Kolokacije in časovni trendi

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

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

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

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

1 Uvod

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

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

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

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

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

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

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

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

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

2 Metodologija

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

2.1 Priprava podatkov

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

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

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

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

Tabela 1: Deleži kolokacijskih kandidatov po frekvenčnih rangih.

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

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

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

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

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

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

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

2.2 Statistična obdelava

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

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

2.2.1 Naklon linearne regresije

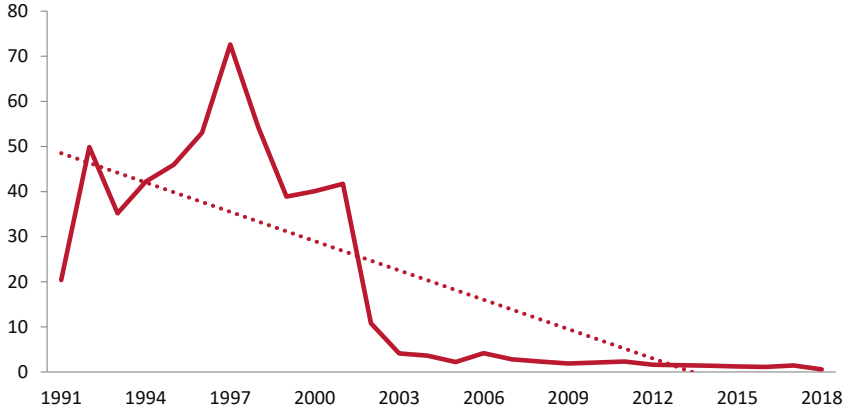
Za vsako kolokacijo smo vzeli nabor njenih relativnih pogostosti iz korpusa Gigafida 2.0 po letih in na podatkih zmodelirali linearno

5 https://orange.biolab.si/

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

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

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



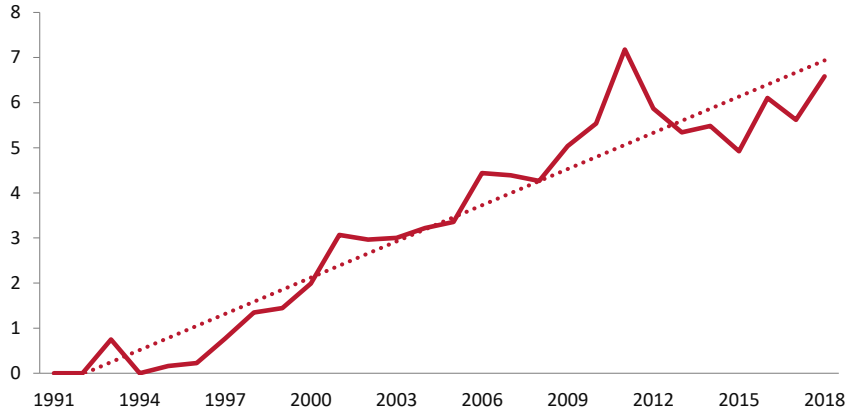
Slika 2: Relativna pogostost kolokacije *nemška marka* po letih v korpusu Gigafida 2.0.

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

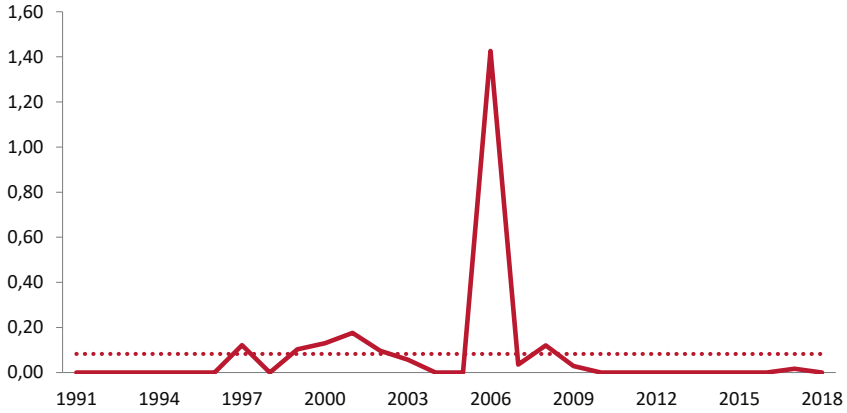
2.2.2 Koeficient določenosti

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

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



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



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

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

2.2.3 Razmerje med maksimalno in povprečno relativno pogostostjo

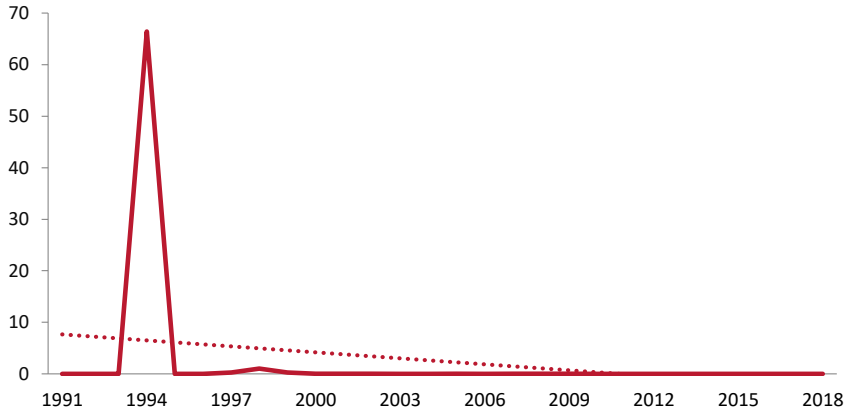
Tretja mera, ki smo jo uporabili za spremljanje trendov rabe kolokacij, je razmerje med maksimalno in povprečno relativno pogostostjo (*m*). Izračunamo ga po spodnji formuli, pri čemer je f_{rmax} najvišja letna relativna pogostost dane kolokacije, f_{ri} letna relativna pogostost, i_0 začetno leto in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

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

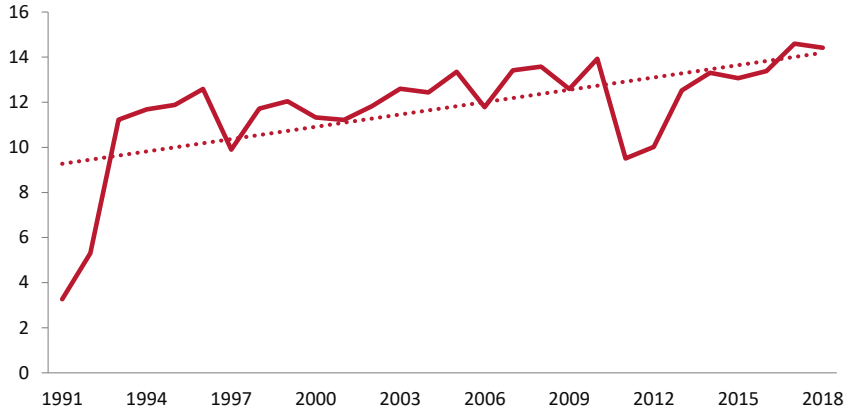
Če je pogostost rabe kolokacije v jeziku skozi čas povsem konstantna, je vrednost *m* enaka 1 (maksimalna in povprečna relativna pogostost sta v tem primeru enaki). Višja kot je maksimalna relativna pogostost kolokacije in bolj kot odstopa od povprečja, višja je vrednost *m*. Višja vrednost m torej nakazuje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

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

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



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

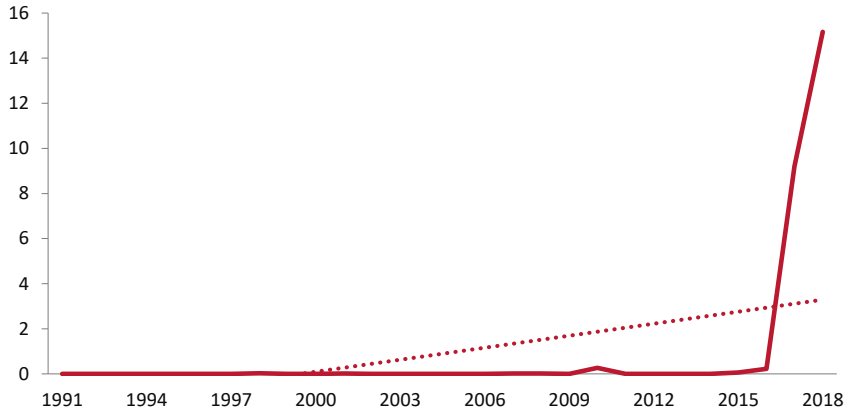


Slika 6: Relativna pogostost kolokacije *pomemben del* po letih v korpusu Gigafida 2.0.

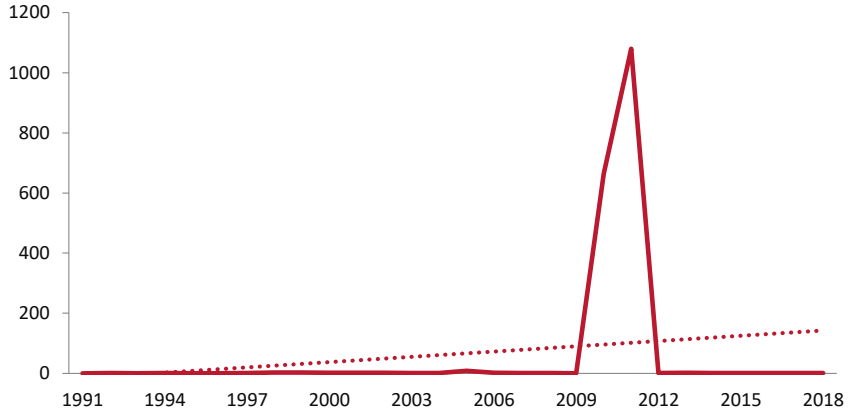
tako izrazito od povprečja, saj so letne relativne pogostosti medsebojno relativno primerljive (Slika 6).

2.2.4 Količnik nedavne rasti

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



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



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

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

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

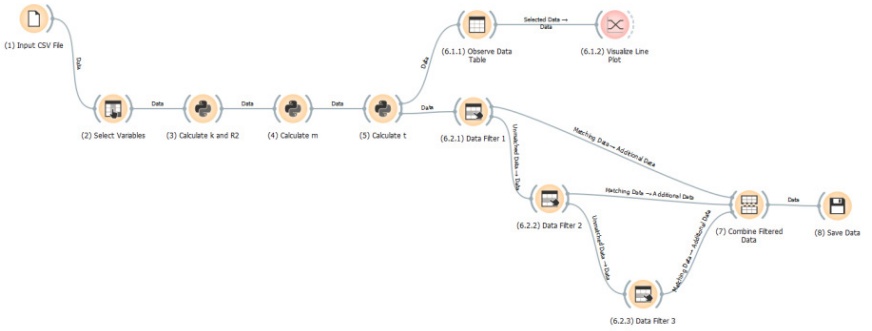
Višje kot so relativne pogostosti kolokacije v zadnjih treh opazovanih letih in manjše kot so njene relativne pogostosti v predhodnih opazovanih letih, višja je vrednost količnika *t*. Višji količnik nedavne rasti torej nakazuje, da raba kolokacije v zadnjih letih strmo narašča. Primer kolokacije z visokim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

2.3 Prototip delotoka za spremljanje kolokacijskih trendov

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

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

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



Slika 9: Delotok COLTREND v okolju Orange.

3 Analiza

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

3.1 Celostni pogled na izluščene podatke

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

ali

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Tabela 10: Primeri ustaljenih kolokacij v korpusu Gigafida 2.0.

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

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

3.2 Analiza na nivoju posameznih iztočnic

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

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

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

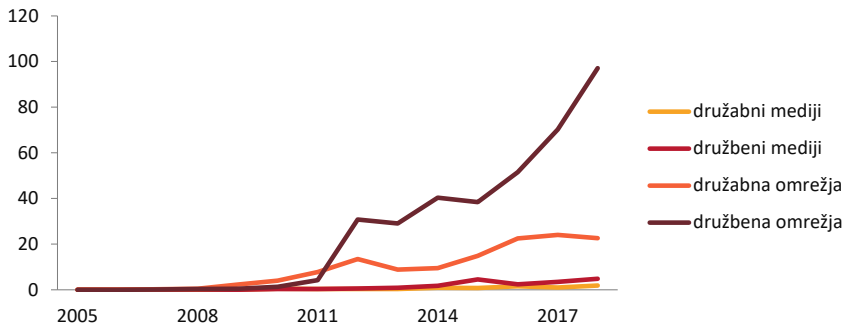
Kolokacija	Absolutna pogostost	k	R^2	m	t
avtonomna vožnja	618	0,084	0,319	8,091	13,689
avtonomna vozila	565	0,077	0,334	8,339	11,495
avtonomni avtomobili	169	0,023	0,359	5,295	4,337
avtonomna cona	391	0,017	0,134	4,637	1,489
avtonomna regija	281	0,009	0,087	2,746	1,433
avtonomna skupnost	180	0,010	0,351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

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

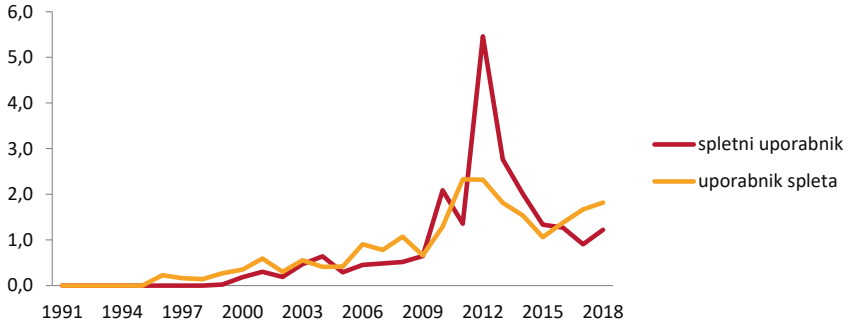
Tabela 12: Konkurenčne kolokacije *družbeni mediji*, *družabni mediji*, *družabna omrežja* in *družbena omrežja* v korpusu Gigafida 2.0.

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

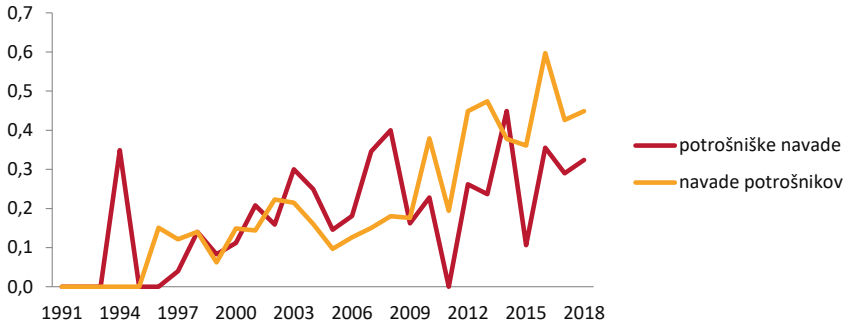
Tako po pogostosti kot tudi po količniku nedavne rasti *t* in količniku *m* izstopa kolokacija *družbena omrežja*. Vizualizacija razporeditve relativnih pogostosti vseh štirih kandidatov (Slika 10) pokaže, da je do leta 2011 prednjačila kolokacija *družabna omrežja*, zatem pa so jo s strmo rastjo, ki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



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



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

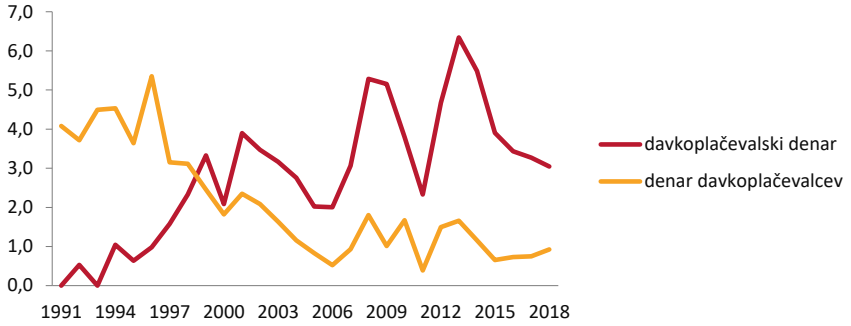


Slika 12: Relativne pogostosti konkurenčnih kolokacij *potrošniške navade* in *navade potrošnikov* v korpusu Gigafida 2.0.

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

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

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



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

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

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

4 Diskusija

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

4.1 Izdelava novih slovarskih virov

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

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

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

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

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

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

4.2 Posodabljanje obstoječih slovarjev

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

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

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

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

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

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

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

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

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

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

4.3 Kolokacijski trendi in slovarski uporabniki

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

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

9 https://www.dwds.de/

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

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

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

5 Zaključek

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

10 https://www.dictionary.com/

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

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

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

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

Zahvala

Projekt *Kolokacije kot temelj jezikovnega opisa: semantični in časovni vidiki* (J6-8255), projekt *Nova slovnica sodobne standardne slovenščine: viri in metode* (J6-8256) in raziskovalni program št. P6-0411 (*Jezikovni viri in*

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

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

Reference

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Herbst, T. (1996): What are Collocations: Sandy Beaches or False Teeth. *English Studies 4*: 379–93.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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