

# BESEDNE SKICE ZA SLOVENŠČINO. KRITIČNI POGLED

Olga Yeroshina Pobirk, Petra Zaranšek, Simon Šuster

Trojina, zavod za uporabno slovenistiko, Škofja Loka

UDK 811.163.6'374:811.163.6'322.3'324

Prispevek obravnava ozadje delovanja Besedne skice, ki je ena temeljnih funkcij orodja Sketch Engine, z vidika leksikografskega dela. Avtorji poleg opisa besedne skice kritično presojujejo vlogo, ki jo imajo pri izdelavi in prikazu besednih skic označenost korpusa, njegova besedilna sestava, slovnična razmerja, določena v slovnici besednih skic, in nenazadnje tudi nastavitve pri prikazu kolokatorjev.

besedna skica, označenost korpusa, sestava korpusa, slovnična razmerja, statistični izračuni

The article discusses the functioning of Word Sketch, one of the basic functions of the Sketch Engine web-based application, particularly with regard to lexicographic work. In addition to describing an individual example, the authors assess the role that other factors play in the creation and display of Word Sketches, such as corpus annotation, corpus composition, the grammatical relations that are determined in the sketch grammar, and the settings regarding the arrangement of the displayed collocates.

Word Sketch, corpus annotation, corpus composition, grammatical relations, statistical scores

## 1 Uvod

Uporaba korpusov za opis jezika v leksikografske in druge raziskovalne namene tudi v slovenskem prostoru ni več novost, kar potrjuje vedno več študij, ki obravnavajo to tematico. Korpus nudi obilico jezikovnih podatkov, vendar potrebujemo za njihovo obdelavo in interpretacijo zmogljiva ter »pametna« orodja. Primer takšnega sistema je večfunkcionalno orodje Sketch Engine (v nadaljevanju: SkE), ki omogoča različne prikaze in načine obdelave podatkov v jezikovnem korpusu.

SkE je orodje za prikaz in analizo jezikovnih podatkov v besedilnem korpusu, ki obsega naslednje osnovne funkcije, predvidene za vsebinsko delo z izbranim korpusom: Konkordance (*Concordance*), Seznam besed (*Word List*), Besedna skica (*Word Sketch*), Tezaver (*Thesaurus*), Primerjalna skica (*Sketch-Diff*). Od naštetih sodita med dve bistveni funkciji

Konkordance in Besedna skica, ki je tudi predmet tega prispevka. V nadaljevanju bo najprej na kratko predstavljeno delovanje in namen besednih skic, nato pa bomo osvetlili ozadje njihovega delovanja, ki ga je treba vsaj do določene mere poznati, da znamo pravilno interpretirati dobljene podatke.

## 2 Besedna skica

Besedna skica predstavlja velik korak v smislu pomoči pri leksikografskem delu, odkar so prišli v rabo korpusi. Besedna skica prikazuje leksikalni profil izbrane iztočnice s podatki o njenem tipičnem sobesedilnem okolju (Gantar idr. 2009: 33). Gre za avtomatske, na korpusu temelječe sežetke slovničnega in kolokacijskega vedenja neke besede (Krek, Kilgarrieff 2006), ki leksikografu dajejo hiter pregled nad dogajanjem okoli iztočnice. Leksikografu na ta način pomagajo tako pri pomenskem

razdvoumljanju kot pri odločanju o tem, katere informacije naj tudi neposredno predstavi v slovarju oz. bazi, ki ga/jo sestavlja (Atkins, Rundell 2008: 110).

## 2.1 Sestava besedne skice

Slika 1 prikazuje del besedne skice za lemo *ogenj*.

Številka 39.083 nad besedno skico pove število pojavitev iskane leme, tj. besede v vseh njenih oblikah, v izbranem korpusu. Iz prvega stolpca z oznako *a\_modifier* je mogoče razbrati, da je iztočnica *ogenj* v korpusu Fida-PLUS 8.961-krat rabljena kot samostalniško jedro v kombinaciji s pridevnikom. Spodaj sledi seznam pridevniških kolokatorjev, ki jo najpogosteje modificirajo in z njo tvorijo samostalniško besedno zvezo (npr. *taborni ogenj*). Stolpci *prec\_z-d*, *prec\_na-d* in *prec\_v-d* prikazujejo besede, ki se običajno pojavljajo

pred predlogi *z*, *na*, *v*, ki skupaj z iztočnico tvorijo predložno besedno zvezo (npr. *odstaviti z ognja*). V stolpcu *gen\_2* so naštetih samostalniški kolokatorji, ki skupaj z iztočnico v rodilniku tvorijo samostalniško besedno zvezo (npr. *prekinitev ognja*). Stolpec z oznako *is\_obj4* vsebuje glagole, ob katerih se *ogenj* pojavlja kot samostalnik v tožilniku (npr. *zakuriti ogenj*). Podobno stolpec *is\_subj* navaja glagole, ki so najpogosteje rabljeni ob iztočnici v imenovalniku, a ker je zgornja meja absolutne pogostosti v želji po bolj strnjeni besedni skici nastavljena na 50 pojavitev, imamo le en glagol, ki presega to mejo – *goreti (ogenj gori)*. V stolpcu *prec\_prep* so naštetih najpogostejši predlogi, ki se pojavljajo ob iztočnici *ogenj*.

S klikom na »podčrtano številko« (Slika 1), je mogoče dostopati neposredno do konkor-danc za izbrani kolokator. Številčne vrednosti desno od kolokatorjev pomenijo naslednje:

### ogenj Fida PLUS 620m freq = 39083

<b>a modifier</b>	<b>8961 1.3</b>	<b>prec z-d</b>	<b>2208 5.0</b>	<b>prec v-d</b>	<b>2769 3.0</b>	<b>prec prep</b>	<b>10019 2.7</b>	<b>gen 2</b>	<b>3074 1.8</b>
<input type="checkbox"/> taboren	580 96.24	<input type="checkbox"/> odstaviti	474 75.87	<input type="checkbox"/> železo	235 65.21	<input type="checkbox"/> z	2552 34.13	<input type="checkbox"/> prekinitev	706 81.98
<input type="checkbox"/> navzkrižen	487 83.4	<input type="checkbox"/> igranje	53 34.38	<input type="checkbox"/> roka	360 47.81	<input type="checkbox"/> ob	509 30.09	<input type="checkbox"/> bruhaletc	125 81.19
<input type="checkbox"/> zmeren	491 68.16	<input type="checkbox"/> igra	138 32.5	<input type="checkbox"/> znajti	154 35.15	<input type="checkbox"/> nad	212 29.71	<input type="checkbox"/> požiralec	108 73.22
<input type="checkbox"/> šibek	557 65.35	<input type="checkbox"/> igrati	194 32.1	<input type="checkbox"/> vreči	99 33.3	<input type="checkbox"/> v	3115 28.12	<input type="checkbox"/> prasketanje	61 60.19
<input type="checkbox"/> blag	227 56.11	<input type="checkbox"/> vzeti	128 30.11	<input type="checkbox"/> poslati	157 30.26	<input type="checkbox"/> okrog	70 25.98	<input type="checkbox"/> prižiganje	71 56.05
<input type="checkbox"/> olimpijski	460 51.53	<input type="checkbox"/> boj	56 26.96	<input type="checkbox"/> pošiljati	64 29.59	<input type="checkbox"/> pred	324 24.69	<input type="checkbox"/> kurjenje	81 52.74
<input type="checkbox"/> topniški	82 50.9	<input type="checkbox"/> umakniti	51 25.55	<input type="checkbox"/> biti	267 19.4	<input type="checkbox"/> proti	165 23.86	<input type="checkbox"/> bruhanje	66 49.15
<input type="checkbox"/> majhen	784 47.4	<input type="checkbox"/> biti	93 10.34	<input type="checkbox"/> imeti	169 11.02	<input type="checkbox"/> na	1547 23.7	<input type="checkbox"/> ustavitev	62 43.49
<input type="checkbox"/> peklenski	102 47.04		>>		>>	<input type="checkbox"/> okoli	53 19.98	<input type="checkbox"/> vladavina	65 40.9
<input type="checkbox"/> stražen	50 46.92					<input type="checkbox"/> skozi	56 18.64	<input type="checkbox"/> bog	71 37.23
<input type="checkbox"/> odprt	300 44.87	<b>prec na-d</b>	<b>2628 4.9</b>	<b>is obj4</b>	<b>3185 2.9</b>	<input type="checkbox"/> kot	198 18.12	<input type="checkbox"/> element	82 33.14
<input type="checkbox"/> močen	398 42.43	<input type="checkbox"/> olje	776 70.79	<input type="checkbox"/> zakuriti	279 71.92	<input type="checkbox"/> brez	94 15.53	>>	
<input type="checkbox"/> večeren	124 38.96	<input type="checkbox"/> kuhati	335 54.83	<input type="checkbox"/> zanetiti	320 70.05	<input type="checkbox"/> zaradi	89 13.46		
<input type="checkbox"/> počasen	76 33.58	<input type="checkbox"/> kozica	80 50.6	<input type="checkbox"/> pogasiti	411 69.47	<input type="checkbox"/> iz	199 13.4	<b>is subj</b>	<b>1557 1.8</b>
<input type="checkbox"/> prijateljski	72 32.93	<input type="checkbox"/> dušiti	62 36.86	<input type="checkbox"/> bruhati	154 56.44	<input type="checkbox"/> od	144 13.34	<input type="checkbox"/> goreti	71 39.89
<input type="checkbox"/> svet	75 27.42	<input type="checkbox"/> minuta	60 28.91	<input type="checkbox"/> podtakniti	163 55.19	<input type="checkbox"/> k	77 13.06	>>	
<input type="checkbox"/> vroč	54 24.98	<input type="checkbox"/> postaviti	111 24.58	<input type="checkbox"/> prižgati	218 52.01	<input type="checkbox"/> o	73 5.49		
<input type="checkbox"/> živ	55 24.0	<input type="checkbox"/> biti	59 5.6	<input type="checkbox"/> kuriti	95 49.31	<input type="checkbox"/> pri	56 3.53		
<input type="checkbox"/> nizek	57 19.34		>>	<input type="checkbox"/> gasiti	53 39.8	<input type="checkbox"/> za	179 2.89		
<input type="checkbox"/> notranji	56 18.95			<input type="checkbox"/> ugasniti	62 35.4	<input type="checkbox"/> po	71 2.37		
<input type="checkbox"/> srednji	51 18.18			<input type="checkbox"/> zmanjšati	160 34.02		>>		
<input type="checkbox"/> velik	192 15.75			<input type="checkbox"/> opaziti	53 20.18				
	>>				>>				

Slika 1: Del besedne skice za lemo *ogenj*.

»podčrtana številka« pove absolutno pogostost kolokatorja ob izbrani iztočnici, številka desno od nje pa izpostavljenost posameznega kolokatorja. Izpostavljenost (ang. *salience score*) je statistična vrednost, ki daje podatek o tem, kako močna je vez med kolokatorjem in izbrano iztočnico – večja ko je vrednost izpostavljenosti, močnejša je vez (prav tam).<sup>1</sup> Kolokatorji so v našem primeru razvrščeni glede na izračun izpostavljenosti in ne po absolutni pogostosti, zato se npr. sicer dva izmed najpogostejših glagolov *biti* in *imeti* v stolpcu *prec\_v-d* nahajata na zadnjem mestu.

Pri razbiranju in interpretaciji podatkov, pridobljenih z besedno skico, je poleg poznavanja same uporabe orodja zelo pomembno tudi, da uporabnik pozna ozadje njegovega delovanja, zato bodo v nadaljevanju predstavljeni dejavniki, ki odločilno vplivajo na dobljene rezultate: besedilna sestava korpusa, naložena v SkE, označenost korpusa in slovnicična razmerja, definirana v datoteki s slovnicičnimi razmerji.

### 3 Ozadje delovanja besednih skic

#### 3.1 Slovnica besednih skic

Za funkcije Besedna skica, Primerjalna skica in Tezaver potrebujemo *slovnico besednih skic* (ang. *Sketch Grammar*, dalje SBS), ki je nabor slovnicičnih razmerij, za katera želimo, da so prikazana v besednih skicah. Ta razmerja so definirana glede na zakonitosti v konkretnem jeziku, namen jezikovnih poizvedovanj ter nabor oznak, uporabljenih pri označevanju posameznega korpusa. Slovnico, ki je osnova za izdelavo besednih skic za slovenščino na 621-milijonskem korpusu FidaPLUS, je leta 2006 napisal Simon Krek (Krek, Kilgarriff 2006).

Kot primer slovnicičnega razmerja za korpus FidaPLUS navajamo recipročno zvezo pridevnika in samostalnika:

\*DUAL

=a\_modifier/modifies

2:[tag="P.\*"] [tag="P.\*" | word="," | word="se" | word="si"]{0,5} 1:[tag="S.\*"]

Razmerje je sestavljeno iz elementov, ki se začnejo in končajo z oglatimi oklepaji [], predstavljajo pa potencialne pojavnice v realnem besedilu. Znotraj elementov in med njimi so uporabljeni znaki iz regularnih izrazov „,“|\*{} v kombinaciji z oblikoskladenjskimi oznakami, uvedenimi z atributom *tag*, ali dejanskimi koščki besedila, uvedenimi z atributom *word*. Zgoraj opisano razmerje predvideva naslednjo situacijo: na prvem mestu se mora pojaviti pridevnik v katerikoli obliki, sledi mu od nič do pet mest, na katerih se lahko pojavijo pridevnik, vejica, *se* ali *si*, na koncu pa mora biti samostalnik, ponovno v katerikoli obliki. S številkami 1 in 2 označimo ključna elementa, tj. iskano iztočnico (št. 1) in element, ki ga želimo prikazati v obliki kolokatorja (št. 2). Recipročnost razmerja (*DUAL*) pomeni, da če je npr. naša iztočnica pridevnik (št. 2), bodo v skici prikazani vsi samostalniški kolokatorji (št. 1), in obratno, če nas zanima samostalnik, bodo prikazani vsi pridevniški kolokatorji. Recipročnost je upoštevana tudi pri poimenovanju razmerja.

Rezultat tega razmerja je en stolpec besedne skice za posamezni samostalnik ali pridevnik (iskano lemo), ki za samostalniško lemo poda najpogostejše pridevnike, ki jo modificirajo, za pridevniško lemo pa najpogostejše samostalnike, ki ji sledijo. Na sliki 2 so tako na levi navedeni najpogostejši pridevniški kolokatorji za lemo *ogenj*, na desni pa najpogostejši samostalniški kolokatorji za lemo *navzkrižen*, ki so razvrščeni glede na izpostavljenost.

<sup>1</sup> Statistična izpostavljenost je prilagojena oblika Diceovega koeficienta in v grobem temelji na seštevku sopojavljanja besed v nekem slovnicičnem razmerju. Natančna enačba je dostopna na <http://trac.sketchengine.co.uk/attachment/wiki/SkE/DocsIndex/ske-stat.pdf?format=raw>. (Dostop 8. 7. 2009.)

## ogelj

Fida PLUS 620m freq = 39083

a	modifier	8961	1.3
<input type="checkbox"/>	taboren	580	96.24
<input type="checkbox"/>	navzkrižen	487	83.4
<input type="checkbox"/>	zmeren	491	68.16
<input type="checkbox"/>	šibek	557	65.35
<input type="checkbox"/>	blag	227	56.11
<input type="checkbox"/>	olimpijski	460	51.53
<input type="checkbox"/>	topniški	82	50.9
<input type="checkbox"/>	majhen	784	47.4
<input type="checkbox"/>	peklenki	102	47.04

## navzkrižen

Fida PLUS 620m freq = 21

modifies	2352	6.9	
<input type="checkbox"/>	ogelj	487	71.39
<input type="checkbox"/>	skladnost	176	64.05
<input type="checkbox"/>	zasiševanje	110	63.88
<input type="checkbox"/>	lastništvo	108	47.4
<input type="checkbox"/>	zasišanje	72	43.32
<input type="checkbox"/>	poizvedba	51	42.5
<input type="checkbox"/>	interes	131	37.85
<input type="checkbox"/>	reakcija	64	36.95
<input type="checkbox"/>	povezava	57	26.69

Slika 2: Vzorec rezultatov za recipročno razmerje pridevnik-samostalnik.

Na dan pisanja tega članka je v slovenski SBS definiranih 18 slovničnih razmerij, osem izmed teh je recipročnih.<sup>2</sup> Trenutno je v pripravi nova verzija SBS, ki bo obsegala več slovničnih razmerij (vključena bodo recimo razmerja z vezniki), tokrat s slovenskimi poimenovanji in izboljšavami, ki so se pokazale za smiselne pri delu s SkE v začetni fazi gradnje leksikalne baze za slovenščino v okviru projekta Sporazumevanje v slovenskem jeziku.<sup>3</sup>

Priprava SBS je jezikovnotehnološki postopek, ki ni vezan samo na poznavanje tipičnih jezikovnih pojavov v konkretnem jeziku in nabor oznak, uporabljenih v posameznem korpusu, temveč je pri tem zelo ali celo predvsem pomemben namen, za katerega želimo orodje oz. funkcijo izrabljati. Večina slovničnih razmerij v slovenski SBS, tako trenutni kot nastajajoči, je naravnana izrazito leksikografsko, se pravi za namen izdelave leksikalne baze za slovenščino. Slovnična razmerja so torej definirana glede na vnaprej določene vzorce, za katere predpostavljamo, da so v (slovenskem) jeziku tipični in posledično zanimivi za izdelavo slovarjev in podobnih leksikografskih del.

Zgornji pristop k definiranju slovničnih razmerij seveda ni edini možen. Kot primer

manj vnaprej določujočega pristopa navajamo del skice iz poskusnega 20-milijonskega dela korpusa FidaPLUS za lemo *mesto* (Slika 3).

V ozadju te skice je slovnica, ki je nastala po vzoru slovaške slovnice besednih skic, ki jo je napisal Vladimir Benko in jo je Simon Krek poskusno prilagodil za slovenščino. Primerjava razmerij, opredeljenih na osnovi slovaške SBS, z razmerji iz slovenske SBS, ki je trenutno v rabi za 621-milijonski korpus FidaPLUS, pokaže, da so v slovaškem primeru razmerja zastavljena ohlapnejše. Iskani kolokatorji namreč niso ločeni glede na besedne vrste ali druge slovnične/jezikovne oznake, ampak so v skupni stolpec združeni samo glede na mesto pojavljanja v relaciji do iskane leme.

Na sliki 3 so v levem stolpcu prikazani izrazi, ki se pojavljajo pred povezavo predloga za in iskano lemo *mesto*, v desnem pa izrazi, ki

## mesto Fida PLUS 20m SLOVASKI freq = 20946

Y za-d X	1543	1.3	v-d Y X	2438	1.3
boj	63	34.65	nov	506	44.85
tekma	81	30.5	središče	145	44.62
dežuren	16	30.32	glaven	78	30.44
potegovati	23	29.4	knežji	10	24.79
kandidat	27	23.8	bližina	28	23.66
napredovati	12	21.05	osvojit	27	22.4
razpis	18	20.31	okolica	21	20.29
boriti	12	19.32	zasesti	16	20.03
služba	16	15.67	velik	73	19.19
veljati	11	11.9	center	30	18.4
da	12	8.98	večen	10	17.08
iti	8	5.84	francoški	14	16.32
le	11	5.74	rojsten	10	16.16
v	79	5.49	star	26	15.29
in	18	4.56	svet	8	13.06
se	37	4.07	naš	33	13.0
biti	115	2.82	nemški	13	12.99
še	8	1.1	bližnji	9	12.71
tudi	9	0.44	majhen	15	11.49
			ta	90	10.68
			italijanski	8	10.53
			drug	30	9.91
			nek	11	9.17
			sam	15	8.85
			del	17	8.22

Slika 3: Del skice za lemo *mesto*.

<sup>2</sup> Več o drugih vrstah razmerij glej v Krek, Kilgarriff 2006.

<sup>3</sup> www.slovenscina.eu.

se pojavljajo med predlogom *v* in iskano lemo *mesto*. Iskana lema je zmeraj označena z oznako *X*, njena sobesedilna okolica, ki jo želimo prikazati, pa z oznako *Y*. Ker razmerje ni omejeno na nobeno od besednih vrst, se med njimi pomešano znajdejo samostalniki, glagoli, pridevniki, členki, vezniki, predlogi in zaimki.

Ker se poskuša oddaljiti od obstoječih zakonitosti in predvidevati čim manj, tak pristop po eni strani pušča odprtih več možnosti, da odkrijemo obrobnejše, morda nepričakovane pojave v jeziku, po drugi strani pa je prav zaradi tega potrebno več pozornosti in časa za pregled in interpretacijo množice nediferenciranih jezikovnih podatkov, zaradi česar je v praksi, predvsem v splošnokoleksikografskih projektih,<sup>4</sup> kjer so časovni in finančni okviri jasno zastavljeni, tak pristop manj primeren.

### 3.2 Označenost korpusa

Za iskanje in povezavo besed v slovnična razmerja, ki so nato predstavljena v besednih

skicah, mora SkE vedeti, kako najde besede, ki so v nekem razmerju. Če npr. razmerje predvideva zvezo pridevnika in samostalnika, potrebujemo oblikoskladenjsko označen korpus, kakršen je FidaPLUS (gl. nadaljevanje). Strogo gledano je sicer edini predpogoj za vključitev korpusa v orodje SkE tokenizirnost, tj. razdeljenost besedila na pojavnice, kar pomeni, da bi lahko zelo osnovna oblika SBS delovala že na korpusu, ki ni ne lematiziran ne oblikoskladenjsko označen, seveda pa bi tako dobili zelo osiromašene in leksikografsko ne prav uporabne rezultate.

Ker so besedne skice za slovenščino izdelane na osnovi oblikoskladenjskih oznak, je torej zelo pomembna pravilna označenost korpusa. FidaPLUS je bila označena avtomatsko, kar s seboj prinaša določen odstotek napak, saj je slovenščina zaradi obremenjene morfologije, pri kateri prihaja do prekrivnosti končnic, zelo težaven jezik za označevanje.<sup>5</sup> Posledica tega je pojavljanje napak v besednih skicah, kjer se lahko v nekem razmerju znajdejo izrazi, ki tja ne sodijo (Slika 4).

Primeri na sliki 4 prikazujejo napačno uvrščene kolokatorje, ki so posledica napačno pripisane leme in oblikoskladenjske oznake, saj je samostalniški kolokator prepoznan kot glagol in je tudi umeščen v stolpec z razmerjem *pridevnik-polnopomenski glagol*.<sup>6</sup> Spodnje okno

glasben		Fida PLUS 620m freq = 149951	
post verb: 2563 0.1			
<input type="checkbox"/> spotiti	549 85.35		
<input type="checkbox"/> voščiti	405 73.9		
<input type="checkbox"/> večeriti	136 62.93		
<input type="checkbox"/> zvrstiti	213 54.88		
<input type="checkbox"/> podlagati	61 42.51		
<input type="checkbox"/> skupiniti	69 36.92		

  

Dnevnik, časopis	19.20 Kulturni utrip, ponovitev - 19.50	Glasbeni spoti - 20.00 Pop mix z Andrejo, glasbena
Dnevnik, časopis	živo iz studia Televizije Celje - 21.00	Glasbeni spoti - 21.30 Oglasi - 21.35 Ko karte govorijo
Dnevnik, časopis	živo iz studia Televizije Celje - 21.35	Glasbeni spoti - 22.05 Oglasi - 22.10 Viša tedna
časopis	Rock TV 22.00 Legalizacija marihuane 22.37	Glasbeni spoti 22.40 Poročila: Objektiv Gorenjske
časopis	zdravnik, ponovitev 18.40 Santa Barbara 19.30	Glasbeni spoti 20.00 Detektiva, 1. del angleške
časopis	vsakogar, ponovitev 18.40 Santa Barbara 19.30	Glasbeni spoti 20.00 Lutkovno gledališče, prenos
časopis	ponovitev filma 18.00 Avto drom MMTV 19.30	Glasbeni spoti 20.00 Očetje in sinovi, ameriški
časopis	prijatelji, ponovitev 19.00 Kuhajmo skupaj 19.30	Glasbeni spoti 20.00 Beverly Hills Buntz, ameriška

Slika 4: Napačna lematizacija kolokatorjev *spot*, *voščilo*, *večer*, *zvrst*, *podlaga* in *skupina*.

- 4 Z izrazom *splošnokoleksikografski projekti* so mišljeni projekti, usmerjeni v opis splošnega jezika.
- 5 O tem, da je slovenščina v primerjavi z angleščino težaven jezik za označevanje, priča tudi dejstvo, da slovenski nabor oblikoskladenjskih oznak obsega okoli 2.000 elementov (Erjavec, Krek 2008: 51). Za primerjavo, standardni nabor britanskega nacionalnega korpusa vsebuje okoli 160 oznak (Označevalni sistem CLAWS).
- 6 »Spoti« je mogoče razumeti kot samostalnik v množini in kot obliko glagola *spotiti*.

s konkordancami potrjuje, da gre v resnici za razmerje *pridevnik-samostalnik*, k napaki pa je najverjetneje prispevalo tudi pomanjkanje konteksta, ki pomaga pri avtomatskem razdvoumljanju.

Naslednji izsek besedne skice za glagol *pobirati* (Slika 5) prikazuje razmerje *samostalnik v imenovalniku-polnopomenski glagol*, v katerem želimo najti osebkke, ki se tipično pojavljajo z zadevnim glagolom. Na vrhu zadetkov smo označili kolokatorja, ki sta sicer pravilno lematizirana, vendar napačno označena kot samostalnika v imenovalniku namesto kot samostalnika v tožilniku, kar potrdijo konkordance v spodnjem oknu.

Nekatera mesta s pogosto napačno označenimi besedami so:

a) pridevnik ženskega spola v tožilniku v krepkem tisku je napačno označen kot prislov (... označili ... za »**pogumno** žensko in veliko državnico.«);

b) samostalnik srednjega spola je napačno označen kot glagolska oblika v tretji osebi srednjega spola (*Kakovost slike je osnovno **merilo** pri nakupu digitalnega fotoaparata.*);

c) samostalnik v imenovalniku ali tožilniku množine je napačno označen kot samostalnik v rodilniku ednine (*Nekje sem zasledila, da **ženske** že po naravi težje prenašamo vročino ...; Dobro poznam ta občutek: **gledaš** mlade **ženske**, toda one te sploh ne opazijo.*);

d) samostalnik moškega spola v tožilniku je napačno označen kot samostalnik moškega spola v imenovalniku (***Prispevek** pobira javno komunalno podjetje ...*);

e) pridevnik moškega spola v imenovalniku je napačno označen kot samostalnik srednjega spola v rodilniku množine (*Pričakovali bi, da bo sin **vesel** takšne materine požrtvovalnosti.*).

Orodje SkE sicer prepozna tudi odvisnostnoskladenjsko razčlenjeni korpus, za kakršnega je mogoče napisati še kompleksnejšo in s tem natančnejšo SBS od trenutno veljavne za slovenščino, ki upošteva le oblikoskladenjske oznake (Krek, Kilgarriff 2006).<sup>7</sup> V tem primeru so tudi rezultati, dobljeni z besedno skico, pravilnejši, saj so prav odvisnostnoskladenjska razmerja ta, ki bi jih v besedne skice radi zajeli, vendar trenutno nimamo dovolj natančnega sredstva za njihovo razločevanje. Če bi lahko odvisnostnoskladenjske oznake uporabili vzporedno z oblikoskladenjskimi, bi to zelo zmanjšalo število napak zaradi napačne lematizacije.

**pobirati** Fida PLUS 620m freq = 11487

has subj	1675	7.6
<input type="checkbox"/> denar	69	26.29
<input type="checkbox"/> država	106	25.51
<input type="checkbox"/> davek	34	24.55
<input type="checkbox"/> suša	16	23.9
<input type="checkbox"/> davkarija	9	23.82

Delo, časopis	toliko kot pri nogometnem prvenstvu. Denar	pobira	Karmen. Naj zmaga najboljši, C s se glasi
revija	lahko » z lahkoto in z namenom « sploh še	pobira	javni denar za neko poslanstvo, ki ga preprosto
revija	organizirane kriminalne skupine, ki revežem	pobirajo	denar z obljubami, da jih bodo spravile
revija	tolarjev. </p><p> < Navaden > Denar sicer	pobira	žalsko komunalno podjetje, vendar je strogo
Dnevnik, časopis	www.dnevnik.si. (bš </p><p> Kdo staršem	pobira	denar? </p><p> Osupel sem v Dnevniku (8.

Slika 5: Napačna prepoznavna sklonov kolokatorjev.

7 Za korpus FidaPLUS, kot je v SkE, velja nabor oznak MULTTEXT-East v.3, v prihodnje pa bo uporabljen izboljšani nabor oznak JOS v kombinaciji z natančnejšim označevalnikom (Erjavec, Krek 2008).

### 3.3 Besedilna sestava korpusa

Ena temeljnih reči, ki najbolj vpliva na to, kakšne rezultate bo dala besedna skica, je besedilna sestava korpusa, saj analiza ne more pokazati ničesar drugega kot preoblikovano sliko tega, kar je bilo vanjo poslano. Pod besedilno sestavo razumemo vrste besedil, ločene glede na zvrst (leposlovje, stvarna besedila) in prenosnik (govorno, časopis, internet itn.), njihov obseg ter razpršenost virov znotraj posamezne vrste besedil.

FidaPLUS kot referenčni korpus vsebuje zelo raznolika besedila, med njimi tudi televizijske sporede, male oglase ipd. Tovrstna besedila

zaradi svoje ponavljajoče se narave vsebujejo veliko istovrstnih zadetkov, poleg tega pa v njih tudi primanjkuje konteksta, ki bi pomagal razdvojniti, zaradi česar so včasih rezultati, dobljeni iz takšnih besedil, manj relevantni za splošnokoleksikografske namene. Danes se tovrstna besedila (vključno s cenami delnic, križankami, športnimi rezultati, kazali ipd.) vsaj pri leksikografskih korpusih običajno odstrani (Atkins, Rundell 2008: 85, Kilgarriff, Rundell, Dhonnchadha 2006).

Na sliki 6 so označeni kolokatorji, ki, kot potrjujejo konkordance, izvirajo predvsem iz omenjenih besedil.

## glasben Fida PLUS 620m freq = 149951

modifies	161833	7.0	adj_modified	3432	2.1
<input type="checkbox"/> lestvica	12610	86.02	<input type="checkbox"/> zabavno	217	69.37
<input type="checkbox"/> oddaja	9747	70.51	<input type="checkbox"/> slovensko	201	56.87
<input type="checkbox"/> poslušalnica	339	67.28	<input type="checkbox"/> srednje	174	53.25
<input type="checkbox"/> jutranjica	349	66.34	<input type="checkbox"/> domače	129	51.8
<input type="checkbox"/> motorček	485	64.71	<input type="checkbox"/> literarno	48	48.72
<input type="checkbox"/> šola	15271	61.99	<input type="checkbox"/> samostojno	60	42.25
<input type="checkbox"/> utrineSlovenija	152	58.56	<input type="checkbox"/> humoristično	17	41.14
<input type="checkbox"/> utrinekSlovenija	128	56.58	<input type="checkbox"/> ljudsko	25	39.17
<input type="checkbox"/> utrinek	1063	56.29	<input type="checkbox"/> nizko	54	38.99
<input type="checkbox"/> spremljava	807	55.21	<input type="checkbox"/> veliko	269	38.56
<input type="checkbox"/> scena	2255	55.02	<input type="checkbox"/> narodnozabavno	12	37.98
<input type="checkbox"/> video	1552	54.33	<input type="checkbox"/> razvedrilno	12	36.27
<input type="checkbox"/> zvrst	1000	54.07	<input type="checkbox"/> nekaj	120	36.25
<input type="checkbox"/> želja	3733	53.04	<input type="checkbox"/> bogato	47	35.98
<input type="checkbox"/> pedagog	651	51.81	<input type="checkbox"/> popularno	20	35.94
<input type="checkbox"/> cocktail	154	51.48	<input type="checkbox"/> osnovno	27	34.57

Dnevnik, časopis	prodaja - 11.00 TV shop predstavlja - 13.30 Glasbeni motorček, ponovitev mladinske oddaje -
Dnevnik, časopis	, ponovitev avstralske nanizanke - 19.05 Glasbeni motorček, mladinska oddaja - 19.30 Risanke
Dnevnik, časopis	prodaja - 11.00 TV shop predstavlja - 13.30 Glasbeni motorček, ponovitev mladinske oddaje -
Dnevnik, časopis	, ponovitev avstralske nanizanke - 19.05 Glasbeni motorček, mladinska oddaja - 19.30 Risanke
Dnevnik, časopis	prodaja - 11.00 TV shop predstavlja - 13.30 Glasbeni motorček, ponovitev mladinske oddaje -
Dnevnik, časopis	, ponovitev avstralske nanizanke - 19.05 Glasbeni motorček, mladinska oddaja - 19.30 Risanke
časopis	22.40 Halifax, nani </p><p> TV 3 </p><p> 14.00 Glasbeni motorček - 16.30 Glasba brez meja - 17.30
časopis	Popotovanja z Janinom - 18.00 Bonanca - 19.05 Glasbeni motorček - 19.30 Risanke - 20.00 Poslednja
časopis	Južni Brooklyn. </p><p> TV 3 </p><p> 13.30 Glasbeni motorček - 17.00 Poletni utrip - 18.00

Slika 6: Prikaz tipičnih kolokatorjev iz televizijskih sporedov.

Poleg zgoraj omenjenega na dobljeno skico odločilno vpliva tudi odločitev snovalcev korpusa glede velikosti posameznih besedilnih komponent v korpusu (npr. knjižni proti časopisnemu prenosniku). S ponazoritvijo na sliki 7 želimo pokazati, kako se s spremembo besedilne sestave korpusa spremeni tudi besedna skica.

### glasben

Podkorpus revij		Podkorpus časopisov	
modifies	30167 3.9	modifies	110134 5.8
<input type="checkbox"/> lestevica	7648 10.06	<input checked="" type="checkbox"/> oddaja	9451 9.95
<input type="checkbox"/> scena	640 6.62	<input type="checkbox"/> lestevica	4951 9.43
<input type="checkbox"/> šola	1765 6.32	<input type="checkbox"/> šola	13259 9.23
<input checked="" type="checkbox"/> industrija	597 6.22	<input checked="" type="checkbox"/> želja	3559 8.61
<input checked="" type="checkbox"/> zvrst	320 5.9	<input checked="" type="checkbox"/> video	1479 7.98
<input checked="" type="checkbox"/> datoteka	263 5.86	<input type="checkbox"/> scena	1604 7.94
<input type="checkbox"/> skupina	950 5.55	<input type="checkbox"/> skupina	4507 7.8
<input checked="" type="checkbox"/> kariera	280 5.4	<input checked="" type="checkbox"/> večer	1709 7.76
<input checked="" type="checkbox"/> plošča	321 5.33	<input checked="" type="checkbox"/> utrinek	1054 7.63
<input checked="" type="checkbox"/> spremljava	191 5.2	<input checked="" type="checkbox"/> tv	1495 7.53
<input checked="" type="checkbox"/> festival	311 5.15	<input checked="" type="checkbox"/> utrip	868 7.25
<input checked="" type="checkbox"/> založba	227 5.15	<input checked="" type="checkbox"/> gost	1276 7.17
<input checked="" type="checkbox"/> okus	192 4.86	<input checked="" type="checkbox"/> uganka	762 7.16

Slika 7: Različni kolokatorji glede na podkorpusa revij in časopisov.

Prikazan je vzorec prvih trinajstih zadetkov za razmerje *pridevnik-samostalnik* za lemo *glasben* v dveh podkorpusih: s črno so označeni tisti kolokatorji, ki jih ne najdemo v drugem podkorpusu. Situacija je v tem primeru seveda skrajna, saj imamo na eni strani le podkorpus vseh časopisnih besedil iz korpusa FidaPLUS, na drugi strani pa vsa revijalna besedila iz korpusa, referenčni korpus pa je kombinacija obojega. Kljub temu je razvidno, da z izdelavo dveh podkorpusov oz. tudi z manj izrazito spremembo deležev v referenčnem korpusu sprožimo spremembe v razporeditvi kolokatorjev v besednih skicah. Tako lahko glede na sestavo korpusa FidaPLUS – 65 odstotkov besedil nosi oznako časopisno, 23 odstotkov oznako revijalno –<sup>8</sup> sklepamo, da bi se s krčenjem časopisnega dela korpusa zadetki iz televizijskih sporedov znašli *nižje* na seznamu kolokatorjev.

Televizijski sporedi se namreč pogosteje pojavljajo v časopisnem kot pa v revijalnem gradivu, o čemer pričajo podatki o izvoru levo od konkordanc na sliki 6.

### 3.3.1 Razpršenost virov

Kot že omenjeno, je za to, da dobimo relevantne podatke, pomembno tudi, da so viri čim bolj razpršeni. Besedna skica ne nudi podatkov o razpršenosti virov, v katerih se pojavljajo kolokatorji, je pa ta podatek mogoče najti v konkordancah.

Za kolokator *kredit* (Slika 8), ki je del razmerja *samostalnik-predlog* 'glede', ob vpogledu v konkordance ugotovimo, da vsi zadetki izhajajo iz enega vira. Kadar so viri za zadetek nerazpršeni, kolokator za splošnoleksikografske namene najverjetneje ni relevanten.

### 3.4 Statistično razvrščanje zadetkov v besedni skici

Poleg besedilne sestave in označenosti korpusa, na kar kot uporabniki nimamo vpliva, je za rezultate, ki jih dobimo v besedni skici, pomembno tudi to, kakšne nastavitve izberemo pri uporabi orodja. Vsakič ko želimo izdelati besedno skico, izberemo tudi možnosti prikaza, kar presega zgolj število izpisanih kolokatorjev ipd. (Gantar idr. 2009: 34), ampak zadeva tudi določitev tega, kolikšni naj bosta vrednosti najnižje absolutne pogostosti kolokatorjev in statistične izpostavljenosti oz. po katerem od statističnih izračunov naj bodo kolokatorji v skici razvrščeni.

Na sliki 9 prikazujemo mesta, kjer se skici za isto iskano lemo razlikujeta glede na izbrani statistični izračun.

Z besednih skic za lemo *obdobje*, okrajšanih na 16 kolokatorjev, je v slovnčnih razmerjih *samostalnik-samostalnik* v *rodilniku* in *pridevnik-samostalnik* mogoče razbrati, da se skici precej razlikujeta – s črnimi kvadrati so

<sup>8</sup> V novem referenčnem korpusu slovenščine, ki nastaja v okviru projekta Sporazumevanje v slovenskem jeziku, so besedilna razmerja zastavljena drugače, predvsem je načrtovan manjši obseg časopisne komponente (Logar Berginc, Šuster 2009).



**vprašanje** Fida PLUS 620m freq = 331775

a modifier	123348 2.2	post glede-d	1006 25.6	post o-d	10554 9.9
<input type="checkbox"/> nerešen	1554 72.33	<input type="checkbox"/> kredit	43 35.99	<input type="checkbox"/> smiselnost	155 55.67
<input type="checkbox"/> referenčni	1737 70.62	<input type="checkbox"/> usposabljanje	17 23.75	<input type="checkbox"/> veroizpoved	134 54.86
<input type="checkbox"/> pogost	3672 70.23	<input type="checkbox"/> meja	19 16.37	<input type="checkbox"/> smisel	148 33.45
<input type="checkbox"/> odprt	5049 69.51	<input type="checkbox"/> financiranje	9 14.48	<input type="checkbox"/> umestnost	13 29.86
<input type="checkbox"/> zastavljen	1708 69.32	<input type="checkbox"/> zdravje	10 13.83	<input type="checkbox"/> primernost	32 29.35
<input type="checkbox"/> mandaten	827 66.59	<input type="checkbox"/> pravilo	9 12.42	<input type="checkbox"/> upravičenost	36 29.08
<input type="checkbox"/> nagrađen	2464 64.56		>>	<input type="checkbox"/> dopustnost	15 28.98

FPsplet.internet Jožef Kvas Odgovor župana na svetrniško vprašanje glede stanovanjskih kreditov Ob dnevu samostojnosti

FPsplet.internet Jožef Kvas Odgovor župana na svetrniško vprašanje glede stanovanjskih kreditov Ob dnevu samostojnosti

FPsplet.internet Jožef Kvas Odgovor župana na svetrniško vprašanje glede stanovanjskih kreditov Ob dnevu samostojnosti

FPsplet.internet Jožef Kvas Odgovor župana na svetrniško vprašanje glede stanovanjskih kreditov Ob dnevu samostojnosti

FPsplet.internet Jožef Kvas Odgovor župana na svetrniško vprašanje glede stanovanjskih kreditov Ob dnevu samostojnosti

FPsplet.internet Jožef Kvas Odgovor župana na svetrniško vprašanje glede stanovanjskih kreditov Ob dnevu samostojnosti

FPsplet.internet Jožef Kvas Odgovor župana na svetrniško vprašanje glede stanovanjskih kreditov Ob dnevu samostojnosti

FPsplet.internet Jožef Kvas Odgovor župana na svetrniško vprašanje glede stanovanjskih kreditov Ob dnevu samostojnosti

FPsplet.internet Jožef Kvas Odgovor župana na svetrniško vprašanje glede stanovanjskih kreditov Ob dnevu samostojnosti

Slika 8: Prikaz kolokatorja in konkordanc iz istovrstnih virov.

**obdobje** Fida PLUS 620m freq = 159035

Izračun absolutne pogostosti		Izračun statistične izpostavljenosti	
a modifier	100616 3.1	a modifier	100616 3.1
<input type="checkbox"/> prehodni	5810 92.95	<input type="checkbox"/> prehodni	5810 92.95
<input type="checkbox"/> zadnji	5793 51.77	<input type="checkbox"/> sušen	1565 77.49
<input type="checkbox"/> dolg	5269 58.19	<input type="checkbox"/> mandaten	1029 72.98
<input checked="" type="checkbox"/> nov	2661 24.43	<input type="checkbox"/> pripravljen	1398 69.38
<input type="checkbox"/> časoven	2526 68.96	<input type="checkbox"/> časoven	2526 68.96
<input type="checkbox"/> lanski	2493 53.81	<input checked="" type="checkbox"/> obračunski	515 65.96
<input type="checkbox"/> življenjski	2428 55.36	<input checked="" type="checkbox"/> petleten	690 58.21
<input checked="" type="checkbox"/> določen	2389 50.37	<input type="checkbox"/> dolg	5269 58.19
<input checked="" type="checkbox"/> različen	2129 37.99	<input checked="" type="checkbox"/> štirileten	672 57.43
<input checked="" type="checkbox"/> naslednji	1979 42.99	<input type="checkbox"/> življenjski	2428 55.36
<input checked="" type="checkbox"/> kratek	1723 45.01	<input type="checkbox"/> lanski	2493 53.81
<input type="checkbox"/> sušen	1565 77.49	<input checked="" type="checkbox"/> primerjalni	498 52.44
<input type="checkbox"/> pripravljen	1398 69.38	<input checked="" type="checkbox"/> starosten	836 52.02
<input type="checkbox"/> pretekli	1361 50.74	<input type="checkbox"/> zadnji	5793 51.77
<input type="checkbox"/> mandaten	1029 72.98	<input checked="" type="checkbox"/> povojen	719 51.44
<input type="checkbox"/> prihodni	972 33.13	<input checked="" type="checkbox"/> desetleten	435 51.21
gen. l		gen. l	
a modifier	16561 2.1	a modifier	16561 2.1
<input checked="" type="checkbox"/> leto	669 24.46	<input type="checkbox"/> tranzicija	275 58.34
<input checked="" type="checkbox"/> življenje	391 28.68	<input type="checkbox"/> odraščanje	170 53.42
<input type="checkbox"/> tranzicija	275 58.34	<input type="checkbox"/> mirovanje	155 52.36
<input checked="" type="checkbox"/> razvoj	237 23.92	<input checked="" type="checkbox"/> januarjulija	24 47.87
<input checked="" type="checkbox"/> rast	209 32.81	<input type="checkbox"/> renesansa	115 47.66
<input checked="" type="checkbox"/> delovanje	174 25.13	<input checked="" type="checkbox"/> puberteta	92 47.65
<input type="checkbox"/> odraščanje	170 53.42	<input checked="" type="checkbox"/> barok	87 46.43
<input type="checkbox"/> mirovanje	155 52.36	<input checked="" type="checkbox"/> dinastija	107 43.78
<input type="checkbox"/> prilagajanje	145 40.31	<input checked="" type="checkbox"/> osamosvajanje	81 42.71
<input checked="" type="checkbox"/> priprava	145 23.11	<input checked="" type="checkbox"/> cvetenje	102 42.51
<input type="checkbox"/> vladavina	138 41.54	<input type="checkbox"/> vladavina	138 41.54
<input checked="" type="checkbox"/> oda	134 33.58	<input type="checkbox"/> razcvet	117 40.45
<input checked="" type="checkbox"/> mir	133 27.61	<input type="checkbox"/> prilagajanje	145 40.31
<input type="checkbox"/> razcvet	117 40.45	<input checked="" type="checkbox"/> razsvetljenje	46 40.13
<input type="checkbox"/> renesansa	115 47.66	<input checked="" type="checkbox"/> adolescenca	45 38.93
<input checked="" type="checkbox"/> zgodovina	115 21.08	<input checked="" type="checkbox"/> najstništvo	21 38.69

Slika 9: Primerjava rezultatov glede na dve vrsti statističnega izračuna za lemo obdobje v dveh slovnčnih razmerjih.

označeni kolokatorji, ki jih skica pri drugem statističnem izračunu ne pokaže, s črtami so povezani kolokatorji, pri katerih je porazdeljenost precej različna. Beseda, ki je v korpusu relativno redka, a se pogosto pojavlja prav z iskano lemo, bo imela zelo visoko vrednost statistične izpostavljenosti. Tako se na sliki v drugem stolpcu znajde kolokacija *obračunsko obdobje*, ki se v korpusu pojavi 515-krat, kar je premalo, da bi se uvrstila na vrh seznama zgolj po absolutni frekvenci (levi stolpec). Sklenemo lahko, da pri prvem izračunu najdemo pri vrhu bolj splošne kolokatorje (*zadnji, dolg, nov*), pri drugem pa kolokatorje, ki so sami zase redkejši v korpusu (*sušen, mandaten, pripravljen*) (gl. tudi Atkins, Rundell 2008: 110). Iz vsega tega je mogoče zaključiti, da je izbira statističnega izračuna in ostalih nastavitev, enako kot večina drugih stvari, prav tako odvisna od namena uporabe.

#### 4 Zaključek

V prispevku smo predstavili možnosti, ki jih ponuja uporaba besednih skic pri raziskovanju jezika, s poudarkom na leksikografskem delu. Splošni način njihove uporabe in zlasti njihova neprecenljiva vrednost pri leksikografskem delu sta bila v preteklosti že opisana, zato smo v tem prispevku predvsem opozorili na uporabniku manj očitne vidike in ozadje njihovega delovanja, ki pa pomembno vplivajo na dobljene rezultate. To na prvi stopnji obsega besedilno sestavo in označenost korpusa ter razpršenost virov, ki morajo biti dobro premišljeni, da odsevajo čim bolj realno podobo jezika. Dobljeni rezultati so nadalje odvisni od načina obdelave podatkov, kjer igra ključno vlogo slovnica besednih skic, in čisto na koncu tudi od uporabnika samega, ki na rezultate vpliva z izbiro nastavitvev orodja. Neizpodbitno dejstvo je namreč še vedno to, da kljub neizmernemu delu, ki ga z uporabo tega in podobnih orodij za nas opravi računalnik, za vnosom podatkov, potrebnih za nastanek besednih skic in jezikovnotehnoloških aplikacij nasploh, še

zmeraj stoji človek, kar pomeni, da to, kar surovim oziroma neobdelanim jezikovnim podatkom pripišemo, tudi (v nekoliko spremenjeni obliki) iz njih dobimo nazaj.

#### Literatura in spletni viri

- ATKINS, B. T. Sue, RUNDELL, Michael, 2008: *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- ERJAVEC, Tomaž, KREK, Simon, 2008: Oblikoskladenske specifikacije in označeni korpusi JOS. Erjavec, Tomaž, Gros, Jerneja (ur.): *Konferenca Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. 49–53.
- GANTAR, Polona idr., 2009: *Specifikacije za izdelavo leksikalne baze za slovenščino: opis analize referenčnega korpusa*: <<http://www.slovenscina.eu/Vsebine/Sl/Kazalniki/Kazalnik5/Kazalnik5.pdf>>. (Dostop 8. 7. 2009.)
- KILGARRIFF, Adam, RUNDELL, Michael, UÍ DHONNCHADHA, Elaine, 2006: Efficient corpus development for lexicography: building the New Corpus for Ireland. *Language Resources and Evaluation Journal* 40/2. 127–152.
- KREK, Simon, KILGARRIFF, Adam, 2006: Slovene word sketches. Erjavec, Tomaž, Gros, Jerneja (ur.): *Konferenca Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. 62–67.
- LOGAR BERGINČ, Nataša, ŠUSTER, Simon (v tisku): Gradnja novega korpusa slovenščine. *Jezik in slovnica* 54: 3–4.
- Označevalni sistem CLAWS*: <<http://ucrel.lancs.ac.uk/claws/>>. (Dostop 3. 7. 2009.)
- Sketch Engine beta*: <<http://beta.sketchengine.co.uk/>>. (Dostop 3. 7. 2009.)