Leksika na spletu (in kje jo iskati)

Mija Michelizza

Inštitut za slovenski jezik Frana Ramovša ZRC SAZU
Novi trg 4, 1000 Ljubljana
mmija@zrc-sazu.si

Povzetek

V prispevku primerjamo korpus besedil blogov in Wikipedije s korpusom Nova beseda (ki spletnih besedil ne vsebuje) in ugotavljamo, kateri tipi novejše leksike se pojavljajo na spletu oz. konkretno v obeh omenjenih zbirkah besedil. Uporabniki zaradi nepoznavanja jezikovnih pripomočkov na spletu pa tudi zaradi neažurnosti slovarjev in besedilnih korpusov za razna jezikovna in jezikoslovna vprašanja pogosto uporabljajo spletne iskalnike, zato v prispevku prikažemo, kaj vpliva na razvrstitev zadetkov v spletnih iskalnikih, ter se vprašamo, kaj nam lahko pove podatek o številu zadetkov. Na konkretnih primerih si ogledamo, na kakšne težave lahko s tovrstnim iskanjem naletimo, in opozarjamo, na kaj naj bodo (predvsem jezikoslovci v svojih raziskavah) pozorni.

Lexicon on the Web (and Where to Search for it)

In the article, we compare a corpus of blog and Wikipedia texts to a corpus Nova beseda (which does not contain web texts). Moreover, we determine which types of modern lexicon appear on the Web, namely in both text collections mentioned above. Users often use web search engines to answer language and linguistic questions either due to their lack of familiarity with online language tools or unupdated dictionaries and corpora. We therefore show what affects the ranking of hits in web serach engines, and explore what the information on number of hits can indicate. We use actual examples to examine what kind of problems can be encountered in searching with web search engines, and point out what (especially linguists in their research) to pay attention to.

1 Uvod

Mnoge jezikoslovne raziskave danes nastanejo s pomočjo besedilnih korpusov in trend, ki se kaže, je, da se v referenčnih korpusih besedilom pisnega in govornega prenosnika vse bolj dodaja tudi besedila elektronskega prenosnika. Največji korpus pisnih besedil za slovenščino Gigafida vključuje 16 % internetnih besedil, in sicer besedila novičarskih portalov in predstavitvene strani podjetij ter državnih, pedagoških, raziskovalnih, kulturnih ipd. ustanov. Pred tem je imel manjši delež besedil z interneta (1,24 %) tudi referenčni korpus FidaPLUS, ki je sedaj skoraj v celoti vključen v Gigafido. Poskusno je bilo v besedilni korpus FIDA vključenih enajst dokumentov (ali 20.999 pojavnic) spletnih besedil že leta 1998 (Logar Berginc et al., 2012). Vključevanje spletnih in internetnih besedil v besedilne korpuse je običajna praksa tudi v tujini (prim. Michelizza, 2011; Logar Berginc in Ljubešić, 2013), mnogi so tudi korpusi spletnih besedil (npr. korpusi WaC (ang. Web as Corpus), za slovenščino slWaC). Za jezikoslovne raziskave (kot tudi za običajne poizvedbe o jeziku) pa so pogosto v uporabi tudi spletni iskalniki.

2 Namen članka

V članku želimo s primerjavo korpusa besedil blogov in Wikipedije s korpusom Nova beseda (ki spletnih besedil ne vsebuje) ugotoviti, kateri tipi novejše leksike se pojavljajo na spletu oz. konkretno v obeh omenjenih skupinah besedil. Ker uporabniki za jezikovna in jezikoslovna iskanja pogosto uporabljajo spletne iskalnike, v prispevku prikažemo, kaj vpliva na razvrstitev zadetkov in kako nezanesljiva je informacija o številu zadetkov. S pomočjo konkretnih primerov večinoma iz primerjave korpusov dobljene leksike opozorimo na težave, s katerimi se lahko srečamo. Čeprav je iskanje s spletnimi iskalniki zelo enostavno in hitro, pa so informacije, ki jih na ta način dobimo, pogosto zelo omejene, česar se moramo uporabniki, zlasti pa jezikoslovci v svojih raziskavah, zavedati.

3 Leksika na spletu

Z namenom preveriti leksikalne razlike med obema zbirkama besedil, smo analizirali manjši korpus blogov in Wikipedije (skupaj okrog 500.000 pojavnic) ter ga primerjali s besedilnim korpusom Nova beseda, ki spletnih besedil ne vsebuje. V raziskavi Michelizza (2011) je bilo s primerjavo besed omenjenih korpusov dobljenih 1058 novih leksemov (545 iz blogov in 513 iz Wikipedije), torej takih, ki se v Novi besedi niso pojavili. Te lekseme smo razdelili v devet skupin[1] ter nekatere izmed njih v nadaljevanju uporabili tudi za ponazoritev težav pri iskanju na spletnih iskalnikih.

(1) Nov leksem za novo predmetnost: V podkorpusu besedil z Wikipedije takih primerov ni bilo, v podkorpusu blogov pa zasledimo le manjši del tovrstnih leksemov (npr. *flipanje*, *skrolanje*, *tvit*, *bail-out*).

(2) Novoopomenjeni leksem: Novoopomenjenih leksemov s primerjavo pojavnic dveh korpusov ni

[1] Delitev temelji na skupinah novejše slovenske leksike z vidika njenega generiranja, kot jih predstavlja Gložančev (2009), na koncu pa sta dodani še dve skupini.

mogoče iskati – vseeno je bilo med analizo najdenih nekaj primerov (*izgubljenček* 'zguba', *vsipan* 'pijan' ipd.).

(3) Nov leksem kot slovenska ustreznica za novejšo prevzeto: V analiziranem gradivu najdemo izraz *čivkač* v pomenu družabnega omrežja Twitter.

(4) Nov leksem kot posledica determinologizacije: Tovrstni novi izrazi so bili najdeni samo pri analizi besedil Wikipedije (npr. *melanopsin*, *dimetrodon*, *pleziozaver* ipd.).

(5) Novotvorjenke:

a) z lastnoimensko podstavo (npr. *janšegrad*, *murkovalec*, *barbivic*, *zdrnovškati se*, *nebondovsko*, *barbibutelj* ipd.),

b) z občnoimensko podstavo, ki je zaznamovana glede na kvalifikator v SSKJ (npr. *fejstnost*, *aufbiksati*, *šlatalec* ipd.),

c) novotvorjenke, ki izkazujejo zaznamovanost zaradi nenavadnih, manj običajnih obrazil (npr. *floskularjenje*, *premožnjakar* ipd.),

č) novotvorjenke, ki so zaznamovane tako zaradi podstave kot zaradi obrazila (npr. *futrač*, *glupača*, *frajerišenje*, *pifling* ipd.),

d) pomanjševalnice (npr. *blogec*, *zdrsek*, *lišajček* ipd.),

e) novotvorjenke, ki so poenobesedeni leksemi besednih zvez (iz polnopomenskih besed so nastale zloženke, iz predložnih zvez pa izpeljanke iz predložne zveze), nekatere izmed njih imajo frazeološki pomen (npr. *navrstnež*, *prvožogaški*, *pralnomožganski*, *polnoriten*, *vžepljivost* ipd.).

(6) (Fonetično)-oblikoslovna prevzetost tujih leksemov: Mnogi novi leksemi te skupine so posledica želje po stilnem učinkovanju v besedilih, nepoznavanja citatnega zapisa ali pa gre za primere, ko izraz v slovenščini (še) ne obstaja oz. ni vsesplošno uveljavljen (npr. *autlet*, *goodi*, *rumor*, *inboks* ipd.).

(7) Poobčnobesedenje lastnoimenskega izhodišča: Do poobčnobesedenja lastnoimenskih izhodišč prihaja pri imenih znamk in industrijskih izdelkov (npr. *skajpanje*, *uggice*, *pofotošopati*, *salomonke*, *martenske*) ali pa gre za poobčnobesedenje osebnih imen (npr. *potrč*, *golubić*, *anderlić* ipd.).

(8) Lastnoimenske novosti: V besedilih Wikipedije je delež novosti s področja lastnoimenskih poimenovanj precejšen. Navajamo nekaj primerov lastnoimenskih poimenovanj za zemljepisna imena: *Bernissart* (valonska občina v Belgiji), *Vajots Dzor* (armenska provinca); znane osebnosti: *Bellincione* (Dantejev oče), *Biondetti* (dirkač formule 1); viskije: *Balvenie*, *Bunnahabhain*, *Glengoyne* ipd. Hkrati pa je treba opozoriti, da so zaradi prevajanja geselskih člankov Wikipedije iz drugih jezikov ta imena pogosto zapisana citatno ali z mednarodno prečrkovalno različico, npr. za *Djalalabad* (ali *Jalalabad*) v slovenščini uporabljamo tudi zapis *Džalalabad*. Posebej je treba omeniti še skupino leksemov, ki jih uvrščamo med besedne igre in grafološke inovacije (npr. *pešhonda* (*...čaka me še cca. 20min. "pešhonde" do doma :()*, *Stovodiček* (*Tale arhitekt Stovodiček (Hundertwasser je Čeh drugače) mi je zelo všeč zaradi njegovega rekla, da če je lepo krivo, je tudi lepo.*) ipd.).

Analiza korpusa izbranih besedil z blogov in Wikipedije v primerjavi s korpusom Nova beseda pokaže, da se največ razlik na področju leksike (izvzemši stalne besedne zveze in frazeologijo) kaže pri blogih v obliki novotvorjenk, ki so pogosto priložnostnice,[2] pa tudi kot poobčnobesedenja lastnoimenskih izhodišč, s čimer se izkazuje želja po inovativnosti in ekspresivnosti ter stilnem učinkovanju v izražanju. Na Wikipediji je že zaradi same narave besedil izpostavljen vidik terminološkosti oz. determinologizacije ter lastnoimenskih novosti. Večine teh besed v obstoječih slovarjih ne bi našli (izjema je beseda *tvit*, ki jo najdemo v SNB in SSKJ[2]), če pa bi nas zanimalo kaj več o kaki od teh besed, bi si verjetno marsikdo pomagal tudi s spletnimi iskalniki.

4 Jezikovno in jezikoslovno iskanje s spletnimi iskalniki

Slovarji (vsaj za jezike z manjšim številom govorcev) niso vedno najbolj ažurni z objavo nove leksike, zato moramo uporabniki večkrat uporabiti druge pripomočke. Poleg besedilnih korpusov (ki pa imajo podobno težavo kot slovarji, saj je bila npr. zadnja posodobitev referenčnega korpusa za slovenščino leta 2012, zadnja besedila v njem pa so iz leta 2011) za jezikovne in jezikoslovne težave, analize in raziskave tako običajni uporabniki kot tudi jezikoslovci pogosto iščejo s pomočjo spletnih iskalnikov (Michelizza, 2011). Tu pa je nujno opozoriti na posebno previdnost pri interpretaciji rezultatov.

Interpretacija števila zadetkov in razvrstitev

Spletni iskalnik deluje na podlagi t. i. dvofaznega algoritma: ko vpišemo iskani izraz,

[2] V tej raziskavi dobljeno leksiko smo iskali še s pomočjo spletnih iskalnikov in mnoge izmed novotvorjenk se v spletnem iskalniku Google pojavijo samo enkrat (gre za ista besedila, ki so bila zajeta v korpus besedil blogov in Wikipedije).

najprej pajki poiščejo vse spletne strani, ki ta izraz vsebujejo, v drugi fazi pa program te zadetke razvrsti (Oblak in Petrič, 2005). Razvrstitev spletnih iskalnikov ni naključna in je odvisna od različnih dejavnikov (Gatto, 2009; Lana, 2004): (1) priljubljenost strani (merjena s številom drugih strani, ki se povezujejo na to stran); (2) pojavnost iskane besede ali besedne zveze na strani (kolikokrat se pojavlja, kje se pojavlja: če je beseda v naslovu, podnaslovu, v oznakah (ang. *Tags*) ter v hiperpovezavah, bo imela stran višjo razvrstitev); (3) geografski izvor poizvedovanja (višje bodo razvrščene strani, ki so geografsko gledano bližje uporabniku); (4) komercialni vidik (sponzorirane spletne strani so vedno na vrhu oz. na drugače izpostavljenem mestu iskalnika); (5) omejen čas iskanja (par stotink ali tisočink sekunde za obdelavo; ko iskalniki dosežejo omejeni čas, se poizvedba zaključi in rezultati so poslani uporabniku). Gre zgolj za nekatere izmed dejavnikov, ki vplivajo na razvrstitev, saj so ti algoritmi bolj kot ne skrivnost. Popolno razkritje algoritmov pri enem iskalniku bi lahko povzročilo vzpon drugega. Pri Googlu je ena pomembnejših tehnologij za razvrščanje zadetkov PageRank,[3] ki ne prešteva neposrednih povezav, ampak povezavo s strani A na stran B interpretira kot glas strani A za stran B. Nato oceni pomembnost strani glede na število prejetih glasov.[4] Zaradi omejenega časa iskanja in zaradi vse večjega števila podatkov na spletu, je tudi vse več strani, ki jih spletni iskalniki ne prikažejo. Starejše, neažurirane strani izpadejo, vse pomembnejša za razvrstitev v spletnih iskalnikih je tudi prilagoditev spletnih strani za mobilne naprave. Kot bomo videli v nadaljevanju, lahko v spletnem iskalniku Google npr. leta 2015 za identično iskanje najdemo manj pojavitev kot leta 2011, zato se je pri analizi spletnih besedil nujno zavedati relativnosti frekvenčnih podatkov na spletnih iskalnikih.

Omenili smo že, da je razvrščanje zadetkov v spletnih iskalnikih pogosto sponzorirano. Pojavili so se ponudniki t. i. optimizacije spletnih strani, ki naročniku pomagajo, da se uvrsti čim višje pri razvrščanju zadetkov v spletnih iskalnikih. Znano pa je, da uporabniki pogosto po liniji najmanjšega odpora pogledajo le prvih nekaj zadetkov v iskalniku. Pri spletnem korpusu projekta CUCWeb npr. opozarjajo na tovrstni šum, ki ga je mogoče opaziti v spletnih iskalnikih, saj so nekateri uporabniki spleta razvili posebne programe, ki pretentajo iskalnike, da so nekatere spletne strani višje razvrščene pri rezultatih iskanja, kot si dejansko zaslužijo. Predvideva se, da 8 % vsega, kar najdejo spletni iskalniki, spada med tovrstni šum. Uporabniki naredijo osnovno spletno stran in zraven še veliko drugih strani, s katerih naredijo povezavo na osnovno stran, in si na ta način povišajo razvrstitev na iskalniku (Fetterly et al., 2004; v Boleda et al., 2006).[5]

24. oktobra 2009 je takratni predsednik vlade RS Borut Pahor v enem izmed javnih nastopov uporabil besedo *krucefiks* in logična posledica je bila, da se je ta oblika na spletu razširila, čeprav so novičarski spletni portali navajali v SSKJ in SP 2001 uslovarjeno različico *krucifiks*. Še v začetku leta 2011 je iskalnik *Google* z omejenim iskanjem na slovenščino našel približno 4.500 zadetkov za *krucifiks*, medtem ko se je *krucefiks* pojavil kar 34.700-krat. Iskalnik pa je ob rezultatih vseeno priporočal zapis *krucifiks*.

Slika 1: Iskanje besede *krucefiks* v spletnem iskalniku Google leta 2011.

Stanje v začetku leta 2015 je bilo diametralno nasprotno, saj smo v Googlu (ponovno z omejitvijo na slovenščino) našli 5.280 pojavitev *krucefiksa*, *krucifiks* pa je imel kar 369.000 pojavitev. Konec leta 2015 z enakim iskanjem spet dobimo drugačno sliko. *Krucifiks* ima 3.540 zadetkov, *krucefiks* pa 6.670, pri čemer nas pri iskanju *krucifiks* vpraša, če smo morda mislili *krucefiks*. Google pogosto menja iskalne algoritme in kot kaže, je tudi v vmesnem času med temi iskanji prišlo do večjih sprememb, za jezikoslovca, ki bi želel ugotoviti prevladujočo

[3] Poimenovana je po ustanovitelju Googla in izumitelju algoritmov Larryju Pageu.

[4] Pred nastankom Googla so iskalniki delovali s pomočjo preprostih algoritmov, ki so temeljili na ključnih besedah – to pa je bilo precej enostavno zlorabiti. Še posebej pornografska industrija je začela izkoriščati to pomanjkljivost iskalnikov. Pogoste iskane besede so skrili po vsej svoji strani, npr. v drobnem tisku na beli podlagi. Leta 1998 so bili rezultati iskanja za poizvedbo avtomobil na takrat priljubljenem spletnem iskalniku Lycos večinoma pornografske spletne strani (Battelle, 2010).

[5] Meja med tovrstnim legalnim in nelegalnim početjem je pogosto nejasna. V okviru optimizacije spletnih strani sta se uveljavili poimenovanji črni klobuki in beli klobuki (ang. *Black Hats* in *White Hats*). Gre za izraza, ki izhajata iz žargona Divjega zahoda oz. kavbojskih filmov, kjer so bele klobuke nosili dobri, črne pa slabi kavboji (Battelle, 2010). V sklop nedovoljenih postopkov optimizacije npr. štejemo, (1) ko je v kodo HTML dodano besedilo, ki ga zaznajo iskalniki, na zaslonu pa ni vidno; (2) ko obstajata različni strani za iskalnike in za uporabnike in (3) ko se za optimizacijo spletnih strani uporablja program, ki sam proizvaja vsebino, zanimivo za iskalnike (<http://www.seo-blog.com/hats.php>, 22. november 2010).

obliko v rabi na spletu, pa bi lahko bila informacija zavajujoča.[6]

	krucifiks	*krucefiks*
začetek leta 2011	4.500	34.700
začetek leta 2015	369.000	5.280
konec leta 2015	3.540	6.670

Tabela 1: Pojavitve za *krucifiks* in *krucefiks* v iskalniku Google.

V času nogometnega svetovnega prvenstva v Južnoafriški republiki leta 2010 se je veliko pisalo o navijaškem glasbilu, t. i. *vuvuzeli*.[7] Za primer smo vzeli iskanje v spletnem iskalniku Google (z omejitvijo na slovenščino), kjer smo izbrali časovno obdobje po meri. Število pojavitev besede v času svetovnega prvenstva, torej med 11. junijem in 11. julijem 2010, in v istem obdobju leto poprej (med 11. junijem in 11. julijem 2009) smo primerjali konec leta 2010 in konec leta 2015 (gl. tabelo 2). Tako pri iskanju leta 2010, kot tudi leta 2015 je frekvenca višja v mesecu, ko je potekalo svetovno prvenstvo, vendar glede na izrazito manjše število zadetkov v obeh izbranih časovnih obdobjih, sklepamo, da iskalnik v letu 2015 med prikazom zadetkov upošteva le še redke spletne strani iz leta 2010.

vuvuzela	11. 6.–11. 7. 2010	11. 6.–11. 7. 2009
konec leta 2010	21.300	301
konec leta 2015	45	3

Tabela 2: Pojavitve za *vuvuzelo* v iskalniku Google.

Nelematiziranost spletnih iskalnikov

Težave z nelematiziranostjo spletnih iskalnikov omenja že Kilgarriff (2006), konkretno pa jih pokažemo z iskanjem primerov novih leksemov, ki smo jih obravnavali v 3. poglavju. Če v iskalnik vpišemo[8] osnovno obliko nekaterih besed, ki smo jih našli v korpusu blogov in Wikipedije, spletni iskalnik Google ne najde zadetkov. Taki primeri so: *murkovalec* (v korpusu se pojavi v obliki *murkovalca*), *barbivic* (*barbivice*), *zdrnovškati se* (*se zdrnovška*), *barbibutelj* (*barbibutlja*), *premožnjakar* (*premožnjakarjev*), *vžepljivost* (*vžepljivostjo*) in *pofotošopati* (*bi pofotošopal*). Čeprav je iskalnik Google splošno razširjen in med uporabniki najbolj priljubljen tudi za jezikovne poizvedbe (Michelizza, 2011), smo preverili zgornje primere še v iskalniku Najdi.si, ki prav tako ne najde zadetkov za osnovne oblike zgoraj obravnavanih iskanih besed.

Zapis z veliko in malo začetnico

Iskalniki tudi ne ločujejo zadetkov, zapisanih z veliko in malo začetnico. V preteklosti je tako razlikovanje omogočal iskalnik AltaVista, ki se je kasneje združil z Yahoojem in zato izgubil prenekatere prednosti, ki jih je prej nudil jezikoslovcem: poleg že omenjenega razlikovanja male in velike začetnice še iskanje posebnih znakov, vseboval je tudi določeno jezikoslovno znanje, kot je npr. enačenje nemškega ß in ss. Kasneje se je iskanje združilo z oglaševanjem, kar je nedvomno vplivalo tudi na mnoge odločitve o možnostih iskanja (Fletcher, 2007). Če želimo v Googlu iskati prevzeta tuja leksema *goodi* in *rumor*, se med zadetki pojavi tako zapis z veliko kot z malo začetnico, ki pa jih uporabniki ne moremo ločiti in moramo zato pregledati vse zadetke oz. toliko, kolikor se nam zdi potrebno. Podoben primer so zgledi poobčnobesedenja osebnih imen (*potrč*, *golubić*, *anderlić*), ki jih je v spletnih iskalnikih, ki nimajo možnosti iskanja z ločevanjem male in velike začetnice, praktično nemogoče iskati.

Zapis skupaj, narazen in z vezajem

V podobno težavnem položaju se znajdemo, če želimo izvedeti, kako se piše določena beseda – skupaj, narazen ali z vezajem. V rezultatih iskanja dobimo zadetke, ki zanemarjajo presledek (npr. pri iskanju *peshonda* dobimo zadetke *peshonda*, *peš honda* in *peš-honda*). Tiste, ki so pisani skupaj, sicer lahko najdemo s pomočjo t. i. Boolovega operatorja. V iskalnik vpišemo poizvedbo »*peshonda -peš -honda*«, kar pomeni, da bo iskalnik izločil vse zadetke, ki vsebujejo *peš* in *honda*, torej bo poiskal samo tiste, zapisane skupaj.

Tujejezični elementi

Čeprav lahko v iskalniku Google nastavimo iskanje po slovenskih spletnih straneh in po spletnih straneh v slovenščini, pa so pri iskanju tujejezičnih leksemov na tak način precejšnje težave. Pri primeru *rumor* najde Google z omejitvijo na slovenščino 20.900 zadetkov, ki pa že na prvi strani niso vsi v slovenščini. Podobno je pri primeru *bail-out* (22.600 zadetkov), kjer se pojavi še težava zapisa skupaj, narazen in z vezajem. Če v iskalniku Google iščemo besedo *tvit*, sicer najde kar 30.900 rezultatov, vendar nas hkrati vpraša »*Ste morda mislili tweet*«, kar bi lahko kakega uporabnika usmerilo v tujejezični zapis. Podoben predlog dobimo, ko v iskalno okence vpišemo *martenske* (1.330 zadetkov). Google nam predlaga *martinske*

[6] Katera od obeh prevladuje, težko rečemo, v Gigafidi je npr. razmerje med njima *krucifiks* 129 : *krucefiks* 70, za leto 2010 *krucifiks* 49 : *krucefiks* 48.

[7] Avgusta 2010 se je beseda *vuvuzela* prvič pojavila v oxfordskem slovarju (angl. *Oxford Dictionary of English*), v slovenščini je bila prvič vključena v SNB, ki je izšel leta 2012, čez dve leti pa je bila dodana tudi v SSKJ[2].

[8] Vsa nadaljnja iskanja so bila izvedena konec leta 2015.

(10.500 zadetkov). Čeprav je poimenovanje nastalo iz znamke Dr. Martens, bi lahko iz rezultatov sklepali, da se je v slovenščini uveljavilo poimenovanje *martinske* in ne *martenske*. Vendar pa podrobnejši pregled pokaže, da so med zadetki pri iskanju *martinske* tudi spletne strani, ki omenjajo slovaško smučišče Martinské Hole, martinske jedi ('jedi, ki jih jemo za martinovo'), martinske peči,[9] Martinsko jamo ipd. Kateri izmed izrazov (*martinske* ali *martenske*) se je v slovenščini (bolj) uveljavil, bi bilo treba natančneje preučiti in verjetno bi za to potrebovali drugo orodje (uravnoteženi korpus). Na primeru iskanja *martinske*, smo videli, da je iskalnik zanemaril tudi zahtevo po iskanju brez diakritičnega znamenja in med zadetke uvrstil *Martinské Hole*. Konec leta 2010 in v začetku leta 2011 so bile pri Googlu podobne težave z iskanjem besed s šumevci oz. brez njih v slovenščini (npr. pri iskanju besednih oblik *mizi* in *miži*). Konec leta 2015 pa iskalnik v omenjenem primeru že razmeroma dobro ločuje med zadetki, ki so zapisani s šumevcem oz. brez njega.

Seveda pa so nam spletni iskalniki lahko v veliko pomoč tako pri iskanju osnovnih jezikovnih podatkov, kot tudi pri zahtevnejšem jezikoslovnem raziskovanju. Med zadetki (navadno precej visoko uvrščeni) so pogosto različni spletni slovarji, ki lahko uporabniku pomagajo posredno (prek iskalnika) pri razreševanju jezikovnih težav (Lorentzen in Theilgaard, 2012). Za primer vzemimo v skupino novoopomenjenih izrazov uvrščeni leksem *vsipan*, ki ga lahko v enakem pomenu kot v podkorpusu blogov najdemo tudi med množico izrazov za pomen 'pijanost' na Prostem slovarju žive slovenščine Razvezani jezik (<http://razvezanijezik.org/?page=pijanost>, 24. oktober 2015), do katerega nas usmeri prav spletni iskalnik. Iskalnik nas lahko privede tudi do informacije o zapisu v lastnoimenske novosti uvrščenega *Djalalabada*. Na drugi strani zadetkov lahko najdemo zapis *Džalalabad*, ki je v Slovarju slovenskih eksonimov na portalu Termania.

Navkljub enostavnosti uporabe in vsem drugim prednostim, ki jih spletni iskalniki omogočajo, pa je nujno, da se uporabniki, predvsem pa jezikoslovci pri svojem delu omejitev in možnosti teh pripomočkov zavedamo in jih upoštevamo pri interpretaciji informacij. Pri spremljanju novejše leksike na spletu bo oblikovanje (in sprotno dopolnjevanje) specializiranega korpusa spletnih besedil, ki se nam obeta v sklopu projekta Janes, več kot dobrodošlo, za potrebe uslovarjanja novih, pogosto čez noč uveljavljenih gesel (tudi na račun spletne rabe jezika) pa je lahko uporaben Sprotni slovar slovenskega jezika, ki nastaja v sklopu portala Fran <www.fran.si>. S pomočjo neuspešnih poizvedb po slovarjih na spletni strani Inštituta za slovenski jezik Frana Ramovša ZRC SAZU <http://bos.zrc-sazu.si>, vključuje besedje, ki v drugih splošnih slovarjih še ni vključeno. Gre za metodo »log-files«.[10]

5 Zaključek

S primerjavo besedilnega korpusa Nova beseda in korpusov besedil blogov in Wikipedije smo pokazali, da na ta način pridobljeno besedje s spleta predstavlja predvsem novotvorjenke, ki so pogosto priložnostnice in poobčnobesedenja lastnoimenskih izhodišč. Pogoste so še lastnoimenske novosti in determinologizirani leksemi. Za iskanje tudi takih leksemov pogosto uporabljamo spletne iskalnike, ki pa za jezikoslovca prinašajo zelo omejene, včasih tudi zavajajoče informacije, česar se moramo zavedati. Prav zaradi tega je nujno oblikovanje večjega korpusa spletnega jezika, novosti pa je treba v slovarski obliki tudi sproti beležiti.

6 Literatura

Battelle, John. 2010. Iskanje. Kako so Google in njegovi tekmeci na novo napisali pravila poslovanja. Ljubljana: Pasadena.

Boleda, Gemma et al. 2006. CUCWeb: a catalan corpus built from the Web. Proceedings of the 2nd International Workshop on Web as Corpus. Trento. 19–26.

Fletcher, William H. 2007. Concordancing the web: promise and problems, tools and techniques. Corpus Linguistics and the Web (ur. M. Hundt et al.). Amsterdam: Rodopi. 25–45.

Gatto, Maristella. 2009. From Body to Web. An Introduction to the Web as Corpus. Bari: Editori Laterza.

Gložančev, Alenka, Jakopin, Primož, Michelizza, Mija, Uršič, Lučka, Žele, Andreja. 2009. Novejša slovenska leksika (v povezavi s spletnimi jezikovnimi viri). Ljubljana: Založba ZRC, ZRC SAZU.

Kilgarriff, Adam. 2006. Googleology is bad science. Computational linguistics, Volume 1, number 1. 147–151.

Lana, Maurizio. 2004. Il testo nel computer. Dal web all' analisi dei testi. Torino: Bollati Boringhieri.

Lorentzen, Henrik, Theilgaard, Liisa. 2012. Online dictionaries – how do users find them and what do they do once they have? Proceedings of the 15th EURALEX International Congress. 7-11

[9] Pridevnik *martinski* najdemo že v SSKJ (*martínski* -a -o prid. (î) metal., v zvezah: martinski postopek *postopek za pridobivanje jekla v martinovki*; martinska peč *martinovka*; martinsko jeklo *jeklo, ki se pridobiva v martinovki*).

[10] Tema je bila 25. maja 2015 obravnavana na Lingvističnem krožku Filozofske fakultete v Ljubljani. Predstavili so jo Primož Jakopin, Helena Dobrovoljc in Aleksandra Bizjak Končar.

August 2012. Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo. 654–660.

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

Logar Berginc, Nataša, Ljubešić, Nikola. 2013. Gigafida in slWaC: Tematska primerjava. Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave, št. 1. 78–110.

Michelizza, Mija. 2011. Vloga in pomen spletnih besedil v slovenščini. Doktorska disertacija. Ljubljana: Univerza v Ljubljani, Filozofska fakulteta.

Oblak, Tanja, Petrič, Gregor. 2005. Splet kot medij in mediji na spletu. Ljubljana: Univerza v Ljubljani, Fakulteta za družbene vede.