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