Baza kolokacijskega slovarja slovenskega jezika

Simon Krek,[+]* Polona Gantar,[†]* Iztok Kosem,[‡†] Vojko Gorjanc,[†] Cyprian Laskowski[‡*]

[+] Laboratorij za umetno inteligenco, Institut »Jožef Stefan«, Jamova 29, 1000 Ljubljana
* Center za jezikovne vire in tehnologije, Univerza v Ljubljani, Večna pot 113, 1000 Ljubljana
simon.krek@guest.arnes.si
† Oddelek za prevajalstvo, Filozofska fakulteta, Univerza v Ljubljani,
Aškerčeva 2, 1000 Ljubljana
apolonija.gantar@ff.uni-ljs.si, vojko.gorjanc@ff.uni-lj.si
‡ Center za uporabno jezikoslovje, Trojina
Partizanska cesta 5, 4220 Škofja Loka
iztok.kosem@trojina.si, cyp@trojina.si

Povzetek

V prispevku so opisani postopki izdelave Baze kolokacijskega slovarja slovenskega jezika, ki predstavlja samostojno fazo v procesu izdelave Kolokacijskega slovarja slovenskega jezika. Opis zajema prilagoditev že obstoječih kolokacijskih podatkov v predhodno izdelani Leksikalni bazi za slovenščino in se osredotoča na dopolnitev z avtomatsko izluščenimi podatki iz korpusa. V osrednjem delu prispevka je opisan nabor avtomatsko izluščenih podatkov, ki jih trenutno vsebuje kolokacijska baza, ter njihov prikaz v spletnem vmesniku. Prispevek zaključuje analiza evalvacije uporabniškega vmesnika in načrti za nadaljnje delo.

Slovene Collocations Dictionary Database

The paper describes the compilation of Slovene Collocations Dictionary Database which represents a separate stage in the process of compiling the Collocations Dictionary of Slovene. The described process includes the adaptation of the existing collocations data in the Slovene Lexical Database and focuses on the upgrade with the automatically extracted data from the corpus. The central part of the paper describes the new data collection currently included in the collocations database, and its visualisation in the web interface. We conclude with the presentation of evaluation results, both of the data and the user interface, and with the plans for future work.

1 Uvod

V prispevku opišemo izdelavo Baze kolokacijskega slovarja slovenskega jezika (BKSSJ), ki predstavlja del procesa izdelave Kolokacijskega slovarja slovenskega jezika (KSSJ).[1] Ta bo ob zaključku vseboval 5.000 gesel in bo prosto dostopen na spletu. Ideja je povezana z dejstvom, da kolokacijski slovar za slovenščino ne obstaja in da so kolokacijski podatki, ki so strukturirani v obstoječih slovarjih v t. i. iztržkih in slovarskih zgledih, nedostopni kot računalniško procesljiva baza, hkrati pa ne odražajo stanja v sodobnem slovenskem jeziku. Potreba po kolokacijskem slovarju za slovenščino je bila v jezikovni skupnosti že večkrat izražena (Čibej et al., 2015). BKSSJ trenutno vsebuje 2.500 gesel, v načrtu je razširitev na približno 50.000 iztočnic.

Kolokacijski slovarji (Rundell et al., 2010; Cleveland-Marwick et al., 2013) se selijo na splet (prim. Roth, 2013), hkrati pa zaradi vse bolj naprednih orodij ter statistik za luščenje kolokabilno izstopajočih sopojavitev nastajajo avtomatsko generirani kolokacijski slovarji (Baisa in Suchomel, 2014; Kallas et al., 2015). Ti so potem lahko kot taki predstavljeni uporabnikom ali pa služijo kot osnova za izdelavo kolokacijskih in podobnih slovarjev (gl. Kallas et al., 2015).

2 Od leksikalne baze do kolokacijskega slovarja

Osnovo KSSJ sestavljajo kolokacijski podatki, vključeni v Leksikalno bazo za slovenščino (LBS; Gantar, 2015), in avtomatsko izluščeni kolokacijsko relevantni podatki iz korpusa Gigafida (BKSSJ). Končni slovar bo vključeval večnivojski opis kolokacij, ki ga bo mogoče vključiti v obsežnejšo slovarsko bazo, namenjeno izdelavi slovarja sodobnega slovenskega jezika (SSSJ; Gorjanc et al., 2015), hkrati pa bo uporabnikom na voljo kot samostojen prosto dostopen kolokacijski slovar. Glede na to da proces izdelave SSSJ predvideva več med seboj soodvisnih leksikografskih faz (Gantar et al., 2015), pri izdelavi KSSJ preizkušamo in nadgrajujemo tako proces avtomatskega luščenja podatkov iz korpusa kot tudi možnosti vključevanja množičenja pri čiščenju in urejanju avtomatsko izluščenih podatkov.

Prvi del izdelave KSSJ je bila priprava in ureditev podatkov v LBS za izdelavo druge verzije, tj. LBS2. V ta namen smo podatke iz LBS obogatili z (a) metapodatki, kot so identifikator leme, podatek o frekvenci leme v korpusu, podatek o izvoru, tj. ali je bila iztočnica v LBS izdelana ročno ali že na podlagi avtomatsko izluščenih podatkov, (b) dodali XML-oznake posameznim elementom sheme, npr. identifikator kolokatorja, pridobljenega iz Sloleksa, in oznake za besedno vrsto ter (c) odstranili napake, ki so nastale pri ročni izdelavi gesel v LBS.

[1] Projekt se izvaja v okviru raziskovalnega programa Slovenski jezik – bazične, kontrastivne in aplikativne raziskave ter infrastrukturnih programov Centra za jezikovne vire in tehnologije Univerze v Ljubljani in Centra za uporabno jezikoslovje na zavodu Trojina: http://www.cjvt.si/kssj/.

Drugi del izdelave KSSJ je zajemal avtomatsko luščenje kolokacijskih podatkov. Ta proces je bil na manjšem obsegu lem preizkušen že pri izdelavi LBS (Kosem et al., 2012). Avtomatsko izluščeni podatki in podatki iz LBS2 so bili združeni v osnovo za kolokacijski slovar. V tem postopku smo najprej avtomatsko izluščili kolokacijske podatke, tj. lemo in kolokatorje za omejen nabor kolokacijsko relevantnih skladenjskih struktur in pripadajoče korpusne zglede, za 2.500 v LBS že obstoječih gesel ter jih združili z na novo izluščenimi podatki za dodatnih 2.500 lem. Postopek je zahteval več prilagoditev, ki smo jih opravili v postopku postprocesiranja, kot npr. poenotenje poimenovanj skladenjskih struktur, prepoznavo struktur pri glagolih, pripravo podatkov po novi shemi XML, vzpostavljanje povezav med kolokacijami v LBS in njihovimi zgledi, pripisovanje novih avtomatskih podatkov LBS geslom na podlagi obstoječih pomenskih členitev ipd. Za luščenje smo uporabili orodje Sketch Engine (Kilgarriff et al., 2004) in za slovenščino prilagojeno aplikacijo GDEX za izbiro dobrih korpusnih zgledov (Kosem et al., 2011).

Združeni podatki iz 2.500 gesel LBS2 in 2.500 novih avtomatsko izluščenih gesel torej predstavljajo osnovo, pripravljeno za nadaljnjo obdelavo, v okviru katere predvidevamo uporabo množičenja (Fišer et al., 2015), zlasti pri razporejanju kolokacij pod ustrezne pomene ter pri čiščenju podatkov, ki ga ni bilo mogoče izvesti v fazi luščenja in postprocesiranja.

3 Baza kolokacijskega slovarja slovenskega jezika (BKKSJ)

V tem delu prispevka opišemo BKSSJ, avtomatsko izluščenih 2.500 gesel iz korpusa Gigafida z izbranim naborom skladenjskih relacij, skupaj s korpusnimi zgledi. Avtomatsko izluščeni podatki so bili v postopku postprocesiranja dodatno prilagojeni, predvsem glede ujemanja po spolu, sklonu in številu, kot samostojna baza pa so dostopni tudi v spletnem vmesniku.[2]

BKSSJ ima večdelno vlogo, saj predstavlja podatkovno osnovo za gesla v KSSJ, je samostojna zbirka tako za analizo in izboljšavo in nadgradnjo avtomatskih metod luščenja podatkov kot za preizkušanje množičenja, in je navsezadnje tudi spletno dostopen vir za različne uporabnike.

3.1 BKSSJ kot zbirka

Baza z 2.500 iztočnicami (1.000 samostalnikov, 750 glagolov, 625 pridevnikov in 125 prislovov) vsebuje 2.310.100 kolokacij v 72.117 skladenjskih strukturah, za vsako kolokacijo je tipično izločenih tudi pet zgledov rabe iz korpusa. Izhajajoč iz metodologije, uporabljene pri izdelavi LBS, je bilo izluščenih 528 različnih skladenjskih struktur, od tega pri samostalnikih 192, pri glagolih 147, pri pridevnikih 107 in pri prislovih 82. V Tabeli 1 navajamo prvih pet najpogostejših.

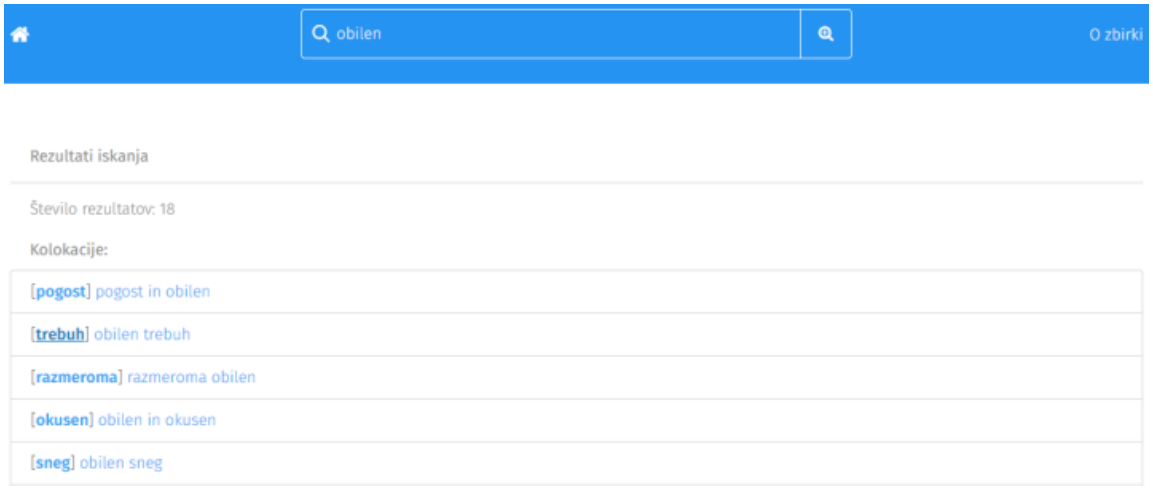
Ker gre za strojno luščenje, BKSSJ vsebuje tudi napake, ki izhajajo deloma iz jezikoslovnega označevanja korpusa Gigafida, deloma iz luščenja v orodju Sketch Engine. Primer prvega je kolokacija *kraj kolesa* namesto *kraja kolesa* v strukturi *SBZ0 + sbz2,* primer drugega je *bolezen očija* v enaki strukturi. Pri kolokatorju *oči* sta možni dve lemi: *oči* (eden od staršev) ali *oči* (množinski samostalnik – *oko*). V postprocesiranju je bila izbrana rodilniška oblika prve leme (*očija*) namesto rodilniške oblike druge leme (*oči* = bolezen oči).

Odpravo napak lahko izvajamo na dveh ravneh. Kot prvo, z izboljševanjem postopka avtomatskega luščenja, tj. z izboljšavo slovnice besednih skic, orodja GDEX za luščenje dobrih zgledov in navsezadnje tudi oblikoskladenjskega označevalnika za slovenščino. Tovrstne izboljšave potrebujejo bolj sistematičen pristop in obsežnejši pregled napak, zaradi česar smo se tudi lotili evalvacije podatkov v BKSSJ z vidika potencialnih uporabnikov (glej 3.3). Po drugi strani se napake lahko odpravljajo z izboljševanjem postopka postprocesiranja. Na primer, za zgornji primer *bolezen očija* in podobne primere lahko pravo lemo identificiramo z analizo oblik, ki se pojavljajo v izluščenih zgledih.

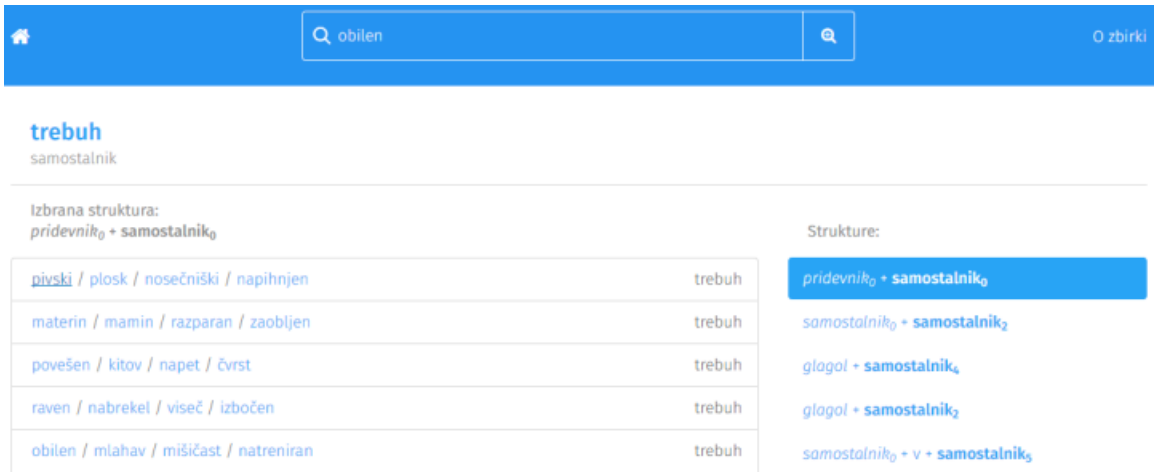
št.	**struktura**	**opis**	**primer kolokacije**	**število struktur**
1	sbz0 SBZ2	samostalnik v poljubnem sklonu + samostalnik v rodilniku	[štafeta, eliksir, vrelec, kult] mladosti	1.783
2	GBZ sbz4	glagol + samostalnik v tožilniku	priznati [premoč, krivdo, zmoto, neodvisnost]	1.672
3	PBZ0 sbz0	pridevnik v poljubnem sklonu + samostalnik v poljubnem sklonu	mlada [generacija, ženska, družina, igralka]	1.609
4	GBZ sbz2	glagol + samostalnik v rodilniku	priznati [imunitete, očetovstva, krivde, neodvisnosti]	1.598
5	GBZ z sbz6	glagol + z + samostalnik v orodniku	priznati z [nasmeškom, grenkobo, obžalovanjem, nasmehom]	1.193

Tabela 1: Najpogostejše skladenjske strukture v BKSSJ.

[2] Pri izdelavi vmesnika so sodelovali: Rok Rejc, Gašper Uršič, Simon Krek, Iztok Kosem, Polona Gantar, Vojko Gorjanc. Spletna stran: http://bkssj.cjvt.si/.



Slika 1: Prikaz kolokacij, ki vključujejo iskano besedo.



Slika 2: Prikaz kolokacij znotraj izbranega kolokatorja, ki je v bazi obdelan kot iztočnica.

Slika 3: Prikaz korpusnih zgledov za izbrano kolokacijo.

3.2 BKSSJ v spletnem vmesniku

Spletni vmesnik, ki omogoča iskanje po celotni bazi, je bil zasnovan v okviru diplomske naloge na Naravoslovnotehniški fakulteti (Uršič, 2015) ter študentskega dela na Fakulteti za računalništvo in informatiko UL. Zadetki pri iskanju so ločeni na prikaz iztočnic, ki vsebujejo kolokacije za iskano lemo in na prikaz seznama kolokacij (zadnje prikazuje Slika 1 zgoraj), pri čemer je pri seznamu kolokacij vedno poudarjeno izpisan kolokator, ki je v bazi predstavljen kot iztočnica.

Pri prikazu slovarskega gesla je mogoče kolokacije filtrirati po skladenjskih strukturah, ki so prikazane na desni strani vmesnika (Slika 2), ob kliku na posamezno kolokacijo pa so na novi strani prikazani pripadajoči izluščeni zgledi iz korpusa (Slika 3).

3.3 Evalvacija BKSSJ

Z namenom izboljšati uporabniško izkušnjo in optimizirati prikaz izluščenih podatkov je bila že v tej fazi izvedena kratka anketa med potencialnimi uporabniki BKSSJ oz. bodočega KSSJ.[3] Poleg osnovnih podatkov o anketirancih (smer študija; starost, spol) in o izbiri naprave za dostop do BKSSJ (namizni ali prenosni računalnik, tablica, telefon) je bil osnovni namen ankete zbrati predvsem opažanja, predloge in želje, ki se nanašajo na organizacijo in postavitev podatkov na strani, navigacijo med razdelki in vizualizacijo podatkov. Odgovore anketirancev smo razdelili v 5 kategorij, ki jih na kratko povzemamo v nadaljevanju.

(1) Razvrščanje kolokacij. Večina anketirancev je opozorila na preobsežnost seznamov kolokacij, ki se izpišejo za iskano besedo v primeru, ko ta v bazi ni prikazana kot iztočnica (prim. Sliko 1). Predlagane so predvsem naslednje možnosti razvrščanja: po pogostnosti, po obliki oz. strukturi kolokacije, po abecedi iztočnice, po pomenskih sklopih in po besedni vrsti iztočnice. Predlagane so tudi kombinacije razvrščanja, npr. najprej po strukturi, znotraj posamezne strukture pa po pogostnosti. Taka razvrstitev se zdi smiselna zlasti z vidika relevantnosti in intuitivne prepoznavnosti tipičnih sopojavitev, hkrati pa bi bilo mogoče na ta način potisniti napačno izluščene sopojavitve, ki se jim v postopku avtomatizacije ni mogoče povsem izogniti, na konec obsežnih seznamov.

(2) Prikaz podatkov na strani. Največ nejasnosti in neintuitivnosti pri branju podatkov povzroča razmerje med prikazom iskane besede kot kolokatorja (kadar ni iztočnice; Slika 1) in prikazom iskane besede kot iztočnice (Slika 2). Ker je trenutno v bazi le 2.500 iztočnic, je večina iskanih besed v bazi prikazana le, če se pojavljajo kot kolokatorji pri v bazi obstoječih iztočnicah, pri čemer te kolokacije za iskano besedo po pričakovanju niso najbolj tipične. Ta problem bo delno odpravljen z razširitvijo baze na 50.000 iztočnic in s katerim od zgoraj predlaganih načinov razvrščanja. Anketirance je pri prikazu podatkov na strani motilo tudi preveč praznega prostora, postavitev menija za filtriranje struktur na desno stran zaslona, preveč zapleten (metajezikoslovni) zapis struktur, prikaz kolokatorjev v nizih za poševnicami namesto v stolpcih ipd. Načeloma so bili anketiranci zadovoljni z barvnimi kombinacijami, velikostjo fonta in številom izpisanih zgledov.

(3) Navigacija po strani/geslu. Kot dobro rešitev so anketiranci izpostavili spustni meni s predlogi tipičnih sopojavitev, kritizirali pa so premikanje po straneh med obsežnimi seznami kolokacij na prvem nivoju in preobsežne sezname struktur na drugem nivoju. V ta namen so bile predlagane rešitve združevanja struktur pod opisne razlage, npr. kakšen je (+ iskani samostalnik); kaj počnemo z (+ iskani samostalnik) oz. združevanje struktur s posameznimi konkretnimi predlogi pod skupno oznako "predlog", npr. glagol + predlog + samostalnik$_5$, ki bi vsebovala strukture z vsemi predlogi (v, na, po, pri itd.), ki se vežejo s samostalnikom v mestniku.

(4) (Ne)ustreznost kolokacij. Po pričakovanju je anketirance zmotila neustreznost prikazanih podatkov, vezana pretežno na napake pri avtomatskem luščenju kolokacij iz korpusa. Izpostavljeni so bili predvsem primeri, kjer je prikazana kolokacija oz. kolokator del širše (ustaljene) zveze, npr. *mali sneg ← nekaj malega snega; hud sneg ← v hudem snegu; zdrav duh ← zdrav duh v zdravem telesu.* Prav tako prihaja pri nekaterih prikazanih sopojavitvah za naključne kombinacije, ki sicer formalno ustrezajo skladenjski strukturi, vendar v besedilu pripadajo dvema različnima skladenjskima strukturama, npr. *sanjati v ligi* (gbz v SBZ5: sanjati o čem + zmaga v ligi). Poleg omenjenega, so anketiranci opozorili še na napačno lematizacijo, neustrezno obliko kolokatorja, npr. osnovnik namesto presežnik: *lep med lepotico ← najlepša med lepoticami*; razlikovanje med glagoli s in brez *se*/*si* ter med zanikanimi in nezanikanimi glagoli, npr. *spomniti zabavo ← spomniti se zabave*; priznati [premoč, krivdo] in ne priznati [imunitete, očetovstva, krivde], na napake v zapisu struktur, npr. s + prislovom namesto "s prislovom"; na vrstni red znotraj kolokacije, npr. zamenjava med osebkom in predmetom v imenovalniku: *poskrbeti vreme ← vreme poskrbi za kaj*; neločevanje med stalnimi zvezami in kolokacijami, npr. *reševanje/jabolko/zgladitev spora* (SZ: jabolko spora), vključevanje lastnoimenskih kolokatorjev, npr. *trgovina v Citypark/v Globokem/v Žireh* itd., ki so s kolokacijskega vidika manj relevantni.

(5) (Ne)uporabnost. Med razlogi, ki delajo bazo v tej fazi manj uporabno, so anketiranci izpostavili predvsem premajhno strukturiranost in izčiščenost podatkov, preveliko količino podatkov in nerazločevalnost med relevantnimi (pogostimi) in nerelevantnimi oz. manj relevatnimi kolokacijami. Med manjkajočimi podatki so našteli predvsem: podatke o (ne)standardnosti oz.

[3] Za namene evalvacije BKSSJ sta bili izdelani dve anketi: prva je bila namenjena študentom (prevajalstvo, novinarstvo, računalništvo idr.), druga pa sodelavcem pri projektu priprave novega slovarja sodobne slovenščine. V času priprave prispevka sta bili anketi še aktivni, zato so v prispevku upoštevani vmesni rezultati, in sicer zgolj ustrezno rešene ankete (od skupno 243 le 92 anket) oz. ankete, ki vsebujejo komentarje (35 anket). Anketi sta dostopni na https://www.1ka.si/a/94327 in https://www.1ka.si/a/92168.

zaznamovanosti kolokacije; pogostnosti kolokacije; pomenu iztočnice; med manjkajočimi funkcionalnostmi pa možnost določiteve števila besed v okolici iskane besede in dostop do baze za tekstovno rudarjenje.

4 Zaključek in nadaljnje delo

Nadaljnje delo bo potekalo v dveh smereh: pri izdelavi KSSJ bo poudarek na izdelavi pomenske členitve za leme, ki ne izhajajo iz LBS, vzpostavitvi procesa množičenja ter evalvaciji rezultatov, pridobljenih v tem procesu. Pri izdelavi BKSSJ predvidevamo množični izvoz podatkov iz korpusa (približno 50.000 iztočnic) in izboljšanje njihove vizualizacije na spletu

Poleg tega preučujemo tudi izrabo dodatnih funkcionalnosti v orodju Sketch Engine, kot sta gručenje (ang. clustering) in najpogostejši besedni niz (ang. longest comment match), za namene izboljšave postopka avtomatskega luščenja. Prvi namreč kaže dober potencial za grupiranje (semantično) podobnih kolokacij, kar je koristno tako za vizualizacijo BKSSJ (ena od preferenc uporabnikov pri evalvaciji) kot za samo pomensko členitev pri izdelavi končnih gesel v KSSJ, medtem ko najpogostejši besedni niz lahko služi kot dodatna informacija pri postprocesiranju izluščenih podatkov.

5 Literatura

Vit Baisa in Vit Suchomel. 2014. SkELL: Web Interface for English Language Learning. *Eighth Workshop on Recent Advances in Slavonic Natural Language Processing*, str. 63–70. Brno: Tribun EU.

Karen Cleveland-Marwick et al. (ur.). 2013. Longman collocations dictionary and thesaurus. Harlow, Essex: Pearson Education.

Jaka Čibej, Vojko Gorjanc in Damjan Popič. 2015. Vloga jezikovnih vprašanj prevajalcev pri načrtovanju novega enojezičnega slovarja. V: V. Gorjanc, P. Gantar, I. Kosem in S. Krek, ur., *Slovar sodobne slovenščine: problemi in rešitve*, str. 168–181. Ljubljana: Znanstvena založba Filozofske fakultete.

Darja Fišer, Jaka Čibej, Kaja Dobrovoljc, Polona Gantar, Iztok Kosem, Špela Arhar Holdt, Damjan Popič in Tomaž Erjavec. 2015. Množičenje za slovar sodobnega slovenskega jezika. V: V. Gorjanc, P. Gantar, I. Kosem in S. Krek, ur., *Slovar sodobne slovenščine: problemi in rešitve*, str. 566–586. Ljubljana: Znanstvena založba Filozofske fakultete.

Polona Gantar. 2015. *Leksikografski opis slovenščine v digitalnem okolju*. Ljubljana: Znanstvena založba Filozofske fakultete UL. http://www.ff.uni-lj.si/Portals/0/Dokumenti/ZnanstvenaZalozba/e-knjige/Leksikografski.pdf.

Polona Gantar, Iztok Kosem in Simon Krek. 2015. Leksikografski proces pri izdelavi spletnega slovarja sodobnega slovenskega jezika. V: V. Gorjanc, P. Gantar, I. Kosem in S. Krek, ur., *Slovar sodobne slovenščine: problemi in rešitve*, str. 280–297. Ljubljana: Znanstvena založba Filozofske fakultete.

Vojko Gorjanc, Polona Gantar, Iztok Kosem in Simon Krek, ur., 2015. *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete.

Jelena Kallas, Adam Kilgarriff, Kristina Koppel, Elgar Kudritski, Margit Langemets, Jan Michelfeit, Maria Tuulik in Ülle Viks. 2015. Automatic generation of the Estonian Collocations Dictionary database. V: I. Kosem, M. Jakubiček, J. Kallas, S. Krek, ur., *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference, 11-13 August 2015*. Herstmonceux Castle, United Kingdom, str. 1–20. Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd.

Adam Kilgarriff, Pavel Rychlý, Pavel Smrz, David Tugwell. 2004. The Sketch Engine. V: G. Williams, S. Vessier, ur., *Proceedings of the Eleventh EURALEX International Congress.* EURALEX 2004 Lorient, France July 6-10, 2004, str. 105–116. Lorient: Universite de Bretagne - sud.

Iztok Kosem, Polona Gantar in Simon Krek. 2012. Avtomatsko luščenje leksikalnih podatkov iz korpusa. V: T. Erjavec, J. Žganec Gros, ur., *Zbornik 15. mednarodne multikonference Informacijska družba - IS 2012, zvezek C,* str. 117–122. Ljubljana: Institut Jožef Stefan.

Iztok Kosem, Miloš Husák in Diana Mccarthy. 2011. GDEX for Slovene. V: I. Kosem, K. Kosem, ur., *Electronic Lexicography in the 21st Century: New applications for new users. Proceedings of eLex 2011.* Bled, 10-12 November 2011, str. 151–159. Ljubljana: Trojina, zavod za uporabno slovenistiko.

Tobias Roth. 2013. Going Online with a German Collocations Dictionary. V: I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets, M. Tuulik, ur., *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17–19 October 2013,* str. 152–163. Tallinn, Estonia. http://eki.ee/elex2013/proceedings/eLex2013_11_Roth.pdf.

Michael Rundell et al. (ur.). 2010. *Macmillan collocations dictionary*. Oxford: Macmillan Education.

Gašper Uršič. 2015. *Oblikovanje uporabniškega vmesnika za spletni kolokacijski slovar*. Diplomsko delo. Ljubljana: Univerza v Ljubljani, Naravoslovnotehniška fakulteta.