Opis modela za pridobivanje in strukturiranje kolokacijskih podatkov iz korpusa

Simon KREK
Institut »Jožef Stefan«, Filozofska fakulteta Univerze v Ljubljani,
simon.krek@ijs.si

Polona GANTAR
Filozofska fakulteta Univerze v Ljubljani, apolonija.gantar@ff.uni-lj.si

Iztok KOSEM
Filozofska fakulteta Univerze v Ljubljani, iztok.kosem@ff.uni-lj.si

Kaja DOBROVOLJC
Filozofska fakulteta Univerze v Ljubljani, Institut »Jožef Stefan«,
kaja.dobrovoljc@ff.uni-lj.si

Abstract

This paper describes a method for extracting collocation data from text corpora based on a formal definition of syntactic structures, which takes into account not only POS-tagging level of annotation but also syntactic parsing (syntactic treebank model), and introduces the possibility of controlling the canonical form of extracted collocations based on statistical data on forms with different properties in the corpus. Specifically, we describe the results of the extraction from the syntactically tagged Gigafida 2.1 corpus. Using the new method, 4,002,918 collocation candidates in 81 syntactic structures were extracted. We evaluate the extracted data sample in more detail, mainly in relation to the properties that affect the extraction of canonical forms: definiteness in adjectival collocations, grammatical number in noun collocations, comparison in adjectival and adverbial collocations, and letter case (uppercase and lowercase) in canonical forms. The conclusion highlights the potential of the methodology used

for the grammatical description of collocation and phrasal syntax, and the possibilities for improving the model in the process of compilation of the Slovene Digital Dictionary Database.

Ključne besede: kolokacije, strojno prepoznavanje kolokacij v korpusu, digitalna kolokacijska baza

Keywords: collocations, discovering collocations in corpora, digital collocation database

1 Uvod

Razvoj obsežnih besedilnih zbirk in orodij za njihovo kompleksno obdelavo je v zadnjih treh desetletjih omogočil razvoj različnih metod, ki omogočajo avtomatsko pridobivanje večbesednih enot iz korpusov, predvsem za izdelavo slovarskih virov, za računalniško obdelavo naravnega jezika ter za izdelavo različnih jezikovnih aplikacij.

Kolokacije so zaradi svoje pretežno binarne zgradbe, pretežne zastopanosti leksikalnih elementov in njihovega statistično izstopajočega sopojavljanja za razliko od kompleksnejših večbesednih enot, kot so različni tipi frazeoloških enot, ki poleg strukturne ustaljenosti predpostavljajo tudi določeno semantično celovitost, deležne več pozornosti pri razvoju mehanizmov za avtomatsko luščenje (Ramisch 2020, Ramisch et al. 2020).[1] Mehanizmi luščenja večbesednih enot tipično izkoriščajo mehanizem, ki prepoznava zaporedja leksikalnih enot na podlagi njihove oblikoskladenjske označenosti v korpusu in statističnih mer, ki določajo vrednosti sopojavljanja. Najbolj prepoznaven in uveljavljen model, predvsem na področju leksikografije, je model besednih skic v orodju Sketch Engine, ki deluje na podlagi slovnice besednih skic ter lematiziranega in oblikoslovno označenega korpusa.[2] V okviru projekta NSSSS – Nova slovnica sodobne standardne slovenščine: viri in metode (ARRS J6-8256) – je

1 S spletnim servisom elexiFinder z iskalnim pogojem »collocation« in »extraction« lahko najdemo 306 prispevkov: https://bit.ly/3smDBj7.

2 Sistem besednih skic za slovenščino (Krek in Kilgarriff 2006) je bil v okviru projekta SSJ (Krek 2015) že uporabljen pri izdelavi Leksikalne baze za slovenščino (Gantar 2015) in pri izdelavi Kolokacijskega slovarja sodobne slovenščine (Kosem et al. 2018).

bil naš namen izdelati metodologijo za strojno luščenje kolokacijskih podatkov iz korpusa Gigafida, ki nadgrajuje obstoječi sistem, temelječ na slovnici besednih skic za slovenščino (Krek in Kilgarriff 2006, Krek 2015, Gantar 2015, Kosem et al. 2018). Sistem smo nadgradili na podlagi predpostavke, da je spiske (enobesednih ali večbesednih) kolokacijskih kandidatov mogoče uspešneje strojno izluščiti iz skladenjsko razčlenjenega korpusa, in sicer na podlagi označenih odvisnostnih povezav ter lastnosti pojavnic na izvoru ter cilju.

V prispevku opišemo metodologijo strojnega luščenja kolokacij iz korpusa Gigafida 2.1 na podlagi definiranih strukturnih in skladenjskih razmerij znotraj besedne zveze ter z upoštevanjem statističnih parametrov pri izpisu kolokacije kot celote. Najprej predstavimo postopek luščenja ter bazo izluščenih kolokacij (Krek et al. 2021). Nato ocenimo izluščene podatke na podlagi kvantitativnih in kvalitativnih jezikoslovnih analiz. V zaključku izpostavimo možnosti, ki jih za slovnični opis kolokativnosti in besednozvezne skladnje prinaša uporabljena metodologija in odprto dostopni empirični podatki, ter možnosti za izboljšave modela pri izgradnji Digitalne slovarske baze za slovenščino.

2 Strojno luščenje kolokacij iz korpusa

V razdelku opišemo formalni zapis kolokacijskih struktur v datoteki formata XML (2.1), ki predstavlja osrednji del nove metodologije za luščenje kolokacij. Najpomembnejši del opisa je vsebovan v definiciji skladenjskih struktur (2.2), ki je sestavljen iz opisa komponent kolokacije, skladenjskih povezav med njimi ter različnih omejitev glede na (a) identifikacijo komponent v korpusu ter (b) izpis končnih kanoničnih oblik kolokacije. V zadnjem delu razdelka (2.3) opišemo še postopek strojnega luščenja kolokacij iz korpusa na podlagi predlaganega sistema.

2.1 Formalni zapis kolokacij

Za potrebe luščenja na podlagi nove metodologije je bilo treba najprej natančneje definirati, kaj opredeljujemo s pojmom kolokacija, kar je

opisano v prispevku Gantar et al. (2021). Pri definiranju oblikoskladenjske zgradbe smo ob ponovno premišljenem konceptu kolokacije izhajali iz predhodno definiranih gramatičnih relacij v orodju Word Sketches za slovenščino (Krek 2015). Uporabi oblikoskladenjskega nivoja označevanja smo na novo pridružili še nivo skladenjskega razčlenjevanja, pri katerem smo definirali odvisnostna skladenjska razmerja znotraj kolokacije. Statistične in frekvenčne podatke smo upoštevali tako na ravni leme kot tudi kolokacije kot celote, kar se je pokazalo kot ustrezen postopek že v predhodnih avtomatskih luščenjih kolokacij iz korpusa (Gantar et al. 2016). Hkrati smo frekvenčne podatke upoštevali tudi pri določanju reprezentacijske, končne oblike kolokacije, tj. oblike, v kateri naj bi bila kolokacija zastopana tudi v slovarju. V procesu izdelave novega formalizma za luščenje kolokacij je bila večina kolokacijskih struktur, ki so bile upoštevane v Leksikalni bazi, prevedena iz formalizma v orodju Sketch Engine v nov formalizem. Novi formalizem se od tistega v orodju Sketch Engine razlikuje v tem, da:

- namesto jezika Corpus Query Language (CQL), ki upošteva oblikoskladenjske oznake, uporablja lasten sistem za definiranje omejitev pri poljubnem nivoju označevanja, od besednih vrst in njihovih lastnosti, skladenjskih povezav in njihovih oznak, konkretnih leksikalnih elementov, ter drugih nivojev označevanja, ki bi jih lahko uporabili kdaj kasneje, npr. za označevanje semantičnih vlog, semantičnih tipov itd.;
- so v novem sistemu izbrane glagolske strukture med seboj eksplicitno ločene glede na zanikanje (izraženo z nikalnim členkom ali glagolsko) in povratnost (izraženo s prostim glagolskim morfemom ali povratnim zaimkom);
- se za razliko od sistema v orodju Sketch Engine identifikacijske številke in poimenovanja struktur ne razlikujejo glede na to, ali je izhodišče prvi ali drugi kolokator v kolokaciji;
- so poimenovanja oz. oznake struktur spremenjena tako, da neposredno odražajo razlikovalne lastnosti posamičnih komponent na ravni besednih vrst in lastnosti po sistemu oznak MULTEXT-East/JOS (glej Tabelo 2);

- je predvsem za potrebe avtomatizacije postopka luščenja poleg omejitev (angl. *restriction*), kar s CQL omogočajo besedne skice, mogoče tudi določiti, katera od oblik posamezne komponente (besede), ki jo najdemo v korpusu, naj bo izpisana v konkretni kolokaciji, glede na možnosti znotraj predvidene kanonične oblike kolokacije pri konkretni strukturi (angl. *representation*);

Vseh kolokacijskih struktur v sistemu DSB je (trenutno) 82, od tega po parih kolokatorjev šest takih, ki upoštevajo zanikanje (Tabela 1), 25 z izraženo povratnostjo ter 26 kombinacij s predložnimi zvezami.[3] Enako kot pri luščenju z orodjem Sketch Engine kolokatorji pripadajo štirim besednim vrstam: samostalnikom, glagolom, pridevnikom in prislovom.

Tabela 1: Leksikalno-gramatične lastnosti komponent v kolokacijskih strukturah.

Kolokator-1	Kolokator-2	Zanikanje	Povratnost	Predlog	Skupaj
glagol	glagol	4	7		11
glagol	samostalnik	2	10	10	20
glagol	pridevnik		2		4
samostalnik	glagol	2	3		6
samostalnik	samostalnik			5	11
samostalnik	pridevnik				1
samostalnik	prislov			1	1
pridevnik	glagol		1		2
pridevnik	samostalnik			5	10
pridevnik	pridevnik				1
pridevnik	prislov				1
prislov	glagol		2		4
prislov	samostalnik			5	7
prislov	pridevnik				1
prislov	prislov				2
Skupaj		6	25	26	82

Za govoreče oznake uporabljamo kratko kombinacijo upoštevanih oblikoskladenjskih kategorij in lastnosti po sistemu MTE/JOS

3 Pri zanikanju in povratnosti pri štetju v Tabeli 1 ne upoštevamo mesta ali števila takih elementov v strukturi.

(Erjavec et al. 2010a, Erjavec et al. 2010b), pri čemer je za jezikoslovno rabo ključna berljiva oznaka kolokacijske strukture, za računalniško rabo pa identifikacijska številka. Za posamične komponente v govorečih oznakah uporabljamo 22 različnih kombinacij, in sicer v Tabeli 2 navedene kategorije in lastnosti (v zadnjem stolpcu navajamo seštevek, kolikokrat je bila komponenta uporabljena v oznakah v vseh 82 strukturah):

Tabela 2: Kategorije komponent v kolokacijskih strukturah po sistemu oznak MULTEXT-East/JOS.

Št.	Komponenta	Kategorija	Lastnost-1	Lastnost-2	Število
1	d	predlog			26
2	gg	glagol	glavni		41
3	ggm	glagol	glavni	namenilnik	2
4	ggn	glagol	glavni	nedoločnik	14
5	ggz	glagol	glavni	zanikani	1
6	gp	glagol	pomožni		2
7	l	členek			8
8	p0	pridevnik	vsi skloni		13
9	p1	pridevnik	imenovalnik		5
10	p2	pridevnik	rodilnik		1
11	p4	pridevnik	tožilnik		2
12	r	prislov			18
13	s0	samostalnik	vsi skloni		20
14	s1	samostalnik	imenovalnik		8
15	s2	samostalnik	rodilnik		13
16	s3	samostalnik	dajalnik		9
17	s4	samostalnik	tožilnik		7
18	s5	samostalnik	mestnik		5
19	s6	samostalnik	orodnik		5
20	vd	veznik	podredni		4
21	vp	veznik	priredni		4
22	zp	zaimek	povratni		27

Glede na zaporedja komponent, ki nastopajo v kolokacijskih strukturah, lahko za lažje razumevanje njihove kombinacije razporedimo v devet stolpcev, pri čemer upoštevamo pozicijo komponente v

kanoničnih oblikah kolokacije, tj. vnaprej določenih izpisih kolokacij glede na strukturo:

Tabela 3: Zaporedje komponent v kolokacijskih strukturah.

Stolpec	Opis	Komponente
1	nikalni členek 1	l
2	kolokator 1	gg, ggz, p0, p1, p2, r, s0, s1
3	povratni zaimek 1	zp
4	veznik	vd, vp
5	predlog	d
6	nikalni členek 2	l
7	pomožni glagol	gp
8	kolokator 2	gg, ggm, ggn, p0, p1, p4, r, s0, s1, s2, s3, s4, s5, s6
9	povratni zaimek 2	zp

Celotno listo 82 kolokacijskih struktur navajamo v Prilogi. V Tabeli 4 spodaj kot primer navajamo izbor desetih struktur, prvih pet glede na število izluščenih kolokacij, preostalih pet za potrebe prikaza oznak v vseh ostalih devetih stolpcih/kategorijah:

Tabela 4: Kolokacijske strukture glede na zastopane kategorije in število izluščenih primerov.

ID	Oznaka	Zgled	1	2	3	4	5	6	7	8	9	Št. kolokacij
34	p0-s0	svetovno prvenstvo		p0						s0		720.605
53	s0-s2	direktor podjetja		s0						s2		518.199
70	s0-gg	raziskava pokaže		s0						gg		385.018
23	gg-s4	podpisati pogodbo		gg						s4		270.965
15	gg-d-s5	imeti v mislih		gg			d			s5		235.771
30	p0-vp-p0	domač in tuj		p0		vp				p0		32.127
77	s1-gp-s1	nogomet je šport		s1					gp	s1		26.520
72	s0-l-gg	trditev ne drži		s0				l		gg		19.400
95	l-gg-zp-ggn	ne uspeti se uvrstiti	l	gg	zp					ggn		479
94	gg-zp-ggn-zp	odločiti se vrniti se		gg	zp					ggn	zp	5

Za izdelavo algoritma za samodejno luščenje kolokacij iz korpusa smo izdelali formalizem zapisa vseh potrebnih informacij v formatu XML. Ta omogoča kasnejše prilagajanje, dodajanje ali odvzemanje

struktur pri nadaljnjih luščenjih kolokacij. V nadaljevanju formalizem podrobneje opišemo.

2.2 Definicija skladenjskih struktur

V okviru spodaj je kot zgled naveden celoten zapis najpogostejše izluščene strukture z oznako p0-s0 (ID 34), ki definira samostalniško jedro, ki ga modificira pridevnik:

```
<syntactic_structure id="34" label="p0-s0" type="collocation">
  <!-- example: bela zastava / rdeča jagoda -->
  <system type="JOS">
    <components order="fixed">
      <component cid="1" type="core" label="p0"/>
      <component cid="2" type="core" label="s0"/>
      <component cid="3" type="other" status="forbidden"/>
    </components>
    <dependencies>
      <dependency from="2" to="1" label="dol" order="to-from"/>
      <dependency from="#" to="2" label="#"/>
      <dependency from="1" to="3" label="vez"/>
    </dependencies>
    <definition>
      <component cid="1">
        <restriction type="morphology">
          <feature POS="adjective"/>
        </restriction>
        <representation>
          <feature rendition="word_form"/>
          <feature selection="agreement" msd="gender+number+case"
            head_cid="2"/>
        </representation>
      </component>
      <component cid="2">
        <restriction type="morphology">
          <feature POS="noun"/>
        </restriction>
        <representation>
          <feature rendition="word_form"/>
          <feature selection="msd" case="nominative"/>
        </representation>
      </component>
      <component cid="3"/>
    </definition>
  </system>
</syntactic_structure>
```

Primer 1: Zapis najpogostejše izluščene strukture z oznako p0-s0 (ID 34).

Posamično skladenjsko strukturo definira element <syntactic_structure>, ki predvideva tri obvezne atribute. Ti vsebujejo:

- identifikacijsko številko strukture: @id
- govorečo oznako strukture: @label
- tip strukture:[4] @type

Definicija strukture se opira na specifične nabore oznak in sisteme označevanja korpusov, zato na prvem nivoju pod strukturo v elementu <system> definiramo sistem označevanja, ki ga bomo upoštevali. Ta vsebuje atribut @type, katerega vrednost definira izbrani sistem označevanja. V okviru projekta NSSSS smo na ravni oblikoskladenjskega in skladenjskega označevanja korpusa Gigafida 2.1 uporabili sistem oznak JOS oz. MULTEXT-East, tako na oblikoskladenjski kot na skladenjski ravni.

Znotraj specifičnega sistema označevanja nadalje definiramo tri ločene skupine informacij:

- posamezne besede oz. elemente, ki sestavljajo kolokacijo – komponente,
- povezave med elementi na skladenjskem nivoju – odvisnostno drevo,
- omejitve in druge informacije, ki jih rabimo za izpis kolokacij – definicija strukture.

2.2.1 Komponente

Komponente so definirane v elementu <components>, ki vsebuje atribut @order. Ta lahko vsebuje vrednosti 'fixed' in 'variable'. Z atributom določamo, ali pri strojni obravnavi strukture in izpisu komponent upoštevamo njihovo zaporedje, kot je določeno v definiciji strukture, ali upoštevamo stanje, ki smo ga našli v korpusu – torej pri izpisu upoštevamo, kakšno zaporedje komponent pri konkretni kolokaciji prevladuje v večini stavkov iz korpusa. Primer strukture, pri kateri je zaporedje variabilno, je zveza prislova in glagola z oznako r-gg (ID 43), pri kateri bo izpis kolokacije variiral glede na tipično pojavljanje obeh elementov oz. pomenske skupine prislovov, npr. *ostati doma* (gg-r) proti *veliko pomeniti* (r-gg).

4 V prispevku obravnavamo 82 struktur, ki spadajo v type="collocation". Predvidena tipa sta še: type="single" za enobesedne lekseme in type="other" za večbesedne enote.

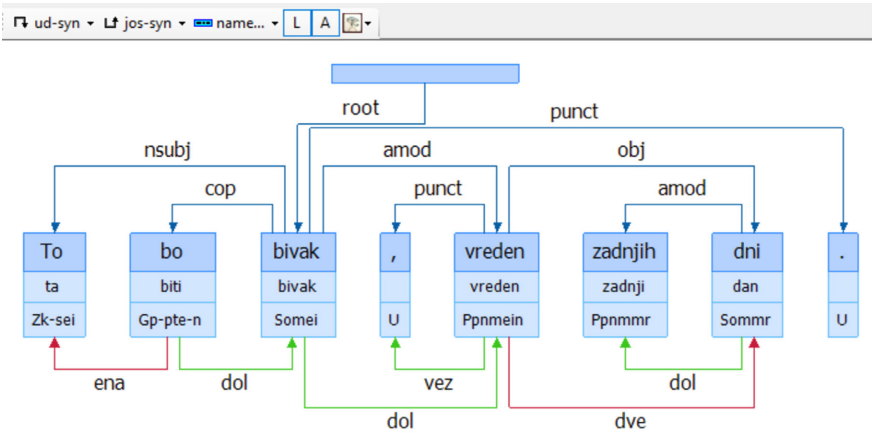
Vse komponente so naštete v (pod)elementih <component>, ki vsebujejo več atributov:

- identifikacijsko številko komponente: @cid,
- govorečo oznako komponente: @label,
- tip komponente: @type,
- status komponente: @status.

V atributu @label ponavljamo informacijo iz celotne oznake strukture, a referiramo le na del, ki definira to specifično komponento. Atribut @type določa jedrnost komponent in lahko vsebuje dve vrednosti: 'core' in 'other'. Jedrne komponente, označene s prvo vrednostjo, so dejanske komponente te kolokacijske strukture, ki so vsebovane v oznaki kolokacije in so tudi vključene v njen izpis. Komponente, označene z 'other', uporabimo v primerih, ko moramo za pravilno identifikacijo kolokacije v določeni strukturi definirati dodatne elemente, ki so bodisi obvezni ali prepovedani. Komponente, ki so v atributu @type opredeljene z vrednostjo 'other', morajo zato vsebovati tudi atribut @status, v katerem sta dovoljeni vrednosti 'obligatory' in 'forbidden'. Prva določa, da se mora komponenta obvezno nahajati v stavku, v katerem smo našli kolokacijo, čeprav te komponente potem ne izpišemo kot del kolokacije. Druga vrednost ima obratno vlogo – v korpusnem stavku se komponenta s statusno vrednostjo 'forbidden', kot je definirana v strukturi, ne sme nahajati.

Za razumevanje sistema skladenjskih struktur in luščenja kolokacij je pomembno dobro poznavanje vloge dodatnih (neizpisanih) komponent, zato podrobneje pojasnjujemo dva primera prepovedanih in obveznih nejedrnih komponent. Komponenta s statusno vlogo 'forbidden' je vključena v strukturo, ki jo kot zgled navajamo zgoraj, zato bomo uporabili kar to.

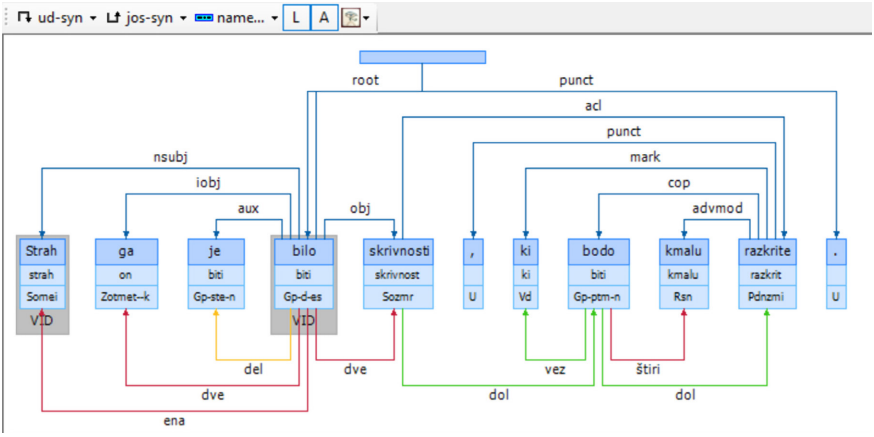
V zgledu iz korpusa ssj500k 2.2 (Slika 1) vidimo, da je samostalniško jedro povedkovega določila (*bivak*) povezano s pridevniškim jedrom odvisnega stavka (*vreden*) s povezavo 'dol' (določilo). Če bi luščenje kolokacij v strukturi p0-s0 omejili zgolj s tem, da mora biti prva jedrna komponenta samostalnik, druga pridevnik, in da sta povezani s povezavo z oznako 'dol', bi izluščili tudi »lažne kolokacije«



Slika 1: Sestavina z vlogo 'forbidden', ki jo opredeljuje oznaka 'vez' na primeru iz korpusa ssj500k 2.2.

(**vreden bivak*), česar pa ne želimo. Zato dodatno prepovemo povezavo med pridevnikom in vezniškim elementom (ločilo, veznik itd.), ki jo opredeljuje oznaka 'vez'.

Komponenta s statusno vlogo 'obligatory' je vključena v strukturo p0-r (ID 85), ki jo v prikazu struktur (Priloga) zastopa zgled *[biti] znan danes*. S to strukturo iščemo kolokacije, v katerih pridevnik nastopa v vlogi povedkovega določila (*biti znan*), dodan pa je tudi prislov (časa, kraja itd.) v vlogi prislovnega določila. Potreben je torej povezovalni element, tj. glagol *biti*, ki pa ga v kolokaciji ne izpisujemo.



Slika 2: Glagol *biti* (oblika: *bodo*) z vlogo 'obligatory' na primeru iz korpusa ssj500k 2.2.

V strukturi p0-r torej zahtevamo, da se v stavku pojavlja glagol *biti*, ki je povezan tako s pridevnikom (*razkrit*) kot s prislovom (*kmalu*), tako kot je prikazano na Sliki 2. Kot rezultat pa bosta v kolokaciji izpisana zgolj pridevnik in prislov, v tem primeru kolokacija *razkrit kmalu*.

2.2.2 Skladenjske povezave

Naslednjo večjo enoto opisa strukture predstavlja element <dependencies>, ki opredeljuje skladenjske povezave med komponentami. V (pod)elementih <dependency>, katerih število mora ustrezati številu komponent, so obvezni trije atributi (@from, @to, @label). Možen je še dodaten (opcijski) atribut @order:

- izvor povezave odvisnostnega drevesa (po sistemu MTE/JOS): @from,
- cilj povezave odvisnostnega drevesa (po sistemu MTE/JOS): @to,
- oznaka povezave (po sistemu MTE/JOS): @label,
- vrstni red povezanih komponent: @order.

Zadnji atribut @order z dovoljenimi vrednostmi 'to-from', 'from-to' ali privzeto vrednostjo 'any' določa, ali se morata komponenti, ki sta povezani s to odvisnostno povezavo, v stavku nahajati v specifičnem besednem redu ali ne. V primeru strukture ID 34, ki jo navajamo zgoraj, uporaba atributa @order pomeni, da se mora pridevnik v stavku dejansko nahajati pred samostalniškim jedrom kot levi prilastek, da bi kolokacijo prepoznali kot ustrezajočo tej strukturi. Znak #, uporabljen kot vrednost v atributih @from in @label, pomeni, da ne želimo omejevati, iz katerega elementa vodi povezava v drevesnici ali katera oznaka opredeljuje povezavo. Nadomešča torej katerikoli izvor ali oznako povezave.

2.2.3 Omejitve in izpis

Najbolj obsežen del formalnega opisa strukture predstavlja element <definition>, v katerem za posamezne komponente določamo

njihove omejitve pri iskanju v korpusu <restriction> in variable pri izpisu najdenih kolokacij <representation>. Element <representation> vsebujejo samo komponente, ki so opredeljene kot jedrne ('core') in so dejansko vključene v izpis kolokacije.

Element <restriction>, ki opredeljuje omejitve, vsebuje atribut @type, ki določa na katerem označevalnem nivoju bomo našli podatke o omejitvah. Trenutno sta v uporabi vrednosti 'morphology' in 'lexis'. Prva vrednost določa, da se bodo omejitve nanašale na oblikoskladenjski nivo označevanja v korpusu. Druga vrednost pomeni, da se pri identifikaciji komponente omejujemo na konkretne pojavnice, bodisi na ravni besedne oblike ali leme, kot jo najdemo v korpusu. Primer take rabe so variante veznika *kot*, *kakor*, *ko* v strukturi p0-vd-s1 (ID 32), s katero iščemo pridevniške komparacije (*čist kot solza*). Če pri omejitvah izberemo oblikoskladenjski nivo označevanja, omejitve glede kategorij in lastnosti navajamo v elementu <feature>, kot atribute pa uporabimo kategorije iz nabora oznak, z vnaprej predvidenimi vrednostmi. Primer, ki ga navajamo spodaj, opredeljuje omejitev na ravni kategorije (POS) z vrednostjo 'adjective', kar pomeni, da se kot rezultat luščenja na mestu te komponente v kolokaciji lahko pojavlja zgolj beseda, ki je v korpusu na ravni oblikoskladenjskega označevanja opredeljena kot pridevnik:

```
<feature POS="adjective"/>
```

Enako opredeljujemo vse druge kategorije in lastnosti, v našem primeru po sistemu MTE/JOS. Če želimo znotraj posamezne lastnosti dovoliti več vrednosti, to lahko naredimo z uporabo pokončnice, npr.

```
<feature case="genitive|accusative"/>
```

Če v atributu @type uporabimo vrednost 'lexis', bomo konkretne vrednosti oz. besede, ki jih identificiramo v korpusu, navedli v atributih @lemma ali @word_form, kot na primer v prej navedenem zgledu:

```
<feature lemma="kot|kakor|ko"/>
```

Element <representation> opredeljuje variable pri izpisu najdenih kolokacij. Te bomo prav tako našli v elementu <feature>, vendar z drugačnimi atributi. Z atributom @rendition določamo, kakšen tip informacije bomo uporabili pri izpisu. Vrednosti 'lemma' in 'word_form' opredelita, da bomo uporabili bodisi lemo ali eno od besednih oblik komponente, kot jih najdemo v korpusu. Vrednost 'lexis' v atributu @rendition pomeni, da bomo uporabili element, ki ga (morda) v korpusu nismo našli, vendar ga v vsakem primeru hočemo izpisati na mestu komponente v kolokaciji. Za konkretno ubeseditev tega elementa uporabimo atribut @string s poljubnim nizom črk, ki se potem izpiše v kolokaciji. Primer take rabe so negacijske strukture, pri katerih v vsakem primeru želimo, da se izpiše nikalni členek *ne*, čeprav bi bil npr. v korpusu pogostejši *ni* ali zanikane osebne oblike glagola *biti*.

Nadalje v elementu <feature> z atributom @selection (v kombinaciji z atributom @rendition) izbiramo, katero od možnih besednih oblik, ki jih na mestu te komponente najdemo v korpusu, izpišemo v kolokaciji. Vrednosti, ki so predvidene v atributu @selection so: 'all', 'msd' ali 'agreement'. Prva ('all') pomeni, da izpišemo vse oblike komponente, ki jih najdemo v korpusu. To je koristno denimo v primeru povratnih zaimkov, ki imajo v različnih kombinacijah možni obliki *se* in *si* in če v korpusu najdemo obe, ju v kolokaciji tudi izpišemo s poševnico – *izogibati se/si pogovoru*.

Vrednost 'msd' v atributu @selection uporabimo v primeru, če želimo natančneje opredeliti, katero od najdenih oblik izpišemo, glede na njene oblikoskladenjske lastnosti. Posamične lastnosti v istem elementu opredelimo s kombinacijo lastnosti in njene vrednosti, npr.

```
<feature selection="msd" case="nominative"/>
```

Zapis pomeni, da želimo, naj algoritem izpiše (najpogostejšo) imenovalniško obliko besede, ki jo je našel v korpusu.

Vrednost ‘agreement’ v atributu @selection uporabimo v primeru, če želimo, da se izpisana oblika komponente v določenih lastnostih ujema z istimi lastnostmi, opredeljenimi v drugi komponenti, kar opredelimo v atributih @msd in @head_cid. Prvi atribut opredeljuje lastnosti, ki se morajo ujemati, drugi referira na identifikacijsko številko komponente, ki vsebuje lastnosti, ki jih pri ujemanju upoštevamo. Primer:

```
<feature selection="agreement" msd="gender+number+case"
head_cid="2"/>
```

Primer opredeljuje, da se morata obe komponenti ujemati v spolu, sklonu in številu.

Z opisanimi formalnimi elementi (v kombinaciji s kategorijami, lastnostmi in vrednostmi v izbranem označevalnem sistemu) opredeljujemo vseh 82 kolokacijskih struktur, s katerimi smo iz korpusa Gigafida 2.1 izluščili skupaj nekaj več kot 4 milijone kolokacij, kar opišemo v nadaljevanju.

2.3 Postopek strojnega luščenja kolokacijskih podatkov iz korpusa Gigafida 2.1

Za avtomatsko luščenje kolokacijskih kandidatov smo uporabili leta 2018 objavljeni in nadgrajeni korpus Gigafida 2.0 (Krek et al. 2020), ki med drugim prinaša izboljšave na ravni lematizacije ter oblikoskladenjskega označevanja, izločitev nestandardnih besedil, nadgradnjo korpusa s podreprezentiranimi in sodobnejšimi besedili. Verzija korpusa Gigafida 2.1, ki je bila uporabljena za luščenje kolokacij, vsebuje tudi dodatni nivo skladenjskega razčlenjevanja, označevanje s semantičnimi vlogami ter prepoznavanje imenskih entitet. Predvidevali smo, da bo izboljšanje zanesljivosti označevanja pomembno vplivalo na ustreznost izluščenih kolokacijskih kandidatov povsod, kjer je njihova ustreznost povezana s specifikami na ravni leme, besedne vrste in določenih drugih že omenjenih slovničnih kategorij.

Končna baza kolokacijskih podatkov (Krek et al. 2021) vsebuje 4.002.918 kolokacij, avtomatsko izluščenih iz korpusa Gigafida 2.1 na podlagi definicije 82 kolokacijskih struktur. Najmanjša frekvenca enot v bazi je 10, izluščenih kolokacij z manjšo frekvenco nismo vključili v bazo. Ta je razdeljena po strukturah v 81 datotek v tabelaričnem formatu, z vejico kot separatorjem (format CSV). V bazi je ena datoteka manj, kot je število struktur, ker struktura ID-97 (l-gg--zp-ggn-zp, *ne bati se pokazati se*) ni dala rezultatov s kolokacijami nad frekvenco 10. Vsem kolokacijam so pripisani naslednji podatki v 26 stolpcih:

Tabela 5: Vrste podatkov v bazi kolokacijskih podatkov za posamezno kolokacijsko strukturo.

Stolpec	Naslov stolpca	Opis
1	Structure_ID	identifikacijska številka strukture
2	C1_Lemma	izpis leme prve komponente
3	C1_Representative_form	izpis oblike prve komponente (glede na definicijo strukture)
4	C1_RF_msd	oblikoskladenjska oznaka oblike prve komponente
5	C1_RF_scenario	scenarij izpisa oblike prve komponente
6	C1_Distribution	število različnih kolokacij, ki vsebujejo lemo komponente C1 (znotraj strukture)
7	C1_lemma_structure_frequency	število korpusnih stavkov s kolokacijami, ki vsebujejo lemo komponente C1 (znotraj strukture)
8	C2_Lemma	ENAKE INFORMACIJE ZA KOMPONENTE C2/3/4/5
…	…	…
21	Colocation_ID	identifikacijska številka kolokacije
22	Joint_representative_form_fixed	izpis kanonične oblike kolokacije (glede na strukturo)
23	Joint_representative_form_variable	izpis najpogostejše oblike kolokacije (glede na besedni red)
24	Frequency	frekvenca kolokacije
25	logDice_core	izračun jakosti kolokacije (logDice)
26	Distinct_forms	število različnih oblik kolokacije

Vsebina stolpcev z enostavnejšimi informacijami (identifikacijska številka, lema itd.) ne potrebuje dodatnega pojasnila, podrobneje pojasnjujemo naslednje tipe informacij:

1. Stolpec 5: C1_RF_scenario

Kot opisujemo zgoraj, obliko izpisa (*representation*) posameznih komponent v kolokaciji določajo tri možnosti: (1) izpiše se osnovna oblika komponente, tj. lema, kot jo najdemo v korpusu; (2) izpiše se specifična oblika komponente, ki je določena z dodatnimi pogoji, npr. mora biti v določenem sklonu; (3) izpiše se specifična oblika komponente, ki je določena z ujemanjem z drugo komponento po lastnostih, npr. v spolu, sklonu in številu. Če je bil predvideni scenarij izpolnjen, je v stolpcu 5 navedena vrednost 'ok'. Če zaradi različnih razlogov ni mogoče najti oz. navesti oblike, ki je predvidena v izpisu, se na mestu komponente v kolokaciji izpiše osnovna oblika, vrednost v stolpcu 5 pa je v tem primeru 'lemma_fallback'. Tipični razlog za tak scenarij je situacija, da kolokacija predvideva ujemanje oblik pri dveh komponentah, vendar v korpusnih primerih nismo našli ustrezne oblike za komponento, ki se mora ujemati, npr. v sklonu.

2. Stolpec 6: C1_Distribution

Stolpec za vsako komponento vsebuje seštevek različnih kolokacij, v katerih se (a) znotraj iste strukture pojavlja kot (b) ista komponenta, tj. z isto vrednostjo atributa @cid. Navajamo preprost primer – če imamo pri strukturi p0-s0 naslednje tri kolokacije: *rdeča jagoda* (Collocation_id 1), *rdeč avto* (Collocation_id 2), *moder avto* (Collocation_id 3), iz izračuna lahko vidimo, da se lemi *rdeč* in *avto* pojavljata pri več kolokacijah, *jagoda* in *moder* pa samo pri eni:

- 1, rdeča jagoda, C1_distribution = 2, C2_distribution = 1
- 2, rdeč avto, C1_distribution = 2, C2_distribution = 1
- 2, moder avto, C1_distribution = 1, C2_distribution = 2

3. Stolpec 7: C1_lemma_structure_frequency

V stolpcu je naveden seštevek korpusnih frekvenc, torej najdenih instanc kolokacije v korpusu (stolpec Frequency), vseh kolokacij v strukturi, v katerih se pojavi lema C1.

4. **Stolpca 22 in 23: Joint_representative_form_fixed in Joint_representative_form_variable**

V stolpcih 22 in 23 sta izpisani dve obliki kolokacije. Prva (stolpec 22) upošteva kanonično obliko kolokacije, kot je glede na zaporedje komponent predvidena v strukturi. Druga (stolpec 23) upošteva stanje, ki smo ga našli v korpusu – komponente so navedene v zaporedju, ki je najpogostejše v korpusu. S tem mehanizmom pri nekaterih strukturah pridemo do naravnejših kanoničnih oblik, kot smo prej navedli v primeru strukture r-gg (ID 43), pri kateri bo v stolpcu 23 pri eni kolokaciji navedena oblika gg-r (*ostati doma*), v drugi pa r-gg (*veliko pomeniti*). V stolpcu 22 bosta v obeh primerih navedeni kanonični obliki, ki ju predvideva struktura: *doma ostati* in *veliko pomeniti*.

5. **Stolpec 25: logDice_core**

Vsaka struktura ima opredeljena dva kolokatorja, ki sta označena s type=core in sta polnopomenski besedi (<feature POS=«adjective|noun|verb|adverb«/>). V primeru spodaj navajamo par struktur z odebeljenimi jedrnimi polnopomenskimi besedami:

- p0-s0: **rdeča jagoda**
- p0-s2: biti **obtožen utaje**
- s0-gp-p1: **rezultati** so **dobri**
- s0-d-s5: **otok** ob **obali**
- gg-d-s4: **biti** na **voljo**

Za izračun kolokabilnosti med obema jedrnima besedama potrebujemo naslednje podatke:

- $\mathbf{f_x}$ = pogostost prve jedrne besede v celotnem korpusu (leme z besedno vrsto)
- $\mathbf{f_y}$ = pogostost druge jedrne besede v korpusu (leme z besedno vrsto)
- $\mathbf{f_{xy}}$ = pogostost dane kolokacije (frekvenca Collocation_id)
- **N** = število vseh besed oz. pojavnic v korpusu

Za dani Collocation_id izračunamo mero logDice_core po formuli:

$$\text{logDice\_core} = 14 + \log_2 \frac{2f_{xy}}{f_x + f_y}$$

6. Stolpec 26: Distinct_forms

Stolpec 26 vsebuje izračun, v koliko različnih oblikah (ne glede na veliko ali malo začetnico oz. velike ali male črke) se v korpusu pojavlja dana kolokacija (Collocation_id), npr.:

- rdeča jagoda, rdeče jagode, rdečim jagodam, Rdeča jagoda → 3 različne oblike
- rdeča jagoda, rdeče jagode, rdeča jagoda, rdeča jagoda → 2 različni obliki

V nadaljevanju se posvetimo opisu osnovnih podatkov o izluščenih kolokacijah ter nekaterim pomembnejšim prednostim, ki jih omogoča nova metoda.

3 Jezikoslovni vidiki opisa baze kolokacijskih podatkov

V tretjem razdelku obravnavamo izbrane jezikoslovne teme, ki so zanimive za analizo pri izluščenih kolokacijah, med njimi (ne) določnost oblik pri pridevniških kolokacijah (3.1), slovnično število (dvojina/množina proti ednini) pri samostalniških kolokacijah (3.2), stopnjevanje (osnovnik proti primerniku in presežniku) pri pridevniških in prislovnih kolokacijah (3.3), ter zapis z velikimi in malimi črkami (3.4).

Baza strojno izluščenih kolokacijskih kandidatov bo v prihodnosti služila tako neposredno za nadgradnjo obstoječega Kolokacijskega slovarja sodobne slovenščine (Kosem et al. 2019), kot posredno za potrebe Slovarja sodobnega slovenskega jezika (Gorjanc 2015) ter kot empirična osnova slovničnih analiz skladenjskih pojavov. Novo metodo smo uporabili tudi pri določanju razmerij med enotami v stalnih besednih zvezah (Gantar 2021a) in za analizo sintagmatskih razmerij med leksikalnimi enotami v vezljivostnih vzorcih (Gantar 2021b).

Za potrebe jezikoslovne evalvacije so bili na voljo izluščeni kumulativni podatki za kolokacijske kandidate za 88 lem z minimalno frekvenco vsaj dveh pojavitev, torej je bilo obravnavanih kolokacij več kot jih za omenjene leme vsebuje baza, pri kateri je frekvenčna meja 10 pojavitev. Glede na predhodno metodologijo luščenja je za evalvacijo zanimiv predvsem reprezentacijski del definicije, kar podrobneje opišemo v nadaljevanju. Možnost nadzora nad izpisom kolokacije pomeni, da pri izbranih kolokacijskih elementih lahko dopustimo variabilnost, ki pri konkretnih kolokacijskih kandidatih odraža dejansko stanje v korpusu. V primeru izbranih 82 struktur je bila variabilnost dopuščena na ravni:

- določnih (ali nedoločnih) imenovalniških oblik pridevnika za moški spol ednine – na primer: namesto privzete kombinacije *solaten bife* je prevladujoč izpis z določno obliko *solatni bife*, ki ustrezno nakazuje, da gre pretežno za (terminološko) kulinarično rabo;
- upoštevanja slovničnega števila pri kolokacijah s samostalniki – na primer: pri (glagolski) kolokaciji *ne briti si nog* izpis kaže, da je množinska oblika *nog* pogostejša, kot bi sicer bila privzeta *ne briti si noge*;
- upoštevanja stopnjevanja pri pridevnikih in prislovih – na primer: pri pridevnikih privzeta oblika kolokacije *dober v panogi* postane smiselna, če izluščimo presežniško obliko *najboljši v panogi*, podobno pri prislovni kombinaciji *čedalje glasneje* s primernikom (namesto privzete oblike *čedalje glasno*);
- zapisa z malimi ali velikimi črkami – na primer: izluščena kolokacija *ljubljanska Drama* kaže, da gre med korpusnimi zadetki pretežno za gledališko ustanovo.[5]

Ugotovitve podrobneje opisujemo po omenjenih sklopih (prim. Pori in Kosem 2021).

5 Kolokacijski zgledi so pri vseh kategorijah izpisani v obliki, ki je bila izluščena iz korpusa, zato tudi pri kategoriji, v kateri analiziramo (ne)določnost pridevnikov, najdemo zgled *Zajtrkovalni bife*, ker je bil iz gradiva izluščen pretežno v obliki z veliko začetnico. Enako velja za kolokacijo *Najcenejši aranžmaji* v kategoriji primernik/presežnik itd.

3.1 Določnost pri pridevniških kolokacijah

Z novo metodo je mogoče ustrezneje izpostaviti razmerje med določnimi in nedoločnimi oblikami pridevnika, kot se kažejo v realni rabi – pri čemer se na tem mestu ne spuščamo podrobneje v vprašanje izražanja pomenskih kategorij vrstnosti in svojilnosti, ki so lahko oblikovno prekrivne z določnimi oz. nedoločnimi oblikami (prim. Gantar in Gorjanc 2015). V Tabeli 6 navajamo prvih 30 kolokacijskih kandidatov, ki so razvrščeni po meri logDice in filtrirani glede na:

- oblikoskladenjsko oznako (pridevniški element mora izkazovati lastnosti: moški spol, ednina, imenovalnik),
- izkazano razliko med pripisano korpusno lemo (ki je glede na leksikonsko konvencijo vedno v nedoločni obliki, če ta obstaja) in izpisano obliko pridevnika,
- korpusno frekvenco najmanj 10 pojavitev (meja, uporabljena v kolokacijski bazi),
- pojavljanje posamezne komponente v najmanj dveh kolokacijah.

Z omenjenimi filtri pridobimo zadostno raznolikost elementov za analizo.

Po pričakovanju gre pogosto za termine z določenega področja, pri katerih je določna oblika oz. vrstnost pričakovana, npr. *etilni alkohol, akutni sindrom, avtomatični stabilizator, akutni hepatitis* itd. Zraven lahko štejemo tudi poimenovanja živali in rastlin: *kodrasti pelikan, kodrasti ohrovt, dolgoživi bor* itd.

Z določno obliko pridevnika se izpisuje tudi precej stalnih zvez oz. izrazov, ki so hkrati v terminološki rabi na določenem področju in del splošnega besedišča, npr. *tuji jezik, letni dopust, materni jezik, solatni bife, samopostrežni bife, kolektivni dopust, neplačani dopust* itd.

Metoda, uporabljena v predhodnih luščenjih (prim. Krek 2006), je pri pridevniških elementih v podobnih strukturah omogočala le zanašanje na leme, kar je v zgornjih zgledih privedlo do izvoza »nenaravnih« kolokacij, denimo: *etilen alkohol, tuj jezik, metilen alkohol, leten dopust, materen jezik, solaten bife, akuten sindrom, knjižen*

Tabela 6: Prvih 30 kolokacijskih kandidatov po meri logDice glede na izkazan zapis določnosti/vrstnosti pri pridevniški komponenti kolokacije.

Št.	Kolokacija	GF2.1 (F)	Št. oblik	logDice	Kategorija
1	etilni alkohol	188	7	10,99092	termin (medicina, kulinarika)
2	tuji jezik	17.563	43	10,66132	stalna zveza (jezikoslovje)
3	metilni alkohol	113	5	10,24701	termin (medicina, kulinarika)
4	letni dopust	4.787	20	10,22507	stalna zveza (pravo, ekonomija)
5	materni jezik	4.106	27	9,85621	stalna zveza (jezikoslovje)
6	solatni bife	103	13	9,37135	stalna zveza (kulinarika)
7	akutni sindrom	272	10	9,25843	termin (medicina)
8	knjižni jezik	3.289	24	9,24809	stalna zveza (jezikoslovje)
9	kandirani ananas	20	4	9,10882	stalna zveza (kulinarika)
10	samopostrežni bife	93	12	8,98553	stalna zveza (kulinarika)
11	avtomatični stabilizator	60	10	8,97874	termin (ekonomija)
12	kolektivni dopust	1.015	19	8,91916	stalna zveza (pravo, ekonomija)
13	skupni jezik	6.387	21	8,76027	frazeologija
14	akutni hepatitis	95	9	8,71401	termin (medicina)
15	znakovni jezik	1.662	16	8,57126	stalna zveza (jezikoslovje)
16	uradni jezik	3.722	25	8,55647	stalna zveza (jezikoslovje)
17	neplačani dopust	15	8	8,45259	stalna zveza (pravo, ekonomija)
18	kodrasti pelikan	338	15	8,4337	živalska vrsta
19	Zajtrkovalni bife	11	4	8,4164	stalna zveza (kulinarika)
20	kodrasti ohrovt	24	3	8,40133	rastlinska vrsta
21	akutni infarkt	30	7	8,38873	termin (medicina)
22	bakreni kotliček	117	10	8,2979	vrstnost / lastnost
23	alkoholni kis	46	10	8,25414	stalna zveza (kulinarika)
24	dobrodelni bazar	192	5	8,25405	vrstnost / lastnost
25	poletni dopust	301	15	8,19414	vrstnost / lastnost
26	prisilni dopust	1.028	17	8,08666	stalna zveza (pravo, ekonomija)
27	pogovorni jezik	501	15	8,08655	stalna zveza (jezikoslovje)
28	pritlikavi bor	1.247	18	8,05079	rastlinska vrsta
29	akutni bronhitis	28	7	7,9812	termin (medicina)
30	dolgoživi bor	51	8	7,82243	rastlinska vrsta

jezik, kandiran ananas, samopostrežen bife, avtomatičen stabilizator, kolektiven dopust, skupen jezik, akuten hepatitis, znakoven jezik, uraden jezik, neplačan dopust, kodrast pelikan, zajtrkovalen bife, kodrast ohrovt, akuten infarkt, alkoholen kis, poleten dopust, prisilen dopust, pogovoren jezik, pritlikav bor, akuten bronhitis, dolgoživ bor.

Sprejemljivi sta verjetno obe obliki kolokacij, v katerih je pridevnik mogoče dojemati bodisi v smislu izražanja vrste ali lastnosti: *bakren kotliček, dobrodelen bazar*. Vendar tudi v teh dveh primerih prevlada določne oblike v korpusnih podatkih nakazuje, da bi bila ta oblika morda lahko primernejša za slovarsko obliko iztočnice. Kot zadnji je zanimiv primer kolokacije *skupni jezik*, ki je v resnici del frazeološke enote *najti skupen/skupni jezik* (priti do kompromisne rešitve). V tem primeru se po obdelavi kolokacija umakne v frazeološko enoto, te pa imajo svojo notranjo logiko glede izbire kanoničnih oblik (prim. Gantar 2021a).

Sklenemo lahko, da pri vprašanju izbire oblik pridevniške (ne) določnosti dopuščanje variabilnosti prinaša predvidene rezultate.

3.2 Slovnično število pri samostalniških kolokacijah

Pri samostalniških komponentah je v večini struktur dopuščena variabilnost glede slovničnega števila. To pomeni, da je izbira glede edninske, dvojinske ali množinske oblike samostalnika prepuščena ugotovljeni korpusni frekvenci, ne glede na predvideni sklon ali druge lastnosti. Spodaj navajamo prvih 30 kolokacij iz nabora 88 iztočnic, pri katerih je bila pri (kateremkoli) samostalniku izpisana množinska oblika. Razvrščene so po meri logDice in filtrirane po lastnosti množina pri samostalniku, frekvenci najmanj 10, v korpusu pa morajo izkazati najmanj tri oblike.

Poleg napačno izluščenega lastnega imena so hitro opazne kolokacije, ki opozarjajo na frazeološkost: *briti norce (iz koga/česa), brusiti (si) kremplje, (brez) dlake na jeziku, (držati) jezik za zobmi, oprijeti se (česa) kot (zadnje) bilke*. V teh primerih načeloma lahko pričakujemo, da so množinske oblike upravičene, vendar imajo te enote svojo logiko in pri njih večinoma lahko pričakujemo tudi

Tabela 7: Prvih 30 kolokacijskih kandidatov po meri logDice glede na izkazan zapis množinske oblike pri samostalniški komponenti kolokacije.

Št.	Kolokacija	GF2.1 (F)	Št. oblik	logDice	Kategorija
1	briti norce	563	40	13,49133	frazeologija
2	ovratnica proti bolham	25	7	12,92961	ok – da
3	alkoholne pijače	9.140	24	12,64524	ok – nevtralno
4	brusiti si/se kremplje	49	10	12,53789	frazeologija
5	Bajke in povesti	142	10	12,45278	lastno ime
6	oprijeti se kot bilke	26	9	12,23283	frazeologija
7	dlake na jeziku	3.643	7	12,22364	frazeologija
8	prisluškovati pogovorom	666	44	12,12997	ok – nevtralno
9	barvan z barvili	20	7	12,00905	ok – nevtralno
10	drama s talci	251	7	11,93812	ok – da
11	jezik za zobmi	493	5	11,75747	frazeologija
12	alkohol in droge	1.052	15	11,68357	ok – nevtralno
13	brinove jagode	761	8	11,64435	ok – nevtralno
14	priloga k jedem	186	12	11,61657	ok – nevtralno
15	grozdne jagode	863	14	11,59745	ok – nevtralno
16	droge in alkohol	748	16	11,47025	ok – nevtralno
17	priloga jedem	65	4	11,42804	ok – nevtralno
18	babice z vnučki	24	9	11,39514	ok – nevtralno
19	ne briti si nog	11	6	11,39334	ok – da
20	travne bilke	558	16	11,3707	ok – nevtralno
21	aranžmaji iz cvetja	60	9	11,12389	ok – nevtralno
22	prisluhi arbitru	15	3	11,03037	ok – da
23	kotli na biomaso	199	12	11,02698	ok – ne
24	alkohol in mamila	616	11	11,02019	ok – da
25	aluminijasta platišča	649	17	10,98494	ok – nevtralno
26	aplikacija za telefone	461	18	10,94286	ok – nevtralno
27	počitnice in dopusti	217	15	10,90396	ok – nevtralno
28	stopalke so aluminijaste	18	4	10,88452	ok – da
29	oprijeti se bilke	47	7	10,88216	frazeologija
30	kitara s strunami	23	7	10,86178	ok – da

precejšnjo variantnost (prim. Gantar 2021a). Preostale lahko razdelimo na tri kategorije – kolokacije, pri katerih je množinska oblika (a) upravičena ali nujna; (b) neupravičena ali napačna; (c) morda bolj pogosta, vendar bi lahko pričakovali, da bo slovarska oblika v ednini. Pri tistih, ki smo jih uvrstili pod kategorijo (a), lahko preverimo upravičenost z navedbo edninske oblike: *alkohol in mamilo, ne briti si noge, ovratnica proti bolhi, prisluh arbitru, stopalka je aluminijasta, kitara s struno*. Upravičenost množinske oblike verjetno ni na povsem enaki ravni pri vseh navedenih (*ne briti si noge* proti *stopalka je aluminijasta*), vendar se zdi, da je tehtnica močno nagnjena na stran upravičenosti. Nasprotno se v enem od primerov zdi, da je množinska oblika povsem neupravičena in predpostavimo lahko, da je to zaradi terminološkosti: *kotli na biomaso*. Največja je skupina (c), pri kateri bi morda prej pričakovali edninsko obliko, množinska pa ni izrazito moteča. Podobno kot v primeru kategorije (a) lahko upravičenost preverimo z navedbo edninske oblike: *alkohol in droga, alkoholna pijača, aluminijasto platišče, aplikacija za telefon, aranžma iz cvetja, babica z vnučkom, barvan z barvilom, brinova jagoda, droga in alkohol, grozdna jagoda, priloga jedi, priloga k jedi, prisluškovati pogovoru, travna bilka, počitnice in dopust*.

Na nekoliko manjšem naboru preverimo tudi izluščene dvojinske oblike – uporabljeni so bili enaki filtri kot v primeru množine, z dodanim kriterijem logDice = najmanj 5. Kot vidimo v spodnji Tabeli 8, pri 88 izbranih geslih na vrhu nabora (razvrščenega po logDice) pravzaprav ni upravičenih dvojinskih oblik.

Če preverimo širši nabor izluščenih dvojinskih oblik iz cele kolokacijske baze, je sicer mogoče najti primere, pri katerih bi bil izpis dvojinske oblike upravičen, zlasti v primeru parnih organov ali v podobnih parnih situacijah*: ledvici odpovesta, uiti med nogama, enojajčni dvojčici* itd. Sklenemo lahko, da kljub v korpusu izkazani prevladujoči množinski (ali dvojinski) obliki izpostavitev množinske oblike večinoma ni upravičena. Statistični kriteriji za oženje nabora, ki bi izpostavil zgolj kategorijo (a) iz gornje analize, ostaja naloga v okviru nadaljnjega dela.

Tabela 8: Prvih 14 kolokacijskih kandidatov po meri logDice glede na izkazan zapis dvojinske oblike pri samostalniški komponenti kolokacije.

Št.	Kolokacija	GF2.1 (F)	Št. oblik	logDice	KATEGORIJA
1	kitari in ojačevalec	22	9	8,13976	ok – ne
2	gorilnika na biomaso	10	4	7,56204	ok – ne
3	vlogi iz drame	16	8	7,3752	ok – ne
4	bolnišnici v Soboti	107	7	7,30834	ok – ne
5	panogi rudarstva	14	3	7,26529	ok – ne
6	pošiljki z blagom	14	6	6,92624	ok – ne
7	zmečkani jagodi	27	6	6,66875	ok – ne
8	babici in prijateljica	13	7	5,86006	ok – ne
9	aparaturi za bolnišnico	11	4	5,72472	ok – ne
10	posojilna aranžmaja	20	6	5,71633	ok – ne
11	aluminijasta zavitka	14	5	5,61383	ok – ne
12	jezika Unije	25	3	5,52257	ok – ne
13	prispevka v jeziku	34	8	5,1183	ok – ne
14	Bolnišnici v Kabulu	15	5	5,06317	ok – ne

3.3 Stopnjevanje pri pridevniških in prislovnih kolokacijah

Pri pridevniku in prislovu variabilnost preverjamo tudi na ravni stopnjevanja – torej če so v korpusu v konkretni kolokaciji prevladujoče primerniške in presežniške oblike, v primerjavi z osnovnikom, ki je tudi privzeta oblika leme pri pridevnikih in prislovih. V primeru stopnjevanja gre za nekoliko drugačno oceno izluščenih oblik. Uporabljamo samo dve kategoriji: ‘da’ in ‘pomen’. V prvem primeru ugotavljamo, da osnovnik do te mere že na prvi ravno spremeni pomen kolokacije, da je presežniška ali primerniška oblika nujna. V drugem primeru pa se na ravni izolirane kolokacije zdi, da bi lahko izpisovali kombinacijo z osnovnikom, vendar je od primera do primera treba preverjati odtenke pomena. Ker imamo štiri kombinacije primernikov in presežnikov pri pridevniku in prislovu, tokrat izpisujemo po 15 kolokacij, razvrščenih po meri logDice, s standardnimi filtri.

Pridevnik, stopnja – presežnik:

Tabela 9: Prvih 15 kolokacijskih kandidatov po meri logDice glede na izkazan zapis preseżniške oblike pri pridevniški komponenti kolokacije.

Št.	Kolokacija	GF2.1 (F)	Št. oblik	logDice	Kategorija
1	najbližja bolnišnica	234	11	6,39696	ok – da
2	najboljši v panogi	46	13	6,36637	ok – da
3	najbližji bife	29	5	5,23296	ok – da
4	Najcenejši aranžmaji	44	17	5,08256	ok – pomen
5	najpopularnejša aplikacija	34	12	4,95107	ok – pomen
6	najrazličnejše blago	525	12	4,88738	ok – pomen
7	najdražje blago	58	10	4,45775	ok – pomen
8	najgloblja intima	36	9	4,42627	ok – pomen
9	najbližja obala	36	8	4,18586	ok – da
10	najmočnejša panoga	149	22	4,17684	ok – pomen
11	najdražji aranžma	37	19	4,12104	ok – pomen
12	najljubša kitara	18	8	3,96283	ok – da
13	najproduktivnejša panoga	14	6	3,85275	ok – pomen
14	najhitrejše panoge	86	15	3,83533	ok – pomen
15	najenostavnejši alkohol	11	6	3,82769	ok – pomen

Pridevnik, stopnja – primernik:

Tabela 10: Prvih 15 kolokacijskih kandidatov po meri logDice glede na izkazan zapis primerniške oblike pri pridevniški komponenti kolokacije.

Št.	Kolokacija	GF2.1 (F)	Št. oblik	logDice	Kategorija
1	nevarnejši od alkohola	12	4	7,73243	ok – da
2	krajši dopust	544	28	6,27491	ok – pomen
3	višji v panogi	12	4	6,20965	ok – pomen
4	daljši dopust	721	42	6,14763	ok – pomen
5	večji v panogi	23	14	6,04732	ok – pomen
6	zgodnejša civilizacija	20	10	5,22835	ok – pomen
7	raznovrstnejše aplikacije	24	6	4,77585	ok – pomen
8	vrednejše blago	49	10	4,64079	ok – pomen
9	požrešnejša aplikacija	12	5	4,57426	ok – pomen
10	manjše bolnišnice	233	20	3,91739	ok – pomen

Št.	Kolokacija	GF2.1 (F)	Št. oblik	logDice	Kategorija
11	podrobnejše informiranje	11	6	3,6403	ok – pomen
12	poznejša drama	18	8	2,90871	ok – da
13	nižji alkohol	28	13	2,35271	ok – pomen
14	lažja embalaža	10	7	2,02679	ok – pomen
15	višji alkoholi	72	20	1,93463	ok – pomen

Prislov, stopnja – presežnik:

Tabela 11: Prvih 15 kolokacijskih kandidatov po meri logDice glede na izkazan zapis presežniške oblike pri prislovni komponenti kolokacije.

Št.	Kolokacija	GF2.1 (F)	Št. oblik	logDice	Kategorija
1	najglasneje se/si omenjati	99	25	8,52083	ok – pomen
2	najglasneje kričati	202	35	8,32226	ok – pomen
3	najglasneje vzklikati	181	22	8,2651	ok – pomen
4	najodločneje in najglasneje	29	5	8,16101	ok – pomen
5	najraje brati	321	31	7,59477	ok – pomen
6	največ investirati	104	13	6,21955	ok – pomen
7	najbolj mučiti	194	13	5,98804	ok – pomen
8	največ prihraniti	90	18	5,95843	ok – pomen
9	največ brati	71	12	5,15924	ok – pomen
10	najglasneje završati	14	6	4,61861	ok – pomen
11	najglasneje rohneti	13	7	4,54657	ok – pomen
12	najglasneje napadati	16	8	4,37703	ok – pomen
13	največkrat brati	27	13	4,3344	ok – pomen
14	najglasneje rigati	10	8	4,18542	ok – pomen
15	najglasneje rjuti	10	6	4,18238	ok – pomen

Prislov, stopnja – primernik:

Tabela 12: Prvih 15 kolokacijskih kandidatov po meri logDice glede na izkazan zapis primerniške oblike pri prislovni komponenti kolokacije.

Št.	Kolokacija	GF2.1 (F)	Št. oblik	logDice	Kategorija
1	bližje k obali	16	3	9,69166	ok – pomen
2	glasneje se/si pritoževati	220	25	9,37821	ok – pomen

Št.	Kolokacija	GF2.1 (F)	Št. oblik	logDice	Kategorija
3	glasneje opozarjati	732	33	9,09676	ok – pomen
4	glasneje se/si govoriti	302	18	9,00254	ok – pomen
5	pogosteje in glasno	49	9	8,93084	ok – pomen
6	več v jezikih	718	4	8,60427	nekolokacija
7	glasneje izražati	217	23	8,18625	ok – pomen
8	glasneje se oglašati	63	17	8,13513	ok – pomen
9	dlje od obale	58	7	8,03044	ok – pomen
10	glasneje in dolgo	18	4	7,96169	ok – pomen
11	glasneje se spraševati	137	20	7,79925	ok – pomen
12	glasneje govoriti	836	55	7,74834	ok – pomen
13	glasneje slišati	20	7	7,74026	ok – pomen
14	glasneje zavpiti	123	21	7,69155	ok – pomen
15	glasneje napovedovati	158	16	7,27307	ok – pomen

Kot je razvidno iz Tabel 11 in 12, je bilo pri 88 iztočnicah izluščenih razmeroma malo kolokacij, pri katerih je nujno treba uporabiti primerniško ali presežniško obliko. Večinoma so te povezane s pridevniškimi, ki se redko uporabljajo (npr. *blizek*), ali pa je med obema oblikama izrazita pomenska razlika. Na primer: *blizka bolnišnica, dober v panogi, blizek bife, blizka obala, ljuba kitara, nevaren od alkohola, pozna drama*. Zdi se, da primerniške in presežniške oblike ne bi bile moteče, vendarle pa bi bilo s stališča luščenja tipičnih kolokacij problematično, če bi zaradi neizrazite večine obeh neosnovnih oblik umanjkala kolokacija z nestopnjevano obliko. Analiza torej kaže, da bi bilo bolj ustrezno, če bi pri luščenju upoštevali presežniške in primerniške oblike samo v primerih, ko osnovnih oblik sploh ne bi našli v korpusu.

3.4 Zapis z malimi ali velikimi črkami

Pri vseh izluščenih komponentah dopuščamo variantnost tudi na ravni zapisa z velikimi in/ali malimi črkami. S tem dobimo vpogled v realni prevladujoči zapis v korpusu, ki kaže zanimive rezultate. V Tabeli 13 za 88 iztočnic navajamo 30 najpogostejših kolokacij, pri katerih je ena od komponent (prevladujoče) zapisana z veliko

začetnico ali z velikimi črkami. Tokrat je tabela razvrščena po absolutnih frekvencah iz korpusa Gigafida 2.1. Filtriramo tudi po številu oblik – najmanj 3.

Tabela 13: Prvih 30 kolokacijskih kandidatov po absolutni frekvenci glede na izkazan zapis z veliko začetnico ali z velikimi črkami pri kateri od komponent kolokacije.

Št.	Kolokacija	GF2.1 (F)	Št. oblik	logDice	Kategorija
1	Splošna bolnišnica	9.606	34	10,59992	ime ustanove
2	Psihiatrična bolnišnica	4.034	20	10,6693	ime ustanove
3	Sobotna priloga	3.952	20	10,90602	ime publikacije
4	ljubljanska Drama	3.581	23	8,39662	ime ustanove
5	Slonokoščena obala	3.521	14	11,29249	zemljepisno ime
6	Jugoslovanska armada	2.355	17	10,02901	ime ustanove
7	Rdeča armada	2.137	18	8,56928	ime ustanove
8	Slovar jezika	1.786	27	10,18376	ime publikacije
9	Azurna obala	1.561	10	10,20049	zemljepisno ime
10	priloga Dela	1.553	15	7,06795	ime publikacije
11	Teden drame	1.239	22	8,79442	ime dogodka
12	Mala drama	1.219	20	7,37536	ime ustanove
13	Romantična drama	1.169	17	9,3298	ime žanra
14	Komična drama	994	13	9,57607	ime žanra
15	mariborska Drama	889	14	7,54256	ime ustanove
16	Severna obala	860	19	7,56564	zemljepisno ime
17	bolnišnica Jesenice	759	12	10,85457	ime ustanove
18	obala ZDA	710	8	8,85223	zemljepisno ime
19	Delova priloga	650	8	10,19272	ime publikacije
20	Irska armada	637	8	9,27752	ime ustanove
21	Biografska drama	636	14	9,07941	ime žanra
22	Inštitut za jezik	621	17	8,82975	ime ustanove
23	oder Drame	587	14	9,84566	ime ustanove
24	Program v jeziku	583	21	8,78801	ime publikcije
25	Kriminalna drama	559	11	8,43731	ime žanra
26	Novinarsko razsodišče	528	15	7,464	ime ustanove
27	Akcijska drama	437	12	8,00553	ime žanra
28	Aplikacija omogoča	409	19	7,34898	ne-ime
29	obala Amerike	402	15	8,27683	zemljepisno ime
30	Center za informiranje	394	18	7,41138	ime ustanove

Po pričakovanju prevladujejo imena ustanov, publikacij, zemljepisna imena, pogosta so tudi imena žanrov, dogodkov, na listi se pojavljata tudi ena kolokacija, ki ni ime (*Aplikacija omogoča*). Beleženje zapisa z velikimi ali malimi črkami je koristno predvsem zato, ker na očiten način opozarja, da pri izluščeni kolokaciji ne gre za splošno besedišče, temveč za takšna ali drugačna lastna imena, ki jih ne želimo vključiti v slovarske baze ali analize kolokacijskih podatkov.

4 Zaključek

V prispevku smo opisali nov postopek luščenja kolokacijskih kandidatov iz poljubnega korpusa. Novi formalizem za luščenje kolokacij upošteva poljubne nivoje korpusnih oznak, za kar uporablja lasten (generičen) sistem za definiranje omejitev na kateremkoli nivoju označevanja, od besednih vrst in njihovih lastnosti, skladenjskih povezav in njihovih oznak, konkretnih leksikalnih elementov, ter drugih nivojev označevanja, npr. za označevanje semantičnih vlog, semantičnih tipov itd. Za potrebe avtomatizacije postopka luščenja je v novem sistemu poleg omejitev, pri katerih upoštevamo poljubni nivo oznak v korpusu, mogoče tudi določiti, katera od oblik posamezne komponente, ki jo najdemo v korpusu, naj bo izpisana v konkretni kolokaciji, glede na možnosti znotraj predvidene kanonične oblike kolokacije pri konkretni kolokacijski strukturi.

V drugem delu članka smo izpostavili nekatere elemente variabilnosti pri izpisu kolokacij, ki jih omogoča novi sistem. Ti vključujejo: razmerje med določnimi in nedoločnimi oblikami pridevnika v moškem spolu ednine imenovalnika; edninske, dvojinske ali množinske oblike samostalnika; stopnjevanje (primernik, presežnik) pri pridevniku in prislovu; zapis z velikimi in malimi črkami pri vseh elementih kolokacij. Analiza kaže, da je možnost upravljanja z izpisanimi oblikami koristna, vendar bi bilo treba v večini primerov zvišati prag oz. dodatno opredeliti parametre za upoštevanje teh pojavov pri izpisu kolokacij.

5 Nadaljnje delo

Pri načrtovanju nadaljnjega dela se kažejo predvsem naslednje prioritete:

1. Nadgradnja kolokacijskih struktur z binarnih na t. i. razširjene kolokacije. V obstoječih 82 skladenjskih strukturah upoštevamo zgolj binarne kolokacije. V kolokacijah je v nekaterih primerih smiselno izpostaviti tudi dodatne elemente, pri čemer je osnovna binarna kolokacija ohranjena, kljub temu pa dodatni element eksplicitno navedemo. Na primer: *govoriti jezik → govoriti [angleški, francoski, ...] jezik*. Z naborom skladenjskih struktur je nastavljen sistem, ki omogoča kombiniranje obstoječih struktur v kompleksnejši nabor, ki upošteva tudi identifikacijo razširjenih kolokacij.
2. Upoštevanje statističnih podatkov o razpršenosti po virih oz. žanrih. Statističnim podatkom, ki jih v obstoječem sistemu pripisujemo izluščenim kolokacijam, je mogoče dodati tudi metabesedilne podatke iz korpusa, kot je npr. besedilna razpršenost (podatek o številu različnih besedil, v katerih se kolokacija pojavi) ali razpršenost po posameznih virih (npr. če je kolokacija omejena na časnik Delo ipd.). Podobno je mogoče upoštevati tudi časovno dimenzijo, kar pomeni, da poleg distribucije po žanrih oziroma virih upoštevamo tudi razpršenost glede na posamezno leto, česar trenutna statistika ne ponuja.
3. Natančnejša določitev parametrov za obliko izpisa kolokacij: kot je pokazala analiza, je možnost upravljanja z izpisom oblik kolokacije pomemben mehanizem, ki pripomore k temu, da lahko avtomatsko luščimo kolokacije v naravnejši obliki. Mehanizem je smiselno nadgraditi z natančnejšimi opredelitvami, kdaj se dodatne lastnosti dejansko upoštevajo in kdaj ne.
4. Upoštevanje drugih ravni označevanja: v času trajanja projekta NSSSS je semantično označevanje korpusov (prepoznavanje imenskih entitet, semantičnih tipov, semantičnih shem/okvirov, strojno prepoznavanje pomenov, wikifikacija itd.) doživelo precejšen napredek, predvsem z uvajanjem novih

tehnologij – globokih nevronskih mrež. To pomeni, da je pri nadaljnjem delu treba upoštevati tudi naslednji – semantični – nivo označevanja, ki bo po vsej verjetnosti prinesel še boljše rezultate, predvsem pri sestavljanju kolokacij v gruče, ki jih potem lahko pripišemo ustreznemu slovarskemu pomenu.

Zahvala

Prispevek je nastal v okviru raziskovalnega projekta Nova slovnica sodobne standardne slovenščine: viri in metode (J6-8256) ter v okviru programskih skupin Slovenski jezik – bazične, kontrastivne in aplikativne raziskave (P6-0215) in Jezikovni viri in tehnologije za slovenski jezik (P6-0411), ki jih financira Agencija za raziskovalno dejavnost Republike Slovenije.

Reference

Erjavec, T., Krek, S., Arhar, Š., Fišer, D., Ledinek, N., Saksida, A., Sivec, B. in Trebar, B. (2010a). Oblikoskladenjske specifikacije JOS V1.1. Dostopno prek: http://nl.ijs.si/jos/msd/html-sl/index.html.

Erjavec, T., Fišer, D., Krek, S. in Ledinek, N. (2010b). The JOS Linguistically Tagged Corpus of Slovene. V N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, D. Tapias (ur.), *LREC 2010: Proceedings of the Seventh International Conference on Language Resources and Evaluation* (str. 1806–1809). European Language Resources Association. Dostopno prek: http://www.lrec-conf.org/proceedings/lrec2010/pdf/139_Paper.pdf.

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

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.

Gantar, P. (2015). *Leksikografski opis slovenščine v digitalnem okolju*. Ljubljana: Trojina, zavod za uporabno slovenistiko. E-izdaja (2018). Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/62/138/2602-1.

Gantar, P. in Gorjanc, V. (2015). Obrazilo -en/-ni v slovarski obravnavi pridevnikov. V M. Smolej (ur.), *Slovnica in slovar: aktualni jezikovni*

opis, *Obdobja 34* (str. 233–241). Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: https://centerslo.si/wp-content/uploads/2015/11/34_1-Gantar-Gor.pdf.

Gantar, P. (2021a). Zapis kanonične oblike frazeoloških enot v Leksikonu večbesednih enot za slovenščino. V Š. Arhar Holdt (ur.), *Nova slovnica sodobne standardne slovenščine: viri in metode* (str. 198–230). Ljubljana: Znanstvena založba Filozofske fakultete.

Gantar, P. (2021b). Strojno berljiv Vezljivostni leksikon slovenskih glagolov. V Š. Arhar Holdt (ur.), *Nova slovnica sodobne standardne slovenščine: viri in metode* (str. 259–297). Ljubljana: Znanstvena založba Filozofske fakultete.

Gorjanc, V., Gantar, P., Kosem, I. in Krek, S. (2015) Predgovor. V V. Gorjanc, P. Gantar, I. Kosem in S. Krek (ur.), *Slovar sodobne slovenščine: problemi in rešitve* (str. 9–12). Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/15/47/478-1.

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

Kosem, I., Gantar, P., Krek, S., Arhar Holdt, Š., Čibej, J., Laskowski, C. A., 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., Gantar, P., Kosem, I., Dobrovoljc, K., Arhar Holdt, Š., Čibej, J., Laskowski, C. A., 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.

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.), *LREC 2020: Twelfth International Conference on Language Resources and Evaluation*: *conference proceedings* (str. 3340–3345). Pariz: European Language Resources Association. Dostopno prek: http://www.lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf.

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* (str. 358–378). Ljubljana: Znanstvena založba Filozofske fakultete. Dostopno prek: https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/15/47/520-1.

Krek, S. in Kilgarriff, A. (2006). Slovene word sketches. V T. Erjavec in J. Žganec Gros (ur.), *Jezikovne tehnologije: zbornik 9. mednarodne multikonference Informacijska družba IS 2006* (str. 62–67). Ljubljana: Institut Jožef Stefan. Dostopno prek: http://nl.ijs.si/is-ltc06/proc/12_Krek.pdf.

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* (str. 43–77). Ljubljana: Znanstvena založba Filozofske fakultete.

Ramisch, C. (2020). Computational phraseology discovery in corpora with the MWETOOLKIT. V G. Corpas Pastor in J-P Colson (ur.), *Computational Phraseology* (str. 111–134). Amsterdam; Philadelphia: John Benjamins Publishing. https://doi.org/10.1075/ivitra.24.

Ramisch, C., Savary, A., Guillaume, B., Waszczuk, J., Candito, M., Vaidya, A., Barbu Mititelu, V., Bhatia, A., Iñurrieta, U., Giouli, V., Güngör, T., Jiang, M., Lichte, T., Liebeskind, C., Monti, J., Ramisch, R., Stymne, S., Walsh, A., Xu, H. (2020). Edition 1.2 of the PARSEME Shared Task on Semi-supervised Identification of Verbal Multiword Expressions. V S. Markantonatou, J. McCrae, J. Mitrović, C. Tiberius, C. Ramisch, A. Vaidya, P. Osenova in A. Savary (ur.), *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons* (str. 107–118). Dostopno prek: https://aclanthology.org/2020.mwe-1.14.pdf.

Priloga: Nabor struktur

ID	Oznaka	Zgled	1	2	3	4	5	6	7	8	9	Št. kolokacij
34	p0-s0	svetovno prvenstvo		p0						s0		720.605
53	s0-s2	direktor podjetja		s0						s2		518.199
70	s0-gg	raziskava pokaže		s0						gg		385.018
23	gg-s4	podpisati pogodbo		gg						s4		270.965
15	gg-d-s5	imeti v mislih		gg			d			s5		235.771
43	r-gg	dobro poznati		r						gg		176.804
106	s0-vp-s0	sadje in zelenjava		s0		vp				s0		175.994
52	s0-d-s5	razmere na trgu		s0			d			s5		172.684
14	gg-d-s4	odgovoriti na vprašanje		gg			d			s4		122.875
51	s0-d-s4	odgovor na vprašanje		s0			d			s4		95.407
57	s0-gp-p1	odločitev je sprejeta		s0					gp	p1		94.762
71	s0-zp-gg	nesreča se zgodi		s0	zp					gg		91.004
16	gg-d-s6	začeti z delom		gg			d			s6		83.300
13	gg-d-s2	priti do zmage		gg			d			s2		68.925
46	r-p0	zelo pomemben		r						p0		61.175
50	s0-d-s6	ravnanje z odpadki		s0			d			s6		60.876
81	r-zp-gg	dobro se znajti		r	zp					gg		60.334
89	gg-zp-d-s5	znajti se v položaju		gg	zp		d			s5		57.958
48	s0-d-s2	dostop do informacij		s0			d			s2		47.461
22	gg-s3	pomagati ljudem		gg						s3		34.757
88	gg-zp-d-s4	uvrstiti se v finale		gg	zp		d			s4		33.743
30	p0-vp-p0	domač in tuj		p0		vp				p0		32.127
90	gg-zp-d-s6	ukvarjati se s športom		gg	zp		d			s6		27.580
77	s1-gp-s1	nogomet je šport		s1					gp	s1		26.520
47	r-s2	nekaj časa		r						s2		22.664
12	gg-ggn	morati plačati		gg						ggn		20.277
74	l-gg-s2	ne dobiti odgovora	l	gg						s2		19.734
72	s0-l-gg	trditev ne drži		s0				l		gg		19.400
85	p0-r	[biti] znan danes		p0						r		18.425
27	p0-d-s4	izvoljen za predsednika		p0			d			s4		17.344
55	r-r	pretežno oblačno		r						r		16.969
54	s0-s3	pomoč otrokom		s0						s3		13.952

ID	Oznaka	Zgled	1	2	3	4	5	6	7	8	9	Št. kolokacij
76	s1-s1	države članice		s1						s1		13.393
69	gg-zp-s4	vzeti si čas		gg	zp					s4		13.224
29	p0-d-s6	določen z zakonom		p0			d			s6		12.407
86	gg-zp-d-s2	vrniti se z dopusta		gg	zp		d			s2		11.643
17	gg-d-s3	povabiti k sodelovanju		gg			d			s3		9.899
108	gg-zp-s2	lotiti se dela		gg	zp					s2		9.212
68	gg-zp-s3	odzvati se vabilu		gg	zp					s3		9.107
25	l-gg-ggn	ne smeti pozabiti	l	gg						ggn		8.639
82	gg-vd-s0	navesti kot razlog		gg		vd				s0		8.160
28	p0-d-s5	zaposlen v podjetju		p0			d			s5		7.492
93	gg-ggn-zp	začeti ukvarjati se		gg						ggn	zp	7.331
18	gg-p1	ostati nespremenjen		gg						p1		7.204
26	p0-d-s2	sestavljen iz delov		p0			d			s2		6.783
49	s0-d-s3	boj proti korupciji		s0			d			s3		6.028
40	r-d-s5	takoj na začetku		r			d			s5		5.982
36	p0-s3	[biti] namenjen otrokom		p0						s3		4.948
41	r-d-s4	pozno v noč		r			d			s4		4.668
38	r-d-s2	daleč od resnice		r			d			s2		4.321
107	gg-vp-gg	brati in pisati		gg		vp				gg		4.015
98	r-ggn	[biti] moč videti		r						ggn		3.828
73	s0-zp-l-gg	ljudje se ne zavedajo		s0	zp			l		gg		3.790
42	r-d-s6	malo pred polnočjo		r			d			s6		3.536
96	l-gg-ggn-zp	ne smeti privoščiti si	l	gg						ggn	zp	3.154
44	r-vp-r	bolj ali manj		r		vp				r		2.818
100	p1-ggn	[biti] sposoben doseči		p1						ggn		2.470
92	gg-zp-ggn	odločiti se narediti		gg	zp					ggn		2.330
24	ggz-s2	ne imeti težav		ggz						s2		2.233
83	gg-zp-vd-s0	boriti se kot lev		gg	zp	vd				s0		2.135
87	gg-zp-d-s3	cepiti se proti gripi		gg	zp		d			s3		2.132
45	r-vd-s1	manj kot polovica		r		vd				s1		2.015
35	p0-s2	[biti] deležen pozornosti		p0						s2		1.940
78	gg-zp-p1	vrniti se zdrav		gg	zp					p1		1.828
102	s0-ggn	priložnost videti		s0						ggn		1.691

ID	Oznaka	Zgled	1	2	3	4	5	6	7	8	9	Št. kolokacij
32	p0-vd-s1	čist kot solza		p0		vd				s1		1.346
19	gg-p4	pustiti ravnodušnega		gg						p4		1.183
75	l-gg-zp-s2	ne delati si utvar	l	gg	zp					s2		1.037
104	gg-ggm	iti spat		gg						ggm		936
84	s1-vd-s1	država kot lastnik		s1		vd				s1		914
91	s0-d-r	načrt za letos		s0			d			r		780
99	r-ggn-zp	[biti] bolje izogniti se		r						ggn	zp	592
31	p0-d-s3	povabljen k sodelovanju		p0			d			s3		494
95	l-gg-zp-ggn	ne uspeti se uvrstiti	l	gg	zp					ggn		479
101	p1-ggn-zp	[biti] pripravljen pogovarjati se		p1						ggn	zp	354
80	gg-zp-p4	počutiti se varnega		gg	zp					p4		295
39	r-d-s3	nazaj k naravi		r			d			s3		207
105	gg-ggm-zp	priti ogledat si		gg						ggm	zp	187
103	s0-ggn-zp	pravica seznaniti se		s0						ggn	zp	125
37	p2-s2	[biti] slabše kakovosti		p2						s2		19
94	gg-zp-ggn-zp	odločiti se vrniti se		gg	zp					ggn	zp	5
97	l-gg-zp-ggn-zp	ne bati se pokazati se	l	gg	zp					ggn	zp	0
		Skupaj										**4.002.918**