodelujoči po spominu ne

naštejejo (pa tudi v primeru, ko jih vidijo pred sabo, ne izbirajo za prva mesta slovarja).[21]

Preferenčne tendence sodelujočih se kažejo tako na ravni posameznih kolokatorjev kot kolokacijskih struktur. Pri slednjih je zanimivo, da v nasprotju s korpusnimi statistikami sodelujoči za prva mesta slovarja redko izbirajo predložne zveze: preferenca so dvodelne zveze s polnopomenskimi elementi. Podatki tudi pokažejo, da je za obravnavani iztočnici skupna frekvenca kolokacij v korpusu skladnejša s preferencami sodelujočih kot število kolokacij v posamezni strukturi, kar je pri nadgradnji slovarja mogoče uporabiti za (dinamično) prioretiziranje struktur znotraj slovarskih iztočnic.

Tudi pri razvrščanju posameznih kolokacij se korpusna pogostnost izkaže kot mera, ki je skladnejša s preferencami sodelujočih, še bolj pa v prid upoštevanju pogostnosti priča dejstvo, da so jo skoraj vsi sodelujoči opredelili kot kriterij, po katerem so kolokacije izbirali. Pri tem je seveda treba razumeti, da gre za *predvideno* pogostnost, ki je glede na rezultate raziskave tesno povezana s splošnostjo, kot je bilo izpostavljeno že zgoraj. Poudariti je treba, da takšna homogenost pri izbiri kateregakoli od ponujenih kriterijev ni bila pričakovana, je pa zato toliko bolj pomembna za nadaljnje premisleke. Rezultat namreč posredno nakazuje, da sodelujoči jagodnega izbora ne vidijo kot didaktično priložnost za usvajanje manj znanega besedišča, ampak se preferenca pravzaprav bliža korpusno osnovanemu jezikovnemu opisu, ki na prvo mesto postavlja pogoste (tudi razumljive, splošne – lahko bi torej rekli *znane*) kolokacije, s tem pa (najbrž) kompenzira tudi za pomenske informacije, ki jih slovar trenutno prinaša le mestoma. Redkejše, zaznamovane kolokacije so posledično stvar usmerjenega iskanja po slovarju, bodisi na klik prek jagodnega izbora ali pa neposredno prek iskalnega okenca. Dokaj intuitivno se zdi, da uporabniki za izhodišče želijo empirično določeno, stabilno,

21 Vključitev terminološkega gradiva je seveda povezana s slovaropisnimi odločitvami in pristopi. Za Kolokacijski slovar sodobne slovenščine je značilno, da nastane v prvem koraku strojno, nato se ročno nadgrajuje. Gesla, ki imajo ročno urejeno pomensko členitev (kot velja za geslo *belina*), zahtevajo ponazoritev vsakega identificiranega pomena – tudi terminološkega – s kolokacijami, kar (lahko) dodatno spodbudi vključitev v korpusu redkejšega gradiva v slovar.

prototipsko podatkovno jedro, možnosti za nadgradnjo jezikovnega znanja (svojega ali v znanja učencev) pa raje naslavljajo sami, saj so te potrebe individualne, jezikovna vprašanja pa kontekstno odvisna. Pri nadaljnjem razvoju bi bilo to mogoče upoštevati in uporabnikom ponuditi več kontrole nad urejanjem podatkov za didaktične namene, bodisi v slovarju bodisi v specializiranih orodjih v podporo jezikovnemu razvoju. Vendar bi bilo v prvem koraku koristno zagotoviti dodatne uporabniške raziskave, mdr. empirično definirati različne scenarije uporabe Kolokacijskega slovarja sodobne slovenščine.

Kljub kompleksnosti vprašalnika ocenjujemo, da je za dani namen dobro zasnovan in pridobljeni podatki primerni za nadaljnje statistične analize. Za veljavnost raziskave je bilo nujno zagotoviti, da sodelujoči ustrezno interpretirajo pojem kolokacije. Anketa vključuje preverbo razumevanja dveh vrst, kar se je izkazalo za ustrezno. Pregled metapodatkov kaže, da so bile ciljne uporabniške skupine pri anketiranju ustrezno zajete in da podatki omogočajo primerjalne analize za cca. 415 anket. V nadaljevanju bi bilo dragoceno statistično analizirati in opredeliti morebitne korelacije med kriteriji za razvrščanje ter starostjo, statusom ali poklicnim interesom sodelujočih. Takšen vpogled bi omogočil boljše razumevanje preferenc in potreb posameznih uporabniških skupin in zagotovil morebitne adaptacije podatkovnega prikaza tudi na ravni slovarskega vmesnika.

Zahvala

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

Reference

Arhar Holdt, Š., Kosem, I. in Gantar, P. (2016): Dictionary user typology: the Slovenian case. V T. Margalitadze in G. Meladze (ur.): *Lexicography and linguistic diversity: proceedings of the XVII EURALEX International*

Congress: 179–187. Tbilisi: Ivane Javakhishvili Tbilisi State University. Dostopno prek: https://euralex.org/publications/dictionary-user-typology-the-slovenian-case/ (9. 3. 2021).

Arhar Holdt, Š., Logar, N., Pori, E. in Kosem, I. (2020): "Game of Words": Play the Game, Clean the Database. V Z. Gavriilidou, M. Mitsiaki in A. Fliatouras (ur.): *Lexicography for inclusion: EURALEX XIX: Congress of the European Association for Lexicography*: *Vol. 1*: 41–49. Dostopno prek: https://euralex2020.gr/wp-content/uploads/2020/11/EURALEX2020_ProceedingsBook-p041-049.pdf (9. 3. 2021).

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

Bergenholtz, H. (2011): Access to and presentation of needs-adapted data in monofunctional internet dictionaries. V P. A. Fuertes-Olivera in H. Bergenholtz (ur.): *E-lexicography: the Internet, digital initiatives and lexicography*: 30–45. London in New York: Continuum.

Bogaards, P. (2003): Uses and users of dictionaries. V P. Van Sterkenburg (ur.): *A Practical Guide to Lexicography*: 26–33. Amsterdam in Philadelphia: John Benjamins.

Fuertes-Olivera, P. A. in Tarp, S. (2014): *Theory and practice of specialised online dictionaries: Lexicography versus terminography*. Berlin in New York: de Gruyter.

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

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. https://doi.org/10.1093/ijl/ecw014.

Gorjanc, V., Gantar, P., Kosem, I. in Krek, S. (ur.) (2017): *Slovar sodobne slovenščine: problemi in rešitve* (1. izd.). Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani. Dostopno prek: https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/book/15 (9. 3. 2021).

Groves, M. R., Fowler, F. J. Jr., Couper, M. P., Lepkowski, J. M., Singer, E. in Tourangeau, R. (2004): *Survey methodology*. Hoboken (NJ): J. Wiley.

Kallas, J., Kilgarriff, A., Koppel, K., Kudritski, E, Langemets, M., Michelfeit, J., Tuulik, M. in Viks, Ü. (2015): Automatic generation of the Estonian

Collocations Dictionary database. 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*: 11–20. Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd. Dostopno prek: https://elex.link/elex2015/proceedings/eLex_2015_01_Kallas+etal.pdf (9. 3. 2021).

Kosem, I., Krek, S., Gantar, P., Arhar Holdt, Š., Čibej, J. in Laskowski, C. A. (2018): Collocations dictionary of modern Slovene. V J. Čibej, V. Gorjanc, I. Kosem in S. Krek (ur.): *Proceedings of the 18th EURALEX International Congress: lexicography in global contexts*: 989–997. Ljubljana: Ljubljana University Press, Faculty of Arts. Dostopno prek: https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/3000-1 (9. 3. 2021).

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

Kosem, I., Lew, R., Müller-Spitzer, C., Ribeiro Silveira, M., Wolfer, S., Dorn, A., Gurrutxaga, A., ... Nesi, H. (2018). The image of the monolingual dictionary across Europe: Results of the European survey of dictionary use and culture. *International Journal of Lexicography,* 32 (1): 92–114. doi: 10.1093/ijl/ecy022.

Krek, S., Arhar Holdt, Š., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešić, N., Kosem, I. in Dobrovoljc, K. (2020): Gigafida 2.0: the reference corpus of written standard Slovene. V N. Calzolari (ur.): *Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*: 3340–3345. Paris: ELRA – European Language Resources Association. Dostopno prek: http://www.lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf (9. 3. 2021).

Lew, R. in de Schryver, G.-M. (2014): Dictionary users in the digital revolution. *International Journal of Lexicography*, 27 (4): 341–359.

Logar Berginc, N. (2009): Slovenski splošni in terminološki slovarji: za koga? V M. Stabej (ur.): *Infrastruktura slovenščine in slovenistike. Obdobja 28*: 225–231. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani. Dostopno prek: https://centerslo.si/simpozij-obdobja/zborniki/obdobja-28/ (9. 3. 2021).

Logar, N., Grčar, M., Brakus, M., Erjavec, T., Arhar Holdt, Š. in Krek, S. (2012): *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES:*

gradnja, vsebina, uporaba (1. izd.). Ljubljana: Trojina, zavod za uporabno slovenistiko: Fakulteta za družbene vede.

Lorentzen, H. in Theilgaard, L. (2012): Online dictionaries – how do users find them and what do they do once they have? V R. Vatvedt Fjeld in J. M. Torjusen (ur.): *Proceedings of the 15th EURALEX International Congress*: 654–660. Oslo: Universitetet i Oslo, Institutt for lingvistiske og nordiske studier.

Müller-Spitzer, C. (ur.) (2014): *Using Online Dictionaries*. Berlin in Boston: De Gruyter Mouton.

Pori, E., Čibej, J., Kosem, I. in Arhar Holdt, Š. (2020): The Attitude of Dictionary Users towards Automatically Extracted Collocation Data: A User Study. 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): 168–201. doi: 10.4312/slo2.0.2020.2.168-201.

Pori, E., Čibej, J., Kosem, I. 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.

Rychlý, P. (2008): A Lexicographer-Friendly Association Score. V P. Sojka in A. Horák (ur.): *RASLAN 2008: Recent Advances in Slavonic Natural Language Processing: Second Workshop in Recent Advances in Slavonic Natural language Processing*: 6–9. Brno: Masarykova Univerzita.

Stabej, M. (2009): Slovarji in govorci: kot pes in mačka? *Jezik in slovstvo*, 54 (3–4): 115–138.

Tarp, S. (2009): Reflections on Lexicographical User Research. *Lexikos*, 19 (1): 275–296.

Welker, H. A. (2013a): Methods in Research of Dictionary Use. V R. H. Gouws, U. Heid, W. Schweickard in H. E. Wiegand (ur.): Recent Developments with Focus on Electronic and Computational Lexicography [dodatna številka]. *Dictionaries. An International Encyclopedia of Lexicography*: 540–547. Berlin, New York: De Gruyter Mouton.

Welker, H. A. (2013b): Empirical Research into Dictionary Use since 1990. V R. H. Gouws, U. Heid, W. Schweickard in H. E. Wiegand (ur.): Recent Developments with Focus on Electronic and Computational Lexicography [dodatna številka]. *Dictionaries. An International Encyclopedia of Lexicography*: 531–540. Berlin, New York: De Gruyter Mouton.

Slovenske ontologije semantičnih tipov: samostalniki

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

Eva PORI
Filozofska fakulteta, Univerza v Ljubljani

The paper presents the Slovene Ontology of Semantic Types for nouns (SLONEST-noun), the first of a series of ontologies that will be joined under a name Slovene Ontologies for Semantic Types (SLONEST). First, we make an overview of existing ontologies, especially those that have been used in lexicographic projects for different languages, and determine their relevance for our purposes. We determine that WordNet, LexicoNet and the Estonian ontology are the most relevant for our purposes of preparing a not too detailed ontology of semantic types in terms of subcategory levels, but one that facilitates linking with lexical resources in other languages. Next, we present the SLONEST-noun ontology, which consists of 21 top-level categories and three levels of hierarchical subcategories. The ontology is also available in the CLARIN.SI repository. We describe each semantic type category, its subcategories, provide examples of nouns, and discuss the potential differences and similarities with other ontologies. Importantly, the ontology was developed and evaluated using the collocational data from the Collocations Dictionary of Modern Slovene, and the senses from the Comprehensive Slovene-Hungarian Dictionary, which are being compiled at the Centre for Language Resources and Technologies, University of Ljubljana. We also point out some of the issues we faced with certain (sub)categories and the decisions made. We conclude the paper by making an overview of how our top-level categories compare with those in other ontologies, and outlining plans for the future.

Keywords: ontology, semantic types, nouns, SLONEST, collocations

1 Uvod

Semantični tipi so nadpomenke, ki zastopajo pomen celotne skupine leksikalnih enot in opravljajo funkcijo abstraktnega pomenskega imenovalca. Ker semantični tipi predstavljajo pomenske koncepte v različnih medsebojnih razmerjih, jih je potrebno organizirati v hierarhično ontologijo. Tovrstne leksikalne ontologije so potem uporabne za različne namene v različnih disciplinah, zelo pomembno vlogo igrajo npr. v računalništvu in sorodnih znanostih pri kategorizaciji podatkovnih množic, dragocene pa so tudi pri snovanju leksikografskih opisov pomenskih konceptov, ki jih najdemo v slovarjih.

Semantični tipi so poleg abstrakcije pomenov leksikalnih enot pomembni tudi pri abstrakciji kolokacij oz. kolokatorjev, ki so pogosto uporabljeni kot izhodišče pri identifikaciji pomenov ter so ključni pri oblikovanju pomenskih opisov. Pomeni namreč v večji meri[1] izhajajo iz leksikalnih nizov pogostih kolokatorjev, ki jim je skupna določena semantična lastnost (Stubbs 2002: 449). Semantične tipe tako razumemo kot abstraktne pomenske povezovalce nizov konkretnih po pomenskoskladenjskih lastnostih podobnih kolokatorjev, ki so povezani z definicijami prek nadpomenke (npr. *barva* za semantični niz kolokatorjev *rdeča, modra, zelena* itd.), najbolj tipičnega ali splošnega kolokatorja (npr. *objekt* za semantični niz kolokatorjev *objekt, hlev, šola, hiša, hotel* itd.) ali podobnega pomensko povezanega poimenovanja.

V mednarodnem prostoru so leksikalne ontologije precej uveljavljene, pri čemer se zlasti v računalniškem jezikoslovju teži k uporabi enotne ontologije za različne jezike (takšen primer je predvsem WordNet), medtem ko v leksikografiji najdemo ontologije za različne jezike, ki bodisi izhajajo iz leksikalnih virov za določen jezik (npr. nemški LexicoNet) ali pa so bile izdelane na podlagi tujih, največkrat angleških, ontologij (npr. estonska ontologija).

V slovenskem jezikoslovju oz. leksikografiji trenutno ne obstaja neka uveljavljena leksikalna ontologija, čeprav se v zadnjem času kažejo prizadevanja v to smer, recimo taksonomija semantičnih tipov pri

1 Pomen v kombinaciji z leksikalnimi enotami, ki se pojavljajo v ožjem ali širšem kontekstu, namreč določajo tudi slovnične besede, npr. predlogi in prosti morfemi.

projektu Leksikalne baze za slovenščino (LBS)[2] (Gantar 2009, 2015b; Gantar idr. 2012)[3] oz. na njej temelječem Spletnem slovarju slovenskega jezika,[4] Pojmovnik Sinonimnega slovarja slovenskega jezika, prvi rezultati avtomatskega gručenja kolokacij in posledično nizov za potencialne semantične tipe pa so vidni v Kolokacijskem slovarju sodobne slovenščine (Kosem idr. 2018). Trenutno stanje v slovenskem prostoru torej kaže na potrebo po leksikalni ontologiji semantičnih tipov, ki bi služila kot osnova pri opredeljevanju pomenskih konceptov, pa tudi pri (avtomatskem) gručenju kolokacij in posledično prepoznavi posameznih pomenov. Takšna ontologija bi potem tudi omogočila izboljšanje avtomatskih postopkov in pohitrila izdelavo slovarjev; to pa je ključnega pomena, upoštevajoč dejstvo, da se v Sloveniji trenutno veliko jezikovnih virov izdeluje povsem na novo.

Izdelave Slovenskih ontologij semantičnih tipov (SLONEST) smo se lotili v okviru projekta KOLOS in v tem razdelku predstavljamo ontologijo semantičnih tipov za samostalnike (SLONEST-sam), pri čemer poleg predstavitve same ontologije ponudimo tudi pregled relevantnih tujih in slovenskih ontologij, ki so nam služile kot izhodišče. Zato predstavitev vsakega semantičnega tipa komentiramo tudi z vidika povezljivosti s semantičnimi tipi široko rabljene ontologije WordNet in relevantnih drugih ontologij.

2 Pregled obstoječih relevantnih ontologij

Ontologije najdemo v različnih disciplinah. Ker je bil naš cilj izdelati ontologijo za jezikoslovne oz. leksikografske namene, smo se pri pregledu tujih in domačih ontologij omejili predvsem na tiste, ki so bile uporabljene oz. se uporabljajo pri snovanju jezikovnih virov. Predstavljamo jih v nadaljevanju tega razdelka.

WordNet (Fellbaum 1998) je leksikalna podatkovna zbirka, v kateri osrednjo vlogo igrajo sinseti oz. sinonimski nizi (npr. *mešanica,*

2 Slovarska oz. leksikalna podatkovna zbirka *Leksikalna baza za slovenščino 1.0* (*Slovene lexical database 1.0*) je prosto dostopna in je na voljo na CLARIN.SI: http://hdl.handle.net/11356/1030.

3 Podoben pristop najdemo tudi v eSSKJ: https://fran.si/201/esskj-slovar-slovenskega-knjiznega-jezika

4 http://ssj.slovenscina.eu/spletni-slovar

zmes, asortiman), v katere so urejene iztočnice štirih besednih vrst (samostalniki, pridevniki, glagoli in prislovi). Vsak sinset predstavlja posamezen leksikalni koncept, spremlja pa ga razlaga, pogosto tudi oznaka in primer rabe. Zadnja dostopna različica je 3.1[5] in vsebuje 155.287 literalov (iztočnic), razvrščenih v 117.659 sinsetov. WordNet pa je tudi ontologija, ki vsakemu od sinsetov pripiše semantični tip, imenovan leksikografska mapa (ang. lexicographer file). Semantičnih tipov, ki so na eni sami ravni in torej niso deljeni v podkategorije, je skupaj 45, in sicer 26 samostalniških (npr. Človek, Žival, Čas, Proces), 15 glagolskih (npr. Telo, Sprememba, Komunikacija), 3 pridevniški in 1 prislovni. Tako je recimo zgoraj omenjeni primer sinseta (*mešanica* ipd.) označen s semantičnim tipom Skupinsko ('noun.group').

Slovenska verzija wordneta se imenuje sloWNet (Fišer 2009) in po strukturi ter pristopu sledi angleškemu izvirniku. Tako tudi uporablja zgoraj omenjeno ontologijo semantičnih kategorij. SloWNet je bil izdelan z avtomatskimi postopki, pri čemer so bili uporabljeni dvojezični slovarji, vzporedni korpusi in Wikipedija, kasneje pa so se podatki izboljševali tudi z uporabo metode množičenja. Zadnja verzija sloWNeta je 3.1 (Fišer 2015), vsebuje pa 43.460 sinsetov in 71.803 literalov, od katerih jih je bilo 33.546 ročno potrjenih.[6]

Pomemben vir je tudi FrameNet,[7] prosto dostopen vir, ki je zasnovan kot ontološka leksikalna baza, namenjena tako človeški kot računalniški rabi. Temelji na označenih primerih dejanske rabe, v katerih so besede ali besedne zveze analizirane z vidika njihovih pomenskih in skladenjskih razmerij. Vsak pomen besede je mogoče uvrstiti v svojo pomensko shemo (ang. frame), ta pa je opredeljena s t. i. shemskimi elementi (ang. frame elements), ki poimenujejo različne pomenske vloge. Trenutno je v FrameNetu 1224 pomenskih shem in 10.535 pomenskih elementov (1286 različnih).[8] Vsakemu shemskemu elementu je pripisan tudi semantični tip, ki naj bi med drugim služil za razlikovanje med leksikalnimi enotami, ki

5 http://wordnetweb.princeton.edu/perl/webwn
6 http://hdl.handle.net/11356/1026
7 https://framenet.icsi.berkeley.edu/fndrupal/
8 https://framenet.icsi.berkeley.edu/fndrupal/current_status

so povezane z eno ali več pomenskimi shemami. Tako sta recimo glagola *hvaliti* in *kritizirati* pokrita s shemo Sodba, razlikujeta pa ju semantična tipa pozitivna_sodba in negativna_sodba (Ruppenhofer idr. 2010: 86). V FrameNetu je 110 semantičnih tipov, ki so nadalje pogručeni v 41 krovnih tipov (t. i. supertipov). Precej semantičnih tipov je prekrivnih z WordNetovimi tipi in tipi ostalih ontologij, največje razlike pa so ravno v neontoloških tipih, kot je npr. Pragmatična_funkcija. Poleg tega razvrščenost supertipov ni vedno hierarhično smiselna, npr. pod supertipom Območje (*Region*) najdemo semantična tipa vodno območje (*Body_of_water*) in reliefno obliko (*Landform*), hkrati pa obstaja tudi supertip Vodno območje (*Body_of_water*), v katerem je semantični tip tekoča voda (*Running-water*).

FrameNetu podoben pristop prepoznavanja stavčnih vzorcev uporablja tudi pristop Corpus Pattern Analysis (CPA) (Hanks 2004, 2008; Hanks in Pustejovsky 2005), ki pa se osredotoča na analizo tipičnih pomenskih vzorcev posameznega glagola in manj na medsebojno povezovanje. S pristopom CPA je bil izdelan tudi slovar Pattern Dictionary of English Verbs.[9] Isti pristop je bil uporabljen tudi pri izdelovanju slovarjev za italijanščino (T-PAS), španščino (Verbario) in hrvaščino (CROATPAS). Za naše namene je bistvenega pomena ontologija CPA, ki vsebuje 253 semantičnih tipov za samostalnike, hierarhično razporejenih v 5 krovnih kategorij (Entiteta, Dogajanje, Skupina, Del in Lastnost), ki se delijo v nadaljnje hierarhično urejene podkategorije (do največ osem podravni).

Pristop, podoben CPA, se je uporabil tudi pri izdelavi Leksikalne baze za slovenščino (LBS), v kateri so se konkretnejša poimenovanja semantičnih tipov ročno gručenih kolokacij[10] že uporabila v stavčnih definicijah oz. pomenskih shemah, predvsem glagolskih iztočnic, npr. *RASTLINA **cveti**, ko so RAZMERE ugodne, da lahko požene cvetove*. Za RASTLINO so v tem primeru tipični kolokatorji *rastlina, roža, rožica, cvetlica, drevo* ipd., medtem ko imamo za

9 https://www.pdev.org.uk/

10 Pri gručenju kolokacij prvi kriterij ni bil vedno semantični, temveč formalni, npr. samostalniške kolokatorje pridevniških iztočnic se je najprej gručilo po spolu in šele potem po semantičnih lastnostih (akutna bolezen/levkemija/driska; akutno obolenje; akutni hepatitis/prehlad/bronhitis).

RAZMERE več semantičnih nizov, ki pripadajo eni od treh kategorij: čas (npr. ***cveteti*** *v poletnem/deževnem obdobju*), lokacija (***cveteti*** *na vrtu/polju*) in način (***cveteti*** *v rožnati/beli barvi*). Na podlagi analize pomenskih shem v LBS, torej z uporabo pristopa od spodaj navzgor, je bila izdelana štiristopenjska taksonomija semantičnih tipov, predstavljena v Gantar (2015a), ki pa zaradi manjšega nabora iztočnic v LBS še ne predstavlja celotne ontologije semantičnih tipov za samostalniške leksikalne enote.

Ontologija SIMPLE-CLIPS[11] je nastala v okviru projekta CLIPS (Corpora e Lessici dell'Italiano Parlato e Scritto), katerega cilj je bil izdelati korpuse in leksikone za italijanski jezik, tako govorjeni kot pisni. SIMPLE-CLIPS, ki je v marsičem podobna ontologiji CPA, ima 143 semantičnih tipov za samostalnike, ki so hierarhično razdeljeni v 4 krovne kategorije in nadalje v več podkategorij (do pete ravni). Razporejenost (pod)kategorij je nekoliko drugačna kot pri CPA, npr. Lastnost najdemo pod krovno kategorijo Entiteta, Skupina in Del pa sta obe podkategoriji krovne kategorije Sestavljeno (*Constitutive)*.

LexicoNet je ontologija nemških samostalnikov (Geyken in Schrader 2006) in se uporablja za leksikografske namene pri pripravi Digitalnega slovarja nemškega jezika (DWDS). LexicoNet je hierarhija konceptov in je na krovni ravni razdeljen na konkretne in abstraktne koncepte, potem pa vsako od kategorij nadalje drobi na več podkategorij, ki se lahko spet drobijo (do največ 10 ravni). Kot pišeta Geyken in Schrader (2006), vsebuje LexicoNet približno 90.000 leksikalnih enot, ki temeljijo na pomenih v velikem nemškem enojezičnem slovarju (Wörterbuch der deutschen Gegenwartssprache) in so s koncepti povezane na treh ravneh (tip, vloga in pojavitev), pa tudi z vidika meronimije in holonimije. Pomemben je podatek, da so pomene združevali v eno leksikalno enoto, če za njih niso našli ustreznega koncepta v ontologiji.

Za leksikografske namene je bila izdelana tudi ontologija semantičnih tipov za estonski jezik (Langemets 2010),[12] saj jo pri pripravi slovarskih in ostalih leksikalnih virov uporablja Inštitut za

11 http://webilc.ilc.cnr.it/clips/Ontology.htm

12 Zahvaljujemo se Margit Langemets za prevod ontologije v angleščino.

estonski jezik. Ontologija vsebuje semantične tipe za samostalnike, glagole, pridevnike in prislove in je v marsičem podobna WordNetovim semantičnim kategorijam, tako po številu tipov kot po njihovem poimenovanju. Semantični tipi so razdeljeni v zgolj dve ravni, krovne kategorije in podkategorije.

Prav tako leksikografsko motivirana ontologija je Pojmovnik Sinonimnega slovarja slovenskega jezika (SSSJ) (2016),[13] ki umešča samostalniške sinonimne nize v pojmovne skupine in podskupine. Pojmovnik SSSJ ni prosto dostopen, prav tako nismo zasledili dokumentacije ali znanstvenih publikacij, ki bi ponudile informacije o metodologiji izdelave in podrobnejšo utemeljitev hierarhične delitve. Ontologija sicer vsebuje sedem semantičnih tipov (Abstrakta, Človek, Predmet, Prostor, Rastlina, Snov, Žival), ki se delijo v kategorije in nekatere še v podkategorije. Opazne so številne podobnosti s tujimi ontologijami, čeprav najdemo tudi določene izjeme (npr. Meso kot podkategorija pod Snov; večina drugih ontologij ga ima pod Hrana).

Pregledane ontologije lahko razdelimo v dve skupini. Na eni strani so ontologije z dokaj širokim krovnim naborom semantičnih tipov in z malo ali brez podravnmi (WordNet, sloWNet, FrameNet, estonska ontologija). Na drugi strani so hierarhično zelo razvejane ontologije, navadno z malo krovnimi semantičnimi tipi in več ravnmi (pod)kategorij (CPA, LBS, SIMPLE-CLIPS, LexicoNet, Pojmovnik SSSJ), čeprav je LexicoNet s svojo zelo podrobno kategorizacijo in hierarhično razvejanostjo neke vrste izjema. Pomembno je poudariti, da se pri mnogih ontologijah kaže težnja po čim boljši povezljivosti z WordNetom kot široko uporabljenim virom v mednarodni skupnosti, saj to precej poveča uporabnost ontologije in seveda jezikovnih virov, ki jo uporabljajo. Tako obstajajo tudi študije, ki primerjajo možnosti povezovanja ontologij, npr. Koeva idr. (2018) primerjajo povezljivost kategorij v CPA in WordNetu.

13 https://fran.si/208/sinonimni-slovar

3 Izdelava ontologije semantičnih tipov za slovenščino

Pri pripravi ontologije smo izhajali iz želje, da bi ontologija čim bolj olajšala opredeljevanje semantičnega tipa leksikalnim enotam (enobesednim ali večbesednim). Pri tem smo imeli v mislih tako ročno označevanje kot polavtomatsko, tj. pregledovanje avtomatsko pripisanih podatkov. Na krovni ravni se nam je zaradi zagotovitve čim večje kasnejše povezljivosti s tujimi jezikovnimi viri kot izhodišče zdela najprimernejša ontologija, ki jo uporablja WordNet (in posledično sloWNet), vendar pa smo se zavedali, da za gručenje kolokacij in natančnejše opredeljevanje semantičnih konceptov potrebujemo tudi (pod)kategorije. Pri oblikovanju (pod)kategorij se je tako zdelo smiselno opreti na nekoliko podrobnejše leksikografsko motivirane ontologije, v našem primeru predvsem na nemško LexicoNet in deloma tudi estonsko, seveda pa smo gledali tudi ostale. Pri tem je pomembno izpostaviti, da smo od avtorjev LexicoNeta leta 2019 pridobili povsem zadnjo verzijo ontologije, ki je bila na podlagi evalvacij že nekoliko popravljena.

3.1 Metoda

Za izdelavo obsežne in hierarhično razvejane ontologije smo potrebovali širok nabor različnih leksikalnih enot in z njimi povezanih kolokacij. Ker smo izhajali iz WordNeta, smo za izhodišče vzeli samostalniške leksikalne enote oz. literale iz sloWNeta, ki so že imele pripisane krovne semantične tipe. Ob tem smo se zavedali, da je precej podatkov v sloWNetu avtomatskih, kar je pomenilo, da je bilo treba v prvem koraku potrditi njihovo umeščenost v krovno kategorijo. To potrjevanje je bilo opravljeno na podlagi pregleda kolokacij leksikalnih enot in po potrebi tudi korpusnih zgledov, v pomoč za potrditev ustreznosti prevoda pa so nam bili tudi angleški izvirniki. Čeprav so bile v marsikaterih primerih leksikalne enote same na sebi, brez konteksta, pomensko dovolj jasne, je bilo preverjanje v korpusih in slovarjih ključnega pomena, saj za ponazoritve ontoloških kategorij nismo želeli navajati leksikalnih enot oz. njihovih pomenov, ki se v jeziku ne pojavljajo. Kljub morebitni pojavitvi v korpusih pa smo zaradi očitne

neprimernosti izločali leksikalne enote tujega izvora, zlasti (zaradi avtomatskega postopka neprevedena) angleška poimenovanja (npr. *screwdriver, wake, wester)* in, pri rastlinah, latinska imena.

Naša metoda je bila kvalitativna, saj smo ročno preverili in razvrstili obsežen seznam leksikalnih enot, pri čemer smo kombinirali pristopa od zgoraj navzdol (abstrakcija konceptov) in od spodaj navzgor (abstrakcija nizov kolokacij). Izdelava ontologije semantičnih tipov je bila tako sestavljena iz naslednjih korakov:

1) Za vsak semantični tip se je najprej pripravil osnutek kategorij in podkategorij, ki je bil izdelan na podlagi pregleda hierarhije kategorij istih ali podobnih semantičnih tipov ontologij, omenjenih v Razdelku 2 (zlasti estonske in nemške) in tam navedenih primerov leksikalnih enot. To je tudi pomenilo prevajanje primerov leksikalnih enot iz tujih ontologij v slovenščino. Hkrati smo za pripravo izhodiščne kategorizacije opravili tudi pregled vzorčnega nabora leksikalnih enot iz sloWNeta oz. njihovih kolokacij.
2) Sledilo je razvrščanje in hkrati potrjevanje kategorizacije z obsežnejšim naborom leksikalnih enot. Pri tem koraku so sodelovale tri označevalke-jezikoslovke, pri čemer je vsaka prevzela določene semantične tipe, v primerih kompleksnejših in bolj problematičnih semantičnih tipov pa smo opravili tudi dvojno ali celo trojno označevanje. Označevanje je potekalo v skladu s splošnimi navodili za označevanje vseh semantičnih tipov in smernicami za pripisovanje (pod)kategorij, ki so bile vnaprej izdelane za vsak semantični tip posebej. Med splošnimi navodili velja posebej izpostaviti temeljna vodila:
    a) beleženje alternativnega koncepta: v primerih dvoma med dvema (pod)kategorijama na istem hierarhičnem nivoju;
    b) vpeljava nove semantične (pod)kategorije (v primeru, da to tendenco izkazuje več kandidatov);
    c) beleženje drugih konceptov, ki jih razkrijejo kolokacije leksikalnih enot: pri primerih, ki pripadajo več različnim (pod)kategorijam v hierarhiji (primeri z dvema ali več koncepti);
    d) oprtost opredeljevanja koncepta in pripisovanja semantičnega tipa na sodobne korpusne, slovarske vire in orodja:

Gigafida 2.0, Slovar sopomenk sodobne slovenščine, Kolokacijski slovar sodobne slovenščine, SkE[14] ipd.

Smernice so vključevale natančnejšo opredelitev posameznega semantičnega tipa in opise (pod)kategorij, ponazorjene s konkretnimi primeri. Smernice smo dopolnjevali in nadgrajevali v skladu z dogovori s sprotnih sestankov z označevalkami, ki so bili namenjeni obravnavi problematičnih mest in razreševanju ključnih označevalskih dilem.

3) Po označevanju vsakega semantičnega tipa je bilo treba opraviti še pregled konsistentnosti, preveriti smiselnost zasnovanih (pod)kategorij in sprejeti odločitve o njihovem morebitnem premeščanju, vpeljavi novih (pod)kategorij, vsebinski razširitvi ali preimenovanju posameznih (pod)kategorij ipd. Kot bo predstavljeno v nadaljevanju, pa smo se v določenih primerih odločili za premeščanje kategorije ali več kategorij v drug semantični tip, kar je v nekaterih primerih privedlo celo do opustitve semantičnega tipa oz. združevanja semantičnih tipov. Vse spremembe smo dosledno beležili, po eni strani za namene dokumentacije, po drugi pa zaradi zagotovitve kasnejše povezljivosti z ostalimi ontologijami.
4) V zadnjem koraku smo preizkusili izdelano ontologijo s pripisovanjem semantičnih tipov izbranemu naboru samostalniških gesel, ki jih pripravljamo za drugo verzijo Kolokacijskega slovarja sodobne slovenščine. Skupaj smo označili 1136 konceptov oz. pomenov 675 enobesednih iztočnic, pri čemer smo uporabili 271 različnih kategorij semantičnih tipov.

3.2 SLONEST-sam

Ontologija za samostalnike je zgolj prva v seriji ontologij za različne besedne vrste, ki bodo združene (gnezdene) pod krovnim imenom

14 Funkcija besedna skica (ang. Word Sketch) v orodju Sketch Engine nam lahko precej olajša prepoznavanje pomenov kolokatorjev in razbiranje pomenskih tendenc iz njihove kontekstualne okolice. Preverimo lahko semantično mrežo posamezne leksikalne enote oz. povezovalnost leksikalne enote v razmerju do druge leksikalne enote, kar nam je v pomoč pri pomenskem opredeljevanju oz. umeščanju leksikalnih enot v ustrezni semantični tip.

Slovenske ontologije semantičnih tipov (SLONEST). V tem razdelku sledi prikaz zgradbe slovenskih ontologij semantičnih tipov za samostalniške iztočnice (SLONEST-sam), predstavljenih s kratkim vsebinskim opisom kategorij in podkategorij ter opredeljenih tudi v odnosu do drugih ontologij.

Tabela 1: Seznam semantičnih tipov za samostalniške iztočnice z ID kodo v podatkovni bazi.

ID koda	Semantični tip[15]
01	ČLOVEK
02	TELO
03	ŽIVAL
04	RASTLINA
05	MIKROORGANIZEM
06	GLIVA
07	HRANA
08	SNOV
09	ARTEFAKT
10	PROSTOR
11	OBLIKA
12	POJAV
13	PROCES
14	MERA
15	ČAS
16	ČUSTVO
17	LASTNOST
18	STANJE
19	KOGNICIJA
20	AKTIVNOST
21	SKUPINSKO

Trenutna verzija SLONEST-sam[16] zajema 21 semantičnih tipov oz. konceptov s hierarhično urejenimi semantičnimi kategorijami in

15 Imena vseh krovnih kategorij oz. semantičnih tipov v besedilu zapisujemo s samimi velikimi črkami, pri navajanju njihovih kategorij in podkategorij pa ohranjamo samo veliko začetnico.

16 SLONEST-sam 1.0 je uradno objavljen in prosto dostopen na CLARIN.SI: http://hdl.handle.net/11356/1428.

podkategorijami (do največ četrte ravni), ki znotraj posameznega semantičnega tipa predstavljajo samostojne pomenske koncepte oz. skupine, opredeljene na skupnih pomenskih lastnostih (Tabela 1). Vsakemu semantičnemu tipu smo za lažje spremljanje in čezjezično rabo pripisali tudi ID kodo.

01-ČLOVEK

Semantični tip ČLOVEK se nanaša na leksikalne enote, ki opredeljujejo posameznika (človeka) po njegovih temeljnih značilnostih: telesnih, umskih, vedenjskih in mentalnih lastnostih; sorodstvenih ali nesorodstvenih razmerjih; (ne)poklicnih aktivnostih; po različnih načinih in oblikah nazorske usmeritve ali (ne)pripadnosti (ideološki, politični, družbeni); po družbenem statusu ali pravnem položaju in podobno. Ta tip vključuje tudi ostala človeku podobna mitološka bitja oz. antropomorfne entitete.

Zajema naslednje kategorije:

- Naziv (akademski, naslavljalni, plemiški; vzdevek): *doktor, profesor; gospod, gospodična; grof, kralj;*
- Lastnost (telesna, umska, mentalna, vedenjska, sorodstvena, nesorodstvena, geografska, nazorska ...): *garač, natančnež, neumnež; katoličan; Slovenec; partner, ljubica; strokovnjak; pragmatik, optimist; tabornik;* tudi primeri, ki so v enem od svojih pomenov zaznamovani: *kmet, šminker, boginja, čarovnica;*
- Aktivnost (poklic, funkcija, nosilec aktivnosti): *urednik, vzgojiteljica; minister, župan; šolar; napadalec, ujetnik;*
- Mitologija (pravljična, nadnaravna in bajeslovna bitja, božanstva in duhovi): *boginja, angel, duh.*

Pri semantičnem tipu ČLOVEK v primerjavi z WordNetom in ostalimi ontologijami na krovni ravni obstaja popolna prekrivnost, izjema je le nemški LexicoNet, ki ima mitološka bitja kot povsem ločen tip, na isti ravni kot ČLOVEK.

02-TELO

Semantični tip TELO se nanaša na leksikalne enote, ki označujejo osnovne, sestavne dele človeškega telesa: glava, vrat, trup, okončine (roke, noge), pa tudi notranje in druge (reprodukcijske/spolne) organe, kosti, tkiva in celice, pri čemer dele organov kot fizične dele uvrščamo v isto podkategorijo kot organe, katerih del so (npr. *prekat* = Notranji organi). Kot sestavne dele človeškega telesa pojmujemo tudi površinske elemente telesa (npr. *koža, lasje*) in telesne tekočine ali izločene snovi oz. telesne izločke (npr. *kri, slina*).

Semantični tip TELO označuje leksikalne enote, ki jih opredeljujejo naslednje kategorije:

- Glava ali čutni organ: *obraz, lobanja, lice, brada, usta, jezik, uho, oko;*
- Život (trup): *ženske prsi, materin trebuh;*
- Okončine (s funkcijo prijemanja in premikanja, pa tudi posameznimi, manjšimi deli, ki jih gradijo): *noga, roka; prst, noht, členek;*
- Površina telesa: *moška koža, dolgi lasje, brada, las, brki;*
- Notranji organi (ki opravljajo določeno funkcijo, procese dihanja, presnavljanja ...): *želodec, pljuča*; v to skupino uvrščamo tudi različne tipe krvnih žil: *arterija, kapilara, aorta, vena;*
- Kosti: *okostje, medenica, hrbtenica;*
- Tkiva in celice (krovna (žlezna), oporna (hrustančno, kostno tkivo), mišična, vezivna (maščobno, krvno, kostno, hrustančno tkivo) in živčna tkiva (živčni sistem oz. živčevje): *mišično tkivo, tetiva, celična stena; ščitnica, hipofiza, živec, hrbtenjača;*
- Drugi organi (kamor uvrščamo vse ostale organe, ki jih ne moremo uvrstiti v nobeno od ostalih navedenih kategorij; pogosto gre za reproduktivne oz. spolne organe): *anus, vulva, danka, maternica, nožnica, jajčnik, jajcevod, semenovod;*
- Telesne tekočine in snovi: *kri, znoj, slina, solze;*
- Drugo (kamor uvrščamo leksikalne enote, ki se nanašajo na splošnejša ali krovna/skupna poimenovanja): *telo, vitalni organ, vaskularni sistem, živčevje; dihalna pot, dihalna odprtina.*

Čeprav se leksikalne enote za bolezni ali (bolezenske) tvorbe nanašajo tudi na človeško telo, jih ne uvrščamo v ta tip, ampak sledimo logiki WordNeta in jih uvrščamo v podkategorijo Stanje človeka pri semantičnem tipu STANJE (*krasta, žulj, odrgnina*). Skladno z WordNetom v ta tip tudi ne uvrščamo leksikalnih enot za živalske dele telesa, ki jih opredeljujemo v okviru podkategorije Del telesa pri semantičnem tipu ŽIVAL.

Semantični tip TELO v SLONEST-sam je tako povsem prekriven s Telo ('noun.body') v WordNetu, obstajajo pa večje razlike z nemškim LexicoNetom, ki telo ali del telesa obravnava kot podkategorijo pri Naravna stvar ('Natural thing') pod tipom Fizična stvar ('Physical objects'), in nekoliko manjše z estonsko ontologijo, ki recimo dele živalskega telesa obravnava kar pod Telo.

03-ŽIVAL

Semantični tip ŽIVAL se nanaša na leksikalne enote, ki opredeljujejo žival glede na njeno pripadnost posamezni taksonomski kategoriji (deblu, razredu, redu, družini, rodu ali vrsti) ali (z)gradbenemu tipu, glede na temeljne skupne lastnosti, dele telesa ali mitološke poteze ipd.

Ločimo naslednje kategorije:

- Taksonomija (vretenčarji: sesalci, ptice, plazilci, dvoživke, ribe; nevretenčarji: členonožci, mehkužci, ožigalkarji; črvi in črvom podobne živali ipd.): *opica, kača, žaba;*
- Lastnost: *samica, mladič, kužek, mešanec;*
- Del telesa: *rep, taca, gobček;*
- Mitološka žival: *samorog, zmaj.*

Tudi pri semantičnem tipu ŽIVAL imamo precejšnjo prekrivnost z WordNetom in ostalimi ontologijami. Obstajajo pa razlike v obravnavi mikroorganizmov, za katere ima SLONEST-sam ločen semantični tip (podobno kot LexicoNet), WordNet pa jih obravnava kot živali. Razhajanja z ostalimi ontologijami pa najdemo tudi na ravni podkategorij, kjer SLONEST-sam sledi taksonomski delitvi, medtem

ko nekatere ontologije, npr. estonska, ločeno izpostavijo posamezne skupine, kot so ptice, ribe in insekti, ostale pa umeščajo v skupno podkategorijo Tip.

04-RASTLINA

Semantični tip RASTLINA se nanaša na leksikalne enote, ki opredeljujejo rastlino glede na njeno uvrstitev v posamezno taksonomsko kategorijo (deblo, razred, red, družino, rod ali vrsto) ali (z)gradbeni tip, glede na temeljne skupne ali posamezne lastnosti, dele rastline ali vrsto plodov.

Ločimo naslednje kategorije:

- Taksonomija (semenke, praproti, mahovi): *črni ribez, ringlo; jelenov jezik, goli protovec, šotni mah, jetrenjak;*
- Lastnost: *rastlinica, rožica;*
- Del rastline: *veja, steblo, cvet, iglica;*
- Plod: *češnja, jabolko, hruška.*

Semantični tip RASTLINA je prekriven s kategorijo Rastlina ('noun. plant') v WordNetu, izjema so le glive in lišaji, za katere ima SLONEST--sam ločen semantični tip in se v tem dejansko loči tudi od vseh ostalih ontologij, saj nobena gliv in lišajev ne obravnava ločeno (gl. spodaj).

05-MIKROORGANIZEM

Semantični tip MIKROORGANIZEM zajema leksikalne enote, ki se nanašajo na predstavnike večinoma enoceličnih (lahko tudi mnogoceličnih) organizmov oz. mikroorganizmov (mikrobov): *virusi, bakterije, alge, enocelične rastline* in *enocelične živali.* Kot že omenjeno, mikroorganizme v večini ostalih ontologij najdemo pod ŽIVAL, SLONEST-sam pa jih obravnava ločeno.

06-GLIVA

Semantični tip GLIVA zajema leksikalne enote, ki se nanašajo na predstavnike samostojnega kraljestva gliv: *goba, mušnica*, *lišaj.*

V vseh ostalih ontologijah so glive zajete pod rastlinami. Za ločen semantični tip smo se odločili na podlagi dejstva, da jih ekologi in biologi na podlagi že nekaj desetletij splošno sprejetega koncepta petih kraljestev živega izpostavljajo kot kraljestvo živih bitij, ločeno od rastlin in živali (Whittaker 1969; Podobnik 1985).

07-HRANA

V semantični tip HRANA uvrščamo leksikalne enote, ki se nanašajo na vse vrste jedi (po skupinah izdelkov: meso in mesni izdelki, mlečni izdelki, pekovski izdelki ipd.), tudi pripravljene jedi, namaze in olja, vse vrste pijač (osvežilne, sladke, grenke, alkoholne, brezalkoholne pijače ipd.), začimbe in dodatke, s katerimi začinimo ali izboljšamo okus jedi, pa tudi leksikalne enote za splošnejša ali skupna poimenovanja za obroke in ostalo hrano.

Zajema naslednje kategorije:

- Jed: *golaž, pica; marmelada; rastlinsko olje;*
- Pijača: *sok, malinovec, oranžada; čaj, kava; vino, pivo;*
- Začimbe in dodatki: *sol, kis, koriander, žafran;*
- Drugo: *prehrana, pojedina, obrok; predjed, priloga.*

V osnovi smo sledili logiki WordNeta, ki ima HRANO povsem ločeno ('noun.food'); za razliko od LexicoNeta, ki hrano uvršča v kategorijo Snovi in materiali ('Substances and materials'), znotraj podkategorije Material po funkciji ('Material by function'). Podobno kot WordNet, a drugače kot nekatere druge ontologije, pa smo kemične dodatke in aditive s prehrambeno funkcijo, kot so vitamini, emulgatorji, sredstva za zgoščevanje ipd., uvrstili v semantični tip SNOV.

08-SNOV

Semantični tip SNOV opredeljuje leksikalne enote, ki se nanašajo na snovi naravnega (živalskega in rastlinskega) in umetnega izvora v različnih vrstah agregatnega stanja (trdno, tekoče, plinasto), na različne vrste materiala (gradbeni, odpadni material in ostala sredstva ter surovine) ter kemijske elemente in spojine (elementi, spojine,

kovine, nekovine, zlitine, kemijski simboli in formule, kemijski pojmi). V semantični tip SNOV uvrščamo tudi osnovne (snovne) gradnike človeškega telesa, npr. hormone ali beljakovine.

Zajema naslednje kategorije:

- Naravna: *kamen, les, bombaž; voda, sneg;*
- Kamnine, kristali in minerali (kamor uvrščamo tudi drage in poldrage kamne): *diamant, zemlja, ruda, barit, boksit, pirit;*
- Umetna: *plastika, kevlar;*
- Material: s*teklo, gips, mavec, omet, cement, opeka, plutovina; zemeljski plin; gnoj;*
- Kemijska: *kisik, dušik; H20, kalcijev klorid; aluminij, zlato, Cu, C; NH_4CNO; izotop, atom, kislina; kortizol, trombocit, DNK.*

Glede obravnave tega semantičnega tipa se večina ontologij bolj ali manj ujema, pojavljajo se zgolj manjša odstopanja, med drugim tudi že prej omenjena obravnava kemičnih dodatkov in aditivov. Omeniti velja še material oz. sredstva glede na funkcijo (zdravila, opojne snovi, sredstva za higieno in nego telesa: *antibiotik, milo, šampon* ipd.), ki jih recimo LexicoNet obravnava pod SNOV, a smo v SLONEST-sam sledili WordNetu in jih uvrstili pod ARTEFAKT, v podkategorijo Sredstva ali snovi.

09-ARTEFAKT

Semantični tip ARTEFAKT zajema leksikalne enote, ki označujejo stvari oz. predmete, ki jih je ustvaril ali izdelal človek. Primere uvrščamo v ustrezno kategorijo in podkategorije glede na vrsto, način, funkcijo, položaj, namen ipd.

ARTEFAKT tako predstavljajo naslednje kategorije, ki se nadalje delijo v hierarhično urejene podkategorije:

- Oblačilo (obleka, obutev, pokrivalo, nakit in dodatki) glede na osebo (otroška, ženska, moška), položaj na telesu (zgornja, spodnja), material, funkcijo ali poseben namen: *jopica; nogavice, rokavice; volnena jopa, usnjena jakna; večerna obleka, športne hlače, smučarska jakna; bokserice, tangice; baretka;*

- Tekstilni izdelek: *vzorčasto blago, lanena rjuha, volneno pregrinjalo;*
- Posoda (za shranjevanje): *koš, vedro, zaboj, posoda za omako;*
- Prevozno sredstvo (kopensko, zračno, vodno, vesoljsko): *terenec, tovornjak; letalo, helikopter*; *ladja, jadrnica*; *raketa, vesoljska postaja;*
- Glasbeni inštrument (oz. skupine glasbil glede na način izvajanja (godala, pihala, trobila, ostala glasbila), njihove dele in tudi glasbene pripomočke): *boben, ksilofon; violina, kontrabas; kitara, harfa; flavta, klarinet*; *bariton, krilovka, kornet, trobenta; orgle, klavir; ustnik, struna*, *lok, trzalica;*
- Orožje (ročno, vojaška naprava, municija in ostalo strelivo): *puška, revolver; bomba, raketa, mina, granata;*
- Zgradba (enostavni kompleksi ali funkcionalne zgradbe (za človeka in žival) ter njihovi deli glede na funkcijo in namen (za bivanje, za delo, za storitve, za izvajanje javnih, kulturnih, verskih ali političnih dejavnosti, za hrambo, kot del infrastrukture ipd.)): *stanovanjska hiša, hlev; kuhinja, dnevna soba; tovarna, jeklarna; banka, slaščičarna; šola, občina, župnijski dom; drvarnica, garaža; kolesarska steza;*
- Dokument (tiskano ali pisno gradivo, listine, tudi e-dokumenti ali publikacije, besedila v spletni obliki oz. računalniški dokumenti): *prijavni obrazec, ljubezensko pismo, rokopis romana, izstavljen račun, zdravniški recept; blogerski zapis, internetna objava, tvit, e-sporočilo;*
- Denar (v konkretnem pomenu): *denar, bankovec, dolar, ček;*
- Pohištvo in oprema (ter deli pohištva in opreme): *stol, postelja*; *regal, omara*, *kavna mizica; umivalnik; noga od stola, kljuka od vrat;*
- Umetniški izdelek (umetnine ali umetniške kreacije, tudi konkretna umetniška dela (literarna, gledališka, glasbena ipd.)): *fotografija, skulptura, spomenik; Božanska komedija;*
- Komunikacija (informacijsko-komunikacijska tehnologija (IKT); grafični simboli oz. znaki za matematične pojme, kemijske elemente in druge abstraktne pojme): *H, Br, cm; vezaj, pika, osminka;*

- Naprava (računalniška, pisarniška, komunikacijska, zabavna, hišna in gospodinjska, signalna, svetilna in druge naprave): *računalnik, tiskalnik; radijska postaja; videorekorder; opekač kruha, mikrovalovna pečica; ventilator, semafor, reflektor;*
- Pripomoček (kuhinjski, računalniški in pisarniški, svetilni, športni, merilni, optični, igrače in drugi pripomočki (lepotni, spolni ipd.)): *pokrovka, metla; luknjač; sveča, blazina za vodo, športni obroč; višinomer, kompas; družabna igra; lesena noga, vibrator* ter
- Sredstvo ali snov (farmacevtska, za osebno nego in ostalo): *zdravilo; mamilo; šampon, milo.*

Pri semantičnem tipu ARTEFAKT najdemo nekoliko večja razhajanja med SLONEST-sam in WordNetom. Slednji ima namreč dokumente (npr. *knjiga, publikacija*) pod ločeno kategorijo Komunikacija ('noun.communication'), pri čemer pa razmejitvena linija ni povsem jasna oz. merilo razvrščanja ni enotno in dosledno: vizualne dokumente večinoma zasledimo pod Komunikacija, vendar ne vseh (primeri tipa *fotografija*), akustične in elektronske pa pod Artefakt ('noun.artifact'). Po načelu konkretnosti in konsistentnejše opredelitve kategorije smo vse dokumente (tudi elektronske, vizualne ipd.) uvrstili pod ARTEFAKT.

Druga večja razlika je obravnava denarja in z njim povezanih leksikalnih enot (*denar, ček* ipd.), ki so v SLONEST-sam zastopani s podkategorijo Denar pri ARTEFAKTU, medtem ko jih ima WordNet v ločeni kategoriji Lastnina ('noun.possession'), skupaj z besedami, kot so *zaklad, bogastvo, jamstvo, kredit, darilo* ipd. Dejansko je bila omenjena WordNetova kategorija precej majhna, tj. ni vsebovala veliko predstavnikov, in tudi heterogena, mi pa smo njeno vsebino pokrili z drugimi semantičnimi tipi.

Izpostaviti velja še leksikalne enote za pomen "zgradba ali del zgradbe", za katera ima SLONEST-sam ločeno kategorijo pri ARTEFAKTU, medtem ko številne druge ontologije (npr. estonska; ne pa tudi WordNet) tovrstne leksikalne enote uvrščajo v semantični tip Lokacija ('Location') ali Del ('Part'). Imajo pa po drugi strani mnoge

ontologije pod ARTEFAKT tudi leksikalne enote, ki označujejo skupinska poimenovanja, ki pa jih mi (podobno kot WordNet) obravnavamo v ločenem semantičnem tipu.

10-PROSTOR

Semantični tip PROSTOR zajema leksikalne enote, ki se nanašajo na lokacijo, na (zunanjo) površino in navadno nezamejeno območje, ki se po tem razlikuje od (zamejenega) objekta, zgradbe ali notranjega prostora.

Semantični tip PROSTOR opredeljuje vse, kar se nanaša na kategorije:

- Naravni (vesolje (tudi planeti), zračni prostor, vodno območje, kopno in del kopnega): *nebo; Mars, Jupiter; magnetosfera; slano jezero, meteorski potok; visoka gora, peščen hrib; rt, otok;*
- Geopolitični (podkontinent, država, regija, mesto ali naselje, okrožje, trg ali ulica): *Belgija, Bretanija, Celje, Stritarjeva ulica*;
- Mitološki: *raj, paradiž, pekel;*
- Drugo (ostale leksikalne enote, ki opredeljujejo lego, položaj, smer kraja ali območja): *časovni pas, zemljepisna širina, geocentrična širina, sever, jug.*

Pri oblikovanju semantičnega tipa PROSTOR smo se, za razliko od ostalih semantičnih tipov, precej bolj opirali na nemški LexicoNet in CPA kot na WordNet in tudi estonsko ontologijo. V izhodišču se nam je zdelo smiselneje, da združimo naravna in ne naravna prostorska poimenovanja, ki jih WordNet obravnava v ločenih kategorijah. Tako WordNet uvršča primere, ki se nanašajo na vesoljski in vodni prostor (vesoljski predmeti, nebesna telesa, morja, kopensko vodovje ...), pod Objekt ('noun.object'), enako vse, kar se nanaša na zračni del (*atmosfera, magnetosfera*), pa tudi kopenski del (npr. kontinente), razen če gre za del zemlje v rabi lokacije. Iz tega vidika obstaja skoraj popolna prekrivnost z našo podkategorijo Naravni prostor. Glavni kriterij je torej naravna danost oz. nekaj, kar ni ustvaril človek, vendar pa najdemo v WordNetu nekatere nedoslednosti.

Recimo pod Objekt so umeščeni podatomski delci, kot so *nevtron, elektron* in *hadron*, ne pa tudi *atom* (ki je pod Substance oz. Snov). Mi smo vse delce umestili pod SNOV. Poleg tega najdemo izjeme, kot sta *savana* in *oaza*, ki sta kategorizirani kot Lokacija ('noun.location') in ne Objekt.

Pri kategoriji Lokacija v WordNetu najdemo še več nedoslednosti, posebej v odnosu do Artefakta. Na primer veliko prostorov po namenu, kot so *razstavišče, letališče, dvorišče*, WordNet umešča pod Artefakt, ne pa tudi *tehnološki park, industrijska cona, smetišče*, ki jih najdemo pod Lokacija. Prostore s t. i. lastninsko konotacijo, kot je npr. *zemljišče, posest*, WordNet umešča v ločeno kategorijo Lastnina ('noun.possession'). Mi smo se odločili za omejitev s človekom povezanih prostorskih poimenovanj v okviru tipa PROSTOR na geopolitična, ostala, predvsem namenska, pa poenotili in v skladu s krovno opredelitvijo kategorije uvrstili pod ARTEFAKT.

Omeniti velja še posebnost leksikalnih enot za pomen "zgradba", ki jih mnoge druge ontologije (estonska, FrameNet, CPA, Simple-CLIPS) obravnavajo pod Lokacijo ('Location' ali 'Place'), SLONEST-sam pa sledi WordNetu in LexicoNetu, ki jih obravnavata pod Artefakt.

11-OBLIKA

V semantični tip OBLIKA uvrščamo leksikalne enote, ki opredeljujejo obliko stvari in ostale predmetnosti, geometrijske like in telesa v dvo- ali trirazsežnem prostoru in prostor sam, torej vse, kar nas obdaja, po izgledu/videzu oz. pojavnosti v ravnini ali na površini.

Zajema kategorije, ki označujejo:

- Točko (posamezne, določene manjše dele/elemente prostora (razsežnosti) oziroma manjša mesta na površini): *pika, stikališče, kraj; konusni presek;*
- Črto (množico točk oz. zvezno vrsto točk, tj. premo ali krivo linijo, ki ponazarja gibanje v prostoru ali v ravnini): *elipsa, krožnica, kotna razdalja, diagonala, osnovnica;*

- Površino (dvorazsežni prostor in objekte oz. like kot dele ravnine): *krog, kvadrat, enakostranični trikotnik;*
- Geometrijsko telo (trirazsežni prostor in objekte oz. geometrijska telesa kot dele prostora): *kocka, valj, kvader, elipsoid;*
- Drugo (večpomenske leksikalne enote, ki implicirajo pomen oblike): *reža, ovinek, izboklina.*

Podobno kot WordNet obravnavamo OBLIKO kot samostojno kategorijo. V Pojmovniku SSSJ te kategorije ne zasledimo, mnoge ontologije jo imajo pod Lastnost, LexicoNet pa jo umešča tako na abstraktno kot stvarno raven (krog kot abstraktna oblika ali npr. matematični lik na papirju ali zaslonu), kar na nek način upravičuje potrebo po ločeni kategoriji.

12-POJAV

Semantični tip POJAV označuje čutno zaznavno (nenavadno, specifično) dogajanje oz. fenomen, pri katerem se skozi opazovanje in izkustveno zaznavanje razkriva ali (po)kaže tudi stvar sama na sebi. Zajema leksikalne enote, ki se nanašajo na naravne oz. vremenske pojave, pa tudi na druge pojave ali spremembe (npr. fizikalne), ki se zgodijo ali potekajo neodvisno od človeškega vpliva ali aktivnosti.

V semantični tip POJAV uvrščamo leksikalne enote, ki se nanašajo na kategorije:

- Naravni (ki se nanaša na vreme, podnebje): *sneg, dež, veter, megla, burja, toča; snežinka;*
- Fizikalni (ki se nanaša na fizikalne lastnosti, dogajanje in fizikalne spremembe): *resonanca, prevodnost, vez;*
- Drugo (ostali pojavi): *lesket sonca*, *žuborenje potoka.*

Semantični tip POJAV v SLONEST-sam je prekriven s semantičnim tipom Phenomenon v WordNetu in estonski ontologiji, s pomembno razliko, da SLONEST-sam pod POJAV uvršča tudi naravne dogodke (npr. *potres, mrk*), ki jih ima WordNet v ločeni kategoriji

Dogodek ('noun.event'). Težava je bila namreč v tem, da je bilo razliko med dogodki in pojavi včasih težko opredeliti – dogodki naj bi bili sicer krajše in manj predvidljive narave. Tako je na primer *orkan* v WordNetu Pojav, *izbruh vulkana* pa Dogodek. Nadalje WordNet opredeljuje kategorijo Dogodek kot "nouns denoting natural event" (samostalniki, ki pomenijo naravne dogodke), vključuje pa tudi človeške dogodke, npr. *party* ('zabava'), *celebration* ('proslava'), *match* ('tekma'). Zaradi vseh omenjenih nedoslednosti smo se odločili v SLONEST-sam po eni strani poenotiti naravne pojave (in dogodke) znotraj semantičnega tipa POJAV, človeške dogodke pa obravnavati pod AKTIVNOST.

13-PROCES

Semantični tip PROCES zajema leksikalne enote, ki se nanašajo na proces kot skupek ponavljajočih se ali občasnih dejavnosti, ki medsebojno vplivajo na ustvarjanje rezultata oz. vodijo do rešitve. Temeljna lastnost, ki opredeljuje semantični tip, je procesnost, potek nastajanja, postajanja, spreminjanja stvari, ki se navadno odvija po (vnaprej) določenih pravilih, metodah ali postopkih, nanaša pa se lahko na različna področja in ravni: na področje ekonomije, tudi na socialne, gospodarske in druge z ekonomijo povezane determinante ali ekonomske pojme (prebivalstvo, zaposlenost, človeški kapital, proizvodni proces ...), na področje kemije, človeškega telesa ali poglavitnih življenjskih procesov v človeškem organizmu in naravi/okolju.

V semantični tip PROCES uvrščamo leksikalne enote, ki se nanašajo na naslednje kategorije:

- Telesni: *solzenje, rojevanje, prebava, prehranjevanje;*
- Naravni-fizikalni-kemijski: *cvetenje češnje, vegetativno razmnoževanje, fotosinteza; parna destilacija, filtracija zraka;*
- Ekonomski: *manjšanje prebivalstva, rast prebivalstva, gospodarska rast, globalizacija gospodarstva, decentralizacija financiranja.*

Tudi pri snovanju tega semantičnega tipa in njegovih (pod)kategorij smo se močno oprli na WordNet, je pa treba opozoriti, da WordNet v opisu navaja, da semantični tip vsebuje leksikalne enote, ki se nanašajo na naravne procese, najdemo pa tudi takšna, ki vključujejo človeka (npr. medicinski postopek). A kot že omenjeno, naš glavni kriterij pri tem semantičnem tipu ni človeški izvor oz. vpliv, ampak kompleksnost in v večini primerov daljše obdobje trajanja, skladno s tem smo tudi medicinske procese oz. postopke, ki jih izvaja človek ali je njihov udeleženec, obravnavali v okviru semantičnega tipa AKTIVNOST.

14-MERA

V semantični tip MERA uvrščamo leksikalne enote, ki se nanašajo na različne vrste števil (glavna, cela, algebrska števila) ali posamezne predstavnike vrste števil ter mere. Zajema tudi splošna matematična merska poimenovanja, uradne merske enote in njihove okrajšave.

Semantični tip MERA opredeljujejo temeljne kategorije:

- Števila (različne vrste števil ali splošna poimenovanja za vrste števil): *racionalno število, realno število*; *praštevilo; transcendentno število; največji skupni delitelj, množitelj*; *enice, desetice;*
- Matematične mere (splošna matematična merska poimenovanja, tudi poimenovanja funkcij, konstant): *fi, kvadrat, logaritem, sinus; četrtina, stotina; procent;*
- Enote (vse uradne merske enote, tudi okrajšave uradnih mer ali starejše uradne mere, izlastnoimenska poimenovanja za enote in ostale (dolžinske, utežne, prostorninske) enote): *kvadratni meter, kubični centimeter; kg, g; čevelj, komolec, palec; unča, pud*; *bokal; Celzij, Kelvin; astronomska enota;*
- Valute (tudi starejše, domače in tuje denarne enote): *evro, dolar, frank, drahma, goldinar.*

Pri tem semantičnem tipu smo sledili večini ostalih ontologij, omeniti velja le, da nekatere ontologije (npr. LexicoNet) ne obravnavajo vseh zgoraj naštetih kategorij pod istim krovnim konceptom.

15-ČAS

V semantični tip ČAS uvrščamo leksikalne enote, ki označujejo manjše ali večje enote za čas, daljša časovna obdobja, ki so lahko splošna ali pa specifična, zgodovinsko določena; konkretne navedbe datumov in ur, pa tudi imena različnih praznikov, ki predstavljajo trenutek v času.

Semantični tip ČAS zajema naslednje kategorije:

- Časovna enota: *sekunda, minuta, ura, dan, teden, mesec, leto;*
- Obdobje (splošno ali zgodovinsko daljše časovno obdobje): *mladost, otroštvo, semester, dopust, počitnice*; *bronasta doba, devon, mezolitik, novi vek;*
- Trenutek (v času): *silvester, martinovo, velika noč, božič.*

Pri tem semantičnem tipu ni bistvenih odstopanj med SLO-NEST-sam in ostalimi ontologijami. Omeniti velja zgolj nenavadne umestitve nekaterih leksikalnih enot pri WordNetu v ta semantični tip, npr. *smrtnost* (*nizka raven smrtnosti*).

16-ČUSTVO

Semantični tip ČUSTVO pripisujemo vsem leksikalnim enotam, ki označujejo (večinoma kratkotrajna) duševna in telesna čustvena stanja človeka. Pri opredelitvi te kategorije smo presegli klasično delitev čustev na negativna in pozitivna ter izhajali iz Parrottove tristopenjske hierarhične delitve,[17] pri čemer smo za temelj vzeli primarni nivo bazičnih čustev. Izhajali smo iz kategorij bazičnih oz. enostavnih čustev, ki imajo različno vrednostno komponento – lahko so pozitivna (*ljubezen, veselje, presenečenje*) ali negativna (*presenečenje, jeza, žalost, strah*) – in jih pripisujemo leksikalnim enotam za kompleksna oz. sestavljena čustva.

V semantični tip ČUSTVO uvrščamo leksikalne enote, ki se nanašajo na naslednje kategorije:

17 Izhajamo iz raziskav W. Gerroda Parrotta: *Emotion knowledge: further exploration of a prototype approach* (1987) in *Čustva v socialni psihologiji* (2001), kjer je predstavljena tristopenjska hierarhična delitev čustev na primarna (enostavna) čustva, ki se na sekundarnem in terciarnem nivoju nadalje cepijo na več sestavljenih/kompleksnih čustev.

- pozitivno čustvo Ljubezen (ki je posledica navezovanja, pozitivnega odnosa do drugega, kot je naklonjenost, (spolno) poželenje ali hrepenenje): *prisrčnost, nežnost, privlačnost, sentimentalnost, sočutje; poželenje, želja, strast, hrepenenje;*
- pozitivno čustvo Veselje (občutek popolnega zadovoljstva ali izpolnitve želje): *radost, zadovoljstvo, optimizem, veselje; vznemirjenje, zadovoljstvo, blaženost;*
- pozitivno ali negativno čustvo Presenečenje (ki ga doživljamo, kadar se izkaže, da imajo določene stvari, ljudje ali situacija drugačne lastnosti od pričakovanih): *začudenje, zbeganost;*
- negativno čustvo Jeza (ki se nanaša na občutek nezadovoljstva ali sovražnosti): *vznemirjenost, pretiravanje; bes, zavist, ljubosumje, razočaranje;*
- negativno čustvo Žalost (ki se nanaša na trpljenje, razočaranje, sram ali zanemarjanje): *agonija; krivda, obžalovanje; odtujenost, zavračanje, ponižnost, osamljenost;*
- negativno čustvo Strah (ki nastopi predvsem ob občutku ogroženosti (nas samih, naših vrednot) in nemoči, da bi se zaščitili; lahko pa sproži tudi neobvladljivo željo po umiku): *groza, živčnost, negotovost, panika, strah;*
- Drugo (vsa splošna, krovna ali skupna poimenovanja ali sorodni pojmi za čustva oz. občutja): *afekt, razpoloženje, občutek.*

LexicoNet umešča vsa čustva pod semantični tip LASTNOST, v različnih (prekrivnih) podkategorijah. WordNet ima za čustva tudi ločen semantični tip ('noun.feeling'), kamor umešča občutja in čustva, leksikalnim enotam, ki se nanašajo na občutja, pa pogosto pripisuje tudi pomen za lastnost; npr. *zvestoba* je OBČUTJE (tistega, ki se čuti zvestega) in LASTNOST (tistega, ki je zvest; gledano z zunanje perspektive).

V povezavi s semantičnim tipom ČUSTVO in WordNetom je treba omeniti tudi sicer slabo zastopan tip Motiv ('noun.motive'), v katerem so samostalniki, ki opredeljujejo cilje, npr. *obsession* ('obsesija'), *mania* ('manija'). Večino teh leksikalnih enot smo mi pokrili z drugimi semantičnimi tipi, predvsem ČUSTVO, LASTNOST in STANJE.

17-LASTNOST

Semantični tip LASTNOST označuje leksikalne enote, ki opredeljujejo trajnejše značilnosti (za razliko od kratkotrajnih, kot npr. čustvo), po katerih se posameznik (človek), predmetnost oz. ostale stvari razlikujejo od drugih.

Zajema naslednje kategorije:

- Človeška (osebnostna, tudi telesna, in biološka): *sočutnost, družabnost, ljubeznivost; slokost, debelost, pritlikavost; vitalnost, človeškost, ženskost;*
- Čutnozaznavna (se nanaša na videz in občutenje ter prostorsko razsežnost): *svetlost, barva; valovitost; glasnost; širina, višina, globina;*
- Področna (fizikalna, kemijska in drugo): *radioaktivnost, sila, navor; hidrofilnost, hidrofobnost, kristalna struktura;*
- Splošna (se nanaša na človeka in predmetnost): *škodljivost, nezakonitost, nestalnost, urejenost, snovnost, uporabljivost, pritrjenost.*

Pri semantičnem tipu LASTNOST smo sledili logiki WordNeta ter vanj uvrstili tudi prostorske in vizualne lastnosti, v skupno podkategorijo čutnozaznavnih lastnosti (ki zajema čutnost, vizualnost in prostorsko razsežnost). Pojmovnik SSSJ jih obravnava ločeno oz. v okviru samostojnih podkategorij; npr. barve (*belina*, *sivina*) pod Barva in vizualne lastnosti (*mračnost*) pod Videz. Številne ontologije vključujejo pod LASTNOST tudi Obliko, ki pa jo v SLONEST-sam obravnavamo kot samostojen semantični tip.

18-STANJE

V semantični tip STANJE uvrščamo leksikalne enote, ki opredeljujejo način obstajanja različnih procesov v določenem trenutku: tj. telesno, duhovno/duševno, čustveno in trajnejše bolezensko stanje posameznika (človeka), splošno, telesno, zdravstveno, higiensko oz. bolezensko stanje živalskih ali rastlinskih vrst, stanje posamezne predmetnosti, ki se nanaša na vse ostale (ne človeške) odnose, na

razmerja med posameznimi stvarmi, elementi, količinami, vrednostmi, delom in celoto, ali stanje posameznika v razmerju do drugega, na medčloveške (družbene) odnose/razmerja nasploh oz. položaj človeka v družbi.

Semantični tip STANJE zajema naslednje kategorije:

- Človek (bolezensko, duševno ali telesno): *rak, epilepsija, aids; prijateljstvo, partnerstvo, sorodstveni odnos;*
- Žival: *papigovka, brejost, brezmlečnost;*
- Rastlina: *plesen, ogorelost;*
- Razmerje: *nasprotnost, protislovje, slovnično razmerje;*
- Finance: *bogastvo, revščina, premoženje;*
- Splošno: *celovitost, onesnaženost, razmetanost.*

Semantični tip STANJE v SLONEST-sam združuje tri WordNetove kategorije, tj. Stanje ('noun.state'), Lastnina ('noun.property'), Razmerje ('relation'), pa tudi že omenjeno Motiv. Omeniti velja predvsem primere tipa *bogastvo, premoženje*, ki jih WordNet umešča v ločeno podkategorijo Lastnina, mi pa smo jih uvrstili v podkategorijo STANJE-finance, in primere tipa *denar*, ki so nekoliko problematični z abstraktno-konkretnega vidika in smo jih uvrstili v podkategorijo Denar pri ARTEFAKTU (v konkretnem pomenu). Razmerje, ki ga imamo v SLONEST-sam kot podkategorijo pod STANJE, je sicer ločena kategorija v WordNetu, a raba spet ni dosledna, npr. *(human) relationship* ('(človeški) odnos')) najdemo pod Razmerje, ne pa tudi *prijateljstvo*, ki je uvrščeno pod Stanje.

19-KOGNICIJA

V semantični tip KOGNICIJA uvrščamo leksikalne enote, ki so povezane s človeškim znanjem in se nanašajo na človeške kognitivne procese ali človeške kognitivne procese ali razumske, zaznavne, spoznavne in/ali presojevalne spretnosti in zmožnosti ter ostale človeške umske dejavnosti.

Osnovno vodilo oz. ključen kriterij pri označevanju je kognicija (umskost) in abstraktnost. Prototipični primeri te kategorije so

leksikalne enote tipa *domneva, stališče, mnenje*, *načelo,* pa tudi tiste, ki se neposredno ali posredno nanašajo na vede oz. discipline (*Kepplerjev zakon* = fizika, *ateizem* = religija).

Semantični tip KOGNICIJA zajema naslednje kategorije:

- Umsko-zaznavni procesi in stanja (logične oz. kognitivne operacije; razumski, umski, spoznavni, presojevalni, pa tudi zaznavni človeški procesi, stanja in dejavnosti): *dejstvo, domneva, predsodek, stališče, mnenje; vonj, slušna zaznava;*
- Spretnosti: *jahalna spretnost, melodičen posluh, obvladovanje tehnike;*
- Vede in discipline (področja): *Daltonov zakon, pesništvo, estetika, pediatrija, razvojna psihologija*.

Pri tem semantičnem tipu smo v SLONEST-sam sledili WordNetu, ki ga edini izpostavlja kot samostojno kategorijo. Ostale ontologije uporabljajo ločene (pod)kategorije za umske procese oz. vede in področja.

20-AKTIVNOST

Semantični tip AKTIVNOST opredeljuje leksikalne enote, ki vključujejo človeško aktivnost, potrebno za uresničitev, dosego namena ali cilja, in lahko segajo na različna področja: zaznavno, čutno in čustveno ter kognitivno področje, na področje družbenih aktivnosti, medsebojnih stikov, področje umetnosti, gospodarskih in negospodarskih dejavnosti ter nenazadnje na področje osnovnih (človeških) življenjskih procesov in dejanj (kot je npr. *spanje*).

Zajema naslednje kategorije:

- Telesna (zajema telesno nego in osnovne življenjske procese oz. vitalne funkcije): *tuširanje, umivanje, piling, striženje; dihanje, spanje;*
- Stik (s predmetnostjo ali osebo): *dotikanje, božanje, objemanje; udarjanje;*
- Percepcija (vezana na čutno zaznavanje): *vohanje, poslušanje, gledanje, zaznavanje;*

- Čustvena: *razburjanje, doživljanje; čustveno sprejemanje, toleriranje;*
- Kognicija (se nanaša na logične, kognitivne ali umske operacije oz. aktivnosti): *raziskovanje, načrtovanje;*
- Komunikacija (govorno-pisna, telesna, nečloveška): *pisanje, govorjenje, pohvala, obljuba; odkimavanje, mahanje, mežikanje; lajanje, meketanje;*
- Gibanje/premikanje: *plazenje, skakanje, poskakovanje; vstopanje, izstopanje; zibanje, padanje; prečkanje;*
- Zaužitje (se nanaša na proces pridelave, priprave in uživanja hrane/pijače): *gnojenje, žetev; kuhanje, pečenje; prehranjevanje, pitje, goltanje, žvečenje;*
- Sprememba: *uničenje, čiščenje ulic, pranje avta;*
- Tekmovanje/konflikt: *bojevanje, udarec, soočenje, obračunavanje, pretep;*
- Stvaritev/ustvarjanje: *izdelovanje nakita, izdelava slike; ročno šivanje, tkanje platna, vezenje na platno; restavriranje freske;*
- Lastnina: *kupovanje, prodajanje, trgovanje; prodaja, nakup, odkup, cenitev;*
- Družbene aktivnosti: *pravično vladanje, demokratično vodenje, druženje krajanov, organizacijska podpora, sodelovanje javnosti;*
- Stanje: *bivanje študentov, zimsko mirovanje;*
- Zvočni pojav (če je povzročitelj človek, žival ali naprava): *hreščanje radia, brnenje budilke, frfotanje s krili;*
- Človeški dogodek: *orgelski koncert, sejem gradbeništva; reprezentančna tekma; vojna;*
- Medicinski postopek: *translacija; računalniška tomografija, skupinsko zdravljenje;*
- Dejavnost-negospodarska (šport, ples, igra): *tek, odbojka, nogomet; standardni ples; kartanje;*
- Dejavnost-gospodarska: *svilogojstvo, živinoreja, tesarstvo.*

Semantični tip AKTIVNOST najdemo v vseh ontologijah (pod različnimi imeni), še najbolj se ravno tu razlikuje WordNet, ki ima predvsem sporazumevalne aktivnosti (pisno, govorno, telesno in

nečloveško sporazumevanje) pod Komunikacija.[18] V SLONEST-sam smo se zaradi konkretnosti in pogojenosti s človeško aktivnostjo odločili uvrstiti sporazumevalna dejanja v podkategorijo Komunikacija pri semantičnem tipu AKTIVNOST.

21-SKUPINSKO

Semantični tip SKUPINSKO označuje leksikalne enote, ki se nanašajo na človeka (oz. ljudi), živali ali rastline, posamezno predmetnost, pa tudi tiste, ki se nanašajo na umetnostne, literarne (slogovne) smeri, obdobja, ideološka, religiozna gibanja, družbeno ureditev ali sistem.

V podkategoriji Človek in Človek-organizacija uvrščamo leksikalne enote, ki po principu metonimije zastopajo primarni pomen "ljudje v ustanovi" ali pa primarni pomen "ustanova/organizacija ali podjetje glede na svojo dejavnost". Slednje, primere tipa t. i. samostalniških metonimij (npr. *šola, cerkev, sindikat*), uvrščamo v ločeno podkategorijo Človek-organizacija.

Semantični tip SKUPINSKO zajema kategorije:

- Človek: *študentarija, moštvo, sosedstvo, četverica; kvartet;*
- Človek-organizacija: *ministrstvo, univerza, akademija, klinika;*
- Žival: *pleme, rod, jata, trop, ogrožena vrsta;*
- Rastlina: *goščava, gaj, deževni gozd, tipska vrsta;*
- Artefakt: *ladjevje, komplet kart, promet;*
- Gibanje-sistem: *kubizem, realizem, futurizem; socializem, boljševizem; budizem, islam;*
- Skupno: *par, skupina treh.*

Kljub temu da ta semantični tip najdemo kot samostojno kategorijo v mnogih ontologijah, smo se sprva odločili poskusiti uporabiti podkategorijo Skupinsko pri vsakem od obstoječih podtipov, vendar pa smo se po podrobni analizi in težavah z umeščanjem leksikalnih enot, ki se lahko nanašajo na več semantičnih tipov, raje odločili za samostojen semantični tip z več podkategorijami.

18 Več o kategoriji Komunikacija v WordNetu v razdelku 3.3.

3.3 Izbrani problemi

Pri zasnovi ontologije smo sledili ključnemu izhodiščnemu kriteriju: zagotoviti jasnost in konsistentnost krovne kategorizacije in povezljivost z vidika mednarodnih podatkovnih baz. Izogibali smo se postavitvi kategorizacije, ki bi dovoljevala isti koncept znotraj več podkategorij. Tipičen primer posledice tovrstne odločitve je opustitev krovne kategorije 'noun.tops', ki jo WordNet uporablja za zelo splošne semantične tipe, od katerih so mnogi poimenovanja drugih semantičnih tipov (*človek, rastlina, pojav* ipd.).

S konceptualnimi dilemami smo se soočili tudi na drugih ravneh ontologije. Veliko ontologij, na katere smo se opirali, izhodiščno deli semantične tipe na konkretne in abstraktne; četudi ta delitev v sami ontologiji pogosto ni eksplicirana, jo je mogoče izbrati na podlagi krovnih kategorij. Tako smo tudi pri snovanju SLONEST-sam sledili principu, da bi bilo mogoče koncept leksikalne enote že izhodiščno uvrstiti med konkretnega ali abstraktnega. To je pripeljalo do težav, saj ravno WordNet, ki nam je služil kot izhodišče, pri določenih kategorijah, kot sta npr. KOMUNIKACIJA in KOGNICIJA, meša abstraktne in konkretne koncepte. Naša rešitev je bila, da problematične koncepte raje umestimo pod druge semantične tipe, pri čemer jih po potrebi združujemo v podkategorijah, ki tudi prek svojih poimenovanj ohranjajo povezavo s komunikacijskimi in kognicijskimi lastnostmi.

Pri analizi leksikalnih enot v semantičnem tipu KOMUNIKACIJA smo tako našli leksikalne enote, ki so opredeljevale koncepte različnih oblik verbalnega (govornega, pisnega) in neverbalnega, pa tudi človeškega ali nečloveškega komuniciranja oz. sporazumevanja; leksikalne enote, ki se nanašajo na sistem izmenjave simbolov/oznak ali najrazličnejših informacij med informacijskim virom in sprejemnikom (IKT-sporazumevanje) ipd. Pogosti so bili mejni primeri, ki v abstraktnem pomenu označujejo zvrsti ali vrste besedila ter se nanašajo na različne vede in področja (discipline) (jezikoslovje, književnost, likovna umetnost, glasba ...), zato so bili primerni za uvrstitev v semantični tip KOGNICIJA (pod posamezno vedo ali področje), na konkretni ravni pa se nanašajo na dokument ali umetniško

kreacijo oz. umetniški izdelek in se jih lahko umesti v semantični tip ARTEFAKT, podkategorijo Umetniški izdelek (primeri tipa *tragikomedija, komedija, dramsko besedilo, filmski scenarij, besedilo opere*). Problematični so bili tudi koncepti, ki so izražali neko sporazumevalno človeško ali živalsko aktivnost (tip *pisanje, govorjenje: lajanje, meketanje*), saj je bila z vidika naše opredelitve krovne abstrakcije ustreznejša kategorija AKTIVNOST (podkategorija Komunikacija). Podobno smo v AKTIVNOST premestili koncepte, ki izražajo človeško govorno dejanje (rezultat dejanja), njihove semantično povezane konkretizirane različice, ki pomenijo dokument, pa v semantični tip ARTEFAKT, podkategorijo Dokumenti (tip *izjava, prijava, prošnja, sklep*). Po analizi vseh problematičnih mest in premeščanju leksikalnih enot smo ugotovili, da za KOMUNIKACIJO nimamo konkretnih predstavnikov in smo ta semantični tip opustili.

Precej podobno izkušnjo smo imeli s semantičnim tipom KOGNICIJA, kjer pa smo semantični tip v ontologiji obdržali, smo se pa odločili jasno razločiti med izključno miselno oz. kognitivnimi procesi in stanji, ki so zaradi večje abstraktnosti ostali v KOGNICIJA, ter aktivnostmi, katerih pomen je (lahko) podprt s konkretno (fizično) aktivnostjo človeka (*načrtovanje, projektiranje*), ali rezultati fizičnih in kognitivnih dejanj (tip *načrt, analiza, projekt*) in jih uvrščamo v semantični tip AKTIVNOST (v podkategorijo Kognicija).

Na medkategorialni ravni smo pri semantičnih tipih DOGODEK, POJAV in AKTIVNOST morali jasno opredeliti kriterije za umeščanje leksikalnih enot, saj se je v WordNetu pokazala precejšnja nedoslednost (gl. razdelek o POJAVU). Zaradi potrebe po sistematični ločitvi konceptov, ki so vezani na naravne (vremenske) pojave, od tistih, ki se nanašajo na človeški dogodek (tip: *simfonični koncert, nogometna tekma, obrtni sejem*), smo prve (torej WordNetove naravne dogodke) premestili v podkategorijo Naravni pojav pri POJAVU, druge (človeške dogodke) pa v Človeški dogodek pri semantičnem tipu AKTIVNOST. Pri tem smo izhajali iz osnovnih opredelitev in ključnega pogoja za pomensko določitev oz. umestitev leksikalnih enot v obe omenjeni kategoriji, ki je vezan na prisotnost (AKTIVNOST) ali odsotnost (POJAV) človeške aktivnosti oz. človeškega delovanja.

Vse ostale, ne naravne pojave, ki so pogojeni z aktivnostjo človeka, jih povzročajo živali ali naprave, pri čemer gre večinoma za zvočne pojave, prav tako uvrščamo v semantični tip AKTIVNOST, v posebno podkategorijo Zvočni pojav (*cviljenje zavore, zvonjenje telefona*).

Dileme na znotrajkategorialni ravni so se pojavile pri umeščanju leksikalnih enot v ustrezno podkategorijo znotraj enega semantičnega tipa. Izpostavili bi težave pri kategoriziranju kandidatov v okviru semantičnega tipa ČLOVEK, natančneje pri uvrščanju v podkategoriji Poklic in/ali Nosilec aktivnosti (ki obe pripadata hierarhično višji podkategoriji Aktivnost). Pri tej delitvi smo prvotno sledili LexicoNetu in estonski ontologiji, ki imata ločeni podkategoriji na isti ravni, a je težavnost kategorialnega opredeljevanja kandidatov opozorila na problematičnost takšne kategorizacije. V posameznih mejnih primerih, ki izkazujejo pomensko tendenco v obe smeri, torej se lahko nanašajo bodisi na poklicno bodisi na amatersko/ljubiteljsko aktivnost, namreč ni bilo jasno, v katero od omenjenih podkategorij jih uvrščamo (tip *karikaturist, nogometaš, kmet*). Znotraj krovne podkategorije Aktivnost je bila tako potrebna razmejitev konceptov oz. primerov, ki jih lahko uvrstimo samo v Poklic (tip *farmacevt*) ali samo v Nosilec aktivnosti (tip *interpret*), od primerov, ki izkazujejo tako poklicno (profesionalno) kot amatersko rabo (tip *plavalec* – "športnik; nekdo, ki se poklicno ukvarja s plavanjem" in "nekdo, ki plava"), in jih zato uvrščamo v podkategorijo Poklic-Nosilec aktivnosti.

4 Zaključek

V tem prispevku smo predstavili izdelavo slovenske ontologije semantičnih tipov za samostalnike (SLONEST-sam). Pri snovanju SLONEST-sam smo se oprli na obstoječe mednarodne ontologije, zlasti tiste, ki so v mednarodnem prostoru široko uporabljane. SLONEST--sam resda v določenih delih odstopa od ostalih ontologij, recimo določeni semantični tipi, ki so v drugih ontologijah na hierarhično višjih ravneh, so v SLONEST-sam na nižjih ravneh oz. predstavljajo (pod)kategorije, a vsa takšna odstopanja smo dokumentirali in posledično zagotovili povezljivost med ontologijami, kar je ponazorjeno

v Tabeli 2. Stremenje k takšni povezljivosti bo vsekakor dragoceno pri bodočih prizadevanjih povezovanja slovenskih jezikovnih podatkov s tujimi. Kot primer lahko omenimo aktivnosti v okviru projekta Evropske leksikografske infrastrukture (ELEXIS; Krek idr. 2018, 2019; Pedersen idr. 2018), kjer je med drugim predvideno (pol)avtomatsko medjezikovno povezovanje semantičnih podatkov, ki jih najdemo v slovarjih, tezavrih in podobnih leksikografskih virih.

Kot smo ugotovili, tudi na podlagi eksperimentov označevanja, je za ročno označevanje oz. potrjevanje semantičnih tipov bolje in dejansko nujno imeti dobro zasnovan in utemeljen širok nabor krovnih kategorij, ki že takoj na začetku dovolj jasno razmejujejo splošnejše koncepte in tako omogočajo vsaj osnovno kategorizacijo tudi v primeru dvomov o specifični podkategoriji. Ključno vlogo pri celotnem procesu so odigrale kolokacije kot prva raven kontekstualizacije posameznih pomenskih potencialov besed, saj smo prek njih lahko potrjevali koncepte določenih leksikalnih enot.

V teku je že delo na ontologijah semantičnih tipov za glagole in pridevnike, pri čemer je pomembno poudariti, da se bosta precej oprli na ontologijo semantičnih tipov za samostalnike. Smiselnost takšne navezanosti med ontologijami različnih besednih vrst dobro ponazarja shema semantičnih polj leksikalno-semantične mreže GermaNet,[19] kjer so razvidna tako prekrivanja med tipi besednih vrst kot tudi praznine, t. j. tipi, ki jih najdemo samo pri določenih besednih vrstah. Glavni vodili pri oblikovanju nadaljnjih ontologij SLONEST pa bosta vsekakor ohraniti notranjo povezljivost med ontologijami za različne besedne vrste v slovenščini in zagotoviti nadaljnjo povezljivost z mednarodnimi ontologijami.

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*

19 https://uni-tuebingen.de/en/faculties/faculty-of-humanities/departments/modern-languages/department-of-linguistics/chairs/general-and-computational-linguistics/ressources/lexica/germanet/description/semantic-fields/

tehnologije za slovenski jezik) je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna.

Reference

Bartsch, S. (2004): *Structural and Functional Properties of Collocations in English: A Corpus Study of Lexical and Pragmatic Constraints on Lexical Co-occurrence*. Tübingen: Gunter Narr.

Fellbaum, C. (1998): *WordNet: An Electronic Lexical Database.* Cambridge, MA: MIT Press.

Fišer, D. (2009): sloWNET – slovenski semantični leksikon. V M. Stabej (ur.): *Infrastruktura slovenščine in slovenistike. Obdobja 28*: 145–149. Ljubljana: Znanstvena založba Filozofske fakultete UL.

Fišer, D. (2015): *Semantic lexicon of Slovene sloWNet 3.1*. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1026.

Gantar, P. (2009): Leksikalna baza: vse, kar ste vedno želeli vedeti o jeziku. *Jezik in slovstvo,* 54 (3–4): 69–94. Ljubljana: Slavistično društvo Slovenije. Dostopno prek: http://www.dlib.si (30. 6. 2021).

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: https://www.clarin.si/repository/xmlui/handle/11356/1030 (30. 6. 2021).

Gantar P. (2015a): *Leksikografski opis slovenščine v digitalnem okolju*. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani. Dostopno prek https://www.dlib.si/details/URN:NBN:SI:DOC-C6OT60O0 (30. 6. 2021).

Gantar, P. (2015b): Leksikalna baza za slovenščino: komu, zakaj in kako (naprej)? *Jezikoslovni zapiski,* 17 (2): 77–92. Ljubljana: Inštitut za slovenski jezik Frana Ramovša ZRC SAZU. Dostopno prek: https://ojs.zrc-sazu.si/jz/article/view/2377 (30. 6. 2021).

Geyken, A. in Schrader, N. (2006): LexikoNet – a lexical database based on type and role hierarchies. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy.* European Language Resources Association (ELRA). Dostopno prek: http://www.lrec-conf.org/proceedings/lrec2006/pdf/812_pdf.pdf (30. 6. 2021).

Hanks, P. (2004): Corpus Pattern Analysis. V G. Williams in S. Vessier (ur.): *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004*: 87–97. Lorient: Universite de Bretagne-sud.

Hanks, P. in Pustejovsky, J. (2005): A Pattern Dictionary for Natural Language Processing. *Revue Francaise de linguistique appliquée*, 10 (2): 63–82.

Hanks, P. (2008): Mapping meaning onto use: a Pattern Dictionary of English Verbs. *AACL 2008, Utah.*

Kilgarriff, A., Rychlý, P., Smrz, P. in Tugwell, D. (2004): The Sketch Engine. V G. Williams in S. Vessier (ur.): *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004*: 105–115. Lorient: Universite de Bretagne-sud.

Koeva, S., Dimitrova, T., Stefanova, V. in Hristov, D. (2018): Mapping WordNet Concepts with CPA Ontology. *Proceedings of GWC 2018.* Dostopno prek: http://compling.hss.ntu.edu.sg/events/2018-gwc/pdfs/GWC2018_paper_50.pdf (30. 6. 2021).

Kosem, I., Krek, S., Gantar, P., Arhar Holdt, Š., Čibej, J. in Laskowski, C. (2018): 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. Dostopno prek: http://nl.ijs.si/jtdh18/JTDH-2018-Proceedings.pdf (30. 6. 2021).

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

Kosem, I., Pori, E., Gantar, P., Logar, N., Krek, S., Laskowski, C., Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Gorjanc, V., Klemenc, B. in Ljubešić, N. (2020): *Slovene ontology of semantic types for nouns SLONEST-noun 1.0*. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1428.

Krek, S., Declerck, T., McCrae, J. P. in Wissik, T. (2019): *Towards a Global Lexicographic Infrastructure* [presented at the Language Technology 4 All Conference]. doi: 10.5281/zenodo.3607274.

Krek, S., McCrae, J., Kosem, I., Wissik, T., Tiberius, C., Navigli, R. in Pedersen, B. (2018): European Lexicographic Infrastructure (ELEXIS). V J. Čibej idr. (ur.): *Proceedings of the XVIII EURALEX International Congress on Lexicography in Global Contexts (EURALEX 2018).* doi: 10.5281/zenodo.2599902.

Langemets, M. (2010): *Nimisõna süstemaatiline polüseemia eesti keeles ja selle esitus eesti keelevaras [Systematic Polysemy of Nouns in Estonian and its Lexicographic Treatment in Estonian Language Resources]*. Tallinn: Eesti Keele Sihtasutus.

Parrott, W. (2001): *Emotions in Social Psychology*. Key Readings in Social Psychology. Philadelphia: Psychology Press.

Pedersen, B. S., P., McCrae, J., Tiberius, C., Krek, S. (2018): ELEXIS – a European infrastructure fostering cooperation and information exchange among lexicographical research communities. V F. Bond, T. Kuribayashi, C. Fellbaum in P. Vossen (ur.): *Proceedings of the 9th Global WordNet Conference (GWC 2018), Global Wordnet Association, Singapore.* doi: 10.5281/zenodo.2599954.

Podobnik, A. (1985): Koliko kraljestev živega? *Proteus: ilustriran časopis za poljudno prirodoznanstvo*, 47 (9–10): 334–338.

Pori, E. in Kosem, I. (2018): V iskanju slovarsko relevantne kolokacije na primeru struktur s prislovi. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave,* 6 (2): 154–185.

Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R. in Scheffczyk, J. (2010): *FrameNet II: Extended Theory and Practice.* Dostopno prek https://akb89.github.io/myValencer/framenet_book.pdf (30. 6. 2021).

Shaver, P., Schwartz, J., Kirson, D. in O'Connor, C. (1987): Emotion knowledge: further exploration of a prototype approach. *Journal of Personality and Social Psychology,* 52 (6): 1061–86. doi: 10.1037/0022-3514.52.6.1061.

Snoj, J. idr. (ur.) (2016): *Pojmovnik sinonimnega slovarja.* Dostopno prek: https://fran.si/208/sinonimni-slovar (30. 6. 2021).

Stubbs, M. (2002): Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics,* 7 (2): 215–44.

Whittaker, R. H. (1969): New Concepts of Kingdoms of Organisms. *Science,* (163) 3863: 150–160. doi: 10.1126/science.163.3863.150.

Priloga

Tabela 2: Primerjava krovnih kategorij SLONEST ontologije z različnimi slovenskimi in mednarodnimi ontologijami (Opomba: kategorija v oklepaju pomeni delno ujemanje).

SLONEST	WordNet, sloWNet	Framenet	LexicoNet	CPA	Estonska ontologija	Pojmovnik SSSJ	SIMPLE-CLIPS
ČLOVEK	Human	• Sentient-Human • Animate_being-Sentient • Physical_object-Living_thing	concrete... • Living_beings-Hominids • Mythological_beings	Entity...Human	• Human • Representation	(Človek)	Entity...Human

TELO	Body	Physical_object-Body_part	concrete...
			• Physical_objects...
			  • Body_or_body_part
			  • Cell_or_organ_parts
				Part...Body	Bodypart	Človek-človeško_telo-telesni_organ_del	Constitutive-Part-Body_part

ŽIVAL	Animal	• Living_thing-Animate_being
		• Physical_object-Living_thing
			concrete...
			• Living_beings
			  • Animals
			  • Taxonomic_groups
			• Physical_objects...
			  • Animal_structure
			  • Body_or_body_part
				Entity...Animal	• Animal
					• Bodypart-animal
						(Žival)	Entity...Animal

SLONEST	WordNet, sloWNet	Framenet	LexicoNet	CPA	Estonska ontologija	Pojmovnik SSSJ	SIMPLE-CLIPS
RASTLINA	Plant	Plants	concrete… • Living_beings-Plants • Physical_objects…plant_part	• Entity…Plant • Part…Plant_Part	Plant	(Rastlina)	Entity…Vegetal_entity
MIKRO-ORGANIZEM	Animal		concrete… Living_beings-Microorganisms_and_viruses		Organism		Entity…Microorganism

GLIVA	Plant		concrete…
			• Living_beings
			  • Higher_mushroom
			  • Lichen
					Organism	(Rastlina-goba)	

HRANA	Food	Food	concrete…
			• Materials_and_substances…
			  • Food
			  • Animal-food
				• Entity…
				  • Food
				  • Entity…Beverage
				• Stuff…Beverage
					Food	• Snov
						  • Hrana
						  • Meso
							• Entity…
							  • Food
							  • (Substance)

SNOV	Substance	Physical_entity-Material	concrete…
			(Materials_and_substances)
				• Entity…
				  • Physical_Object-Stuff
				  • Particle
					Material/Substance	(Snov)	(Entity…Substance)

SLONEST	WordNet, sloWNet	Framenet	LexicoNet	CPA	Estonska ontologija	Pojmovnik SSSJ	SIMPLE-CLIPS
ARTEFAKT	• Artifact • Communication • Possession • Location	• Artifact-Structure • Physical_object-artifact	concrete... • (Physical_objects-Artifact) • (Rooms_and_places)	• Entity... • Artifact • Location	• Artefact • Place • Representation	• Predmet • Prostor	• Entity... • Artifact • Location-Building • Representation-information
PROSTOR	• Object • Location	• Physical_object-Location • Absolute_direction_orientation • Body_of_Water • Cardinal • Region	• concrete... (Rooms_and_places) • abstract... (Abstract_spaces)	• Entity... • Location-Natural_Landscape_Feature • Location-Area	• Place • Object	• Prostor-Geomorfološka_pojavnost • Predmet-nebesno_telo • Abstrakta-družbenoorganizacijska_danost	• (Entity... • Location • Physical-object)
OBLIKA	Shape	• (Attribute) • Ontological_type-Attribute	• concrete... Geometric_shapes • abstract... Geometric_shapes Abstract_spaces	Property-Visible_Feature-Shape		• Abstrakta • lastnost • lastnost_človeka	(Entity-property)

SLONEST	WordNet, sloWNet	Framenet	LexicoNet	CPA	Estonska ontologija	Pojmovnik SSSJ	SIMPLE-CLIPS
POJAV	• Phe-nom-enon • Event	• Event • State_of_af-fairs-Event	concrete… • Physical_object-Natural_thing-Sky_or_weather_phenomenon abstract… Event…Natural_event	• Entity…Energy • Eventuality…Process-Weath-er-Event	Phenom-enon	Abstrakta-narav-ni_pojavi	(Entity-Event-Phenomenon)
PROCES	Process		abstract… • (Methods_and_schemes) • (Activities_and_behav-iour)	Eventuality…Pro-cess		(Abstrakta-dejanje)	
MERA	Quantity	• Attribute-Quantity • Ontological_type-Attribute	abstract… • Numbers_and_measures • (Materials_and_sub-stances)	Entity…Numeri-cal_value	Representa-tion	• Abstrakta • količina • merske_enote	(Entity-Repre-sentation)
ČAS	Time	• Attribute-Duration • Relation-Time	abstract… Time_and_periods	• Entity… • Asset-Time_Period • Time_Period • Time_Point	Time	• Abstrakta • čas_kot_ome-jeno_trajanje • čas_kot_točka_v_času	Entity-abstract_entity-time
ČUSTVO	• Feeling • Motive	• Sensory_mo-dality • Pragmatic_function	abstract… Features_and_conditions… feelings	Entity…Psych	Feature	• (Abstrakta • lastnost_člo-veka • stanje_človeka)	• Entity… • Property • Event-Psychologi-cal_event

SLONEST	WordNet, sloWNet	Framenet	LexicoNet	CPA	Estonska ontologija	Pojmovnik SSSJ	SIMPLE-CLIPS

LASTNOST	• Attribute
	• Motive
		• Attribute
		• Relation-Social_relation
			abstract...
			(Features_and_conditions)
				Property	Feature	• Abstrakta
						  • lastnost_človeka
						  • stanje_človeka
							Entity-property

STANJE	• State
	• Property
	• Motive
	• Relation
		• State_of_affairs-State
		• Flexible_orientation
		• Ontological_type
			abstract...
			(Features_and_conditions)
			concrete...
			(Physical_objects...Lesions)
				• Eventuality...
				  • State_of_Affairs
				  • Process
					State	• (Abstrakta
						  • stanje
						  • stanje_človeka)
							• Entity...
							  • Event-State
							  • Organic-object

KOGNICIJA	Cognition	State_of_affairs-Content	abstract...
			• (Communication_means)
			• (Ideas_and_information)
			• (Domains_and_disciplines)
				Entity...Concept	• Abstract
					• Domain
					• Representation
						(Abstrakta_dejavnost)	• (Entity...
							  • Abstract_entity
							  • Representation
							  • Event-Psychological_Event)

SLONEST	WordNet, sloWNet	Framenet	LexicoNet	CPA	Estonska ontologija	Pojmovnik SSSJ	SIMPLE-CLIPS
KOMUNIKACIJA							
AKTIVNOST	• Act • Communication • Event • Process	• Event • Intentional_act • State_of_affairs-Event	abstract… • (Activities_and_behaviour) • (Methods_and_schemes) • (Domains_and_disciplines)	• Eventuality…Activity • Entity…Information_source	• Act • Event	• (Abstrakta • dejanje • dejavnost • dogodek)	• Entity… • Event-Act • Event-Change • Event-Cause_change
SKUPINSKO	Group	• Group • Ontological_type-Group	concrete… • (Living_beings-Hominids) • (Living_beings-Animals) • (Living_beings-Plants) • (Living_beings-Microorganisms_and_viruses) • (Physical_objects-Artifact) abstract… • Cultures_and_social_systems • Form_of_government	Group		• Človek-več_ljudi_kot_celota • Predmet-več_predmetov_kot_celota • Rastlina-več_rastlina_kot_celota • Žival • več_živali_kot_celota • značilna_skupina_živali	• Constitutive-group • (Entity-abstract_entity)

Kolokacije in časovni trendi

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

Jaka ČIBEJ
Filozofska fakulteta, Univerza v Ljubljani; Institut Jožef Stefan

The paper presents the results of a diachronic analysis of collocation use, conducted on the Gigafida 2.0 reference corpus of modern Slovene which contains texts from 1991 to 2018. The analysis focused on identifying new usage (neologisms) or increased usage of words, as well as on detecting different patterns in temporal trends of collocations. We extracted collocations in three different syntactic structures, adjective + noun, verb + noun in accusative, and noun + noun in genitive. Using different statistical measures we wanted to identify patterns in temporal trends of collocations, and describe their relevance for language description purposes, and collocation analysis purposes in general. These calculations, and sample datasets, are also part of the Orange workflow for observing collocation trends (ColTrend), which we uploaded to the CLARIN.SI repository. As large quantity is one of the problematic aspects of analysing collocations, we also looked into the potential of using temporal trends for the identification of corpus noise and lexicographically less relevant collocations. In the discussion section, we focus on the implications of our findings for lexicographic practice, both semantic analysis and the presentation of the information on temporal trends of collocations to the end users.

Keywords: temporal trends, collocation, diachronic analysis, lexicography, ColTrend

1 Uvod

Kolokacije imajo zelo pomembno vlogo v jezikovnem opisu, saj so kolokatorji pogosto uporabljeni kot izhodišče pri identifikaciji pomenov, poleg tega pa so ključni pri oblikovanju pomenskih opisov (Atkins in Rundell 2018). Kot izpostavljajo Gantar idr. (2020), kolokacijo opredeljujejo trije kriteriji: statistični, skladenjski in pomenski. Vsak od teh kriterijev je v večji ali manjši meri prisoten v številnih definicijah kolokacije, ki jih najdemo v literaturi (npr. Hausmann 1989; Church in Hanks 1990; Sinclair 1991; Fontenelle 1994; Herbst 1996; Moon 1998; Atkins in Rundell 2018).

Za pomenski opis je poleg zgoraj omenjenih kriterijev pomemben tudi diahroni vidik oz. časovna razpršenost kolokacij. Kolokatorje tako lahko uporabimo za detekcijo semantičnih sprememb v rabi besed (Geeraerts 1997), zlasti pri zaznavanju semantičnih neologizmov, tj. pri nastanku novih pomenov že obstoječih besed oz. pomenotvorju. Uspešnost implementacije takšnih postopkov so pokazali projekti, kot sta AVIATOR (Renouf 1993) in WebCorpLSE (Kehoe in Gee 2009; Renouf 2009).[1]

Enega od prvih poskusov izrabe kolokacij za zaznavo novih pomenov oz. semantičnih premikov v slovenščini so opravili Pollak idr. (2019), ki so analizirali in kategorizirali kolokacije, tipične za računalniško posredovano komunikacijo (družbena omrežja, forume, bloge ipd.). Glavni fokus raziskave je bil na novem besedišču, pri čemer so kolokacijske kandidate za analizo izluščili s primerjavo specializiranega korpusa in splošnega korpusa; diahroni vidik kolokacij torej ni bil upoštevan. Prepoznali so tri skupine semantično relevantnih kolokacijskih podatkov: kolokacije, katere del je leksikalna enota, ki v slovenskem jeziku še ni bila zabeležena; nove kolokacije obstoječih pomenov besed (ta skupina je bila največja); kolokacije, ki izkazujejo nove pomene obstoječih besed. Prednost prispevka je

1 Nedavni trendi spremljanja semantičnih sprememb v jeziku sicer posvečajo več pozornosti uporabi metod distribucijske semantike (Sagi idr. 2011; Cook idr. 2014; Gulordava in Baroni 2011), pri čemer je leksikološko usmerjen predvsem pristop besednih jezikovnih modelov, ki ga pri svojih raziskavah uporabljajo Heylen idr. (2015). Eno prvo tovrstnih raziskav na slovenskem jeziku sta opravila Fišer in Ljubešić (2016), ki sta preučevala pomenske premike v slovenskih tvitih.

izčrpna analiza izluščenih kolokacij, ena od slabosti pa, da je skladenjski vidik kolokacij precej omejen oz. so analizirani zgolj bigrami samostalniških lem (upoštevana je bila namreč zgolj pozicija takoj pred ali za lemo).

Bigrame so za detekcijo novih pomenov obstoječih besed v danščini uporabili tudi Nimb idr. (2020), ki pa so pri luščenju upoštevali tudi diahrone lastnosti kolokacij oz. bigramov (pojavitve med 2005 in 2018). Pri analizi so se tako osredotočili na bigrame, ki se v 512-miljonskem korpusu niso pojavljali v prvih treh letih in so imeli vsaj 20 pojavitev v naslednjih 11 letih. Rezultati označevanja dveh leksikografov so z vidika semantične relevantnosti podatkov podobni tistim od Pollak idr. (2019), je pa dodana vrednost pri Nimb idr. (2020) prenos ugotovitev v leksikografsko prakso oz. posodobitev slovarja. Kot namreč navajajo Nimb idr. (2020: 122), so rezultati analiz pripeljali ne samo do dodajanja novih pomenov, ustaljenih besednih zvez ali kolokacij, temveč tudi do sprememb obstoječih razlag ali dodajanja zgledov rabe bigramov.

Omenjeni raziskavi nakazujeta samo del potenciala, ki ga imajo lahko podatki o kolokacijah in njihovi rabi skozi čas za namene časovnega opisa. Poudarek je namreč na prepoznavi novejšega besedja oz. besedja, katerega raba narašča. A kot pravi Renouf (2013), kolokacije nam lahko pomagajo spremljati življenje besed, torej jezikovne pojave, kot so rojstvo, povečano rabo (v obliki produktivnosti in kreativnosti) in smrt besed, pa tudi njihovo morebitno oživitev (gl. Žele 2009 za diskusijo o tovrstnih pojavih v slovenskem jeziku; tudi Urbančič 1987).

V okviru nacionalnega projekta KOLOS (*Kolokacije kot temelj jezikovnega opisa: semantični in časovni vidiki*; J6-8255) je bil med drugim predviden tudi razvoj statističnih metod za zaznavanje semantičnih trendov besed v slovenščini s pomočjo kolokacij. V pričujočem prispevku se osredotočamo na ta vidik, torej na časovne trende kolokacij v slovenščini, pri čemer so nas zanimali vsi vidiki diahrone rabe kolokacij, ne samo novejša ali naraščajoča raba. Z različnimi statističnimi merami smo želeli prepoznati vzorce v časovnih trendih in opredeliti njihovo relevantnost za jezikovni opis in analizo

kolokacij nasploh. Ker je eden od pomembnih problemov pri analizi kolokacij tudi njihova količina, nas je zanimalo tudi, ali so vzorci časovnih trendov lahko koristni tudi pri prepoznavi korpusnega šuma in slovarsko nerelevantnih kolokacij. Diskusijski del prispevka je namenjen razmislekom uporabnosti rezultatov za leksikografsko prakso, tako semantično analizo kot direktno predstavitev podatkov o časovnih trendih uporabnikom.

2 Metodologija

Naš eksperiment je bil sestavljen iz treh delov: priprave kolokacijskih podatkov z informacijami o časovni razpršenosti, izbire in izračuna statistik za opazovanje časovnih trendov ter analize.

2.1 Priprava podatkov

Kolokacijske podatke za analizo smo izluščili iz korpusa Gigafida 2.0 (Krek idr. 2019; Krek idr. 2020), ki vsebuje besedila, nastala med leti 1990 in 2018, in je tako primeren za diahrone analize sodobnega slovenskega jezika. V korpusu je opazno manjša količina besed v letih 1990–1995 in 2011, na kar smo bili pozorni pri pripravi podatkov, statistični obdelavi in tolmačenju rezultatov.

Za luščenje kolokacijskih kandidatov smo uporabili najsodobnejšo metodo luščenja kolokacij na skladenjsko razčlenjenih korpusnih podatkih (Krek idr., v tisku).[2] Nova metoda odpravlja kar nekaj težav, ki smo jih zaznali pri evalvaciji kolokacijskih podatkov, izluščenih iz oblikoskladenjsko označenega korpusa (gl. Pori in Kosem 2021). V prvem koraku smo izluščili vse kolokacijske kandidate z vsaj 15 pojavitvami za tri skladenjske strukture: pridevnik + samostalnik (p0-s0), glagol + samostalnik v tožilniku (gg-s4) in samostalnik +

2 V času analize je bila metoda že v fazi razvoja, zato smo uporabili podatke iz prve verzije skripte za luščenje. Kolokacijski podatki, ki so objavljeni v repozitoriju CLARIN.SI, pa so bili izluščeni z drugo, dopolnjeno verzijo skripte. Med prvo in drugo verzijo skripte za luščenje skladenjskih struktur, uporabljenih v naši analizi, ni bilo bistvenih razlik; edina opaznejša izjema je bila ločitev povratnih glagolov v strukturi glagol + samostalnik v tožilniku (gg-s4) v ločeno strukturo glagol + *si/se* + samostalnik v tožilniku (gg-zp-s4). Naši podatki tako v omenjeni strukturi vsebujejo združene podatke obeh omenjenih struktur.

samostalnik v rodilniku (s0-s2).[3] Za vsakega kolokacijskega kandidata smo v ločenih stolpcih navedli identifikacijsko številko kolokacije, prvo lemo, drugo lemo, najpogostejšo obliko kolokacije,[4] skupno pogostost in število različnih morfosintaktičnih oblik, v katerih se je kolokacija pojavljala. Tabela 1 prikazuje deleže kolokacijskih kandidatov po frekvenčnih rangih. Kot lahko vidimo, je daleč največ kolokacijskih kandidatov precej redkih. Čeprav bi z zvišanjem praga pogostosti precej zmanjšali količino podatkov, smo želeli vključiti tudi redkejše kolokacije, predvsem zaradi analiz semantičnih neologizmov.

Tabela 1: Deleži kolokacijskih kandidatov po frekvenčnih rangih.

Pogostost v korpusu	p0-s0	gg-s4	s0-s2
>100.000	10 (<0,01 %)	0 (0,00 %)	2 (<0,01 %)
99.999–10.000	589 (0,11 %)	60 (0,03 %)	125 (0,035 %)
9.999–1.000	11.888 (2,28 %)	2282 (1,15 %)	4378 (1,22 %)
999–100	98.835 (18,95 %)	30.575 (15,41 %)	55.380 (15,39 %)
99–15	410.252 (78,66 %)	165.485 (83,41 %)	300.061 (83,36 %)
Skupaj	**521.574**	**198.402**	**359.946**

V drugem koraku priprave podatkov smo za vsakega kolokacijskega kandidata iz korpusa pridobili podatek o pogostosti po letih, pri čemer sta bila za vsako leto pridobljena podatka o absolutni pogostosti in relativni pogostosti (f_R, tj. pogostosti na milijon besed). Za vsako strukturo se je pripravila ločena datoteka v formatu .CSV (gl. primer v Tabeli 2; zaradi velikega števila stolpcev je prikazan samo del vrstice).

3 Za oznake uporabljamo kratko kombinacijo upoštevanih morfosintaktičnih kategorij in lastnosti po sistemu MTE/JOS (http://nl.ijs.si/jos/msd/html-sl/josMSD-sl.html).

4 Pri najpogostejši obliki je šlo predvsem za zapis (mala/velika začetnica) in število. Za število je bila določena meja 50 % – če je bilo torej dvojinskih ali množinskih pojavitev 50 % ali več od vseh pojavitev kolokacije, se je zapisala množinska oblika kolokacije ali enega od njenih elementov, drugače pa (privzeta) edninska.

Tabela 2: Primer izpisa relativnih pogostosti po letih za kolokacijo *potrebovati soglasje*.

ID	Lema 1	Lema 2	Kolokacija	Pogostost	Oblike	f_R (1990)	...	f_R (2018)
58803	potrebovati	soglasje	potrebovati soglasje	850	27	0	...	0,873

Izluščeni podatki pred analizo niso bili prečiščeni oz. pregledani z vidika njihove relevantnosti za slovarske priročnike, vsebovali pa so tudi korpusni šum oz. napake pri luščenju. Razen dejstva, da bi bilo takšno pregledovanje podatkov pred analizo zelo zamudno, smo dejansko na vseh podatkih želeli preveriti, ali so statistike za analizo diahronih kolokacijskih podatkov lahko koristne tudi za druge namene, npr. čiščenje slabih podatkov ali prepoznavo slovarsko nerelevantnih kolokacij.

2.2 Statistična obdelava

Analizo časovnih trendov rabe kolokacij smo izvedli s pomočjo programske opreme Orange Data Mining,[5] ki omogoča obdelavo in vizualizacijo tabelaričnih podatkov s pomočjo metod podatkovnega rudarjenja in strojnega učenja. V programu je mogoče izdelovati delotoke, ki po korakih obdelajo podatke, jih filtrirajo, razvrščajo in dodatno procesirajo, vsako obdelovalno verigo pa je nato mogoče ponoviti tudi na novih podatkih (npr. isti delotok, ki smo ga pripravili za pregled kolokacij v določenem obdobju, lahko uporabimo tudi za obdelavo podatkov, ki jih pridobimo pozneje v naslednjem obdobju).

Trende v rabi kolokacij smo opazovali s pomočjo štirih statističnih mer: naklon linearne regresije, koeficient določenosti, razmerje med maksimalno in povprečno relativno pogostostjo ter količnik nedavne rasti. Podrobneje jih predstavljamo v nadaljevanju.

2.2.1 Naklon linearne regresije

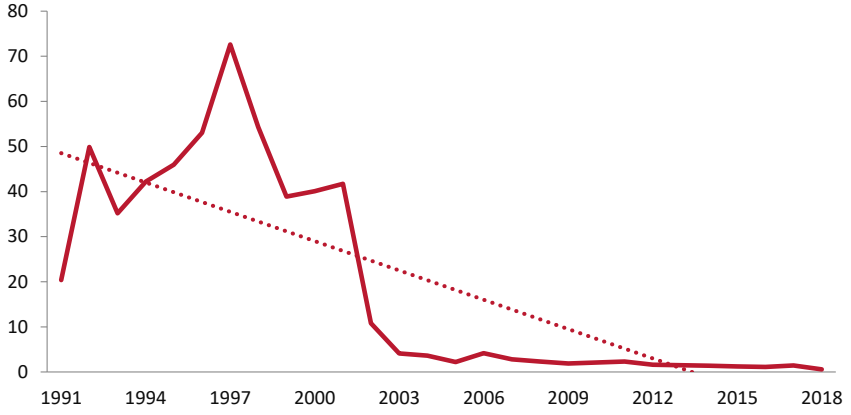
Za vsako kolokacijo smo vzeli nabor njenih relativnih pogostosti iz korpusa Gigafida 2.0 po letih in na podatkih zmodelirali linearno

5 https://orange.biolab.si/

regresijo ter izračunali njen naklon (*k*). Če je naklon negativen, to nakazuje, da se pogostost rabe kolokacije skozi leta zmanjšuje, pozitiven naklon pa pomeni, da je kolokacija v korpusu vedno pogostejša. Pomembna je tudi absolutna vrednost naklona – večja absolutna vrednost namreč nakazuje, da so spremembe v rabi (naraščanje ali padanje) izrazitejše.

Primer kolokacijskega kandidata s pozitivnim naklonom je *spletna stran* (k = 6,83), z negativnim naklonom pa *nemška marka*

Slika 1: Relativna pogostost kolokacije *spletna stran* po letih v korpusu Gigafida 2.0.



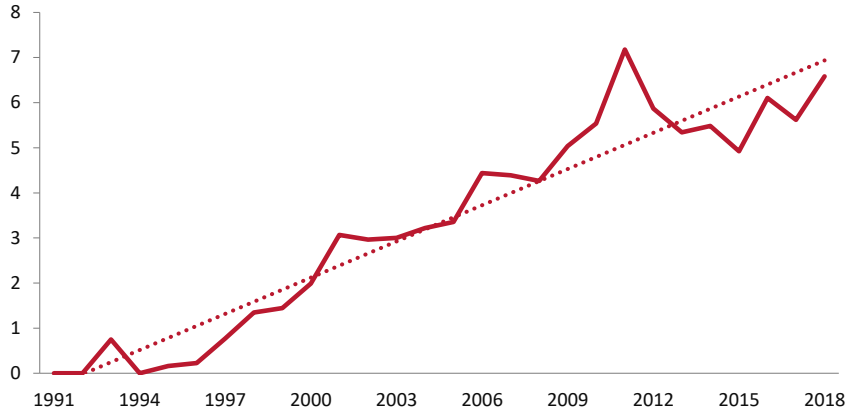
Slika 2: Relativna pogostost kolokacije *nemška marka* po letih v korpusu Gigafida 2.0.

(k = –2,17). Kot prikazujeta Sliki 1 in 2, relativna pogostost kolokacije *spletna stran* z vse pomembnejšo vlogo svetovnega spleta precej enakomerno narašča, pri kolokaciji *nemška marka* pa začne upadati po letu 2002, ko je v Nemčiji prišlo do zamenjave valute z evrom.

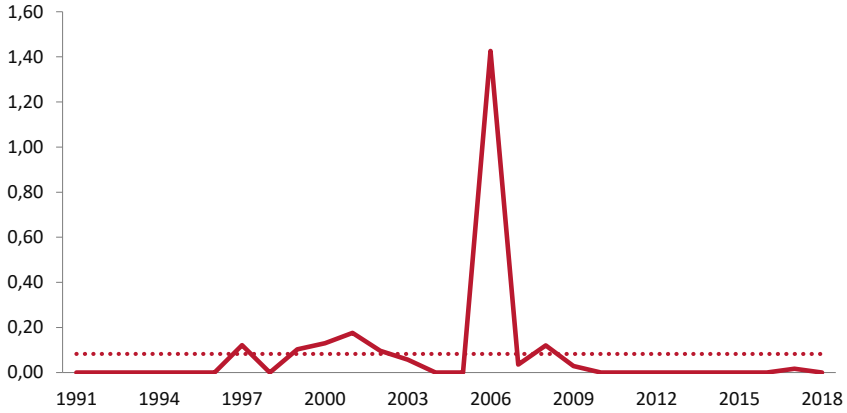
2.2.2 Koeficient določenosti

Izračunali smo tudi koeficient določenosti (R^2), ki meri, kolikšen del skupne variacije podatkov je pojasnjen z linearno regresijo oziroma, bolj poenostavljeno, kako dobro se model linearne regresije prilega vhodnim podatkom. Izkazuje vrednosti med 0 in 1. Ob veliki razpršenosti podatkov (npr. pogosti skoki in padci pogostosti skozi leta) je koeficient določenosti manjši, višji pa je pri bolj konsistentnih podatkih, ki jasneje prikazujejo naraščanje ali padanje (npr. če pogostosti po letih naraščajo ali padajo enakomerno).

Visok koeficient določenosti ima denimo kolokacija *prestižna nagrada* (R^2 = 0,92): iz Slike 3 je razvidno, da se premica linearne regresije dobro prilega letnim relativnim pogostostim. Nasprotno pa ima nizek koeficient določenosti kolokacija *ptujsko igrišče* (R^2 = 3,40 × 10^{-8}): Slika 4 kaže, da gre za po letih neenakomerno razporejeno kolokacijo, ki močno odstopa od modela linearne regresije predvsem zaradi izrazitega vrha v letu 2006 (večina



Slika 3: Relativna pogostost kolokacije *prestižna nagrada* po letih v korpusu Gigafida 2.0.



Slika 4: Relativna pogostost kolokacije *ptujsko igrišče* po letih v korpusu Gigafida 2.0.

konkordanc je iz revij Golf Slovenija in Štajerski tednik iz tega leta in omenjajo ptujsko igrišče za golf).

2.2.3 Razmerje med maksimalno in povprečno relativno pogostostjo

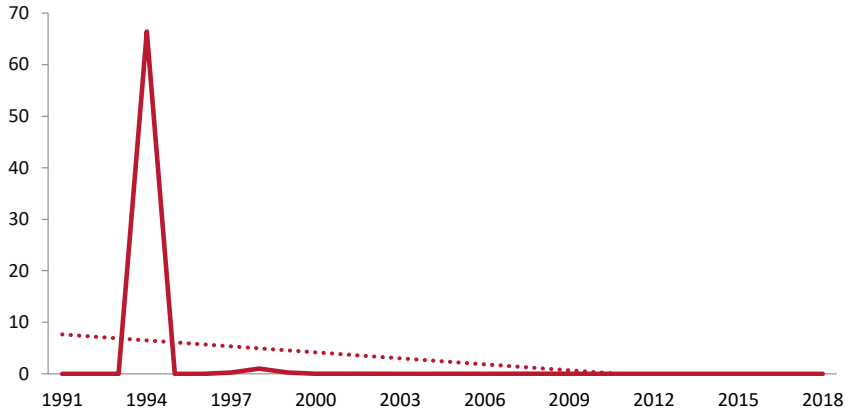
Tretja mera, ki smo jo uporabili za spremljanje trendov rabe kolokacij, je razmerje med maksimalno in povprečno relativno pogostostjo (*m*). Izračunamo ga po spodnji formuli, pri čemer je f_{rmax} najvišja letna relativna pogostost dane kolokacije, f_{ri} letna relativna pogostost, i_0 začetno leto in *n* število let, ki jih opazujemo:

$$m = \frac{f_{r_{max}} + 0{,}1}{\frac{1}{n}\sum_{i_0}^{n} f_{r_i} + 0{,}1}$$

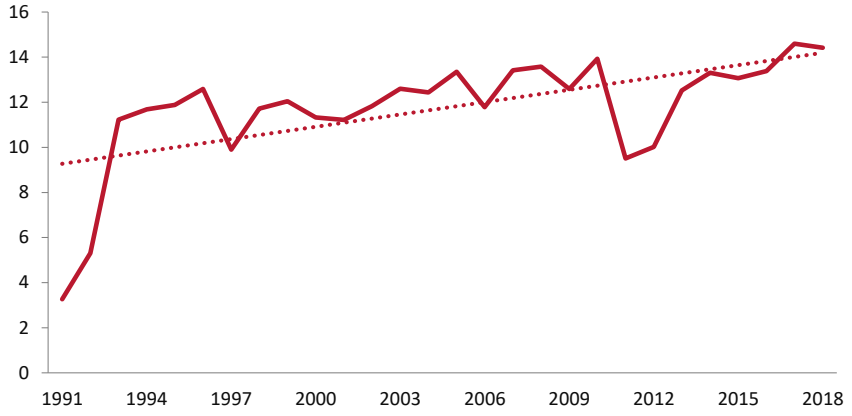
Če je pogostost rabe kolokacije v jeziku skozi čas povsem konstantna, je vrednost *m* enaka 1 (maksimalna in povprečna relativna pogostost sta v tem primeru enaki). Višja kot je maksimalna relativna pogostost kolokacije in bolj kot odstopa od povprečja, višja je vrednost *m*. Višja vrednost m torej nakazuje kolokacijske kandidate z zelo izrazito in nenadno spremembo v pogostosti v primerjavi

s povprečno relativno pogostostjo. Primer tovrstne kolokacije je *predrepna plavut* (m = 26,27): iz Slike 5 je razvidno, da ima kolokacija v letu 1994 zelo izrazit vrh, ki močno odstopa od siceršnjega povprečja. Treba pa je poudariti, da je večina konkordanc iz enega samega vira (*Velika knjiga o ribolovu*), kar nakazuje, da gre bolj verjetno za področno specifično in ne nujno za časovno specifično kolokacijo.

Nizko vrednost *m* opazimo npr. pri kolokaciji *pomemben del* (m = 1,24), pri katerem maksimalna relativna pogostost ne odstopa



Slika 5: Relativna pogostost kolokacije *predrepna plavut* po letih v korpusu Gigafida 2.0.

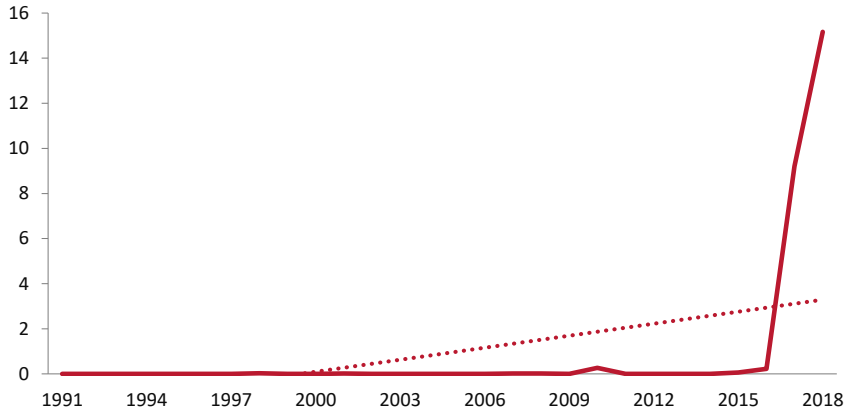


Slika 6: Relativna pogostost kolokacije *pomemben del* po letih v korpusu Gigafida 2.0.

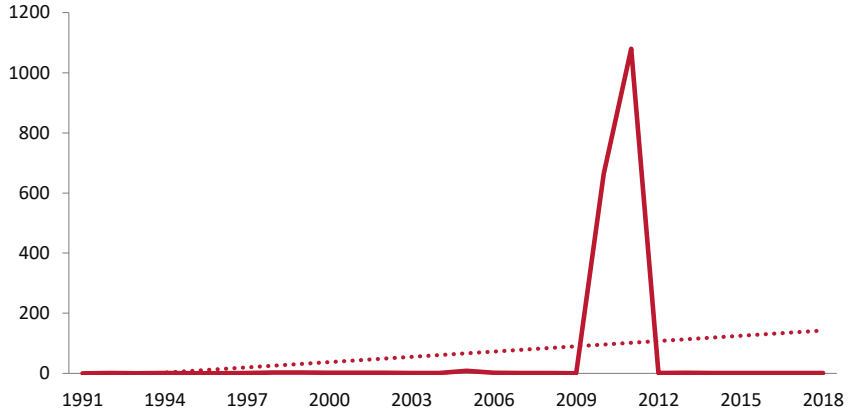
tako izrazito od povprečja, saj so letne relativne pogostosti medsebojno relativno primerljive (Slika 6).

2.2.4 Količnik nedavne rasti

Izračunali smo tudi količnik nedavne rasti (*t*), ki nakazuje, koliko je relativna pogostost kolokacije narasla v zadnjih treh opazovanih letih v primerjavi s povprečno relativno pogostostjo vseh ostalih opazovanih let. Pri tem ima največjo težo zadnje leto, manjši poudarek



Slika 7: Relativna pogostost kolokacije *arbitražna razsodba* po letih v korpusu Gigafida 2.0.



Slika 8: Relativna pogostost kolokacije *tožena stranka* po letih v korpusu Gigafida 2.0.

pa je na predzadnjem letu in letu pred tem. Količnik nedavne rasti smo izračunali po naslednji formuli:

$$t = \frac{f_{r_n} + 0{,}5 \times f_{r_{n-1}} + 0{,}25 \times f_{r_{n-2}} + 0{,}5}{1{,}75 \times \sum_{i_0}^{n-3} f_{r_i} + 0{,}5}$$

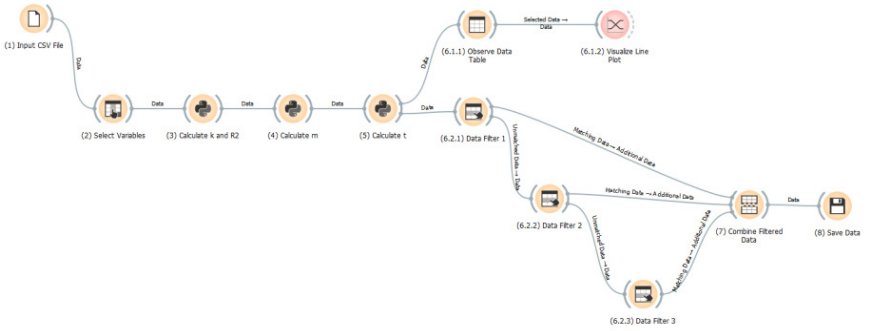
Višje kot so relativne pogostosti kolokacije v zadnjih treh opazovanih letih in manjše kot so njene relativne pogostosti v predhodnih opazovanih letih, višja je vrednost količnika *t*. Višji količnik nedavne rasti torej nakazuje, da raba kolokacije v zadnjih letih strmo narašča. Primer kolokacije z visokim količnikom nedavne rasti je npr. *arbitražna razsodba* (t = 38,46), ki ji je relativna pogostost zaradi razreševanja obmejnega vprašanja med Slovenijo in Hrvaško v zadnjih dveh opazovanih letih v korpusu v primerjavi s prejšnjimi leti skokovito narasla (Slika 7). Nizek količnik nedavne rasti ima npr. kolokacija *tožena stranka* (t = 0,023; gl. Sliko 8), ki v zadnjih letih ne izkazuje nobene rasti, vrh v korpusu pa ima med letoma 2009 in 2012 (tudi v tem primeru gre najbrž za področno specifično kolokacijo, saj pregled konkordanc razkrije, da je večina primerov iz pravnih besedil na spletu, npr. s spletne strani Vrhovnega sodišča Republike Slovenije).

2.3 Prototip delotoka za spremljanje kolokacijskih trendov

Za obdelavo podatkov v orodju Orange Data Mining smo pripravili delotok COLTREND (Slika 9), ki je na voljo tudi v repozitoriju CLARIN.SI in predstavlja neke vrste prototip orodja za spremljanje kolokacijskih trendov. Delotok sestavlja več ločenih skript v programskem jeziku Python, ki vhodne podatke, pripravljene po metodi, opisani v razdelku 2.1, sprocesirajo in opremijo z izračuni vsake od štirih statističnih mer, predstavljenih v razdelku 2.2.

Po opravljenih izračunih se v delotoku avtomatično pripravita dve podatkovni množici, ki se ju lahko izvozi in/ali analizira. Prva podatkovna množica so vsi kolokacijski podatki, opremljeni z dodanimi statističnimi izračuni. Takšna podatkovna množica je ustreznejša za proučevanje kolokacij specifičnih iztočnic, kjer nas zanimajo vse

kolokacije, tudi tiste, ki so z vidika časovnih trendov manj relevantne. Zgolj z vidika časovnih trendov relevantnih kolokacij pa je takšna podatkovna množica prevelika, saj je analiza preveč zamudna; s tem ne mislimo samo na ročno pregledovanje, temveč tudi na dejstvo, da je zaradi velike podatkovne množice nadaljnja obdelava podatkov lahko procesno in časovno potratna. Zato smo delotoku dodali izdelavo še druge podatkovne podmnožice, ki je pripravljena na osnovi vnaprej določenih parametrov statističnih mer. Pragove parametrov je mogoče nastaviti glede na želeni končni nabor kolokacijskih kandidatov z upoštevanjem kapacitet in sredstev, ki so namenjena ročnemu pregledu.



Slika 9: Delotok COLTREND v okolju Orange.

3 Analiza

V tem prispevku se podrobneje osredotočamo na analizo kolokacijske strukture pridevnik + samostalnik (p0-s0) (npr. *rezervni sklad*, *spletna stran*), a je ob tem treba poudariti, da je metodo mogoče na enak način uporabiti ne glede na kolokacijsko strukturo. V razdelku 3.1 najprej predstavimo kvantitativno analizo vseh tovrstnih kandidatov, izluščenih iz Gigafide 2.0, ter manjšega in za analizo relevantnejšega vzorca kandidatov, ki izpolnjujejo vzorčne kriterije. V razdelku 3.2 predstavimo opazovanje trendov rabe kolokacijskih kandidatov na nivoju posamezne iztočnice.

3.1 Celostni pogled na izluščene podatke

Iz Gigafide 2.0 je bilo znotraj kolokacijske strukture p0-s0 izluščenih skupno 521.574 kolokacijskih kandidatov. Od teh jih je 296.819 (57 %) imelo pozitiven naklon, 224.755 (43 %) pa negativnega. Pri nekoliko več kandidatih se torej kaže trend naraščajoče rabe, a je pri tem treba upoštevati tudi dejstvo, da Gigafida 2.0 vsebuje mnogo več besedil iz obdobja po letu 2000 kot iz desetletja prej oz. nasploh večjo količino besedil iz poznejših opazovanih let, zaradi česar linearna regresija morda pri več kandidatih zazna naraščanje.

Naklon je sicer pri večini izluščenih kandidatov zelo nizek, kar nakazuje, da je trend naraščanja ali padanja pri njih minimalen. Kot prikazuje Tabela 3, je pri polovici kandidatov naklon znotraj intervala $-0{,}001 < k < 0{,}001$. Ob takem naklonu se relativna pogostost kolokacije v 28 letih, kolikor jih zajema korpus Gigafida 2.0, v povprečju spremeni za manj kot 0,03 pojavitve na milijon besed, kar je zanemarljiva sprememba. Kolokacijski kandidati z zelo nizkim naklonom lahko sicer spadajo v več različnih scenarijev: (a) kolokacijski kandidati so morda z vidika diahrone analize relevantni, a je zanje v korpusu premalo podatkov, da bi lahko zanesljivo opazovali trend njihove rabe; (b) kolokacijski kandidati so relevantni, a je njihova pogostost skozi vsa leta približno enaka; (c) kolokacijski kandidati niso relevantni, gre za redke pojavitve in šum pri strojnem luščenju.

Tabela 3: Naklon pri kolokacijskih kandidatih z naraščajočim in padajočim trendom v korpusu Gigafida 2.0.

Skupina	Povprečje	Mediana	Minimum	Maksimum	Standardni odklon
Kandidati s pozitivnim naklonom	0,006	0,001	$5{,}697 \times 10^{-21}$	10,636	0,051
Kandidati z negativnim naklonom	−0,007	−0,001	−6,566	$-1{,}044 \times 10^{-21}$	0,040

Iz nabora vseh kandidatov je treba torej izločiti tiste, ki z večjo verjetnostjo spadajo v kategorijo (c). Pri tem se lahko poleg

njihove absolutne pogostosti zanašamo tudi na koeficient določenosti (Tabela 4). Pri večini kandidatov je tudi ta zelo nizek: pri polovici je celo manjši od 0,05, kar nakazuje, da so pojavitve kolokacijskega kandidata v korpusu bodisi redke in sporadične bodisi gre za rabo z zelo izrazitim in kratkotrajnim vrhom ter morebitnim padcem.

Tabela 4: Koeficient določenosti pri kolokacijskih kandidatih z naraščajočim in padajočim trendom v korpusu Gigafida 2.0.

Skupina	Povprečje	Mediana	Minimum	Maksimum	Standardni odklon
Kandidati s pozitivnim naklonom	0,124	0,066	$4{,}792 \times 10^{-36}$	0,923	0,145
Kandidati z negativnim naklonom	0,068	0,041	$1{,}607 \times 10^{-35}$	0,882	0,083
Vsi kandidati	0,010	0,052	$4{,}792 \times 10^{-36}$	0,923	0,125

Količnik *m* je pri polovici izluščenih kandidatov višji od 2 (Tabela 5), kar nakazuje, da moramo v primeru, da iščemo kolokacije z zelo izrazitim vrhom (ki so morda časovno omejene ali področno specifične, kot smo lahko videli v nekaterih primerih v razdelku 2.2), proučiti predvsem kandidate z večjim *m*, ob manjšem *m* pa lahko najdemo splošnejše kandidate z bolj enakomerno razporejeno pojavnostjo v korpusu.

Tabela 5: Razmerje med maksimalno in povprečno relativno pogostostjo pri vseh izluščenih kolokacijskih kandidatih iz korpusa Gigafida 2.0.

Skupina	Povprečje	Mediana	Minimum	Maksimum	Standardni odklon
Vsi kandidati	2,684	2,168	1,101	27,176	1,640

Na podoben način lahko z upoštevanjem količnika nedavne rasti *t* identificiramo tiste kolokacijske kandidate, ki jim je relativna pogostost v zadnjih letih poskočila: v povprečju je vrednost *t* pri vseh izluščenih kandidatih okrog 1 (Tabela 6), kar nakazuje, da v zadnjih

treh letih pri njih ni prišlo do pretirane spremembe v relativni pogostosti. Pri kandidatih z vrednostjo 2 bi npr. pomenilo, da se je njihova relativna pogostost v zadnjih treh letih približno dvakrat povečala glede na povprečje predhodnih let.

Tabela 6: Količnik nedavne rasti (t) pri vseh izluščenih kolokacijskih kandidatih.

Skupina	Povprečje	Mediana	Minimum	Maksimum	Standardni odklon
Vsi kandidati	1,001	0,967	0,022	38,456	0,305

Pri ostalih dveh strukturah se razporeditev vrednosti statističnih mer nekoliko razlikuje: pri strukturi s0-s2 je namreč od 359.946 izluščenih kolokacijskih kandidatov 222.638 (62 %) s pozitivnim naklonom in 137.308 (38 %) z negativnim, podobno tudi pri strukturi gg-s4, kjer je od 198.402 kolokacijskih kandidatov 122.825 (62 %) takšnih s pozitivnim naklonom in 75.577 (38 %) z negativnim.

Manjši in bolj obvladljiv vzorec izluščenih kandidatov lahko torej dobimo, če smiselno nastavimo parametre in filtriramo kandidate, ki so po zgornjih kriterijih za naše namene manj ustrezni. V primeru našega nabora smo tako dobili vzorec 43.562 kandidatov (8 % celotnega nabora iz strukture p0-s0), potem ko smo upoštevali tiste, ki ustrezajo naslednji verigi kriterijev:

$R^2 > 0{,}25$ in $|k| > 0{,}02 \rightarrow$ 13.696 kandidatov
ali

$m > 6{,}00 \rightarrow$ 21.168 kandidatov

ali

$t > 1{,}50 \rightarrow$ 8.698 kandidatov

S tem smo npr. odstranili kandidate, ki izkazujejo zanemarljivo naraščanje in padanje (pri k = 0,02 se denimo v 30 letih relativna

pogostost spremeni le za 0,6 pojavitve na milijon) oz. ki ne izkazujejo zelo izrazitega vrha (pri m = 2,00 je npr. najvišja relativna pogostost le dvakrat večja od povprečne) ali porasta v zadnjem času (pri t = 1,50 je npr. pogostost v zadnjih treh letih približno 50 % večja v primerjavi s predhodnimi leti). Opozoriti je treba, da je treba parametre izbrati glede na razporeditev vrednosti znotraj določene strukture – če npr. naivno upoštevamo parameter t > 1,50 pri strukturi s0-s2, dobimo v vzorcu le 9.674 kandidatov, ker znaša le 3 % celotnega nabora: povprečna vrednost količnika *t* je v primeru te strukture namreč 1,01 s standardnim odklonom 0,33, kar pomeni, da je prag z vrednostjo 1,50 nastavljen nekoliko previsoko. Pragovi torej niso konstantni, temveč določeni glede na razporeditev vrednosti in glede na želeni obseg vzorca.

Ob kvalitativnem pregledu tako dobljenega vzorca lahko kolokacije v grobem strnemo v naslednje kategorije:

Kolokacije z nedavnim izrazitim porastom (primeri so navedeni v Tabeli 7) so tiste, ki jim raba glede na izračune trendov v zadnjem času narašča (zanje sta značilna visoka količnika *m* in *t*). V mnogih primerih gre za kolokacije, ki so vezane na določen aktualen dogodek (*arbitražna razsodba*, *Panamski dokumenti*, *britanski*

Tabela 7: Primeri kolokacij z nedavnim izrazitim porastom v korpusu Gigafida 2.0.

Kolokacija	Absolutna pogostost	k	R^2	m	t
iranski sporazum	407	0,069	0,144	18,126	17,463
Panamski dokumenti	764	0,092	0,137	17,614	10,881
Šarčeva vlada	158	0,029	0,103	16,785	8,882
jedrski sporazum	1088	0,159	0,200	16,226	27,243
britanski referendum	393	0,045	0,124	15,508	4,943
arbitražna razsodba	1175	0,177	0,197	15,379	38,456
begunska kriza	6039	0,718	0,179	14,550	4,002
britanska premierka	1435	0,172	0,265	10,049	19,648
izsiljevalski virusi	380	0,046	0,250	9,892	5,690
avtonomna vozila	565	0,077	0,334	8,339	11,495

referendum, *begunska kriza*, *iranski sporazum*) oz. na konkretne javne osebe (*Šarčeva vlada*), nekatere pa izkazujejo tudi poimenovanja novih konceptov (*avtonomno vozilo*, *izsiljevalski virusi*) ali pa družbene posebnosti (*britanska premierka*, v primeru katere je raba narasla zgolj zaradi dejstva, da prej ženskih premierk v Veliki Britaniji ni bilo). Pri tovrstnih kandidatih še ni povsem jasno, ali bodo utonili v pozabo ali pa se bodo ustalili ali celo ponovno začeli rasti, zato je smiselno opazovati njihovo rabo tudi prihodnjih letih.

Kolokacije s časovno omejeno rabo (Tabela 8) so tiste, ki so v preteklosti že doživele vrh rabe in se zdaj praktično ne pojavljajo več (zanje sta značilna visok *m* in nizek *t*). Podobno kot pri nekaterih kolokacijah z nedavnim porastom gre tudi tukaj za kolokacije, ki so bile vezane na takratno družbeno dogajanje ali dogodek v preteklosti (*Peterletova vlada* z vrhom leta 1992, *vseslovenska vstaja* z vrhom leta 2013) oz. poimenujejo koncepte, ki so bili v preteklem obdobju pogosto obravnavana tema (npr. *prašičja gripa*, *nova gripa* v času pandemije virusa H1N1 leta 2009), ter koncepte, ki so se v vmesnem času preoblikovali ali preimenovali (*občinska skupščina* z vrhom leta 1994).

Tabela 8: Primeri kolokacij s časovno omejeno rabo v korpusu Gigafida 2.0.

Kolokacija	Absolutna pogostost	k	R^2	m	t
Peterletova vlada	648	–0,837	0,213	16,244	0,056
vseslovenska vstaja	758	0,094	0,058	21,016	0,368
nova gripa	1646	0,083	0,027	18,753	0,211
prašičja gripa	641	0,030	0,024	17,789	0,540
občinska skupščina	1484	–0,840	0,145	16,930	0,059

Področno specifične kolokacije (Tabela 9) so glede na uporabljene statistične mere podobne kolokacijam s časovno omejeno rabo, saj je zanje prav tako značilen izrazit vrh v preteklem obdobju, vendar pa pregled konkordanc pokaže, da so značilne za določen žanr ali tematsko področje. Ker je v Gigafidi 2.0 morda samo eno besedilo

s tega področja, to daje vtis, kot da je kolokacija dosegla vrh in nato usahnila. Med tovrstnimi kolokacijskimi kandidati so npr. že omenjena *predrepna plavut* skupaj s kolokacijama *mehke plavutnice* in *repna plavut* (iz *Velike knjige o ribolovu*), *krmilni element* (iz računalniškega priročnika *Access za Windows 95 v uporabi*), *televizijska prodaja* (v veliki večini se pojavlja le v televizijskih sporedih iz 90. let) ter *izpodbijana odločba* in *biometrijski ukrepi* iz pravnih besedil s spletnih strani slovenskih pravosodnih institucij.

Tabela 9: Primeri kolokacij s časovno omejeno rabo v korpusu Gigafida 2.0.

Kolokacija	Absolutna pogostost	k	R^2	m	t
predrepna plavut	446	–0,387	0,064	26,274	0,095
mehke plavutnice	221	–0,211	0,063	25,894	0,162
repna plavut	904	–0,615	0,066	25,575	0,075
krmilni element	510	–0,083	0,035	23,435	0,268
izpodbijana odločba	5782	0,891	0,034	19,285	0,033
biometrijski ukrepi	420	0,046	0,018	23,334	0,334
televizijska prodaja	652	–0,184	0,061	20,390	0,234

Ustaljene kolokacije (Tabela 10) so tiste, pri katerih ni opazen tako izrazit trend naraščanja ali padanja, saj je relativna pogostost skozi

Tabela 10: Primeri ustaljenih kolokacij v korpusu Gigafida 2.0.

Kolokacija	Absolutna pogostost	k	R^2	m	t
pomemben del	14.304	0,182	0,370	1,242	1,245
nadaljnji razvoj	11.362	0,158	0,391	1,246	1,169
izjemen uspeh	3395	0,075	0,467	1,404	1,375
dolgo obdobje	11.324	0,120	0,271	1,273	1,218
mednarodna raven	2272	0,034	0,261	1,319	1,285
rdeča nit	8631	0,196	0,697	1,333	1,285
gonilna sila	4250	0,078	0,406	1,347	1,290
strokovna revija	1955	–0,047	0,527	1,384	0,559

vsa leta primerljiva (zanje sta značilna nizka količnika *m* in *t*). Takšni primeri so npr. *pomemben del*, *nadaljnji razvoj*, *izjemen uspeh*, *dolgo obdobje* in *mednarodna raven*. Lahko pa v tej kategoriji najdemo tudi stalne besedne zveze (*strokovna revija*) in frazeološke enote (*rdeča nit*, *gonilna sila*).

3.2 Analiza na nivoju posameznih iztočnic

Podatki, pripravljeni z delotokom, omogočajo tudi osredotočanje na kolokacije po posameznih iztočnicah, s čimer lahko npr. opazujemo trende pri kolokacijah z novim besediščem oz. z besedami, ki jim je v zadnjem času narasla raba.

S pomočjo pregleda kolokacij z nedavnim porastom pri določeni iztočnici lahko ugotovimo, ali se npr. pri iztočnici uveljavlja nov pomen. Kot primer lahko izpostavimo kolokacije z iztočnico *avtonomen* (Tabela 11): v korpusu najdemo npr. kolokacije z visokim količnikom nedavne rasti *avtonomna vožnja*, *avtonomna vozila* in *avtonomni avtomobili*, pri katerih iztočnica *avtonomen* nastopa v pomenu "samodejen, avtomatski". Druge kolokacije z iztočnico avtonomen v pomenu "samostojen, neodvisen", kot so npr. *avtonomna cona/regija/skupnost/dežela/pokrajina*, imajo precej nižji količnik *t*. Opazovanje iztočnic z veliko količino kolokacij z nedavnim porastom je torej lahko koristno tudi za odkrivanje novih pomenov oz. sprememb pomena.

Tabela 11: Primeri kolokacij z iztočnico *avtonomen* v korpusu Gigafida 2.0.

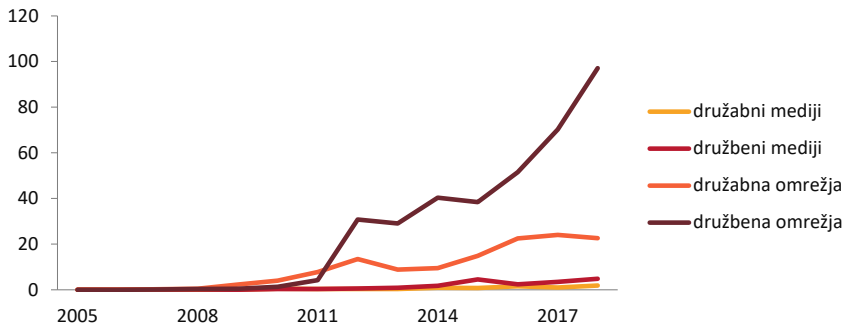
Kolokacija	Absolutna pogostost	k	R^2	m	t
avtonomna vožnja	618	0,084	0,319	8,091	13,689
avtonomna vozila	565	0,077	0,334	8,339	11,495
avtonomni avtomobili	169	0,023	0,359	5,295	4,337
avtonomna cona	391	0,017	0,134	4,637	1,489
avtonomna regija	281	0,009	0,087	2,746	1,433
avtonomna skupnost	180	0,010	0,351	2,469	1,421
avtonomna dežela	187	0,007	0,324	1,734	1,245
avtonomna pokrajina	1152	–0,052	0,085	6,049	1,139

Proučevati je mogoče tudi medsebojno konkurenčnost kolokacij v primerih, ko se za isti koncept v jeziku pojavita dve različici (ali več) in tekmujeta za uveljavitev. Nekaj tovrstnih primerov najdemo npr. pri iztočnicah *družben* in *družaben*, kjer lahko opazujemo trende pri variantah *družbeni mediji*, *družabni mediji*, *družbena omrežja* in *družabna omrežja* (Tabela 12).

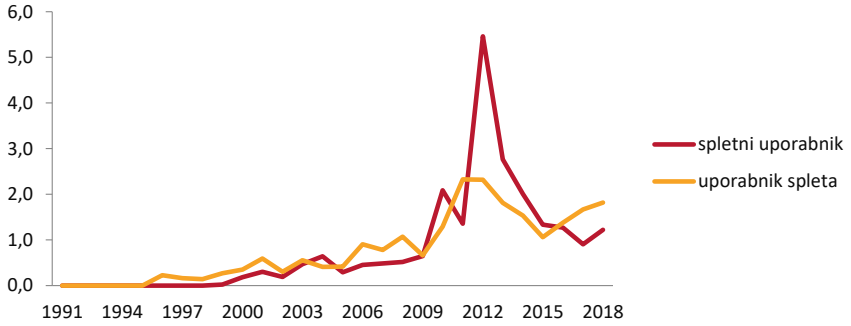
Tabela 12: Konkurenčne kolokacije *družbeni mediji*, *družabni mediji*, *družabna omrežja* in *družbena omrežja* v korpusu Gigafida 2.0.

Kolokacija	Absolutna pogostost	k	R^2	m	t
družabni mediji	343	0,046	0,567	5,199	4,418
družbeni mediji	899	0,119	0,488	6,203	6,898
družbena omrežja	16.364	2,236	0,531	7,417	13,654
družabna omrežja	5689	0,751	0,628	5,070	8,502

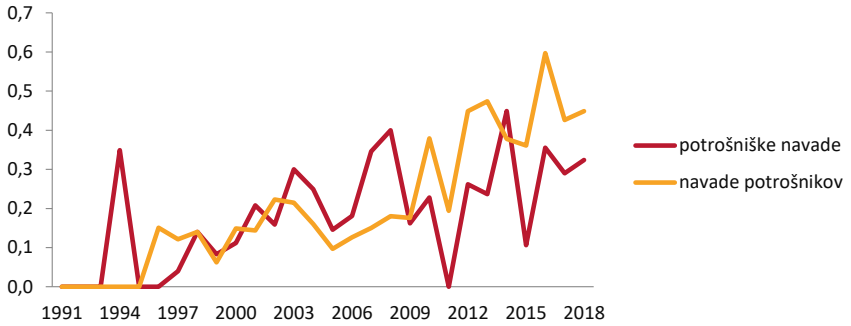
Tako po pogostosti kot tudi po količniku nedavne rasti *t* in količniku *m* izstopa kolokacija *družbena omrežja*. Vizualizacija razporeditve relativnih pogostosti vseh štirih kandidatov (Slika 10) pokaže, da je do leta 2011 prednjačila kolokacija *družabna omrežja*, zatem pa so jo s strmo rastjo, ki se še ni ustavila, izrinila *družbena omrežja*. Ostala kandidata, *družbeni mediji* in *družabni mediji*, se še



Slika 10: Relativne pogostosti konkurenčnih kolokacij *družbeni mediji*, *družabni mediji*, *družabna omrežja* in *družbena omrežja* v korpusu Gigafida 2.0.



Slika 11: Relativne pogostosti konkurenčnih kolokacij *spletni uporabnik* in *uporabnik spleta* v korpusu Gigafida 2.0.

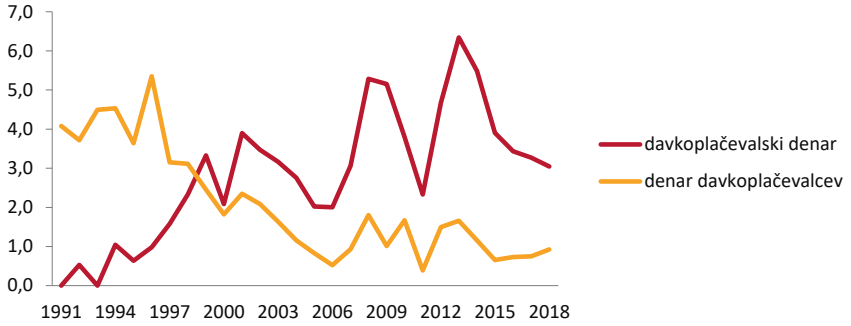


Slika 12: Relativne pogostosti konkurenčnih kolokacij *potrošniške navade* in *navade potrošnikov* v korpusu Gigafida 2.0.

pojavljata, a mnogo redkeje. Poleg statistik za merjenje naraščajoče ali padajoče rabe je torej nujno treba upoštevati tudi relativno pogostost in jo primerjati s potencialnimi konkurenčnimi kandidati.

Medsebojna razmerja diahronih trendov rabe kolokacij lahko opazujemo tudi na medstrukturni ravni. Primer para kolokacij s podobnim trendom rasti sta *spletni uporabnik* in *uporabnik spleta*, kjer izstopa le skokovit porast pri *spletni uporabnik* leta 2012 (Slika 11).

Nekoliko večje razlike zaznamo pri kolokacijah *potrošniške navade* in *navade potrošnikov* (Slika 12), ki tudi izkazujeta naraščajočo rabo, a je pri *navade potrošnikov* rast precej enakomernejša,



Slika 13: Relativne pogostosti konkurenčnih kolokacij *denar davkoplačevalcev* in *davkoplačevalski denar* v korpusu Gigafida 2.0.

medtem ko pri *potrošniških navadah* lahko opazimo nekoliko večja nihanja.

Primer para kolokacij s povsem različnima trendoma rasti sta *denar davkoplačevalcev* in *davkoplačevalski denar* (Slika 13). Tako *denar davkoplačevalcev* izkazuje izrazito padajoč trend, *davkoplačevalski denar* pa naraščajočega.

4 Diskusija

V prejšnjem razdelku smo predstavili več opazovanih vzorcev v časovnih trendih kolokacij, do katerih pridemo z izračunom različnih statističnih mer. V tem razdelku podajamo razmisleke o uporabnosti svojih opažanj za namene jezikoslovnega opisa oz. izdelave leksikalnih (slovarskih) virov, pri čemer izhajamo iz dveh perspektiv oz. situacij: izdelave povsem novih slovarskih virov in posodabljanja obstoječih. V zadnjem delu podajamo še nekaj predlogov za implementacijo rešitev pri predstavitvi podatkov o časovnih trendih slovarskim uporabnikom.

4.1 Izdelava novih slovarskih virov

Sodobni pristopi pri izdelavi leksikalnih virov vse pogosteje izkoriščajo prednosti avtomatskih postopkov, ki ponujajo vse boljše in

zanesljivejše podatke. Zanesljivost kolokacijskih podatkov je tudi za slovenščino že dosegla visoko raven, ki je pripeljala do objave Kolokacijskega slovarja sodobne slovenščine (Kosem idr. 2018), odzivnega slovarja, ki sledi konceptu, da se uporabnikom takoj omogoči dostop do relevantnih, a še neprečiščenih podatkov, ki jih potem leksikografi sproti izboljšujejo in dopolnjujejo.

Zaradi velike količine kolokacijskih podatkov je izboljšanje kakovosti postopkov avtomatskega luščenja kolokacij ključnega pomena tako za leksikografe in uporabnike. Poudarek je torej na ločevanju zrnja od plev in na identifikaciji za konkreten slovar nerelevantnih kolokacij. Upoštevajoč rezultate naše analize lahko rečemo, da takšne kandidate za izločitev iščemo predvsem med kolokacijami z zelo izrazitim vrhom in/ali kolokacijami s časovno omejeno (padajočo) rabo. Za ponazoritev kolokacij z zelo izrazitim vrhom podajamo v Tabeli 13 prvih 30 izluščenih kolokacijskih kandidatov po vrednosti *m* v strukturi s0-s2. Kolokacijski kandidati v krepkem tisku so tisti, ki se v letih 2017 in 2018 še pojavljajo v korpusu, podčrtani pa so primeri napak pri luščenju (napačno prepoznana struktura). Podrobnejša analiza pokaže, da dejansko večina kandidatov (tudi tistih, ki se v jeziku še pojavljajo) ni relevantnih za uvrstitev v slovarske vire. Poraja se le vprašanje relevantnosti področno specifičnih kolokacij (npr. *črta življenja*), ki se pojavljajo med kandidati z izrazitim vrhom, (kot že ugotovljeno v razdelku 3.1); odločitev o njihovi izločitvi/vključitvi je vezana na konkreten slovarski vir, npr. za kolokacijski slovar splošnega slovenskega jezika tudi tovrstne kolokacije niso relevantne. Za ločevanje področno specifičnih kolokacij od tistih s časovno

Tabela 13: Prvih 30 kolokacijskih kandidatov po vrednosti *m* v strukturi s0-s2.

Struktura	Kolokacijski kandidati
samostalnik + samostalnik v rodilniku s0-s2	črta glave, črta usode, črta srca, del bokov, **kazen dinarjev**, **imetniki deležev, član ZDen**, preprečevanje državljanstva, promet proizvodov, zaobljuba svetov, **stran kril**, imetnik SP, **črta življenja**, fototeka ekip, vojak JNA, član ZUstS, enote JNA, član ZVS, **zbor garde**, količina kropa, **točka obrazložitve**, del plavuti, posnemanje Kristusa, član ZAazil, svet Demosa, **Marjan Šarca**, Svet SDZ, virus zika, razglednica vsebineNA, **člani ustanove**

omejeno rabo bi bilo smiselno v prihodnje upoštevati še nekatere druge značilke, npr. razporeditev po besedilnih zvrsteh ter število besedil oz. dokumentov, v katerih je kolokacija v korpusu prisotna. Razmisliti pa je treba tudi, ali je mogoče na avtomatski način ločevati stalne besedne zveze od (slovarsko relevantnih) kolokacij.

4.2 Posodabljanje obstoječih slovarjev

Posodabljanje obstoječih slovarjev lahko vključuje tudi avtomatske postopke, vendar pa je pri že opravljenih semantičnih analizah posameznih leksikalnih enot njihova uporabnost nekoliko manjša. Leksikografi tako lahko dobijo avtomatska opozorila in izluščene podatke, ki jih potem pregledajo in prečistijo. Pri analizi časovnih trendov so za leksikografe tako dragoceni podatki o kolokacijah z izrazitim nedavnim porastom, sploh v primerih, ko gre za povsem nove kolokacije. Najočitnejša uporabnost takšnih kolokacij je nakazovanje novih pomenov besed, kot smo ga zaznali pri pridevniku *avtonomen* (gl. razdelek 3.2). Podoben primer najdemo pri glagolu *deliti*, kjer se po letu 2012 pojavi nova raba (*deliti fotografijo, deliti novico, deliti zgodbo, deliti objavo, deliti posnetek*) v pomenu "objaviti ali posredovati na spletni strani ali družbenem omrežju", ki ga obstoječi slovarji slovenskega jezika še niso zaznali. Uveljavljanje novega pomena in njegovo legitimnost za vključitev v slovarski vir upravičuje večje število kolokacij, ki navadno pripadajo istemu semantičnemu tipu.

Pri majhnem številu kolokacij, ki kažejo na nov pomen, ter njihovih posameznih trendih se pojavi vprašanje dinamike med kolokacijami in stalnimi besednimi zvezami kot samostojnimi leksikalnimi enotami v slovarju. Recimo pri pridevniku *izsiljevalski* se je že leta 2013 pojavila kolokacija *izsiljevalski virus*, leta 2016 pa še *izsiljevalski program* in *izsiljevalska (programska) oprema*. Tako bi v letih 2014–2015 leksikograf kolokaciji *izsiljevalski virus* lahko dodelil status samostojne večbesedne iztočnice, danes pa bi zaradi ostalih zaznanih kolokacij lahko to postala zgolj (ali tudi) kolokacija pri *izsiljevalski*.[6]

6 V našem primeru bi zaradi pogostosti in področne zamejenosti *izsiljevalski virus* ostal stalna zveza, bi bil pa tudi povezan (v slovarju prikazan) z novim pomenom pri *izsiljevalski*.

Pri novih kolokacijah, ki so šele nedavno začele izkazovati naraščajoči časovni trend, se vedno postavlja vprašanje, ali gre za trajen fenomen in ali se bo kolokacija (in mogoče z njo povezan nov pomen) v jeziku uveljavila ali pa gre samo za nekaj začasnega. Medtem ko je bila to mogoče zagata v času tiskanih slovarjev (in elektronskih slovarjev, ki so bili zgolj tiskani slovarji, preneseni na splet), pa v sodobni leksikografiji, ki omogoča hitro odzivnost in dinamičnost slovarskih vsebin, to ne bi smela več biti težava. Dandanes že vidimo, da se besede, ki niso v rabi niti nekaj mesecev (za primer lahko vzamemo izraze, povezane s pandemijo covid-19), že uvršča v slovarje. Seveda se upošteva določene dodatne kriterije, npr. besedilno razpršenost. Zakaj ne bi podobnega počeli tudi s kolokacijami? Navsezadnje je kolokacijo vedno mogoče iz slovarja kasneje odstraniti, če njena raba močno upade.

Zanemariti ne gre tudi pomena novih kolokacij, ki izkazujejo novo rabo, ne pa tudi novih pomenov leksikalnih enot. Eden od takšnih primerov so kolokacije *odprava dioptrije*, *implementirati odločbo, implementirati razsodbo,* ki se v korpusu Gigafida 2.0 prvič pojavijo leta 2015 ali pozneje (v Kolokacijskem slovarju sodobne slovenščine 1.0, ki temelji na korpusu Gigafida 1.0, ki zajema besedila do leta 2011, jih zato ni). Takšne kolokacije so očitni kandidati za vključitev v slovar ob njegovi posodobitvi.

Za posodabljanje slovarskih virov so prav tako relevantni podatki o kolokacijah, ki v jeziku niso nove, a jim je raba v zadnjih letih močno upadla. Takšne kolokacije velikokrat nakazujejo na pešanje rabe pomena, bodisi zaradi družbenih sprememb ali uveljavitve drugih, nadomestnih jezikovnih poimenovanj. Primer so (razširjene) kolokacije,[7] povezane s samostalnikom tolar v pomenu denarne valute (*tolar škode, tolar plače, tolarji kazni, tolarji izgube*). V tem konkretnem primeru je informacija o padajočem časovnem trendu kolokacij, vezanih na pomen iztočnice *tolar*, koristno opozorilo za dodajanje časovne oznake ali prilagoditev razlage, ne pa nujno za izključevanje kolokacij. Podoben primer je *cena impulza,* ki prav tako

7 Gre za kolokacije, ki se vedno pojavljajo z dodatnim eno ali več besednim elementom, v tem primeru s količinskim (npr. *sto milijonov tolarjev škode*).

izkazuje padajoč trend (po letu 2012 se sploh ne pojavi več), je pa kolokacija zanimiva zato, ker izkazuje pomen od *impulz,* ki ga v obstoječih slovarskih virih ni.[8] V primeru, da bi se leksikograf odločil za izpust kolokacije (in pomena) iz slovarja, bi tako pomen v slovenskih slovarskih virih ostal nezabeležen.

Naraščajoči ali padajoči časovni trend (večje) skupine kolokacij znotraj posameznega pomena lahko torej lahko pripelje do sprememb na ravni pomenskega opisa. Je pa tak trend lahko tudi signal za spremembo na ravni pomenske členitve oz. vrstnega red pomenov. V kolikor imamo v slovarski bazi kolokacije sistematično popisane, se takšen proces lahko v veliki meri avtomatizira, razvrščanje pomenov pa postane dinamično – ko raba kolokacij znotraj enega pomena drastično pade, se pomen pomakne nižje v razvrstitvi. Prehod na takšen pristop zahteva opredeljevanje pomenov v slovarski bazi z identifikacijsko številko (ID) in ne zaporedno številko znotraj gesla.

Čeprav z vidika analize časovnih trendov (pogoste) ustaljene kolokacije, ki ne kažejo izrazitih trendov naraščanja ali padanja v rabi, mogoče niso tako zanimive, pa to ne velja za leksikografsko analizo. Med temi kolokacijami so namreč pogosto tudi tiste, ki so najbolj tipične za pomene leksikalnih enot in so posledično najbolj relevantne za vključitev v slovarska gesla.

4.3 Kolokacijski trendi in slovarski uporabniki

V razdelku 4.1 smo že izpostavili vlogo časovnih trendov kolokacij pri pripravi odzivnih slovarjev z avtomatsko izluščenimi podatki. V tem razdelku namenjamo več pozornosti razmisleku direktnega vključevanja podatkov o časovnih trendih v slovarske vmesnike. Podatke o časovnih trendih sicer že najdemo v nekaterih slovarjih, a zgolj za posamezne iztočnice, recimo v nemškem slovarju DWDS (Digitales Wörterbuch der deutschen Sprache),[9] ki ponuja podatke za obdobje

8 Kolokacijo sicer najdemo v Kolokacijskem slovarju sodobne slovenščine pod iztočnicama *impulz* in *cena*.

9 https://www.dwds.de/

70 let, in v Dictionary.com,[10] ki opozarja na besede, ki jim je v zadnjem času raba najbolj narasla.

Ko razmišljamo o smiselnosti vključevanja podatkov o časovnih trendih kolokacij v slovarske vire, se moramo zavedati, da gre za v večini primerov nenujno informacijo, s katero nočemo obremenjevati uporabnikov. Tako se je treba osredotočiti na primere, ko informacija uporabniku koristi pri jezikovni produkciji. Kolokacije, pri katerih bi bilo podatek smiselno prikazati (z diagramom ali oznako), so tako predvsem tiste z izrazito naraščajočim ali padajočim trendom rabe. Mogoče še primernejša skupina so kolokacije, ki so si medsebojno konkurenčne, kot prikazujeta primera *družbeno omrežje* in *davkoplačevalski denar* v razdelku 3.2. Navsezadnje potrebo po sopostavljanju takšnih kolokacijskih variant in variant nasploh potrjujejo tudi številna vprašanja uporabnikov v jezikovnih svetovalnicah[11] in podobnih digitalnih okoljih, kjer uporabniki najpogosteje sprašujejo po primerjavi jezikovnih variant (Arhar Holdt idr. 2015; Arhar Holdt idr. 2017; Čibej 2019).

V virih, kot je Kolokacijski slovar sodobne slovenščine, ki postavlja kolokacije v ospredje in ostale njihove lastnosti, tudi pomene, uporablja kot filtre, je podatek o časovnih trendih mogoče uporabiti tudi na bolj splošni ravni. Uporabnikom se tako lahko ponudi možnost razvrščanja ali filtriranja kolokacij po aktualnosti oz. trendu rabe (kolokacij z najbolj naraščajočo rabo).

5 Zaključek

Podatki o časovnih trendih kolokacij, ki smo jih pridobili z metodami, razvitimi v okviru projekta KOLOS, nam ponujajo dragocen vpogled v njihovo rabo in omogočajo boljše razumevanje obnašanja leksikalnih enot in njihovih pomenov ter jezika nasploh. V prispevku smo pokazali uporabnost različnih statističnih mer (in njihovih kombinacij) pri prepoznavi različnih skupin kolokacij glede na časovni trend. Podatki o naraščajoči, padajoči ali ustaljeni rabi

10 https://www.dictionary.com/

11 Glej npr. številna vprašanja v https://svetovalnica.zrc-sazu.si/tags/sopomenskost.

kolokacij so nadvse uporabni za namene jezikovnega opisa, tako pri pripravi povsem novih virov kot pri posodabljanju obstoječih. Potrebe slovarskih uporabnikov narekujejo tudi iskanje rešitev za učinkovito prikazovanje tovrstnih informacij o kolokacijah neposredno v slovarskih virih.

Naša prihodnja prizadevanja bodo usmerjena v preizkušanje različnih nastavitev parametrov statističnih mer z namenom izdelave čim bolj optimalnih formul za zaznavanje sprememb v rabi kolokacij ter posledično pomenov in leksikalnih enot nasploh. Analize bomo razširili na vse kolokacije, torej tudi na druge skladenjske strukture. Podatke o časovnih trendih nameravamo kombinirati z ostalimi podatki o kolokacijah, npr. s statističnimi merami povezovalnosti, z besedilno razpršenostjo in z morebitno omejenostjo na besedilne tipe. Vse statistične podatke o kolokacijah bomo zbrali v bazi, ki bo po eni strani na voljo leksikografom in jezikoslovcem, po drugi strani pa bo osnova za uporabniško naravnane vire, kot je npr. Jezikovni sledilnik (Kosem idr. 2021). Poleg tega želimo raziskave opraviti tudi na besedilno bolj raznovrstnih podatkih, kar bo omogočila izdelava metakorpusa vseh večjih korpusov slovenskega jezika v okviru projekta *Razvoj slovenščine v digitalnem okolju* (RSDO).[12]

Bistveno sporočilo naše raziskave, pa tudi sorodnih raziskav, kot sta Pollak idr. (2019) in Nimb idr. (2020), je, da se jezik zelo hitro spreminja in da kolokacije ponujajo ključ do spremljanja in popisovanja teh sprememb. Zato je izdelava digitalne slovarske baze z vsemi podatki o slovenščini (tudi kolokacijskimi), ki je načrtovana v projektu RSDO, korak v pravo smer. Le na ta način lahko namreč zagotovimo nenehno ažurnost in posledično kakovostnost slovarskih in ostalih leksikalnih virov sodobnega slovenskega jezika.

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*

12 https://www.cjvt.si/rsdo/

tehnologije za slovenski jezik) je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna.

Reference

Arhar Holdt, Š., Čibej, J., Zwitter Vitez, A. (2015): S pomočjo uporabniških jezikovnih vprašanj in mnenj do boljšega slovarja. V V. Gorjanc, P. Gantar, I. Kosem in S. Krek (ur.): *Slovar sodobne slovenščine: problemi in rešitve (Zbirka Prevodoslovje in uporabno jezikoslovje)*: 196-214. Ljubljana: Znanstvena založba Filozofske fakultete.

Arhar Holdt, Š., Čibej, J., Zwitter Vitez, A. (2017): Value of language-related questions and comments in digital media for lexicographical user research. *International journal of lexicography*, 30 (3): 285-308.

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

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

Cook, P., Rundell, M., Lau L. H. in Baldwin T. (2014): Applying a word-sense induction system to the automatic extraction of dictionary examples. V A. Abel, C. Vettori in N. Ralli (ur.): *Proceedings of the XVI EURALEX International Congress*: 319–328. Bolzano, Italy: EURAC.

Čibej, J. (2019): Končno poročilo o spremljanju uporabniških odzivov, mnenj in načinov uporabniškega vključevanja. Projekt *Slovar sopomenk sodobne slovenščine: Od skupnosti za skupnost*. Ljubljana: Center za jezikovne vire in tehnologije.

Dictionary.com. Dostopno prek: https://www.dictionary.com/ (22. 12. 2020).

Digitales Wörterbuch der deutschen Sprache. Dostopno prek: https://www.dwds.de/ (22. 12. 2020).

Fišer, D. in Ljubešić, N. (2016): Detecting Semantic Shifts in Slovene Twitterese. V A. Horák, P. Rychlý in A. Rambousek (ur.): *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2016*: 1–8.

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

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.

Geeraerts, D. (1997): *Diachronic Prototype Semantics. A Contribution to Historical Lexicology*. Oxford: Clarendon Press.

Gulordava, K. in Baroni, M. (2011): A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. V *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*: 67–71.

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*: 379–93.

Heylen, K. idr. (2015): Monitoring polysemy: Word space models as a tool for large-scale lexical semantic analysis. *Lingua*, 157: 153–72.

Kehoe, A. in Gee, M. (2009): Weaving Web data into a diachronic corpus patchwork. V A. Renouf in A. Kehoe (ur.): *Corpus Linguistics: Refinements & Reassessments*: 255-279. Amsterdam: Rodopi.

Kosem, I., Gantar, P., Krek, S., Arhar Holdt, Š., Čibej, J., Laskowski, C., Pori, E., Klemenc, B., Dobrovoljc, K., Gorjanc, V. in Ljubešić, N. (2018): *Kolokacije 1.0: Kolokacijski slovar sodobnega slovenskega jezika*. Dostopno prek: https://viri.cjvt.si/kolokacije/slv/# (23. 12. 2020).

Kosem, I., Čibej, J., Gantar, P., Arhar Holdt, P., Krek, S., Laskowski, C., Robnik Šikonja, M., Klemenc, B., Dobrovoljc, K., Gorjanc, V., Repar, A. in Ljubešić, N. (ur.) (2021): *Sledilnik 1.0: Jezikovni sledilnik*. Dostopno prek: viri.cjvt.si/sledilnik (12. 2. 2021).

Krek idr. (2019): *Gigafida 2.0: Korpus pisne standardne slovenščine*. Dostopno prek: viri.cjvt.si/gigafida (23. 12. 2020).

Krek, S., Arhar Holdt, Š., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešič, N., Kosem, I. in Dobrovoljc, K. (2020): Gigafida 2.0: The Reference Corpus of Written Standard Slovene. V *Proceedings of the 12th Language Resources and Evaluation Conference*: 3340–3345. European Language Resources Association. Dostopno prek: https://www.aclweb.org/anthology/2020.lrec-1.409 (12. 2. 2021).

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

Nimb, S., Sørensen, N. H. in Lorentzen H. (2020): Updating the dictionary: Semantic change identification based on change in bigrams over time. *Slovenščina 2.0: Empirical, Applied and Interdisciplinary*

Research, 8 (2): 112–138. Dostopno prek: https://doi.org/10.4312/slo2.0.2020.2.112-138 (12. 2. 2021).

Pollak, S., Gantar, P. in Arhar Holdt, Š. (2019): What's New on the Internetz? Extraction and Lexical Categorisation of Collocations in Computer-Mediated Slovene. *International Journal of Lexicography*, 32 (2): 184–206. Dostopno prek: https://doi.org/10.1093/ijl/ecy026 (12. 2. 2021).

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 Univerze v Ljubljani.

Renouf, A. (2009): Corpus Linguistics beyond Google: the WebCorp Linguist's Search Engine in New Paths for Computing Humanists. V R. Siemens in G. Shawver (ur.): *Digital Studies*: Vol 1. The Society for Digital Humanities (SDH).

Renouf, A. (1993): A Word in Time: first findings from dynamic corpus investigation. V J. Aarts, P. de Haan in O. Nelleke (ur.): *English Language Corpora: Design, Analysis and Exploitation*: 279–288. Rodopi, Amsterdam.

Sagi, E., Kaufmann S. in Clark B. (2011): Tracing semantic change with latent semantic analysis. V K. Allan in J. A. Robinson (ur.): *Current Methods in Historical Semantics*: 161–183. De Gruyter Mouton, Berlin, Germany.

Sinclair, J. (1991): *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Urbančič, B. (1987): *O jezikovni kulturi*. Ljubljana: Delavska enotnost.

Žele, A. (2009): Pomenotvorne zmožnosti z vidika (de)terminologizacije (v slovenščini). V N. Ledinek, M. Žagar Karer in M. Humar (ur.): *Terminologija in sodobna terminografija*: 125–139. Ljubljana: Založba ZRC, ZRC SAZU.

Evalvacija uporabniškega vmesnika Kolokacijskega slovarja sodobne slovenščine

Eva PORI
Filozofska fakulteta, Univerza v Ljubljani

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

Jaka ČIBEJ
Filozofska fakulteta, Univerza v Ljubljani; Institut Jožef Stefan

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

The paper focuses on a qualitative analysis of user evaluations of the interface of the Collocations Dictionary of Modern Slovene. The experiment included 40 participants from four target groups: translators and proofreaders, teachers of Slovene as a first language, teachers of Slovene as a second/foreign language, and linguists. The study shows that these groups find the innovative functions offered by the dictionary interface very useful and allow for a more thorough data analysis. On the other hand, a number of features require a visual overhaul to make them more conspicuous and easily perceivable by dictionary users.

The evaluations were carried out using think-aloud protocols (TAP), which turned out to be an efficient method to detect (un)problematic interface elements and yielded a number of constructive findings. These offer a good starting point for further updates to the interface of the Collocations Dictionary and improvements in the evaluation methods for similar studies in the future.

Keywords: collocations dictionary, responsive dictionary, user evaluation, method TAP, dictionary interface

1 Uvod

Odzivni slovarji, med katere sodi tudi Kolokacijski slovar sodobne slovenščine (KSSS) (Kosem idr. 2018), ki se mu posvečamo v tem prispevku, vnašajo v slovaropisni prostor številne novosti, med katerimi so najpomembnejše: strojna priprava jezikovnih podatkov, odprta objava še (jezikoslovno, ročno) nepregledanega slovarja ter njegov postopni, transparentni razvoj v sodelovanju s širšo jezikovno skupnostjo. Odzivni koncept, ki ga je za slovenščino in tudi širše uvedel Slovar sopomenk sodobne slovenščine leta 2018 (Arhar Holdt idr. 2018), je bil v javnosti pozitivno sprejet, vendar predlagane inovacije kljub temu zahtevajo testiranja, evalvacije in nadaljnje razmisleke, pri katerih je treba upoštevati tako strokovno kot uporabniško mnenje. S tega vidika se prispevek posveča odzivnim slovarskim vmesnikom, ki uporabnikom zagotavljajo dostop do klikljivih in medsebojno povezanih jezikovnih informacij, odgovoriti pa morajo tudi na potrebe prikaza in urejanja (filtriranja, grupiranja, izvažanja) strojno pridobljenih podatkov, omogočati njihovo pravilno interpretacijo, sledljivost sprememb ter metodološko premišljeno uporabniško vključevanje in spremljanje.

Slovar sopomenk sodobne slovenščine je bil kot prvi, preizkusni odzivni slovar evalviran z več vidikov, kar pregledno predstavlja prispevek Arhar Holdt in Čibej (2020). Priložnost, da na primerljiv način uporabniško informacijo zberemo tudi za KSSS, je prinesel nacionalni projekt KOLOS (*Kolokacije kot temelj jezikovnega opisa: semantični in časovni vidiki*, ARRS J6-8255), v katerem je bila izvedena tako raziskava, kako v slovarju urediti kolokacijsko gradivo, da bodo na prvem mestu za uporabnike najbolj relevantni rezultati (Arhar Holdt 2021), kot tudi raziskava, kakšen odnos imajo uporabniki do značilnosti strojno izluščenih kolokacij in slovarskega vmesnika, v katerem so dostopne. Idejo raziskave, metodologijo in tisti del rezultatov, ki se osredotoča na uporabniški odnos do napak in pomanjkljivosti v podatkih, predstavlja prispevek Pori idr. (2020). Niso pa še bili predstavljeni rezultati tistega dela raziskave, v kateri so sodelujoči evalvirali slovarski vmesnik. Z ločeno in v smislu

načrtov konkretizirano predstavitvijo teh izsledkov želimo preseči pomanjkljivost uporabniških raziskav, ki povratno informacijo zberejo in opišejo, redko pa jasno opredelijo, kako se je oz. bo slednja upoštevala pri razvoju izdelka.

V metodološkem smislu raziskava temelji na protokolih glasnega razmišljanja, združenih s pristopom polstrukturiranega evalvacijskega intervjuja. Sodelujoči evalvatorji so bili povabljeni, da preklikavajo slovarska gesla po svojem izboru in si na ta način ustvarijo vtis o KSSS, obenem pa glasno izražajo misli, občut(en)ja in težave, ki jih pri tem opazijo. Delo s slovarjem je bilo zvočno in vizualno posneto (snemanje zaslona), prostemu klikanju pa je sledil polstrukturirani del, v katerem je spraševalec opozoril na vmesniške značilnosti, ki jih je sodelujoči morda prezrl (Pori idr. 2020: 198–201). Tak pristop ponuja uvid v tako eksplicitne kot implicitne potrebe sodelujočih, kar vodi do poglobljenega razumevanja načina mišljenja potencialnih slovarskih uporabnikov. Testiranje različnih slovarskih vmesniških funkcij in identifikacije mest za izboljšave pa vodi do izboljšanih slovarskih rešitev.

V nadaljevanju opišemo metodologijo in vzorec sodelujočih, nakar navedemo seznam vmesniških značilnosti oz. elementov, ki so bili pri evalvaciji v središču uporabniške pozornosti. Za vsakega od elementov opredelimo njegov namen v luči odzivnega koncepta, predstavimo rezultate protokola glasnega razmišljanja ter podamo številčne ocene, kako so se sodelujoči (po reprezentiranih uporabniških skupinah) opredeljevali do (ne)problematičnosti elementa. Izsledke evalvacij strnemo v točke za nadaljnji razvoj slovarja, pri čemer naslovimo tudi (ne)uspešnost izvornih slovaropisnih oz. oblikovalskih predpostavk. Na kratko ovrednotimo uporabnost izbrane metode in njeno uspešnost za izbrani namen.

2 Metodologija

2.1 Raziskovalni okvir

Protokoli glasnega razmišljanja (ang. think aloud protocols, krajše TAP) so metodološki pristop, pri katerem udeleženci raziskave ob

izvajanju določene naloge glasno ubesedujejo svojo izkušnjo, težave, mnenja in podobno, raziskovalec pa te informacije beleži in naknadno analizira. Pristop TAP izvira s področja psihologije, znotraj katerega je bil tudi kritično ovrednoten. Glavna opozorila ponujata že Ericsson in Simon (1984), ki odsvetujeta retrospektivne protokole in takšne naloge, pri katerih so pred verbalizacijo potrebni interpretativni kognitivni procesi. Boren in Ramey (2000), ki pristop ocenita v novejši luči, izpostavita kot osnovo tudi jasna navodila, opozarjanje sodelujočih, kadar nastopijo trenutki tišine, sicer pa nevmešavanje v proces, pri analizi pa natančno ločevanje zbranih podatkov od mnenj oz. interpretacij (ibid.: 263).

V naboru metodoloških možnosti za izvedbo slovaropisnih uporabniških raziskav (Welker 2013a, 2013b) se pristop TAP osredotoča na slovarsko rabo, tipično v povezavi z vprašanjem, kako slovarski uporabniki pristopijo k reševanju problema, s kakšnimi morebitnimi težavami se soočajo pri slovarski rabi in kako interpretirajo slovarske podatke. Med sodobnejšimi študijami gre omeniti denimo prispevek Wingate (2002), ki se posveča uporabnosti raznovrstnih slovarskih definicij, monografijo Thumb (2004), ki raziskuje načine iskanja po slovarskih priročnikih in s tem povezane uporabniške težave, Simonsen (2014) se ukvarja s slovarji na mobilnih telefonih, Comeau (2009) pa s TAP opazuje delo leksikografov in ob tem skuša identificirati slovaropisne kompetence. Za utemeljitelja protokolov glasnega razmišljanja na področju razvoja računalniških vmesnikov velja Lewis (1982), v širši praktično-razvojni kontekst pa jih postavljata Lewis in Rieman (1993). V primerjavi z uporabo TAP za družboslovne raziskave je na področju vmesniških evalvacij glavni namen identifikacija močnih in šibkih točk ocenjevanega izdelka s stališča uporabnosti za sodelujočega, ne toliko (karseda objektivno) spremljanje njegovega kognitivnega procesa.

Ker je pristop časovno in finančno potraten (Tarp 2009: 287), na področju slovaropisja že v preteklosti ni bil med najpogosteje izbranimi, v sodobnem času pa ga pogosto nadomeščajo objektivne opazovalne metode, kot sta spremljanje dnevnikov iskanja in sledenje zaslona. Ti pristopi za razliko od TAP pokažejo avtentično, nemoteno

obnašanje slovarskih uporabnikov pri reševanju jezikovne zadrege, vendar ne podajo uporabnikove eksplicitne ocene, opozoril na težave in predlogov za izboljšave posameznih vmesniških značilnosti. Interpretacija, ali je bil uporabnik s slovarsko izkušnjo zadovoljen, ali je razpoložljive funkcionalnosti opazil in jih razumel, ostanejo na strani raziskovalca, prav tako je mogoče o želenih izboljšavah le sklepati. Iz teh razlogov smo se odločili, da za evalvacijo vmesnika uporabimo TAP v kombinaciji s snemanjem ekrana, glasno razmišljanje pa podpremo s polstrukturiranimi vprašanji, ki omogočajo primerjavo zbranih mnenj med različnimi uporabniškimi skupinami.

2.2 Opis raziskave in struktura vzorca

Raziskavo smo zasnovali kot kombinacijo prostega klikanja po slovarju z glasnim razmišljanjem (TAP) in polstrukturiranega, vodenega intervjuja, v katerem so lahko sodelujoči svoje vtise dopolnili in dodatno opredelili. V polstrukturiranem delu raziskave smo uporabnike vsebinsko usmerili in preverjali stopnjo zaznavanja problematičnosti pri izbranih slovarskih geslih: zanimalo nas je, koliko opazijo težave strojnega pridobivanja podatkov, kaj jih pri novem odzivnem konceptu moti in podobno. Izjave sodelujočih so bile zvočno posnete, posneto je bilo tudi njihovo iskanje po slovarju in gibanje miške po ekranu. Posnetki so bili transkribirani in analizirani. Polstrukturirani del intervjuja je skupaj z vprašalnikom predstavljen v Pori idr. (2020).

Kot omenjeno, nas v tem prispevku podrobneje zanimajo rezultati prvega dela evalvacije, v katerem so se uporabniki pri prostem in neusmerjenem sprehodu skozi slovar samoiniciativno opredeljevali do funkcionalnosti in intuitivnosti uporabniškega vmesnika, npr. glede informacije o stopnji izdelanosti gesla (ikona piramide), prisotnosti ali odsotnosti pomenske členitve gesla, gumba s tremi pikami za vstop v posamezno skladenjsko strukturo ipd. V analizi rezultatov smo se z namenom pridobitve podrobnejših, celovitejših in čim bolj realnih ocen usmerili v analize posnetkov zaslona oz. implicitnih podatkov, v neposredni odvisnosti oz. v razmerju do

podatkov, pridobljenih v preostalem delu intervjuja, ki so temeljili na eksplicitno podanih ocenah uporabnikov. Eksplicitna informacija se je pokazala kot koristna dopolnitev implicitni, saj je odpravila morebitne neskladnosti med izrečenim (verbalnim) in neverbalnim, kar je omogočilo lažje razumevanje in ocenjevanje uporabniških mnenj.

Kot je natančneje predstavljeno v Pori idr. (2020: 173), smo ob upoštevanju sheme predvidenih slovarskih uporabnikov (Arhar Holdt idr. 2016) v raziskavo vključili štiri ciljne skupine: prevajalce oz. lektorje, učitelje slovenščine kot prvega jezika, učitelje slovenščine kot drugega/tujega jezika ter jezikoslovce[1]. Nekateri od jezikoslovcev so sicer sodelovali pri posameznih fazah izgradnje KSSS (npr. pri pripravi podatkov, kot svetovalci pri oblikovanju (delov) uporabniškega vmesnika), vendar pa se nam to ni zdelo problematično, temveč kvečjemu dodana vrednost raziskave. Pri pripravi vmesnika so bile namreč številne rešitve kompromisi, pripravljeni v prid domnevni uporabnosti in enostavnosti, niso pa nujno odražale mnenj posameznikov, kar naj bi se pokazalo v naši raziskavi. Poleg tega so od sodelovanja pri pripravi podatkov in vmesnika jezikoslovci imeli priložnost uporabiti KSSS za različne namene (poučevanje, izobraževanja, raziskave ipd.) in so lahko svoje mnenje o vmesniku na podlagi izkušenj že spremenili. Tretji argument za vključitev te skupine je dejstvo, da priprava slovarja temelji na predvidevanjih avtorjev oz. leksikografov o uporabniških potrebah, pri čemer ta predvidevanja niso zbrana na način, da bi jih bilo mogoče primerjati z dejanskim mnenjem uporabnikov. Zbiranje mnenja po enaki metodologiji primerjavo omogoča in (ne)uspešnost predvidevanj, kot bo vidno v rezultatih, tudi osvetljuje.

Skupno je raziskava zajela 40 sodelujočih. Kot je razvidno iz Tabele 1, so bili sodelujoči večinoma osebe, stare med 30 in 50 let, z 10- do 30-letnimi delovnimi izkušnjami, in so prihajali iz različnih slovenskih regij ali (v primeru učiteljev slovenščine kot drugega/tujega jezika) iz tujih držav.

1 Z jezikoslovci imamo v mislih raziskovalce na področju jezikoslovja, nekateri od njih se tudi ukvarjajo s slovaropisjem.

Tabela 1: Prikaz strukture vzorca sodelujočih.

Uporabniška skupina	Vključene institucije	Regija	Starost	Poklicne izkušnje
10 učiteljev slovenščine kot prvega jezika	SŠ Ravne na Koroškem II. gimnazija v MB Ekonomska šola (+gimnazija) Ljubljana	Ljubljanska Podravska Koroška Gorenjska	30–50	10–30 let
10 učiteljev slovenščine kot drugega ali tujega jezika	CSDTJ FF UL	Madžarska Češka Štajerska Ljubljanska Primorska	30–50	10–30 let
10 prevajalcev / lektorjev	SLG Celje samozaposleni samostojni delavec v kulturi	Primorska Dolenjska Savinjska Gorenjska Ljubljanska	30–50	10–30 let
10 jezikoslovcev	CJVT UL FDV UL FF UL samozaposleni	Ljubljanska Štajerska	30–50	10–20 let

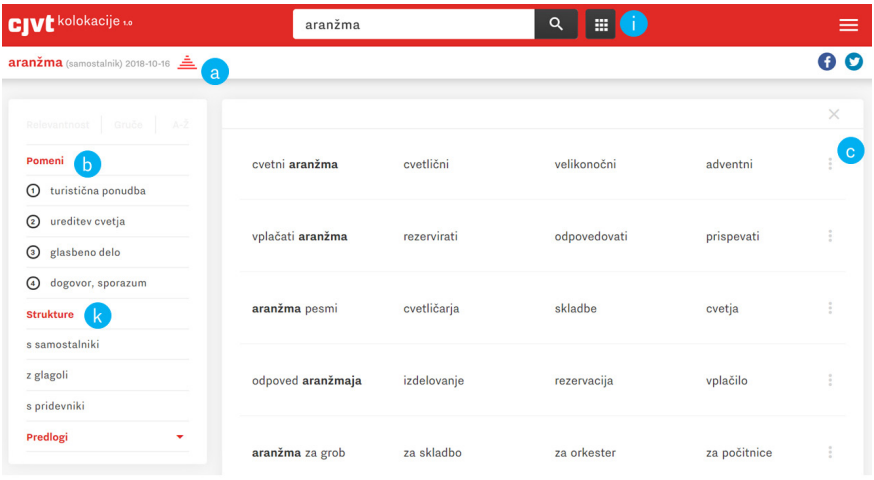
Pred srečanjem smo sodelujoče prosili za pisno privolitev v zvočno snemanje in snemanje zaslona. Posnetke smo najprej transkribirali, nato pa še kategorizirali glede na vsebino izjav (Pori idr. 2020: 176–179), s čimer smo pridobili seznam vmesniških elementov, do katerih so se sodelujoči v TAP najpogosteje opredeljevali. Kot dopolnilo že predstavljenemu predhodnemu delu smo pregledali in dodatno označili še mnenja uporabnikov, ki so bila pri prvotni anotaciji združena pod skupno oznako 'Predlogi sodelujočih' (prim. Tabelo 3 v Pori idr. 2020). V nadaljevanju poglavja navajamo seznam elementov in nekaj metodoloških opozoril glede interpretacije rezultatov, sledijo izsledki za vsakega od obravnavanih elementov posebej.

2.3 Obravnavani vmesniški elementi

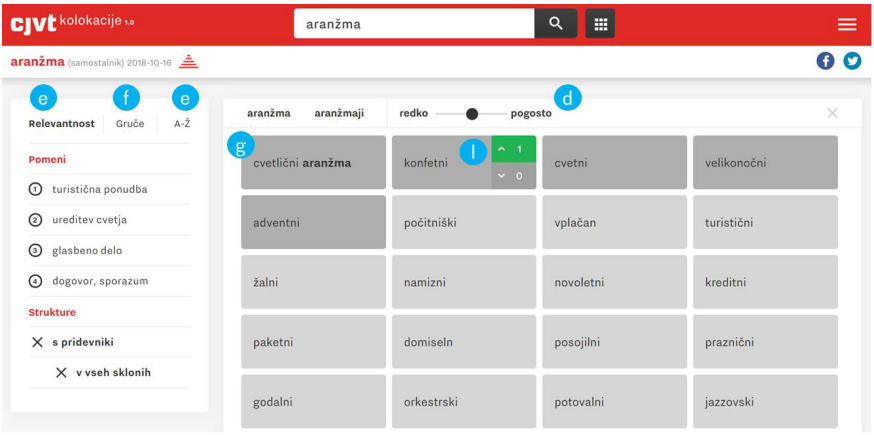
S strani uporabnikov so bili v TAP prepoznani in izpostavljeni predvsem spodaj našteti vmesniški elementi:

a) indikator stopnje gesla
b) pomenska členitev
c) gumb Več
d) pogostnostni filter
e) abecedno in relevantnostno razvrščanje
f) gruče
g) barvna lestvica
h) povezava Gigafida
i) druge povezave
j) zgledi
k) meni
l) uporabniško ocenjevanje

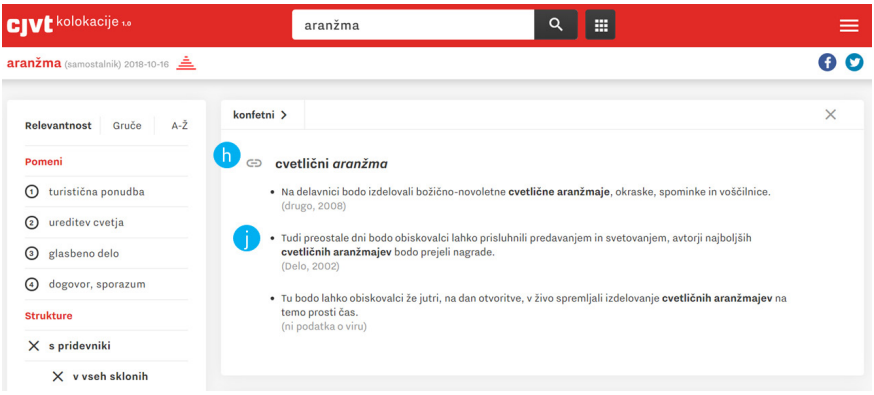
Slike 1, 2 in 3 kažejo elemente v vmesniku KSSS (na primeru gesla *aranžma*).



Slika 1: Izhodiščna stran iztočnice *aranžma* v KSSS.



Slika 2: Razdelek s kolokacijami s strukturo pridevnik + samostalnik za iztočnico *aranžma* v KSSS.



Slika 3: Zgledi za kolokacijo *cvetlični aranžma* v KSSS.

2.4 Analiza uporabniških mnenj

V nadaljevanju opredelimo namen posameznega vmesniškega elementa in podamo izsledke, kako so se sodelujoči (po uporabniških skupinah) opredeljevali do njega. Za lažji kvantitativni pregled so bile posamezne izjave sodelujočih kategorizirane glede na to, ali v zvezi z elementom sodelujoči vidi možne probleme (P) ali ne (NP). Previdnost je bila potrebna pri kategoriziranju uporabniških izjav

tipa 'ni opazno' ali 'ne uporabljam': če je sodelujoči omenil, da določene funkcije ne uporablja, ali pa je bilo na podlagi analiz posnetkov zaslona mogoče ugotoviti, da je ni opazil, v polstrukturiranem delu raziskave pa se je razkrilo, da sodelujoči funkcijo ocenjuje kot pozitivno, tovrstnih izjav nismo ocenjevali kot problematičnih (P), smo jih pa posebej beležili in jih upoštevali pri celostnem vrednotenju uporabniškega mnenja.

Opozoriti je treba, da je za metodo TAP značilna precejšnja stopnja odprtosti: ker izpostavlja individualno komponento posameznika in dopušča prosto izražanje množice raznolikih, nepredvidljivih mnenj, občutenj, povsem enoznačno kategoriziranje in strnjevanje rezultatov ni možno. Kot je v nadaljevanju razvidno iz umanjkanja številčnih ocen glede določenih kategorij pri posameznih uporabniških skupinah, niso vse skupine zaznale ali izpostavile, predvsem pa vrednotile povsem enakih slovarskih značilnosti. Za nekatere, ne pa vse, elemente smo dodatno oceno pridobili iz drugega dela raziskave, odvisno od izpostavitev, ki so bile vezane tudi na poklicni interes ali dejavnost sodelujočega (učitelj, prevajalec, lektor, jezikoslovec).

Omeniti velja tudi specifičnost skupine jezikoslovcev, ki se je po izražanju mnenj in argumentaciji precej razlikovala od ostalih. Analize so pokazale, da je njihove ocene treba kategorizirati drugače, saj vmesniške elemente pogosto ocenjujejo skozi oči predvidenega uporabnika. Primere izjav, kjer sami nečesa niso občutili kot problem, z vidika uporabnika pa ja, smo zato označili z NP ('ni problematično') in P ('je problematično'), torej z dvojno oznako.

3 Interpretativna analiza rezultatov

3.1 Indikator stopnje gesla

Odzivni slovarji se razvijajo postopoma, zato v trenutno objavljeni različici slovarja niso nujno vsa gesla na enaki stopnji ročne (jezikoslovne) pregledanosti. Uporabniki lahko informacijo o stopnji izdelanosti iskanega gesla hitro in pregledno pridobijo iz indikatorja stopnje gesla, tj. iz ikone piramide v levem zgornjem kotu ekrana.

Polnejša piramida pomeni natančnejšo izdelanost, podrobneje pa je razložena v razdelku O viru, ki opisuje zasnovo slovarja.

Tabela 2: Uporabniške odločitve o indikatorju stopnje gesla.

Uporabniška skupina	NP	P	NP in P
Učitelji SLJ1	9	1	-
Učitelji SLJ2	10	-	-
Lektorji/prevajalci	7	3	-
Jezikoslovci	9	1	-

Večina sodelujočih se je do stopenjskega indikatorja opredelila z oceno NP – prisotnost tega elementa v slovarju podpirajo (Tabela 2), zdi se jim dobrodošla in koristna informacija, ki ponudi hiter podatek o (ne)dokončanosti posameznega gesla, posredno pa podaja informacijo o (ne)prisotnosti pomenske členitve.

> [1] *»Piramida, super, odlično. Na eni točki bi se vprašala, koliko je potem to zdaj verodostojno, če še ni izčiščeno, recimo, če ne bi te piramide videla in ne bi vedela, kaj pomeni.«* (lektorica/prevajalka)

Posameznim sodelujočim iz skupine učiteljev slovenščine kot prvega jezika ter nekaj več prevajalcem/lektorjem in jezikoslovcem (Tabela 2) pa se je glede na izjave zdel element problematičen (ocena P) predvsem zaradi neopaznosti. Da je indikator premalo vizualno oz. grafično izpostavljen ali pa njegova funkcija ni takoj jasna, je potrdila tudi kasnejša analiza posnetkov zaslona, saj ga večina sodelujočih samoiniciativno (brez spodbude vodje intervjuja) ni opazila:

> [2] *»Nisem je opazila. Brez opozorila ne bi bila pozorna.«*
> (učiteljica slovenščine kot prvega jezika)

> [3] *»Mislim, opazila sem jo, nisem pa vedela, kaj, kakšno funkcijo ima. To pa ne gre klikat?«* (lektorica/prevajalka)

3.2 Pomenska členitev

Ker je KSSS odzivni slovar, je zgrajen avtomatsko in v izhodišču vsebuje le avtomatsko izluščene podatke iz korpusa. Iztočnice na začetku zato niso ločene po pomenih, saj je na nivoju avtomatskega luščenja težko razlikovati med pomeni. Pomenska členitev, ki je slovarskim iztočnicam dodana v poznejših stopnjah jezikoslovne izdelanosti, uporabnikom omogoča, da filtrirajo pomene, ki jih zanimajo, in na ta način zožijo nabor za njihov kontekst relevantnih kolokacij; opazujejo lahko npr. samo kolokatorje iztočnice *aranžma* v pomenu "turistična ponudba", ne pa v pomenu "glasbeno delo".

Tabela 3: Uporabniške odločitve o pomenski členitvi.

Uporabniška skupina	NP	P	NP in P
Učitelji SLJ1	10	-	-
Učitelji SLJ2	10	-	-
Lektorji/prevajalci	10	-	-
Jezikoslovci	10	-	-

Kot je razvidno iz Tabele 3, so se do pomenske členitve vsi sodelujoči iz vseh uporabniških skupin opredelili pozitivno (NP): označili so jo kot odlično, zelo uporabno in koristno informacijo o pomenskem potencialu posamezne besede, z logično strukturo in pomensko razmejitvijo.

Posebej so izpostavili, da se jim zdi prisotnost pomenske členitve v slovarju nujna: prevajalcem oz. lektorjem bistveno olajša in pohitri delo, učiteljem slovenščine kot prvega jezika predstavlja dobro izhodišče za pripravo najrazličnejših nalog (npr. razvrščanje kolokacij pod ustrezne pomene ali pa določanje pomena posameznim skupinam kolokacij), učiteljem slovenščine kot drugega/tujega jezika pa se zdi koristna, ker tujcem, ki imajo pogosto težave s prepoznavanjem pomena, omogoča lažjo pomensko orientacijo in omejitev na posamezni pomen:

[4] *»Tole se mi pa zdi odlično. Tole, da je razvrščeno po pomenih ... Sploh za naše tujce, da si lahko tole potem omejijo na to, na ta en pomen.«* (učiteljica slovenščine kot drugega/tujega jezika)

[5] *»Tako je lepo strukturo urejeno, ja. Ker prej sem hotel ravno vprašati, a ne bo nič pomena, tega okvirnega pomena.«*
(učitelj slovenščine kot prvega jezika)

[6] *»To, da si potem en tujec, ki nima občutka, kam zdaj to sodi, lahko izvzame tukaj 'ureditev cvetja' in vidi, kaj je povezano s tem, se mi zdi odlično, res.«* (učiteljica slovenščine kot drugega/tujega jezika)

3.3 Gumb Več

Zaradi velike količine podatkov, ki so na začetku avtomatsko izluščeni po konceptu odzivnega slovarja, je zlasti v primeru kolokacij nemogoče prikazati celotni nabor podatkov na eni strani, zlasti ker se pojavljajo znatne razlike med strukturami (npr. pridevnik + samostalnik), od katerih nekatere vsebujejo le nekaj primerov, ostale pa tudi po več deset. Da imajo uporabniki kljub temu lahko reprezentativen pregled nad podatki, smo razdelke, ki se nanašajo na različne strukture kolokacij, skrajšali, da vsebujejo največ štiri primere, do celotnega nabora pa lahko uporabnik dostopa s klikom na gumb Več (ikona treh vertikalnih pik).

Tabela 4: Uporabniške odločitve o funkciji gumba Več.

Uporabniška skupina	NP	P	NP in P
Učitelji SLJ1	10	-	-
Učitelji SLJ2	10	-	-
Lektorji/prevajalci	10	-	-
Jezikoslovci	8	-	2

Znotraj vseh uporabniških skupin je bilo mogoče identificirati pozitivne ocene (NP): funkcijo tega elementa so sodelujoči ocenili kot jasno in logično, možnost izbire, da lahko ostanejo na osnovnem

nivoju iskanja ali pa preidejo na naprednejšo raven, pa se jim zdi odlična:

> [7] *»Se mi zdi, da je to, da imam tukaj na koncu te tri pikice, torej možnost, da nekaj več izvem, odlična, in zdi se mi dovolj pregledno. Mislim, da vem, kaj se dogaja.«* (lektorica/prevajalka)

Posebej izstopata mnenji dveh jezikoslovcev (Tabela 4), ki sta navajala tako argumente za (NP) kot proti (P). Ob svoji pozitivni opredelitvi, da slovar omogoča tudi naprednejše funkcije in višji nivo iskanja informacij, sta izrazila pomisleke glede določenega (ali večinskega) tipa uporabnikov, ki ostaja na osnovnem nivoju iskanja in mu to pravzaprav zadošča:

> [8] *»Tole, po mojem, je nekaj, kar je super. Ampak v resnici pa določen tip uporabnikov ostaja na zelo osnovnem zgornjem nivoju, pri čemer včasih pravijo, da jim je to dovolj. Mogoče je res, da je dovolj tudi osnovni tip iskanja. Jaz pa mislim, da je mogoče to tudi premalo izpostavljeno, zato da bi zares uporabljal naprednejše funkcije.«* (jezikoslovec)

Obenem pa se zdi element premalo vizualno izpostavljen, da bi uporabnika pritegnil k uporabi naprednejših funkcij. Na to ugotovitev je opozorila tudi kasnejša analiza posnetkov zaslona, saj je nekaj sodelujočih element zaznalo šele ob spodbudi vodje intervjuja.

3.4 Pogostnostni filter

Pogostnostni filter uporabnikom tako kot nekatere že prej opisane slovarske funkcije omogoča, da zožijo nabor velikega števila avtomatsko izluščenih podatkov in se osredotočijo samo na določen izsek, ki je relevanten za njihov namen. V tem primeru filter omogoča, da uporabnik izloči tiste besede, ki se pojavljajo bodisi (zelo) redko bodisi (zelo) pogosto. To je zlasti pomembno za učenje slovenščine kot tujega jezika, saj s pogostnostnim filtrom lahko dobimo npr. le tiste tipične kolokacije, ki vsebujejo kolokatorje, ki se v jeziku pojavljajo najpogosteje. Če filtriramo npr. kolokatorje iztočnice *aranžma*

v strukturi pridevnik + samostalnik, dobimo pri zelo pogostem besedišču *počitniški aranžma* (lema *počitniški* ima približno 25.000 zadetkov v korpusu Gigafida 2.0), pri zelo redkem pa *paketni aranžma* (lema *paketen* ima le okrog 2.500 zadetkov, torej približno desetkrat manj).

Tabela 5: Uporabniške odločitve o pogostnostnem filtru.

Uporabniška skupina	NP	P	NP in P
Učitelji SLJ1	9	1	-
Učitelji SLJ2	9	1	-
Lektorji/prevajalci	10	-	-
Jezikoslovci	8	-	2

Večji del sodelujočih iz vseh uporabniških skupin (Tabela 5) se je do kategorije opredelil pozitivno (NP), predvsem z argumentom, da gre za bistveno slovarsko funkcijo, ki uporabniku podaja splošno informacijo o pogostosti določenih besed (v jeziku) in mu pri tem ni treba sklepati po (individualnem in zato pogosto regionalno pogojenem) občutku. Nekaj jih je ob svoji pozitivni oceni podalo tudi predloge o vizualni nadgradnji (gl. razdelek o uporabniških predlogih 3.13).

> [9] *»Ja, to je super. Ker potem lahko jaz tudi pogledam, vidim, kaj je pogosto, kako pogosta je kolokacija v jeziku, ne glede na to, da živim pač v Ljubljani in poznam le osrednji jezik .« (prevajalka)*

Dva jezikoslovca sta ob svoji pozitivni opredelitvi izrazila pomisleke glede (ne)jasnosti logike delovanja filtra. Da je namen in samo delovanje filtra za uporabnika lahko precej konfuzno, saj ni jasno, ali se nanaša na prikazovanje pogostosti kolokatorja ali kolokacije, potrjujeta tudi dve negativni oceni učiteljev (Tabela 5).

> [10] *»Nisem pa siguren zdaj tukaj, redko, pogosto, dilema po moje, da ne veš točno, ali je to povezano s kolokacijo ali s kolokatorjem. Mislim, jaz vem, a ne, ampak če se poskušam postaviti v kožo nekoga, ki se s tem še ni srečal, je tu po moje malo nejasno.«* (jezikoslovec)

[11] *»Aja, tako! Aja, to bi si jaz glih obratno razlagal.«* (učitelj slovenščine kot drugega/tujega jezika)

[12] *»Kaj to pomeni, se pravi, da so te besede, besedne zveze precej v souporabi?«* (učitelj slovenščine kot prvega jezika)

3.5 Abecedno in relevantnostno razvrščanje

Načina razvrščanja po relevantnosti in abecedi uporabnikom omogočata, da kolokacije v določeni strukturi razvrstijo glede na tipičnost, tj. od najznačilnejše kolokacije do najmanj tipične, oz. po abecednem vrstnem redu kolokatorjev, z začetkom pri kolokatorjih, ki se začnejo na a-. Tipičnost je pri razvrščanju po relevantnosti še dodatno nakazana z barvno skalo (gl. razdelek 3.7).

Tabela 6: Uporabniške odločitve o abecednem razvrščanju.

Uporabniška skupina	NP	P	NP in P
Učitelji SLJ1	2	-	-
Učitelji SLJ2	3	-	-
Lektorji/prevajalci	4	-	-
Jezikoslovci	5	-	1

Tabela 7: Uporabniške odločitve o relevantnostnem razvrščanju.

Uporabniška skupina	NP	P	NP in P
Učitelji SLJ1	1	-	-
Učitelji SLJ2	1	-	-
Lektorji/prevajalci	1	-	-
Jezikoslovci	-	-	1

Manjši delež sodelujočih iz vseh uporabniških skupin, kot je prikazano v Tabelah 6 in 7, je samoiniciativno podal pozitivno mnenje (NP) tudi o možnosti uporabe ostalih dveh filtrov, abecednega in relevantnostnega. Da z zaznavanjem in razumevanjem delovanja filtra niso imeli težav, so potrdile tudi analize posnetkov zaslona oz.

implicitno podanih informacij, medtem ko so le-te pri ostalih sodelujočih, ki se do filtrov niso eksplicitno opredelili, pokazale, da ju niso opazili ali pa so ju preprosto ignorirali oz. za obravnavo niso izkazali interesa.

[13] *»Okej, filter A–Ž mi je zelo hitro jasen.«*
(učiteljica slovenščine kot prvega jezika)

[14] *»Aha, tukaj pa nimamo več informacije o pogostosti. Ampak, saj jo tam dobimo. Če hočem, pa potem grem nazaj. Če nekaj najdeš, prehajaš potem med filtri. Dobro. Odlično, po abecedi.«*
(učitelj slovenščine kot drugega/tujega jezika)

Ena jezikoslovka je ob pozitivni opredelitvi, da je relevantnostni filter dobrodošla slovarska funkcija, izrazila tudi pomislek glede (ne)dojemanja razlike med relevantnostjo in pogostnostjo ter s tem vprašljivost razločevanja med funkcijo relevantnostnega in pogostnostnega filtra:

[15] *»Drugim uporabnikom mogoče ne bo takoj jasno, kaj je to relevantnost. Če imamo že na eni strani podatek o pogostnosti, kakšna je pa zdaj povezava med relevantnostjo in pogostnostjo. To mogoče ni čisto intuitivno.«* (jezikoslovka)

3.6 Gruče

Tako kot razvrščanje po relevantnosti in abecedi tudi funkcija gruč uporabnikom omogoča, da spreminjajo vrstni red kolokatorjev, a so pri tej funkciji besede razporejene v različne skupine, in sicer glede na to, kako podobni (po avtomatski metodi) so si konteksti, v katerih se kolokatorji pojavljajo. Če uporabnik gruči npr. kolokatorje iztočnice *aranžma*, dobi v eni skupini *godalni aranžma*, *orkestrski aranžma* in *jazzovski aranžma*, v drugi pa *velikonočni aranžma*, *adventni aranžma* in *praznični aranžma*. Na ta način je mogoče pridobiti pomensko podobne skupine kolokatorjev, še preden je iztočnica v celoti jezikoslovno pregledana in opremljena s pomensko členitvijo.

Tabela 8: Uporabniške odločitve o gručah.

Uporabniška skupina	NP	P	NP in P
Učitelji SLJ1	9	1	-
Učitelji SLJ2	9	1	-
Lektorji/prevajalci	9	1	-
Jezikoslovci	9	1	-

Skoraj vsi sodelujoči (Tabela 8) so se do kategorije opredelili s pozitivno oceno (NP). Lektorji/prevajalci in jezikoslovci menijo, da gre za precej relevantno in dragoceno slovarsko funkcijo, ki podaja informacijo o tematskih sklopih oz. pomenskih razsežnostih same iztočnice ali pomenskem potencialu kolokacije. Še posebej se jim zdi to koristno pri nedokončanih geslih s posledično odsotnostjo pomenskih členitev.

[16] *»Skoraj kot nekakšne sopomenske zveze so te. Ja, je pregledno, ja.«* (učiteljica slovenščine kot prvega jezika)

[17] *»Pogrupira jih glede na čas. 'Ugoden aranžma', glede na ceno, bi se reklo. 'Počitniški', 'turistični', glede na vrsto, ane. Potem pa glede na pridevnike: 'lep', 'domiseln', 'nov', ta, ta, ta, 'cvetlični', glede na vrstne pridevnike, očitno. Kakovostne, vrstne: 'žalni', 'nagrobni', jej. 'Kreditni', 'posojilni', to je bančništvo. Aha, po pomenskem polju. To je pa odlično!«* (lektorica/prevajalka)

[18] *»Potem vidim tako lepo skupino in potem pridem do zgledov, kjer bom rekel, ne vem, 'izvrstna izbira' in mi potem pokaže ta kontekst in lahko potem še celo precej hitro gledam 'izvrstna', 'kakovostna', ko se premikam tukaj zgoraj med kolokatorji. No, in moram reči, da v tem, v tem je meni precej koristno.«* (jezikoslovec)

Kot potencialno problematično so oboji izpostavili dolžino seznama kolokacij oz. količino kolokatorjev v vsaki gruči, predvsem jezikoslovcem bi se z vidika informiranja uporabnika zdelo ustreznejše, če bi bili na osnovni ravni navedeni le reprezentativni predstavniki posamezne skupine/gruče oz. bi bila uporabniku ponujena

možnost prilagoditve izbora semantično bolj ali pa semantično manj relevantnih kolokatorjev.

> [19] *»Tukaj je zelo ključno, koliko prostora dobi ena taka gruča, ne, to je eno, in drugo, kakšen je vrstni red. Če imam gruče zelo lepo oblikovane in imam samo prvih par kolokatorjev v vsaki gruči tukaj napisanih, kar pomeni, da gručo potem odprem, če me res zanima, ne, po moje bi bilo to z vidika informiranja uporabnika veliko boljše.«* (jezikoslovec)

Iz vsake skupine pa je zgolj po en sodelujoči menil (Tabela 8), da funkcija oz. namen te kategorije ni (takoj) jasen in logičen, tudi v razmerju do ostalih filtrov ali elementov razvrščanja.

> [20] *»Kaj je to v bistvu? 'Roka vodje', 'roka Slovenca'? A, gor se moram postaviti, da vidim, ja, okej. Zdaj jaz ne bi vedela, kaj početi najprej. In potem bi me malo zmedlo, ker so tule spodaj pa sive, tiste besede, ki so bolj pogoste.«*
> (učiteljica slovenščine kot drugega/tujega jezika)

Analize posnetkov zaslona v sklopu TAP so pokazale, da je z zaznavanjem kategorije večina sodelujočih imela težave in je brez spodbude vodje intervjuja ob prvem stiku s slovarjem niti ne bi opazila.

3.7 Barvna lestvica

Pri razvrščanju po relevantnosti ali v gruče je uporabniku na voljo tudi barvna informacija o tipičnosti kolokacije – temnejši odtenki sive nakazujejo bolj tipično, svetlejši pa manj tipično kolokacijo.

Tabela 9: Uporabniške odločitve o barvni lestvici.

Uporabniška skupina	NP	P	NP in P
Učitelji SLJ1	9	-	1
Učitelji SLJ2	10	-	-
Lektorji/prevajalci	7	3	-
Jezikoslovci	7	2	1

Večina sodelujočih (Tabela 9) se je do kategorije opredelila pozitivno (NP) in predvsem z argumenti, da je njena funkcija jasna in intuitivna, kar so potrdile tudi implicitne informacije, pridobljene na podlagi analiz posnetkov zaslona.

[21] *»Fino. To sivino opazim, vidim, da so različne sivine, ane. In mislim, da pomeni glede na pogostost, temnejše so pogostejše.«* (učiteljica slovenščine kot drugega/tujega jezika)

Dvema jezikoslovcema se zdi problematična (P) z vidika premalo intenzivnih razlik v sivinah in posledično nezmožnosti mentalnega procesiranja, en jezikoslovec in učiteljica slovenščine kot prvega jezika pa sta ob svoji pozitivni opredelitvi izpostavila morebitno problematičnost glede prevelike količine ali kompleksnosti podatkov, ki so uporabniku ponujeni na enem mestu in mu zato (lahko) ne posredujejo relevantne informacije.

[22] *»Te barve zdaj, ko jih gledam, mislim, nisem ziher, mislim, če jih sploh mentalno kaj dosti procesiram.«* (jezikoslovec)

[23] *»To ja, sivine, ja, ampak v resnici me malo moti v tem smislu, da torej preveč dobim tega, teh podatkov in moram potem razmišljati, v resnici si pa tega ne želim.«*
(učiteljica slovenščine kot prvega jezika)

Temu mnenju se pridružujejo tudi negativne ocene (P) treh prevajalcev/lektorjev (Tabela 9), ki so izpostavili potrebo po možnosti prilagoditve količine podatkov, postopnega širjenja ali ožanja izhodiščnega izbora semantično najbolj relevantnih kolokatorjev, pa tudi možnost podatka o tipičnih kolokacijah že na prvi strani slovarja (gl. tudi razdelek 3.13).

[24] *»Kot uporabniku se mi zdi ta drugi klik preveč. Zato ker je en korak več. Jaz, ko vtipkam, želim čim več informacij takoj, tudi o pogostosti. In na prvi strani, še preden grem v strukturo.«*
(lektorica/prevajalka)

3.8 Povezava Gigafida

Ena od značilnosti odzivnega slovarja je tudi povezljivost z drugimi jezikovnimi viri, kot so npr. korpusi, kar omogoča uporabnikom, da se posvetujejo z večjo količino zgledov iz realne jezikovne rabe. V KSSS ima vsaka kolokacija na voljo tudi povezavo na korpus Gigafida, ki pripelje na konkordance, v katerih se pojavi iskana kolokacija. To omogoča tudi lažje presojanje, ali je določen podatek v slovarju, ki se uporabniku zdi sumljiv, morda šum oz. napaka zaradi avtomatskega procesiranja podatkov.

Tabela 10: Uporabniške odločitve o povezavi na korpus Gigafida.

Uporabniška skupina	NP	P	NP in P
Učitelji SLJ1	8	-	-
Učitelji SLJ2	8	-	-
Lektorji/prevajalci	8	-	-
Jezikoslovci	6	1	1

Vsi sodelujoči, ki so se do te kategorije opredelili, so podali pozitivno oceno (NP). Prisotnost konteksta, možnost vpogleda v širok nabor konkretnih primerov rabe za posamezno iztočnico ali kolokacijo, se jim zdi zelo uporabna in koristna.

> [25] *»Mi je pa všeč, mislim to se mi zdi super, ker potem tudi mi, ko smo začeli delati z Gigafido recimo s tečajniki, smo vedno rekli, če ne veste točno, kako bi se neka beseda uporabljala v slovenščini, poglejte si ostale zadetke, videli boste dejansko rabo, v bistvu ali besedne zveze ali besede v nekem kontekstu, pač to je super.«*
> (učiteljica slovenščine kot drugega/tujega jezika)

Analize posnetkov zaslona so pokazale, da je z zaznavanjem ikone večina sodelujočih imela težave in je brez spodbude vodje intervjuja ob prvem stiku s slovarjem ne bi opazila.

3.9 Druge povezave

Odzivni slovarji kot jezikovni viri, ki so zasnovani za in namenjeni izključno digitalnemu mediju, poleg že omenjenih povezav na korpusne vire (razdelek 3.8) vsebujejo tudi povezave na druge jezikovne vire in (ne)jezikovne spletne strani.

Tabela 11: Uporabniške odločitve o drugih povezavah.

Uporabniška skupina	NP	P	NP in P
Učitelji SLJ1	3	-	-
Učitelji SLJ2	-	-	-
Lektorji/prevajalci	4	-	-
Jezikoslovci	1	-	1

Nekaj sodelujočih, kot je prikazano v Tabeli 11, je samoiniciativno podalo pozitivno mnenje (NP) tudi o možnosti uporabe ostalih povezav, ki jih vključuje in ponuja slovar, npr. Facebook, Twitter ali hiter dostop do ostalih slovarskih in korpusnih virov Centra za jezikovne vire in tehnologije Univerze v Ljubljani prek t. i. gumba za preklop, npr. Sopomenke, Gigafida 2.0 in Sloleks 2.0.

> [26] *»Twitter, Facebook, Sopomenke – super.«*
> (učiteljica slovenščine kot prvega jezika)

Ob eksplicitno izraženih pozitivnih opredelitvah pa so kasnejše analize posnetkov zaslona in (sočasnega) glasnega razmišljanja izpostavile predvsem pomisleke glede (ne)intuitivnosti odprtja in zaprtja okna z viri s klikom na gumb za preklop (namesto s klikom na X).

> [27] *»Kako zdaj to okno s povezavami izgine, a ne ...? Ja, kako to zdaj izgine? No, da vidimo.«* (lektorica/prevajalka)

3.10 Zgledi

V smislu zgledov rabe poleg korpusnih konkordanc, do katerih lahko uporabnik pride s pomočjo povezave na korpus Gigafida, vsebuje

KSSS tudi t. i. dobre zglede, izluščene iz korpusa po sistemu GDEX (Kosem idr. 2011), tj. avtomatsko izluščene povedi, v katerih nastopa kolokacija in ki izpolnjujejo določene kriterije (vsaj en glagol v povedi, brez tujejezičnih besed in URL-naslovov ipd.). Ta izbor uporabnikom ponudi do štiri primere, v katerih je kolokacija uporabljena, in s tem olajša iskanje ustreznega zgleda.

Tabela 12: Uporabniške odločitve o zgledih.

Uporabniška skupina	NP	P	NP in P
Učitelji SLJ1	10	-	-
Učitelji SLJ2	9	1	-
Lektorji/prevajalci	10	-	1
Jezikoslovci	10	-	-

Možnost vpogleda v korpusne zglede oz. ponazoritve z zgledi se zdi skoraj vsem sodelujočim (Tabela 12) dragocena, saj lahko razreši in razdvoumi marsikatero nejasnost, učencem slovenščine kot prvega jezika pomaga pri razvijanju sporočanjske in slogovne zmožnosti, konkretizacija izraza pa je še posebej koristna za učence tujce:

> [28] *»Ravno ti primeri, to mi je najbolj super, ker to sem zelo zelo pogrešala, ja, ker tega je v bistvu v SSKJ-ju zelo malo, ampak tu pa res, v bistvu iz enega gesla dobiš ven ogromno informacij, super, tako da res najdeš karkoli že rabiš – res se mi zdi to zelo uporabna zadevščina.«* (učiteljica slovenščine kot prvega jezika)

> [29] *»Iz konteksta je pa potem vse hitro razvidno, ane. Mislim, v SSKJ-ju to ne, ne bi bilo možno.«* (učiteljica slovenščine kot prvega jezika)

Ena lektorica/prevajalka je ob svoji pozitivni opredelitvi izpostavila tudi pomislek glede (ne)ustreznosti zgledov zaradi njihove nelektoriranosti in morebitnih pravopisnih napak, en učitelj slovenščine kot drugega/tujega jezika pa je menil, da niso dovolj jasni in reprezentativni in da bi bilo treba v nabor vključiti tudi zglede iz strokovnih, ne le splošnih ali publicističnih virov.

[30] *»Pri zgledih mi je všeč, da je neveden vir, da lahko vidim, ali je bila stvar lektorirana. Na podlagi tega lahko tudi ugotovim, pač na podlagi virov lahko ugotovim, ali je šlo za preverjeno ali nepreverjeno besedilo. In tudi če je bilo preverjeno, so včasih lektorji veliki verniki in so vsi podali obliko, ki morda ni ustrezna. Super, če dobim to, vire, potem pa mi je v redu. Ne rabim iti potem gledat v drug slovar recimo. Edino, kar bi jaz v bistvu že pri samih korpusih zelo pazila, od kod črpajo, ane. Če bi se dalo, bi veliko več pač pregledanih in strokovnih virov tukaj noter vključevala.«* (lektorica/prevajalka)

[31] *»Tole je fajn, ker res v kontekst umesti. To mi je res všeč. Potem všeč mi je, da je naveden vir 'Mladina'. Ker vir se mi zdi pomemben tudi za nek občutek verodostojnosti, da potem lažje izbiraš ali pa vidiš res, kje se pojavlja.«*
(učiteljica slovenščine kot drugega/tujega jezika)

Analize posnetkov zaslona v kombinaciji z glasnim razmišljanjem so pokazale, da z zaznavnostjo in uporabo te funkcije večina sodelujočih tudi ni imela težav.

3.11 Meni

Kot že omenjeno v razdelku 3.3, je količina podatkov, ki so na voljo pod iztočnico v KSSS, lahko zelo obsežna (kljub skrajšanim razdelkom, ki ponazarjajo različne kolokacijske strukture in vsebujejo največ štiri kolokacije). S pomočjo menija ima uporabnik možnost, da podatke še dodatno filtrira in s seznama struktur izlušči tiste, ki ga zanimajo, namesto da ročno pregleduje vsak razdelek posebej. Tako lahko npr. s pomočjo menija nastavi, da želi le strukture s pridevniškimi kolokatorji in predlogom 'z' (npr. *okrašen z aranžmajem*) ali s samostalniškimi kolokatorji v rodilniku (*aranžma iz vejevja*, *aranžma iz cvetja*).

Večini sodelujočih (Tabela 13) se zdi struktura menija jasna in pregledna, lektorji/prevajalci so posebej izpostavili odličnost možnosti klasifikacije po besednih vrstah, učitelji slovenščine kot drugega/tujega jezika pa so izrazili navdušenost nad vključenostjo predlogov, z rabo katerih imajo tujci pogosto največ težav, npr. *potovati na [Hrvaško, Kitajsko],* vendar *potovati v [Evropo, Azerbajdžan]*.

Tabela 13: Uporabniške odločitve o meniju.

Uporabniška skupina	NP	P	NP in P
Učitelji SLJ1	9	1	-
Učitelji SLJ2	9	1	-
Lektorji/prevajalci	8	2	-
Jezikoslovci	4	1	5

[32] *»Všeč mi je ta razdelitev, sploh to, da je možno še s skloni povezati, pa tole s predlogi, to se mi zdi ena od boljših možnosti.«* (učiteljica slovenščine kot drugega/tujega jezika)

Ob pozitivnih opredelitvah so kasnejše analize posnetkov zaslona in (sočasnega) glasnega razmišljanja pri jezikoslovcih izpostavile pomisleke glede kompleksnosti in vprašljivosti razumevanja logike delovanja samega filtra. Večji delež jezikoslovcev (Tabela 13) meni, da slovnično izkušen uporabnik z razumevanjem in jasnostjo delovanja filtra ne bi smel imel težav, širši krog uporabnikov pa bi morali ustrezne/pravilne uporabe filtra naučiti oz. jim nuditi dodatno pomoč bodisi v opisni, glasovni ali video obliki; npr. s posnetki, ki na kratko predstavljajo osnovne značilnosti menija ali specifične primere dela z menijem, iskanj po slovarju.

En učitelj slovenščine kot prvega jezika je izpostavil, da se mu zdi logika vklop-izklop izbranega elementa v meniju moteča in pogreša možnost intuitivnejšega preklikavanja med posameznimi pomeni in strukturami.

[33] *»Jaz bi rad kar direktno preklopil nazaj, ne da moram zdaj verjetno tole … aha, tako izklopiti, ja.«* (učitelj slovenščine kot prvega jezika)

Dva učitelja, dva prevajalca in ena jezikoslovka so tudi opozorili na dvojno logiko simbola X ('izberem/potrdim' ali 'ne izberem/ne potrdim'), ki je lahko (za razliko od simbola kljukice) za marsikaterega uporabnika pogosto moteča ali celo konfuzna.

[34] *»Se pravi, jaz sem tukaj zdaj nekaj izklopila ali kaj? A X pomeni …?«* (učiteljica slovenščine kot prvega jezika)

[35] *»Ne vem, ali bi bila bolje kljukica. Saj mogoče z X ni slaba, ko poštekaš to, ko pač začneš to uporabljati. Zdaj, ko imam tu izbiro, rečem: 'aha, tega imam', prej pa v bistvu, ko mi zapre meni, pa ga imam prečrtanega, pa nisem bila stoprocentna.«*
(učiteljica slovenščine kot drugega/tujega jezika)

3.12 Uporabniško ocenjevanje

Odzivni slovar kot izhodiščno avtomatsko zgrajen jezikovni vir vsebuje določeno količino šumnih podatkov, ki najpogosteje nastanejo kot posledica napak pri predhodnem procesiranju besedil, npr. tokenizaciji, lematizaciji in oblikoskladenjskem označevanju (Pori in Kosem 2021), iz katerih so podatki izluščeni. Odzivni jezikovni viri iz tega razloga v razvoj vključujejo tudi uporabniško skupnost, ki lahko s svojim znanjem prispeva k čiščenju podatkovne baze. Pri KSSS imajo uporabniki možnost, da kolokacijo oz. kolokator označijo kot ustrezno (+1) ali neustrezno (–1) ter tako opozorijo na morebitne napake. Njihova mnenja se upoštevajo pri nadgradnji slovarja.

Tabela 14: Uporabniške odločitve o možnosti uporabniškega ocenjevanja.

Uporabniška skupina	NP	P	NP in P
Učitelji SLJ1	8	2	-
Učitelji SLJ2	7	3	-
Lektorji/prevajalci	4	6	-
Jezikoslovci	1	-	9

Pri opredelitvah do možnosti uporabniškega sodelovanja smo identificirali več razhajanj v mnenjih (Tabela 14), in sicer se ocenjevanje kolokacij zdi vsem sodelujočim koristna in dobrodošla informacija, vendar pa so predvsem prevajalci in lektorji izpostavili, da za to pogosto nimajo časa, posamezni učitelji imajo pomisleke, da bi se te funkcije posluževali nekompetentni uporabniki, jezikoslovcem pa se zdi nujna nadgradnja in osmislitev funkcije:

[36] *»Tule imam zelo mešane občutke glede tega. Če bi računali, da bodo to uporabljali samo bolj zahtevni uporabniki, potem se mi to zdi super varianta. Če pa pomislim, da bi tole dala v osnovni šoli otrokom in bi oni tam malce klikali pa se malo igrali, pa po moje lahko kar precej pokvarijo to situacijo tukaj.«*
(učiteljica slovenščine kot prvega jezika)

[37] *»Ja, to se mi zdi vsekakor super, ker jaz velikokrat marsikje opazim kakšno napako pa še kdo drug tudi, ne, od tega, ko berem kakšno spletno stran z novicami pa vidim, da je kaj narobe napisano, ne. Se mi pa nikoli ne ljubi registrirati, da bi potem opozarjal na napako, ne. Mislim, če bi lahko že to naredil, bi včasih to naredil. Tako da to je fino, da je tu tako narejeno, da verjetno uporabnik to lahko takoj opozori.«* (učitelj slovenščine kot drugega/tujega jezika)

[38] *»Pozitivno, ja, edino to, da bi bilo mogoče dobro uporabnike še malo bolj spodbuditi, v tem smislu, da se nekaj zgodi ali pa da imajo možnost nekaj delati s temi ocenami. Tako kot pri Sopomenkah, kjer lahko razvrščajo po njih, a ne ...«* (jezikoslovec)

3.13 Predlogi sodelujočih za konkretno izboljšavo slovarskih funkcij

Ob opredeljevanju do posameznih vmesniških kategorij so uporabniki samoiniciativno predlagali tudi nabor izboljšav, ki se je v manjšem deležu navezoval na nekaj specifičnih predlogov, kot je dodatna informacija o frekvenci kolokatorja oz. kolokacije, možnost izvoza podatkov, opremljenost iztočnic (še posebej homonimnih) z naglasi in izgovorjavo, možnost klika na iztočnico za vrnitev na izhodiščno stran. Večji delež ostalih predlogov, ki jih s kratkimi opisi povzemamo in izpostavljamo v Tabeli 15, pa se je nanašal predvsem na vizualno nadgradnjo posameznih elementov vmesnika, kot je npr. nadgradnja pogostnostnega filtra z barvno lestvico oz. barvnim trakom, grafična izpostavitev ikone piramide v smislu možnosti povečave, intenzivnejših barv ali črt ter dodatne opremljenosti s krajšim naslovom ali opisom, da bi bila v vmesniku opaznejša in da bi bila njena funkcija bolj prepoznavna ipd.:

[39] *»Meni je že na začetku super ta piramida. Ampak jaz bi že na začetku imel piramido, v kateri bi bile vsaj te črtice debelejše, močnejše. Pa sem vedel, zakaj je tam, pa še vedno bi rad imel močnejšo. Ja, tako, da se gor postaviš in dobiš oblaček ali se ti izpiše … Nekaj se mora zgoditi, da te opozori na to.«* (jezikoslovec)

[40] *»Ali pa, da bi vsaj link dali k piramidi, ker ne bodo šli gledat, verjetno, tistega zavihka O viru. Skoraj zagotovo ne. Ampak ja, mogoče bi morala biti res bolj debela piramida.«* (lektorica/prevajalka)

[41] *»Zdaj samo razmišljam, filter redko-pogosto … Mislim, saj mi je logično, da je temna nekako najpogostejša. A bi bilo smiselno tu za črtico spodaj dati temno sivo, srednje pa svetlejšo sivo kot neko, bi rekla, pasico, ki bi se mogoče … Barvni trak, mogoče. Ne vem, to sem zdaj pomislila, ni pa nujno. Se mi zdi, da mogoče, ali pa kaj jaz vem, da bi se barva pike spreminjala. Ne vem, pač bolj ko grem sem, da je temna, bolj ko grem sem, da je svetla. Recimo, mogoče, da se začnem čuditi, zakaj. Pa mi neko informacijo mogoče da.«* (učiteljica slovenščine kot drugega/tujega jezika)

[42] *»Ja, to bi bilo fajn, da bi bilo pač, verjetno vsak ne prepozna tega znaka, ne, da bi bilo napisano Gigafida.«*
(učiteljica slovenščine kot drugega/tujega jezika)

Tabela 15: Nabor uporabniških predlogov za izboljšavo vmesniških elementov.

Vmesniški elementi	Uporabniški predlogi
Indikator stopnje gesla	• grafična izpostavitev in vizualna nadgradnja (odebelitev črt, premik na sredino, dodaten naslov ali kratek napis, možnost povečave)
Pomenski meni	• možnost preklikavanja med pomeni (namesto logike vklop-izklop) • možnost hkratne izbire dveh pomenov
Gumb Več	• grafična izpostavitev ali nadomestitev s puščicami
Pogostnostno razvrščanje	• vizualna nadgradnja (z barvnim trakom oz. barvno lestvico ali spreminjanjem barve pike)
Barvna lestvica	• informacijo ponuditi že na izhodiščni strani • podati informacijo o najbolj tipični strukturi • možnost prilagoditve nabora kolokatorjev

Vmesniški elementi	Uporabniški predlogi
Gruče	• možnost prilagoditve iskanja: skrajšanje seznama kolokacij in vpeljava gumba za širjenje ali ožanje (izhodiščnega) nabora kolokatorjev oz. kolokacij
Povezava Gigafida	• k ikoni dodati napis Gigafida
Druge povezave	• intuitivno zaprtje gumba za preklop med viri s klikom na X
Zgledi	• vključitev zgledov iz strokovne literature
Vizualna podoba	• ocenjevalni kvadratek naj ne prekriva kolokatorja (kjer je pomembna oblika kolokatorja, je glasovanje sekundarnega pomena) • vpeljava intenzivnejših razlik v sivinah
Ostalo	• možnost izvoza podatkov • namesto ponovnega klika lupe pri iskalnem oknu (za vrnitev na izhodiščno stran) vpeljava gumba za klik na iztočnico

4 Ocena metode in nadaljnji razvoj slovarja

Uporabniška evalvacija Kolokacijskega slovarja sodobne slovenščine 1.0 je identificirala odnos do značilnosti, zajetih v prvem (nevodenem in neusmerjenem) tematskem sklopu raziskovalnega intervjuja, ki je temeljil na metodi glasnega razmišljanja. Ocene sodelujočih so bile v veliki meri pozitivne. Pri preverjanju odnosa uporabnikov do inovativnih funkcij slovarskega vmesnika se je pokazalo, da so le-te za kakovostno obravnavo podatkov in učinkovito delo s slovarjem nepogrešljive in dragocene, kar je tudi enotno mnenje vseh uporabniških skupin, ki so sodelovale v predstavljeni kolokacijski uporabniški raziskavi.

Evalvacija je tudi pokazala, katere funkcije slovarskega vmesnika se zdijo uporabnikom ustrezne in koristne ter katere so potrebne premisleka, nadgradnje ali dodatne obravnave. Skladno s predvidevanji se je pokazalo, da so mnenja skupine jezikoslovcev pogosto nekoliko drugačna od mnenj uporabnikov, da se njihovi pomisleki drugim uporabniškim skupinam ne zdijo nujno problematični in na drugi strani ne zajemajo vsega, kar so kot težavo izpostavili uporabniki.

Kot posebej pozitivno so uporabniki izpostavili jasnost delovanja in priročnost posameznih filtrov, prisotnost pomenske členitve, vizualne informacije o stopnji izdelanosti gesla in še zlasti možnost vpogleda v korpusne zglede oz. konkretizacijo jezikovnega izraza v kontekstu realne rabe. Da so posamezne funkcije potrebne vizualne nadgradnje v smislu boljše vpadljivosti in zagotavljanja lažje (uporabniške) zaznavnosti, pa so večinoma razkrili in potrdili šele rezultati analiz posnetkov zaslona (oz. implicitno posredovanih informacij) v sklopu TAP, ki so se izkazali kot izredno pozitivno dopolnilo zvočnim posnetkom (eksplicitno podanim informacijam), saj so razdvoumili marsikatero nejasnost izrečene izjave ali pa razrešili neskladje med (prvotno) podano implicitno in (kasnejšo) eksplicitno informacijo. Metoda TAP v kombinaciji s posnetki zaslona je omogočila izris celotne podobe uporabniškega mnenja ter presojanje skladnosti med verbalno in neverbalno posredovano informacijo, ki je, kot se je potrdilo, v svoji sporočilnosti in izraznosti mnogokrat zgovornejša in na pomembnih mestih podpira, dopolnjuje pomen izrečenega. V tem smislu izbran metodološki postopek ocenjujemo kot zanesljiv, zelo uporaben in uspešen.

Izkušnja z izvedbo evalvacije je razkrila kar nekaj pozitivnih ugotovitev, obenem pa tudi možnosti za izboljšavo metodologije v primeru nadaljnjih podobnih raziskav. Čeprav je zbiranje evalvacijskih ocen, njihovo zapisovanje, kategorizacija in vrednotenje mnenj izredno zamudno, izvajanje poglobljenega intervjuja z glasnim razmišljanjem pa zaradi odprtosti in svobode, ki jo dopušča intervjuvancu, za izvajalca precej zahtevno, so zapisana mnenja zelo dragocena, saj ponujajo možnost uvida v probleme in rešitve, ki pomembno dopolnjujejo razmisleke razvijalcev slovarja.

5 Zaključek

Kvalitativna analiza uporabniške evalvacije vmesnika KSSS se je izkazala za zelo učinkovit način detekcije (ne)problematičnih vmesniških elementov ter predstavlja dobro izhodišče za nadaljnja prizadevanja izboljševanja in nadgradnje kolokacijskega vmesnika v smislu večje

uporabniške prijaznosti in funkcionalnosti. Predstavlja model evalviranja in opredelitev uporabniških problemov, pridobljene ugotovitve, ki razkrivajo tudi možnosti za izboljšavo metodologije, pa bodo koristne za primerljive leksikografske uporabniške raziskave in analize.

Povratne informacije uporabnikov, pridobljene v tej raziskavi, bodo uporabljene pri pripravi naslednje različice KSSS. Zlasti bodo obravnavani vidiki kakovosti, jasnosti in funkcionalnosti posameznih elementov. Glavne načrtovane izboljšave so:

- kakovostnejši nabor kolokacij na prvi strani gesla, kar vključuje rešitve, kot sta npr. več vrstic za kolokacije bolj tipičnih struktur in prikaz le prvih (najbolj tipičnih) kolokacij iz različnih gruč;
- številne (vizualne) nadgradnje oz. izboljšave grafičnih elementov, npr. povečava ikone piramide in uporaba opisa stanja gesla poleg piramide (ali v oblačku ob prehodu s kazalcem), zamenjava ali boljša izpostavitev elementa treh vertikalnih pik, izboljšanje kontrasta sivin in odprava belih polj med ploščicami, izboljšava pogostnostnega filtra ipd.;
- dodajanje novih funkcionalnosti v vmesniku, kot je npr. klikljivost iztočnice za vrnitev na prvo stran gesla in s tem povezana rešitev praznega iskalnega okna za nova iskanja, možnost izvoza podatkov, dodatni filtri za omejevanje pogleda kolokacij (npr. filter za odstranitev/prikaz lastnih imen in lastnoimenskih kolokatorjev);
- preoblikovanje logike delovanja pomenskega menija in najbrž celo njegove postavitve, kar je odvisno tudi od razvoja vmesnikov ostalih virov Centra za jezikovne vire in tehnologije Univerze v Ljubljani;
- osmislitev in nadgradnja uporabniškega ocenjevanja in sodelovanja pri izdelavi gesel, npr. z že uspešno preizkušeno metodo razporejanja zgledov.

Na celostni ravni bo potrebno vložiti precejšen razmislek v rešitve, povezane s količino kolokacijskih podatkov v posameznem geslu in možnosti njihove manipulacije. Je pa evalvacija spet opozorila na že znano dihotomijo med dojemanjem kolokacije kot statističnega fenomena, ki ga opredeljuje jakost povezave med njenimi deli in je

poznan jezikoslovcem, in kot pogostnostnega fenomena, ki ga pri uporabi podatkov pričakujejo/tolmačijo uporabniki. Tu razmišljamo v smeri večjega ekspliciranja informacij, mogoče z neko legendo ali pa celo z (opcijskimi) statističnimi podatki.

Seveda pa bo pomemben del nadgradnje slovarja tudi vsebinski, tj. izdelava novih leksikografsko dokončanih gesel, kar bo pri geslih dodalo oz. izboljšalo tudi vsebine, ki so uporabnikom pomembne, npr. pomenska delitev in gručenje kolokacij.

Na splošno lahko zaključimo, da izsledki raziskave predstavljajo pomembna priporočila in smernice za leksikografsko delo in sodobne jezikovne vire, ki stremijo k uporabniški prijaznosti in čim boljši izkoriščenosti prednosti digitalnih medijev.

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. (2021): Razvrstitev kolokacij v slovarskem vmesniku: uporabniške prioritete. V I. Kosem (ur.): *Kolokacije v slovenščini*: 125–157. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani.

Arhar Holdt, Š., Kosem, I. in Gantar, P. (2016): Dictionary user typology: the Slovenian case. V T. Margalitadze in G. Meladze (ur.): *Lexicography and linguistic diversity: proceedings of the XVII EURALEX International Congress*: 179–187. Tbilisi: Ivane Javakhishvili Tbilisi State University.

Arhar Holdt, Š. in Čibej, J. (2020): Rezultati projekta "Slovar sopomenk sodobne slovenščine: od skupnosti za skupnost". *Zbornik konference Jezikovne tehnologije in digitalna humanistika* 2020: 3–9.

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.

Kosem, I., Krek, S., Gantar, P., Arhar Holdt, Š., Čibej, J. in Laskowski, C. (2018): 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 v Ljubljani.

Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Gantar, A., 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: Znanstvena založba Filozofske fakultete.

Ericsson, K. Anders in Simon, Herbert A. (1984): *Protocol analysis: Verbal reports as data.* The MIT Press.

Kosem, I., Husak, M. in McCarthy, D. (2011): GDEX for Slovene. V I. Kosem, I. in K. Kosem (ur.): *Electronic lexicography in the 21st century: new applications for new users: Proceedings of eLex*: 150–159. Ljubljana: Trojina, Institute for Applied Slovene Studies.

Boren, T. in Ramey, J. (2000): Thinking aloud: Reconciling theory and practice. *IEEE transactions on professional communication,* 43.3: 261–278.

Lewis, C. H. (1982): *Using the "Thinking Aloud" Method In Cognitive Interface Design (Technical report)*. IBM. RC-9265.

Lewis, C. in Rieman, J. (1993): *Task-centered user interface design. A practical introduction*.

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 Univerze v Ljubljani.

Simonsen, H. K. (2014): Mobile Lexicography: A survey of the mobile user situation. V *Proceedings of the XVI EURALEX International Congress: The User in Focus*: 15–19. Bolzano/Bozen: Institute for Specialised Communication and Multilingualism.

Comeau, S. (2009): Sharing the knowledge of lexicographers: Methodology for the extraction of lexicographic abilities. V D. Beck, K. Gerdes, J. Milicevic in A. Polguère (ur.): *Proceedings of the Fourth International Conference on Meaning-Text Theory:* 109–117. Université de Montréal.

Tarp, S. (2009): Reflections on Lexicographical User Research. *Lexikos,* 19 (1): 275–296.

Thumb, J. (2004): Dictionary Look-up Strategies and the Bilingualised Learner's Dictionary. *Lexicographica (Series Maior 117)*. Tübingen: Max Niemeyer.

Welker, H. A. (2013a): Methods in Research of Dictionary Use. V R. H. Gouws, U. Heid, W. Schweickard in H. E. Wiegand (ur.): *Dictionaries. An International Encyclopedia of Lexicography: Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography:* 540–547 . Berlin, New York: Walter de Gruyter.

Welker, H. A. (2013b): Empirical Research into Dictionary Use since 1990. V R. H. Gouws, U. Heid, W. Schweickard in H. E. Wiegand (ur.): *Dictionaries. An International Encyclopedia of Lexicography: Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography:* 531–540. Berlin, New York: Walter de Gruyter.

Wingate, U. (2002): The Effectiveness of Different Learners Dictionaries: An Investigation into the Use of Dictionaries for Reading Comprehension by Intermediate Learners of German. *Lexicographica (Series Maior 112)*. Tübingen: Max Niemeyer.

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

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

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

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

1 Uvod

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

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

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

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

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

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

2 Kolokacije v Slovarju sopomenk sodobne slovenščine

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

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

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

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

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

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

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

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

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

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

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

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

3 Gradivo in metoda

3.1 Gradivo

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3.2 Postopek analize

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

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

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

4 Analiza in diskusija

4.1 Samostalnik

4.1.1 Zveze s pridevnikom kot levim ujemalnim prilastkom

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

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

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

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

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

4.1.2 Predložna zveza s samostalnikom kot desnim prilastkom

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

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

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

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

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

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

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

4.1.4 Glagol s samostalnikom v tožilniku

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

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

4.1.5 Uporabnost podatkov za primerjavo rabe in pomena samostalnikov

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

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

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

4.1.6 Analiza ostalih vzorcev v orodju Sketch Engine

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

4.2 Pridevnik

4.2.1 Zveze s samostalnikom kot jedrom ujemalne zveze

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

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

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

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

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

4.2.2 Predložna zveza s samostalnikom kot desnim prilastkom

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

4.2.3 Zveze pridevnika s pomensko določujočim prislovom

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

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

4.2.4 Uporabnost podatkov za primerjavo rabe in pomena pridevnikov

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

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

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

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

4.2.5 Analiza ostalih vzorcev v orodju Sketch Engine

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

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

4.3 Glagol

4.3.1 Zveze glagola s predlogom, ki mu sledi samostalnik

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

4.3.2 Glagol s samostalnikom v neimenovalniškem sklonu

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

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

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

4.3.3 Zveze glagola s pomensko določujočim prislovom

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

4.3.4 Uporabnost podatkov za primerjavo rabe in pomena glagolov

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

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

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

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

4.3.5 Analiza ostalih vzorcev v orodju Sketch Engine

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

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

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

5 Diskusija in zaključek

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

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

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

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

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

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

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

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

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

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

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

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

Zahvala

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

Reference

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Univerza *v Ljubljani*