Luščenje in jezikoslovna analiza kolokacij iz korpusa Šolar

TADEJA ROZMAN

ŠPELA ARHAR HOLDT

SENJA POLLAK

IZTOK KOSEM

JTDH 2016

Okvir

korpusna leksikalna analiza – usvajanje besedišča

- Špela Arhar Holdt in Tadeja Rozman. 2015. Možnosti uporabe podatkov iz korpusa Šolar za pripravo slovarskih priročnikov.
- Mojca Stritar Kučuk: Napake besedišča. (V: Iztok Kosem idr. 2012.)

Raziskovalno vprašanje

Kaj nam lahko korpus Šolar pove o usvajanju kolokacij

in

kako te informacije uporabiti pri načrtovanju pouka leksike ter pripravi didaktičnih gradiv in jezikovnih priročnikov, namenjenih šolski populaciji?

Raziskovalni problem

- specifike v rabi kolokacij pri mladini v primerjavi z običajno rabo odraslih v splošni rabi
  - ➢ primerjalna analiza korpusov Šolar in Kres

- detekcija težav v rabi kolokacij pri mladini
  - ➢ analiza učiteljskih popravkov v Šolarju

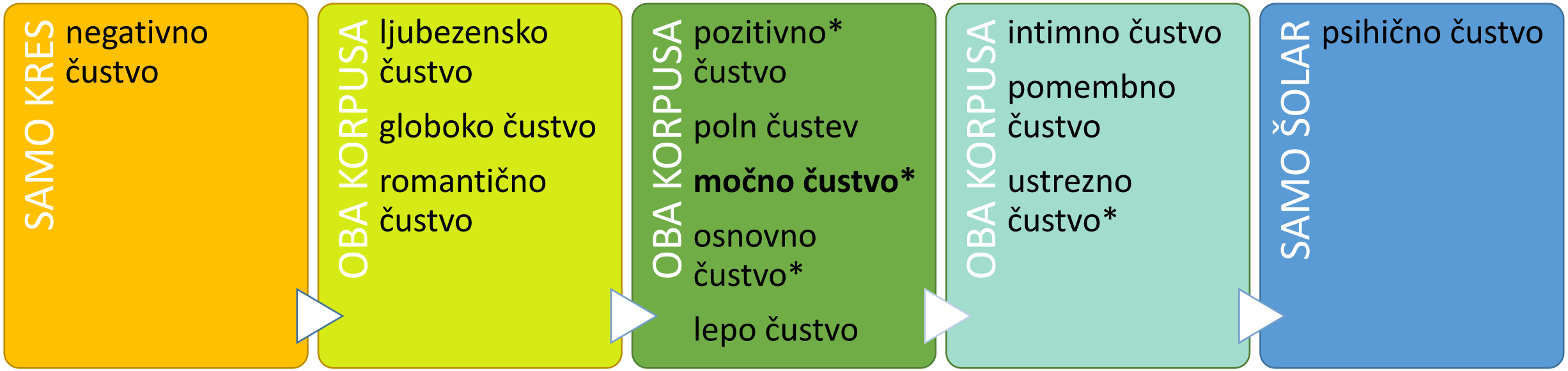
Primerjalno luščenje kolokacij: Šolar vs. Kres

- **Primerjalno luščenje kolokacij** (Pollak in Arhar Holdt, 2015):
  - za seznam besed v korpusu: avtomatski izvoz kolokacij iz orodjaSketchEngine (Kilgariff et al., 2004, API Pollak, 2015): frekvence, kolok. vrednosti (LogDice), povezava na zglede
  - izračun primerjalnih kolokacijskih vrednosti LD-diff (LDŠolar-LDKres), dodatni pogoji (bes. vrsta, minimalne vrednosti, ...)
- **Motivacija**: nove kolokacije, enake oz. primerljive kolokacije, tipične kolokacije vs. redke zveze
- **Izbor za luščenje in analizo**:
  - top 100 samost. v korpusu Šolar (*človek, življenje, ljubezen, otrok, čas* itd.),
  - pozicija pred lemo (in bes. vrsta: pridevnik, glagol ali samostalnik)
  - logDice ≥ 3 in relativna frekvenca ≥ 2 promila (vsaj 2 pojavitvi v Šolarju oz. vsaj 200 v Kresu).

Izluščeni podatki za vzorec [pridevnik + *čustvo*]

Kolokator	Iztočnica	lowUP	LD_diff	logDiceSolar	Frek-Sola	logDiceKres	Frek-Kres	linkSola	linkKres	Pojavljanje
negativen	čustvo	LOWER	-9,089	0,000	0	9,089	263	nolink	http://solar	SAMO_KRES
ljubezenski	čustvo	LOWER	-0,580	7,159	3	7,738	93	http://sola	http://solar	V_OBEH
globok	čustvo	LOWER	0,010	7,685	4	7,677	108	http://sola	http://solar	V_OBEH
romantičen	čustvo	LOWER	0,670	6,712	2	6,040	26	http://sola	http://solar	V_OBEH
pozitiven	čustvo	LOWER	1,500	8,447	8	6,951	70	http://sola	http://solar	V_OBEH
poln	čustvo	LOWER	2,040	6,424	2	4,381	19	http://sola	http://solar	V_OBEH
močen	čustvo	LOWER	2,410	9,764	28	7,354	177	http://sola	http://solar	V_OBEH
osnoven	čustvo	LOWER	2,810	6,438	2	3,625	13	http://sola	http://solar	V_OBEH
lep	čustvo	LOWER	3,210	6,508	4	3,303	18	http://sola	http://solar	V_OBEH
intimen	čustvo	LOWER	3,550	7,433	3	3,884	5	http://sola	http://solar	V_OBEH
pomemben	čustvo	LOWER	4,500	5,701	2	1,200	5	http://sola	http://solar	V_OBEH
ustrezen	čustvo	LOWER	5,540	6,805	2	1,264	2	http://sola	http://solar	V_OBEH
psihičen	čustvo	LOWER	7,359	7,359	3	0,000	0	http://sola	nolink	SAMO_SOLAR

[pridevnik + *čustvo*] glede na LD_Diff



Gim, 4. letnik, NG

Berk je nezmožen **psihičnih čustev** nesvobodo-miselen, neprilagodljiv intelektualec, ki se je kot partizan boril proti Nemcem.

Strok, 4. letnik, MB

Sama dajem mogoče malo večji poudarek **psihičnim čustvom**, ki znajo biti nenadomestljiva, čeprav se zavedam, da je telesna ljubezen prav tako pomembna.

[pridevnik + samostalnik] prekrivno: 1810 zvez izmed 2266 izluščenih (79,9 %)

Prvih 40 glede na frekvenco v korpusu Šolar

šolski spis (frek. 390), šolska naloga, današnji čas, nezakonska mati, družbene razmere, osnovna šola, dobro življenje, naslednji dan, glavni junak, srednja šola, materni jezik, današnji svet, glavni lik, prosti čas, dober prijatelj, prava ljubezen, športni dan, glavna oseba, izgubljeni sin, pravi prijatelj, posmrtno življenje, svetovna vojna, celo življenje, krščanska vera, velik vpliv, spolni odnos, organska potreba, ljubljena oseba, človekova usoda, kulturni dom, nezakonski otrok, očetova smrt, velik problem, mlado dekle, vsakdanje življenje, današnja družba, povodni mož, dolg čas, Cankarjev roman, pisna naloga (frek. 36)

temeljno besedišče

vzgojna vloga slovenščine

[pridevnik + samostalnik] samo v Kresu: 279 zvez izmed 2266 izluščenih (12,3 %)

Prvih 40 glede na logDice

pravna oseba (logDice 11,176), nadaljnje besedilo, fizična oseba, nadzorni svet, gospodarska družba, združene države, svetovalna služba, obveščevalna služba, diplomska naloga, prečiščeno besedilo, delniška družba, zemljiška knjiga, socialno delo, državni svet, občinski svet, odgovorna oseba, posebna potreba, negativno čustvo, lastninska pravica, študentski dom, prednostna naloga, leva stran, škodljivi vpliv, varnostna služba, desna stran, izredna razmera, uradna oseba, tajna služba, državni svet, investicijska družba, delovni čas, terensko delo, prihodnje leto, posredniška družba, zvezna država, ločeno mnenje, brezposelna oseba, mednarodni odnos, ustavna pravica, letalska družba (logDice 8,357)

medpredmet. povezava

neumetnostna besedila

neliterarni eseji

šolski slovar

[pridevnik + samostalnik] samo v Šolarju: 177 zvez izmed 2266 izluščenih (7,8 %)

Prvih 40 glede na logDice

Antigonina sestra (logDice 10,034), Črtomirjev vojak, Kreonov sin, Rebulov roman, Kreonov zakon, Bronjin mož, Antigonin brat, Lojzkin starš, Kreonova žena, Descartesova misel, kiparska naloga*, zavedni del*, Kovačičev roman, Polikarpov odnos, Jazonova žena, Hamletova težava, mrtvaški pot, Polinejkova sestra, zardele oči, Lojzkine oči, aplikativni cilj*, Polikarpovo dejanje, Salomin odnos, Gregorjevo mesto, Sofoklejevo delo, novodobni razum, Bubijevo dekle, Bogomilina želja, Bronjina želja, Odisejevo mnenje, Johanaanova glava, Johannova glava, prerokova glava, edini način, psihično čustvo, edina oseba, Ožbejev oče, rumeni junak, edini prijatelj (logDice 7,241)

Kvalitativna analiza popravkov v korpusu Šolar

- **Luščenje kolokacij**:
  - vsaj en del zveze mora biti popravljen
  - jezikovni popravek mora biti uvrščen v tip 'napaka besedišča'

- **Izbor za luščenje in analizo**:
  - pridevnik + samostalnik (278)
  - samostalnik + samostalnik (271)
  - glagol + samostalnik (274)

[pridevnik + samostalnik] statistika

- skupaj 278 primerov
- šum: 52 primerov (18,7 %)
- relevantnih za analizo = 226 primerov (81,3 %):

    - popravek besedne zveze: 47 primerov (20,8 %); popravek ene besede v zvezi: 179 (79,2 %)
    - esej/spis: 213 (94,2 %), pisni izdelek (učna ura): 13 (5,8 %)
    - samo slovenščina

[pridevnik + samostalnik]

- praviloma nizke frekvence

red prišla *nova*	punca	| *sošolka* po imenu
bivša | *bivše*	punca	| *dekle* . Na vsak na
e stvari. *Mlada*	punca	| *dekleta* | *dekleta*
nlada | *mlado*	punca	| *dekle* nadgrajeval
je o *najstniški*	punci	| *dekletu* Fatimi Isn
njo | *prejšnje*	punco	| *dekle* . Nisi ji hote
i | , kako *lepo*	punco	| *dekle* ima. Julija t

materin	| *materni jezik* in se
materin	| *materni jezik* . Prer
materin	| *materni jezik* lep in
materin	| *materni jezik* . Razl
materinega	| *maternega jezika* .
materinega	| *maternega jezika* .
materinskega	| *maternega jezika* .
materinskem	| *maternem jeziku* in
materinskem	| *maternem jeziku* |
materinskemu	| *materinemu jeziku*
materinski	| *materni jezik* upora
materinski	| *materni jezik* . Zave
materinski	| *materni jezik* in to
materinski	| *materni jezik* . V ce
materinski	| *materni jezik* in to
maternim	| *maternim jezik* | je

[pridevnik + samostalnik] nerelevantni primeri

- vsebinsko narobe, besedilni popravki, nerazumljivi popravki, oblika

Tartuffe je bil duhovnik, ki je odšel k ***bogatemu***	**Argonu**	| *Oregonu* na dom in se naprav
o se mama, ata in Tone odpravili na izlet v ***Kranjsko***	**goro**	| *Rogaško Slatino* . Tam je vide
ɔlj znano mislijo razsvetljenstva, ki jo je zapisal znan	**grški**	| *francoski* ***pisatelj*** Descartes
i upogibalo s tokom. Kreona, ***Antigoninega***	**očeta**	| *strica* , | pa težko uvrstimo
ɔelje v Dobjo vas | , kjer imam pevske vaje. ***Pevske***	**vaje**	| *Te* trajajo do šestih , | in ko
mesto meri 275 km², kar nam pove, da je ***največje***	**mesto**	| *središče* v Sloveniji. V Ljublja
im . Ko *Polonij pove* | *je Polonij povedal* Klavdiju o	**Hamletovi**	| *princevi* ***ljubezni*** do Ofelije,
a stoji §§§ *v* | *na* nekem nam neznanem ***prevoznem***	**sredstvu**	| *avtobusu* , ki bi lahko bil avto
ajvečkrat opazimo športne copate. Je *velik* | ***velika***	**ljubitelj**	| *ljubiteljica* narave in živali. S

[pridevnik + samostalnik] skupine podobnih primerov leksikalnih popravkov

- tipologija popravkov
- tipologija besedišča
  - izrazi s področja slovenščine – ostalo besedišče (pomenska/pojmovna polja, abstraktnost, strokovnost, pogostost rabe …)
- beseda – besedna zveza – mejni primeri

[pridevnik + samostalnik] zaznamovano

- večinoma popravki besed, manj zvez

pedofilija, droge, nezvestoba pa so največkrat ***rizični***	**faktorji**	| *dejavniki* , da se starši od otrok umaknejo, stika z njimi ne
jaz | *Jaz* sem se čisto malo , | — imam *nova* | ***novo***	**frizuro**	| *pričesko* in nekaj lažja sem. Čez dva tedna grem na morje s
e. Mislim, da ji je na začetku *pomembna* | ***pomemben***	**funkcija**	| *položaj* kraljice, ker ne ve, kaj se je v resnici zgodilo z
Ljudje smo bitja odnosov. Za naše ***normalno***	**funkcioniranje**	| *delovanje* potrebujemo drug drugega. Sami pripomoremo k d
e. Romeo je tik pred tem, da spije strup, a v ***zadnjem***	**momentu**	| *trenutku* do vežice pride duhovnik Lorenzo, ki Romea seznan
rali prislužiti kruh. Vzgajala jih je *kot* | *za* ***pogumne***	**možake**	| *ljudi* . To vidimo tudi v besedilu kot je: ” | „ Zdaj greš v bo
gremo skupaj na policijo. Ko smo prispeli na ***najbližjo***	**policijo**	| *policijsko postajo* , smo poiskali dežurnega policista in mu
dolgočasna. Prav gotovo ga tudi to potre in je njegovo	**brezvezno**	| *nesmiselno* ***življenje*** tako še bolj temno in pesimistično. V
e prikupiti s tem, da ga je zavajala in ga peljala na vsa	**živa**	| *mogoča* ***potovanja*** | , misleč, da gre na službeno potovanje

[pridevnik + samostalnik] nenapake

- bolj specifično (vsebinsko natančnejše, pomensko ožje)

oddigrali | *odigrali* predstavo o Romeu in Juliji, ki je bila ***prava***	**drama**	| *tragedija* . *Meni kot oseb*
v različnem času, kar ni bilo značilno za *Antično* | ***antično***	**gledališče**	| *dramatiko* . Tudi v življe
obdobju renesanse. Romeo in Julija sta *glavna* | ***glavni***	**junaka**	| *dram. osebi* . Capuletovi so priredi
bo dobro godilo, če ga dobi s kakšno flašo. ***Celotna***	**knjiga**	| *drama* je dokaj grozna, saj se dog
so | *sta* izpostavljeni dve najbolj temeljni sporočili ***celotne***	**knjige**	| *komedije* . Prvo sporočilo je povez
jaz mislim. Odlomek Krst pri Savici je epsko- ***lirska***	**pesem**	| *pesnitev* , ker *sestavlja* | *vsebuje*
tudi otroka ne. Delo Krst pri Savici je epsko- ***lirska***	**pesem**	| *pesnitev* , ki izpoveduje čustva in g
| *romantike* , ki traja od 1830 do 1848. Delo je epsko- ***lirska***	**pesem**	| *pesnitev* | , zgrajena iz treh delov
oba potrudita za to. Nezakonska mati je *lirska* | ***lirsko***	**pesnitev**	| *delo* , ki spada v obdobje *romantik*
/p> Za konec lahko povem, da je Wiliam Shakespeare ***odličen***	**pisatelj**	| *dramatik* in najboljši pisec tragedi
Romeo in Julija, ljubezenska tragedija odličnega ***renesančnega***	**pisatelja**	| *dramatika* Williama *Shakespeara* |
Temu sonetu *je sledil* | *sledi* Uvod, kjer govori o *sami* | ***samih***	**vojni**	| *bojih* | *bojih med* | *za* pogansko
dramatika Williama Shakespeara je najbolj poznana ***ljubezenska***	**zgodba**	| *drama/* . Je dramsko delo, zgodba
ipravljena storiti marsikaj. Njen značaj spoznavamo skozi ***celotno***	**zgodbo**	| *dramo* in če sprva deluje plehko in

[pridevnik + samostalnik] književnost in jezik

atik Sofokles in	*je* tragično	**dramatsko**	*dramsko* ***besedilo*** .
zvrst oz. odlomek je epsko in		**ilirsko**	*lirsko* ***besedilo*** , to
jena iz treh delov. Prvi del je		**posvetni**	*posvetilni* ***sonet*** Mat
nes imamo vse preveč tujk in		**prenesenih**	*prevzetih* ***besed*** iz r
večinoma uporabljamo sleng,		**privzete**	*prevzete* ***besede*** , p
nska mati spada v liriko in je		**vložena**	*vložna* ***pesem*** . Nast
nska mati spada v liriko in je		**vložena**	*vložna* ***pesem*** . Nast

[pridevnik + samostalnik] oblikovno blizu, pomensko različne

raje rekli država ponekod še vedno obnaša na tak	**diskriminanten**	|	*diskriminatoren* ***način*** . Na
Po končani osnovni šoli , | se je vpisal v srednjo	**električno**	|	*elektro* ***šolo*** §§§ — TOK, v V
e in junakovega ideala. V delu je veliko	**izpovednih**	|	*izpovedanih* ***čustev*** , še po
šestdeset let. Na našo šolo je prišel kot	**književni**	|	*knjižni* ***gost*** , zato , | ker
velik nesmisel, karikira pomembne osebnosti. Za	**Majhne**	|	*male* ***ljudi*** , kamor spada t
” (umobolnici), prav tako tudi *veliki* | *starejši* in	**mali**	|	*mlajši* ***izvedenec*** za metaf
Kreonu privedli Antigono, *kot* | storilko *tega* |	**protizakonskega**	|	*protizakonitega* ***dejanja*** .
ka je bila poražena. Da je bilo to delo Prešernova	**osebnostna**	|	*osebna* ***izpoved*** , bi lahko
Prešeren preusmeri slog predvsem na Črtomirovo	**osebnostno**	|	*osebno* ***raven*** . Zato pravin
cela družina ostane sama? Spomnim se družine iz	**sosedovega**	|	*sosednjega* ***kraja*** , kjer se
rja , | pridejo mladi in stari v Podpeco in gradijo	**snežene**	|	*snežne* ***gradove*** . Imenuje
ključek bi lahko neskončne ure pisali, kaj je prav,	**trda**	|	*trdna* ***načela*** ali prilagajan

[pridevnik + samostalnik] oblikovno blizu, pomensko različne

Gregor je partizan, ki na ***rutinskem***	**sprehodu**	| *obhodu*	opazi mladenko, ki s svojimi kre
ila Črtomirja. Krščanska vera govori o	**posvetnem**	| *posmrtnem* ***življenju*** ,	| — da bosta lal
ljubil Bogomilo in je hotel biti tudi v	**posvetnem**	| *posmrtnem* ***življenju***	znjo. Črtomir se j
opazovalec | , saj je opazil čisto vse	**možne**	| *mogoče* ***podrobnosti***	. Od ženice do nek
i. Poznam en primer ki zaradi ***verske***	**izpovednosti**	| *verske pripadnosti/izpovedi*	starši svoji

[pridevnik + samostalnik] kolokacije

- uporaba pomensko neustrezne kolokacije:

posvetno > posmrtno življenje, iz socialnega reda > sloja, na družbeni ravni > lestvici, na višji ravni > na višjem položaju, filmski trak > filmsko platno

- manj običajna kombinacija besed:

nabrano > naropano bogastvo, imeti v popolni kontroli > pod kontrolo, zadelana > zamašena ušesa, zlati > velik zaslužek, biti vzkipljive jeze > narave, vse mogoče > možne podrobnosti, velika > visoka samopodoba

- mešanje:

verska izpovednost > pripadnost/izpoved, nadmorska gladina > višina

Ocena kvalitativne analize

- razmeroma zamudna
- frekvence relevantnih primerov so nizke, a so podatki pomembni > prispevajo k razumevanju procesov usvajanja „aktivnega“ besedišča > načrtovanje pouka leksike, izbor gradiv za usvajanje besedišča in snovanje priročnikov
- na podlagi trenutnih delnih analiz posploševanje rezultatov še ni mogoče oz. zaključki o procesih usvajanja kolokacij (leksike) ne morejo biti zanesljivi

Kako naprej?

- primerjalno luščenje kolokacij razširiti tudi na manj frekventne samostalnike > primerjava s podatki iz kvalitativne analize
- primerjalno luščenje kolokacij uporabiti na verziji korpusa Šolar brez popravkov (šum!) > statistična analiza
- obe metodi (primerjalno luščenje kolokacij in kvalitativno analizo popravkov) razširiti na druge zveze
- ? analiza po razredih
- ? leksikalne analiza besed (kot osnova)
- ??? nove ideje ???

Reference

Špela Arhar Holdt in Tadeja Rozman. 2015. Možnosti uporabe podatkov iz korpusa Šolar za pripravo slovarskih priročnikov. V: M. Smolej, ur., Obdobja 34: Slovnica in slovar - aktualni jezikovni opis, 1. del, str. 67–74. Znanstvena založba Filozofske fakultete UL. http://centerslo.si/wp-content/uploads/2015/11/34_1-Arhar-Hol-Roz.pdf

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

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

Iztok Kosem, Mojca Stritar Kučuk, Sara Može, Ana Zwitter Vitez, Špela Arhar Holdt in Tadeja Rozman. 2012. *Analiza jezikovnih težav učencev: korpusni pristop*. Trojina, zavod za uporabno slovenistiko.

Nataša, Logar, Miha Grčar, Marko Brakus, Tomaž Erjavec, Špela Arhar Holdt, Simon Krek. 2012: *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Trojina, Fakulteta za družbene vede.

Senja Pollak. 2015. Luščenje kolokacij iz korpusa uporabniških spletnih vsebin. V: M. Smolej, ur., *Obdobja 34: Slovnica in slovar - aktualni jezikovni opis*, 2. del, str. 601–607. Znanstvena založba Filozofske fakultete UL. http://centerslo.si/wp-content/uploads/2015/11/34_2-Pollak.pdf

Senja Pollak in Špela Arhar Holdt. 2015. Identifying corpus-specific collocations: the case of spoken Slovene. V: K. Gajdošová in A. Žáková, ur., *Natural language processing, corpus linguistics, lexicography: proceedings*, S. l., str. 117–125. RAM-Verlag.

Tadeja Rozman, Iztok Kosem, Nataša Pirih Svetina in Ina Ferbežar. 2015. Slovarji in učenje slovenščine. V: V. Gorjanc, P. Gantar, I. Kosem in S. Krek, ur., *Slovar sodobne slovenščine: problemi in rešitve*, str. 150–167. Znanstvena založba Filozofske fakultete UL.

Tadeja Rozman, Irena Krapš Vodopivec, Mojca Stritar Kučuk in Iztok Kosem. 2012. *Empirični pogled na pouk slovenskega jezika*. Trojina, zavod za uporabno slovenistiko.

Marko Stabej. 2011. Jezikovni potrošnik in potrošnica. *Sodobna pedagogika*, 62=128(2), 102–113. Zveza društev pedagoških delavcev Slovenije. http://www.dlib.si/details/URN:NBN:SI:doc-RF8JNPLQ