KRES IN GIGAFIDA KOT KORPUSNA OSNOVA ZA SLOVAR: RAZLIKE IN PODOBNOSTI

[1]Nataša Logar, [2, 3]Nikola Ljubešić, [3]Tomaž Erjavec
[1]Fakulteta za družbene vede, Ljubljana; [2]Filozofski fakultet, Zagreb;
[3]Institut »Jožef Stefan«, Ljubljana

UDK 811.163.6'374.81:81'322.2

Gradnjo stomilijonskega korpusa Kres je narekovala potreba po dopolnitvi korpusa Gigafida z bolj uravnoteženim korpusnim virom sodobne slovenščine. V prispevku prikazujemo taksonomsko in besedilnovrstno zgradbo obeh korpusov ter tematski profil Kresa. Da bi prikazali razlike in podobnosti med Gigafido in Kresom, za katera so si leksikografi enotni, da sta ustrezna vira za nov slovar slovenščine, smo na izbranem naboru samostalnikov primerjali še značilno besedilno okolje, kakršnega dobimo v Besednih skicah na podlagi podatkov iz enega in nato iz drugega korpusa.

korpus, slovenščina, tematsko modeliranje, besedne skice, slovar

The hundred million word corpus Kres was compiled in order to support the Gigafida corpus with a more balanced source of contemporary Slovene. The paper present the taxonomic and text-type composition of both corpora and the topic model of the Kres corpus. In order to illustrate the differences and similarities between Gigafida and Kres, which are appropriate corpus sources for a new dictionary of Slovene, we selected a set of salient nouns and analysed their typical contexts in terms of Word sketches as obtained from both corpora.

corpus, Slovene language, topic modelling, word sketches, dictionary

1 Uvod

Referenčni korpus slovenščine Gigafida (http://www.gigafida.net) z obsegom 1,2 milijarde besed je bil zgrajen leta 2012. Isto leto nekaj mesecev pozneje je nastal še korpus Kres (http://www.korpus-kres.net). Kres je iz Gigafide vzorčeni korpus s 100 milijoni besed, vnaprej načrtovan in uresničen pa je bil zato, ker se je zdelo za raziskovanje slovenščine smiselno oblikovati Gigafidi primerljiv vir z manjšim deležem publicistike ter večjim deležem knjižnih besedil (Logar Berginc idr. 2012).

Ko smo v Logar (2014) razmišljali o Gigafidi in Kresu kot slovarskih virih, smo na vprašanje, ali sta ta dva korpusa ustrezna gradivna osnova za prikaz leksikalne podobe javne pisne slovenščine zadnjih 20 let, odgovorili pritrdilno (prav tam: 10). Sicer pa je bila že v Krek, Kosem, Gantar (2013) kot izhodišče za pripravo geslovnika novega slovarja navedena prav »frekvenčna lista korpusa Gigafida, v kombinaciji z natančno in razmeroma kompleksno statistično obravnavo podatkov iz korpusa Kres, Gos in drugih baz« (prav tam: 24). Zelo podobno je gradivo za novi slovar opredeljeno tudi v Gliha Komac idr. (2015: 4): »Gradivo za izdelavo geslovnika in redakcijsko

obdelavo osrednjih delov posameznih slovarskih sestavkov [...] so korpusni viri, predvsem Gigafida, Kres, Nova beseda in deloma Gos.« Lahko torej zapišemo, da se v letu 2015 ključni slovenski leksikografi o vlogi Gigafide in Kresa pri prihodnjem velikem slovenskem slovarskem podvigu strinjajo med seboj, kar pa v obeh predlogih manjka, je izostritev načinov medsebojnega dopolnjevanja obeh (ali več) korpusov, za katero ta hip manjkajo zlasti (primerjalni) podatki o Kresu. Zato smo se odločili, da izhajajoč iz medsebojnih besedilnovrstnih razmerij ter tematskih profilov Gigafide in Kresa, na manjšem vzorcu preverimo, kako zelo drugačne – ali pa morda vendarle podobne – podatke o leksiki daje Kres, če ga primerjamo z desetkrat večjim referenčnim korpusom, iz katerega je nastal.

2 Gigafida in Kres: medsebojna taksonomska ter besedilna razmerja

Gradnja, vsebina in uporaba obeh korpusov je bila predstavljena že na več mestih (Logar Berginc idr. 2012; Arhar Holdt, Kosem, Logar Berginc 2012; Erjavec, Logar Berginc 2012), zato naj ponovimo le osnovna razmerja:

– v Kres je bilo zajetih 71 % vsakega leposlovnega naslova iz Gigafide in 36 % vseh stvarnih besedil;
– v Kres so iz Gigafide prišla le besedila iz 19 najbolj branih časopisov in 54 najbolj branih revij z lestvice Nacionalne raziskave branosti (2010), in to v enakih deležih, kot so jih ti imeli na imenovani lestvici;
– 5-odstotna kategorija drugo v Kresu zajema podnapise in zapise sej državnega zbora;
– v internetni del Kresa so izmed desetih novičarskih portalov iz Gigafide prišla le besedila z najbolj branih strani 24ur.com, siol.net in rtvslo.si; preostali del prihaja s spletnih strani podjetij in ustanov.

V Tabeli 1 so prikazani število besed in njihovi deleži v obeh korpusih po taksonomiji. Razvidno je (kot že rečeno), da ima Kres v primerjavi z Gigafido predvsem veliko večji delež leposlovja in stvarnih besedil ter manjši obseg časopisov in revij.

Tabela 1: Število besed in njihovi deleži po taksonomiji v Gigafidi in Kresu

Taksonomija	**Gigafida: število besed**	**Gigafida: % besed**	**KRES: število besed**	**KRES: % besed**
tisk	1.001.244.035	**84**	79.830.144	**80**
knjižno	74.356.531	**6**	35.088.699	**35**
leposlovje	23.969.196	**2**	17.030.038	**17**
stvarna besedila	50.387.335	**4**	18.058.661	**18**
periodično	918.936.054	**77**	39.727.239	**40**
časopisi	663.664.965	**56**	19.919.327	**20**
revije	255.271.089	**21**	19.807.912	**20**
drugo	7.951.450	**1**	5.014.206	**5**
internet	185.758.467	**16**	20.001.001	**20**
SKUPAJ	**1.187.002.502**	**100**	**99.831.145**	**100**

3 Kres: tematsko modeliranje

Razlike in podobnosti med dvema korpusoma je mogoče iskati na več načinov. V zadnjih letih sta za ta namen najbolj obširno uporabljani dve metodi: metoda frekvenčnega profila (Rayson, Garside 2000) in metoda tematskega modeliranja (Blei idr. 2003; Sharoff 2010). Po prvi metodi je bila primerjava med Gigafido in Kresom že narejena (Logar Berginc idr. 2012: 95–97), ena od ugotovitev, ki so izhajale iz primerjave logaritemskih verjetnosti lem obeh korpusov, pa je bila (prav tam: 95), da so za Kres med polnopomenskimi besedami značilnejši glagoli (med njimi: *biti*, *reči*, *vprašati*, *vedeti*, *misliti*, *zdeti se*, *videti*), medtem ko so v Gigafidi bolj izraziti samostalniki (*odstotek*, *podjetje*, *milijon*, *direktor*, *banka*, *evro*, *tolar* itn.).

Tudi metodo tematskega modeliranja smo na korpusih slovenščine že preizkusili, in sicer na Gigafidi (Logar Berginc, Ljubešić 2013; Erjavec idr. 2015) ter na obeh različicah korpusa slWaC (Ljubešić, Erjavec 2012; Ljubešić, Erjavec 2014; Logar Berginc, Ljubešić 2013). Tu tovrstnemu pregledu dodajamo še tematsko izkaznico Kresa. Kot smo že zapisali (Logar Berginc, Ljubešić 2013: 89, 90; več o metodi gl. tam):

> Metoda temelji na predpostavki, da je vsak dokument v zbirki nastal iz vsebin z več temami. Vsako temo predstavlja verjetnostna distribucija besed – povedano drugače: za vsako besedo obstaja določena verjetnost, da pripada določeni temi. [...] Rezultat metode sta dve verjetnostni distribuciji: a) verjetnostna distribucija tem za vsak dokument oz. verjetnost, da neki dokument vsebuje določene teme, in b) pogojna verjetnostna distribucija besede pri določeni temi oz. verjetnost posamezne besede, da pripada določeni temi.

Dvajset za Kres najbolj značilnih tem prikazuje Tabela 2 (»teža« v 2. stolpcu kaže razpršenost posamezne teme v korpusu).

Tabela 2: Samostalniške leme, ki z največjo verjetnostjo pripadajo eni temi, in ta tema po teži v Kresu

Tema	**Teža**	**Samostalniška lema**
človek	5,567	dan človek čas leto oče življenje otrok roka mama hiša oči beseda prijatelj mož žena konec svet glava sin
telo, bivanjski prostor	3,364	roka barva noga glava oči vrata obraz las telo soba stran prostor tla obleka miza prst oblačilo stena čas
tv- in radijski program	3,044	leto film glasba oddaja skupina predstava pesem gledališče nagrada festival delo vloga čas koncert dan svet oder tv igralec
jezik, družba, RAZNO	2,878	človek svet življenje beseda čas knjiga jezik način vprašanje delo primer oblika družba pomen zgodovina kultura del moč zgodba
prireditve, čas, prostor	2,866	ura leto dan mesto društvo občina prostor dom center prireditev čas razstava delo hotel ulica sobota obiskovalec hiša cesta
medčloveška razmerja	2,699	otrok življenje človek ženska čas leto družina starš moški odnos ljubezen težava delo partner stvar občutek oseba dan pomoč

pravo	2,685	člen zakon postopek odstavek pravica sodišče organ dan podatek oseba podlaga primer odločba določba pogodba sklad stranka pogoj list
gospodarstvo	2,672	delo razvoj področje podjetje okolje projekt sistem program država leto dejavnost organizacija sredstvo cilj proces storitev prostor trg pomoč
notranja politika	2,593	vlada država predsednik svet zakon republika minister leto predlog stranka član vprašanje zbor odbor zadeva ministrstvo komisija seja poslanec
finance	2,491	leto evro podjetje odstotek milijon cena banka tolar družba denar država plača strošek vrednost mesec račun sredstvo delež prodaja
računalništvo	2,445	e stran podatek slika d številka računalnik ime b n c x naslov m uporabnik program i l r
okolje	2,43	voda pot teren morje leto vrh cesta meter čas gozd reka hiša m dolina zemlja mesto del gora dan
vojna, terorizem, kazniva dejanja	2,4	leto človek vojna država policija dejanje policist kazen zapor dan žrtev čas oblast primer vojska sodišče orožje napad stran
zdravje	2,323	bolezen telo zdravilo dan zdravnik zdravljenje koža bolnik težava leto voda človek snov celica primer bolečina kri rak hrana
hrana	2,218	rastlina voda olje list vino g minuta vrt cvet jed meso zelenjava sladkor vrsta sol okus žlica čas sadje
izobraževanje	2,207	šola delo leto študent učenec program univerza fakulteta znanje učitelj jezik področje študij izobraževanje predmet otrok naloga izpit razred
zgodovina	1,865	leto stoletje mesto cerkev čas delo ime muzej vojna svet del razstava kralj slika človek zgodovina grad življenje hiša
promet, avtomobilizem	1,835	vozilo avtomobil cesta motor avto voznik km vožnja hitrost kolo promet sedež material cena nesreča leto izdelek del gorivo
šport	1,765	tekma leto mesto igra ekipa klub igralec sezona prvenstvo zmaga točka liga prvak tekmovanje minuta konec trener šport dirka
živali, RAZNO	1,377	žival vrsta pes konj riba leto človek ptica voda hrana ladja dan gozd čas krava morje meso vojska vojak

Iz tabele je razvidno, da so v Kresu prisotne različne teme, da so v največjem delu (če seštejemo teže sorodnih vsebin) v njem tematizirani *človek*, *človeško telo* in *bivanjski prostor* ter *medčloveška razmerja*, medtem ko so druge teme del različnih področij in po obsegu nobena posebej ne izstopa. V primerjavi z Gigafido (Tabela 1 v: Erjavec idr. 2015) je v Kresu manj *notranje* in *lokalne politike* ter *športa*. Glede na razlike v taksonomskih deležih v obeh korpusih so torej slednje očitno bolj prisotne v časopisju, prve (*človek*, *človeško telo* itn.) pa v leposlovju in stvarnih besedilih.

4 Gigafida in Kres: razlike v značilnem besedilnem okolju

Na podlagi navedenih ugotovitev ter predhodnih primerjav Kresa z Gigafido in korpusom slWaC smo za primerjalno analizo značilnega besedilnega okolja izbrali 20 samostalnikov, od katerih jih 10 z največjo verjetnostjo pripada (a) temam, bolj značilnim za Kres, preostalih 10 pa (b) temam, bolj značilnim za časopisni del Gigafide (Tabela 9 v: Logar Berginc, Ljubešić 2013). In sicer:

a) *oče*, *mama*, *roka*, *sin* (tema: človek); *obraz*, *soba*, *miza* (tema: telo, bivanjski prostor), *beseda*, *zgodovina*, *kultura* (tema: jezik, družba);
b) *predsednik*, *vlada*, *minister*, *občina* (tema: notranja politika); *podjetje*, *banka*, *vrednost* (tema: finance, gospodarstvo); *prvenstvo*, *prvak*, *trener* (tema: šport).

Zanimalo nas je, kako zelo se značilno besedilno okolje izbranih samostalnikov – in posledično njihov slovarski opis – razlikuje, če podatke zanj najprej pridobimo iz enega in nato iz drugega korpusnega vira. Pri tem je treba opozoriti na to, da možne razlike ne bodo izhajale le iz različnosti besedilnovrstnih deležev obeh korpusov, različnih deležev posameznih besedil v obeh korpusih in različne razporejenosti besed po letih v obeh korpusih, temveč tudi iz njunih izrazito različnih obsegov. Podatke smo pridobili iz Besednih skic orodja Sketch Engine (Kilgarriff idr. 2004), primerjavo pa omejili na prvih 24 kolokatorjev treh za izbrane samostalnike najbolj značilnih zvez s polnopomenskimi besedami (izognili smo se zvezam s pretežno lastnoimenskimi zapolnitvami).

Tabela 3 kaže po eno od teh zvez za samostalnika *občina* in *beseda*. Razvidno je, da je v zvezi *S_osebek_od*, torej v zvezah kot *občina sofinancira*, v Gigafidi in Kresu prekrivnih 15 kolokatorjev, preostalih 9 pa je različnih. Na drugi strani je pri samostalniku *beseda* v zvezi *S_Kakšen?* prekrivnost večja: ta zveza, ki jo zapolnjujejo pridevniški kolokatorji, kaže prekrivnost v 19 primerih (*spremna beseda* itn.) in različnost le v petih.

Tabela 3: Gigafida in Kres: glagolski kolokatorji, značilni za zvezo *S_osebek_od* samostalnika *občina*, in pridevniški kolokatorji, značilni za zvezo *S_Kakšen?* samostalnika *beseda*

OBČINA		**BESEDA**	
Gigafida: *S_osebek_od*	Kres: *S_osebek_od*	Gigafida: *S_Kakšen?*	Kres: *S_Kakšen?*
1. sofinancirati*	**1. sofinancirati**	**1. spremen**	**1. ključen**
2. financirati	**2. financirati**	**2. ključen**	**2. spremen**
3. prispevati	**3. ustanoviti**	**3. božji**	**3. božji**
4. primakniti	**4. prispevati**	**4. izrečen**	**4. zadnji**
5. ustanoviti	**5. odkupiti**	5. poslovilen	**5. uvoden**
6. namenjati	**6. subvencionirati**	**6. uvoden**	**6. prijazen**
7. kriti	**7. objavljati**	**7. lep**	**7. latinski**
8. odkupiti	**8. podpreti**	**8. pisan**	**8. izrečen**
9. praznovati	9. ustanavljati	**9. prijazen**	9. pravi
10. objavljati	10. zadolževati	10. ministrov	**10. pisan**

11. nameniti	**11. primakniti**	**11. oster**	**11. lep**
12. zagotoviti	**12. obveščati**	12. pohvalen	**12. grški**
13. nameravati	**13. praznovati**	**13. glaven**	**13. grd**
14. podpreti	14. črtati	**14. spodbuden**	**14. zapisan**
15. podeljevati	**15. zadolžiti**	**15. zadnji**	**15. spodbuden**
16. načrtovati	16. prirejati	**16. grd**	**16. govorjen**
17. vabiti	**17. nameniti**	**17. latinski**	17. časten
18. pristopiti	**18. namenjati**	18. topel	**18. angleški**
19. zadolžiti	19. razpolagati	**19. govorjen**	**19. glaven**
20. subvencionirati	**20. nameravati**	**20. lasten**	20. izgovorjen
21. obveščati	21. predpisati	**21. grški**	**21. lasten**
22. plačati	22. organizirati	**22. angleški**	22. Jezusov
23. plačevati	23. podeliti	23. sklepen	23. čaroben
24. podpirati	24. razpisovati	**24. zapisan**	**24. oster**

* Krepko so tiskani prekrivni kolokatorji.

Celotna primerjava je pokazala, da je prekrivnost v povprečju 72-odstotna oz. večinoma v razponu med 58 % (kjer je različnih 10 od 24 kolokatorjev) in 83 % (kjer so različni 4 od 24 kolokatorjev). Pričakovali smo, da bo prekrivnost manjša pri samostalnikih, ki z največjo verjetnostjo pripadajo temam, značilnim za Kres, vendar pa se to predvidevanje ni potrdilo, saj je povprečna prekrivnost ne glede na to, katero skupino samostalnikov opazujemo, enaka. Razlike so se pokazale drugje (treba pa bi jih bilo preveriti na širšem gradivu), in sicer pri tipu zveze oz. besedni vrsti kolokatorjev. Namreč: vseh 18 zvez z glagolskimi kolokatorji izkazuje nižjo, tj. povprečno le 65-odstotno prekrivnost med Gigafido in Kresom – z drugimi besedami: če besedno skico ene od zvez samostalnika z glagolskimi kolokatorji v obsegu 24 enot najprej izdelamo na podlagi Gigafide in nato še na podlagi Kresa, se seznama razlikujeta v 8 ali 9 glagolih (kot smo videli pri samostalniku *občina* v Tabeli 3). Na drugi strani je prekrivnost pri samostalniških in pridevniških kolokatorjih višja: pri prvih je 72-odstotna, pri drugih pa sega do 80 % (prim. *besedo* v Tabeli 3).

5 Sklep

To, da so rezultati poizvedb po Kresu drugačni kot v Gigafidi, seveda ni presenetljivo. Posledica različne prisotnosti zlasti časopisnih besedil na eni strani ter leposlovnih in stvarnih besedil na drugi strani so tematska razhajanja, ki se jih bo treba v prihodnje zelo jasno zavedati pri izbiri korpusa kot gradiva za slovar. Tako je npr. šport v Gigafidi 2,3-krat bolj razpršeno prisoten kot v Kresu, zaradi česar bi bil bolj prisoten tudi v slovarskem opisu, če bi ta izhajal samo iz Gigafide, medtem ko npr. za teme, povezane s človekom, velja ravno obratno: v slovarskem opisu bi bile približno enkrat bolj prisotne, če bi ga pripravili samo na osnovi Kresa in ne Gigafide. Te razlike se seveda odražajo tudi na mikroravni, tj. pri kolokacijah kot leksikalno in/ali pragmatično povezanih ponovljivih sopojavitvah vsaj dveh leksikalnih enot, ki sta med seboj v neposrednem skladenjskem razmerju (Barsch 2004). Prikaz kolokacij je

obvezen del vsakega sodobnega korpusnega slovarja, naša – sicer vzorčno zelo omejena – analiza pa je pri glagolskih kolokatorjih pokazala kar tretjinsko razlikovanje besednih skic iz Gigafide in Kresa, kar se sklada z ugotovitvijo o bolj značilni glagolskosti Kresa, ki smo jo zaznali že s primerjavo logaritemskih verjetnosti lem obeh korpusov (Logar Berginc idr. 2012: 95). Besedne skice so sicer orodje, katerega rezultat nato leksikografi nadgradijo še z lastno pomensko-skladenjsko analizo in dopolnilnim pregledom konkordančnih vrstic, vendar pa praksa kaže (Kilgarriff, Kosem 2012), da je zaradi zamudnosti ročnega pregledovanja velike količine podatkov v velikih korpusih zanašanje na avtomatsko pridobljeni povzetek slovničnega in kolokacijskega obnašanja leksike pri izdelavi slovarjev precejšen. V prispevku smo zato poskušali različnost med Gigafido in Kresom, ki sta oba umeščena v Sketch Engine, prvič ujeti v številke in prišli do ugotovitve, da je neprekrivnost tolikšna, da bodo morali uredniki prihodnjega slovarja slovenščine precej natančno opredeliti metodologijo kombiniranega zajema podatkov iz enega, drugega ali obeh virov – zgolj naključno preferiranje zdaj enega, zdaj drugega (in še katerega) vira bo namreč precej zakrilo gradivno razvidnost končnih leksikalnih opisov.

Sicer pa kakršnokoli lovljenje ravnotežja med različnimi korpusnimi viri zaradi težnje po večji objektivnosti jezikovnega opisa odpira še eno vprašanje: kateri korpusni viri, zakaj ti in ne še nekateri (ali pa kar vsi) drugi? Že tukajšnja primerjava dveh s skoraj enakim namenom narejenih korpusov je pokazala, da daje en sam, čeprav še tako skrbno pripravljen vzorec jezika lažno monolitno sliko tega izredno heterogenega pojava. V zadnjem desetletju je v slovenskem prostoru nastal obširen nabor različnih korpusov (prim. npr. Erjavec 2013; http://nl.ijs.si). Za prihodnje slovarsko delo zato ni pomembno le vprašanje, kateri korpusi *bodo* slovarsko gradivo in zakaj, temveč tudi vprašanje, kateri korpusi *ne* bodo slovarsko gradivo in zakaj ne.

Analiza je torej znova potrdila, kar smo slutili ob gradnji obeh korpusov: »vpogled tako v Gigafido kot v Kres [bo] relativiziral prehitro sklepanje o celoti in preučevalce sodobnega slovenskega jezika silil v še bolj poglobljeno interpretiranje rezultatov korpusnih iskanj« (Logar Berginc idr. 2012: 97). Dodamo lahko, da nas tak vpogled opozarja tudi na to, da bi v prihodnje lahko imeli več gradivno povsem ustrezno podprtih leksikalnih opisov slovenščine in da bi bili ti lahko povsem legitimno med seboj le deloma prekrivni.

Literatura

ARHAR HOLDT, Špela, KOSEM, Iztok, LOGAR BERGINC, Nataša, 2012: Izdelava korpusa Gigafida in njegovega spletnega vmesnika. Tomaž Erjavec, Jerneja Žganec Gros (ur.): *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Institut »Jožef Stefan«. 16–21.

BARTSCH, Sabine, 2004: *Structural and Functional Properties of Collocations in English: a Corpus Study of Lexical and Pragmatic Constraints of Lexical Co-ocurence*. Tübingen: Narr.

BLEI, David M. idr., 2003: Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3. 993–1022.

ERJAVEC, Tomaž, 2013: Korpusi in konkordančniki na strežniku nl.ijs.si. *Slovenščina 2.0* 1/1. 24–49.

ERJAVEC, Tomaž, FIŠER, Darja, LJUBEŠIĆ, Nikola, LOGAR, Nataša, MIKOLIČ, Vesna, 2015: Nadgradnja Gigafide: spletna besedila. Vojko Gorjanc, Polona Gantar, Iztok Kosem, Simon

Krek (ur.): *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete. 242–260.

ERJAVEC, Tomaž, LJUBEŠIĆ, Nikola, 2014: The slWaC 2.0 Corpus of the Slovene Web. Tomaž Erjavec, Jerneja Žganec Gros (ur.): *Jezikovne tehnologije*. Ljubljana: Institut »Jožef Stefan«. 50–55.

ERJAVEC, Tomaž, LOGAR, Nataša, 2012: Referenčni korpusi slovenskega jezika (cc)Gigafida in (cc)KRES. Tomaž Erjavec, Jerneja Žganec Gros (ur.): *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Institut »Jožef Stefan«. 57–62.

Gigafida: http://www.gigafida.net

GLIHA KOMAC, Nataša idr., 2015: *Osnutek koncepta novega razlagalnega slovarja slovenskega knjižnega jezika*. Ljubljana: Inštitut za slovenski jezik Frana Ramovša, ZRC SAZU. http://www.fran.si/novi-sskj

KILGARRIFF, Adam idr., 2004: The Sketch Engine. *Proceedings of the 11th EURALEX International Congress*. Lorient: Universite de Bretagne-Sud. 105–116.

KILGARRIFF, Adam, KOSEM, Iztok, 2012: Corpus Tools for Lexicographers. Sylviane Granger, Magali Paquot (ur.): *Electronic Lexicography*. Oxford: Oxford University Press. 31–55.

KREK, Simon, KOSEM, Iztok, GANTAR, Polona, 2013: *Predlog za izdelavo Slovarja sodobnega slovenskega jezika*. http://www.sssj.si

Kres: http://korpus-kres.net

LOGAR, Nataša, 2014: Verodostojnost korpusov kot gradivnega vira za slovar. Irena Grahek, Simona Bergoč (ur.): *Novi slovar za 21. stoletje*. Ljubljana: Ministrstvo za kulturo RS. 1–13. http://www.mk.gov.si/fileadmin/mk.gov.si/pageuploads/Ministrstvo/slovenski_jezik/E_zbornik/7-_Natasa_Logar_-prispevek_-za_oddajo.pdf

LOGAR BERGINC, Nataša idr., 2012: *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko, Fakulteta za družbene vede.

LOGAR BERGINC, Nataša, LJUBEŠIĆ, Nikola, 2013: Gigafida in slWaC: tematska primerjava. *Slovenščina 2.0* 1/1. 78–110.

RAYSON, Paul, GARSIDE, Roger, 2000: Comparing Corpora Using Frequency Profiling. *Proceedings of the ACL Workshop on Comparing Corpora*. Hong Kong. 1–6.

SHAROFF, Serge, 2010: Analysing Similarities and Differences between Corpora. Tomaž Erjavec, Jerneja Žganec Gros (ur.): *Zbornik Sedme konference Jezikovne tehnologije*. Ljubljana: Institut »Jožef Stefan«. 5–11.