Iskanja po Korpusu slovenskega jezika FIDA

Vojko Gorjanc[♣], Špela Vintar[♥]

[♣]Univerza v Ljubljani, Filozofska fakulteta
Oddelek za slovanske jezike in književnosti
Aškerčeva 2
SI-1000 Ljubljana
E-pošta: vojko.gorjanc@guest.arnes.si

[♥]Univerza v Ljubljani, Filozofska fakulteta
Oddelek za prevajanje in tolmačenje
Aškerčeva 2
SI-1000 Ljubljana
E-pošta: spela.vintar@guest.arnes.si

Povzetek

Članek predstavi nekatere možnosti iskanja po Korpusu slovenskega jezika FIDA, referenčnem korpusu za slovenščino (http://www.fida.net), in sicer tiste, ki jih omogoča spletni iskalnik ASP32. Ob posameznih zgledih iskanja pokaže na prednosti in nekatere slabosti obstoječih rešitev pri gradnji Korpusa slovenskega jezika FIDA in pri možnostih iskanj, ki jih ponuja Amebisov spletni iskalnik ASP32.

1. Uvod

Korpusi so danes nepogrešljiv vir jezikovnih podatkov. Na njihovi izrabi temelji dobršen del sodobnega jezikoslovja, omogočajo pa tudi razvoj računalniškega jezikoslovja in jezikovnih tehnologij. V leksikologiji in leksikografiji si danes jezikovnih raziskav in izdelave slovarskih del brez uporabe korpusov sploh ne predstavljamo več. Pozitivne tuje izkušnje in potreba po aktualnem gradivu je spodbudila tudi razvoj tovrstnih jezikovnih virov za slovenščino. Predvsem stomilijonski referenčni Korpus slovenskega jezika FIDA[1] je dobra osnova in spodbuda za raziskave sodobne slovenščine.

Korpusi so v jezikoslovne raziskave vnesli dinamiko metodoloških pristopov; velika količina raznovrstnih podatkov nas nenehno potiska v iskanje možnosti njihove obdelave in pridobivanje novih spoznanj o dejanski rabi jezika. Pričujoči članek iz palete možnosti predstavlja nekatere izmed tistih, ki jih omogoča spletni iskalnik ASP32.

2. Orodja za korpusno analizo

Korpusi predstavljajo strukturirane zaloge besedil, ki so sicer dragocen vir raznovrstnih jezikovnih podatkov, vendar so za njihovo analizo potrebna posebna orodja, ki nam omogočajo dostop do korpusnih podatkov, njihovo urejanje in hranjenje. Na tržišču obstaja kar nekaj programskih orodij, pogosto jim pravimo kar konkordančniki, ki omogočajo analizo jezikovnih podatkov, med najbolj znanimi omenimo Wordsmith (Scott 2000), Monoconc (Barlow 1999), CUE (Mason 1997) in Xkwic (Christ 1993). Najpopularnejša sta prav prva dva izmed naštetih; z možnostmi, ki jih ponujata, sta za uporabnika zelo prijazna programa za osnovno analizo besedilnih zbirk, pri čemer konkordančnik Wordsmith omogoča bistveno več obdelav rezultatov iskanj. Dostopnih orodij je danes vse več, tistim, ki delujejo v okolju DOS, pa se vse bolj pridružujejo programi za okensko okolje.

Večina referenčnih korpusov, predvsem komercialnih, pa danes ob dostopu do korpusa ponuja tudi orodje za njihovo analizo, pri BNC je to npr. SARA, ob dostopu do Češkega narodnega korpusa se srečamo s konkordančnikom GCQP, za analizo korpusa FIDA pa je podjetje Amebis razvilo internetsko različico programa ASP32 (Amebisovo skladišče podatkov). Vsi, ki uporabljamo elektronske slovarje DZS, program že poznamo, seveda pa je za analizo korpusa posebej prilagojen; poleg osnovnih iskanj omogoča tudi urejanje rezultatov in njihovo statistično obdelavo

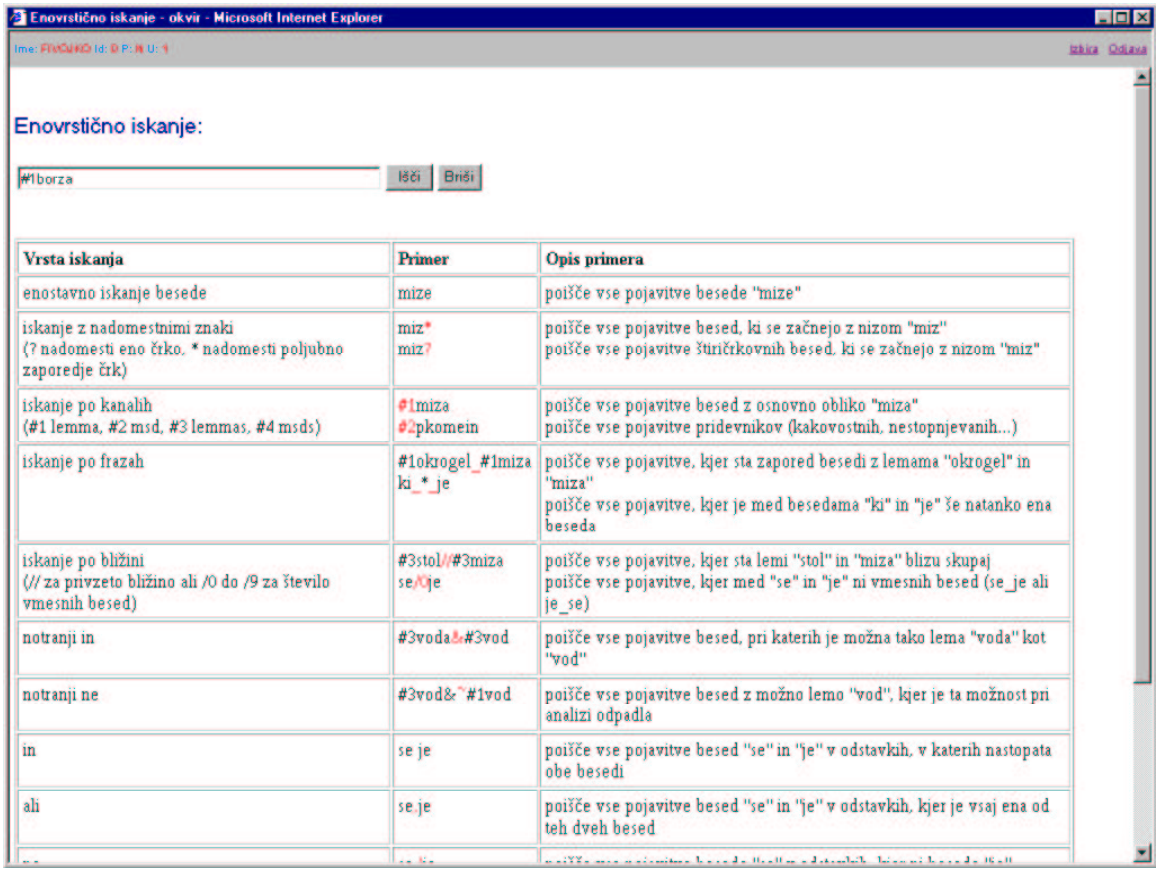
3. Iskanja z Amebisovim ASP32

Pri konkordančnkiu ASP32 v osnovi lahko izbiramo dva tipa iskanja, t. i. enostavno in razširjeno.

[1]V končno obliko Korpusa slovenskega jezika FIDA je iz zaloge besedil FIDA, v kateri je bilo zbranih 60.416 besedil v skupnem obsegu 161.794.040 besed, uvrščenih 29.177 besedil skupnega obsega 103.499.373 besed. Zaloga besedil se bo sproti dopolnjevala, tako da bo za nadaljevanje gradnje korpusa nenehno na voljo aktualno besedilno gradivo. Več o Korpusu slovenskega jezika FIDA v člankih Erjavec, Gorjanc, Stabej (1998) in Gorjanc (1999).

3.1. Enostavno iskanje

Tovrstne iskalne možnosti so vezane na iskanje po vseh besedilih korpusa FIDA. Pri tem lahko uporabimo ustaljene načine iskanja z nadomestnimi znaki, kot sta ? in *, klasičnim iskalnim možnostim pa je dodana serija novih. Iskanje po korpusu je seveda odvisno od samih korpusnih podatkov: nelematizirani in oblikoslovno neoznačeni korpusi omogočajo bistveno manj tipov iskanj oziroma zahtevajo pri morfološko bogatih jezikih veliko večjo iznajdljivost. Lematizacija in oblikoslovne oznake pa omogočajo tudi npr. iskanje po lemi (osnovni obliki besede, pripisani vsaki besedilni pojavitvi določene besede v korpusu) in oblikoslovnih oznakah, t. i. MSD-jih.



Slika 1: Maska enovrstičnega iskanja z ASP32 z razlago osnovnih vrst iskanja.

Lematizacija in oblikoslovno označevanje sta pri morfološko bogatih jezikih izredno zahtevna, zato je v prvi fazi pri tem precej težav. V korpusu FIDA je besedam velikokrat pripisana dvojna lema, zaradi izključno avtomatskega oblikoslovnega označevanja pa še večkrat variantni MSD-ji; šele disambiguacija bo tovrstne probleme v večji meri odpravila.

Delno problem dvojne leme pri iskanju rešuje *notranji ne*, s katerim izključimo iz nabora lem tisto, ki naj pri analizi odpade. Do tovrstnih težav prihaja v veliki meri npr. pri iz pridevnika izpeljanih prislovih, npr. *čudovit : čudovito* (čudovito vreme : čudovito poje). Če želimo v korpusu poiskati samo pridevnik, v iskalni pogoj vpišemo #3čudovit~#1čudovito; iskalnik bo tako izločil vse primere, kjer se v paru oz. nizu lem pojavi lema čudovito.

Sicer pa je pri besednih oblikah, ki imajo več možnih lem, disambiguacijo mogoče »zasilno« opraviti že s statističnimi metodami.

Avtomatsko oblikoslovno označevanje je za slovenščino še vedno problematično. V zadnjih letih je bilo narejenih nekaj pomembnih korakov, tako na področju preskušanja obstoječih programov za oblikoslovno označevanje kot pri razvoju lastnih tehnologij, vendar zaradi prekrivnosti besednih oblik različnih morfosintaktičnih kategorij to še zdaleč ni zadovoljivo rešeno (Džeroski, Erjavec 1998; Zupan 1999).

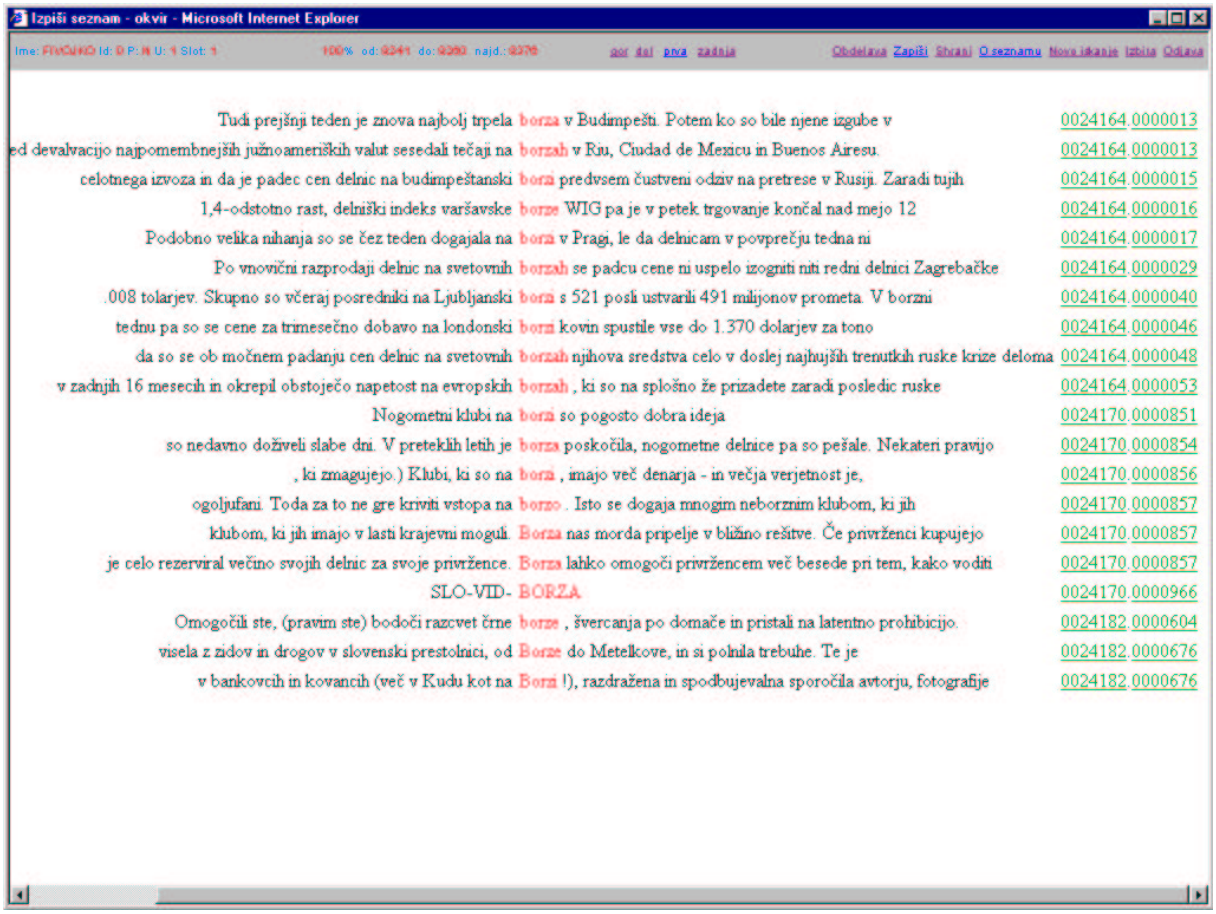
```
<p ID="F0020100.352"><s ID="F0020100.352.1">
<w lemma="na" msd="Dpem,Dpet" lemmas="na"
msds="Dpem,Dpet">Na</w>
<w lemma="ljubljanski"
msd="Pvomeid,Pvometdxn,Pvommi,Pvosdi,Pvosdt,Pvozdi,Pvozdt,Pvozed,
Pvozem" lemmas="ljubljanski"
msds="Pvomeid,Pvometdxn,Pvommi,Pvosdi,Pvosdt,Pvozdi,Pvozdt,Pvoze
d,Pvozem">Ljubljanski</w>
```

```
<w lemma="borza" msd="Sozdi,Sozdt,Sozed,Sozem" lemmas="borza"
msds="Sozdi,Sozdt,Sozed,Sozem">borzi</w>
<w lemma="vrednosten"
msd="Pkomdm,Pkomdr,Pkommm,Pkommr,Pkosdm,Pkosdr,Pkosmm,Pkos
mr,Pkozdm,Pkozdr,Pkozmm,Pkozmr" lemmas="vrednosten"
msds="Pkomdm,Pkomdr,Pkommm,Pkommr,Pkosdm,Pkosdr,Pkosmm,Pko
smr,Pkozdm,Pkozdr,Pkozmm,Pkozmr">vrednostnih</w>
<w lemma="papir" msd="Somdr,Sommr" lemmas="papir"
msds="Somdr,Sommr">papirjev</w>
<w lemma="biti jesti on" msd="Gvps3exxn Gpps3exxnxxxxxn
Zo3zerxxdxs" lemmas="biti jesti on" msds="Gvps3exxn
Gpps3exxnxxxxxn Zo3zerxxdxs">je</w>
<w lemma="vrednost" msd="Sozei,Sozet" lemmas="vrednost"
msds="Sozei,Sozet">vrednost</w>
<w lemma="indeks" msd="Somdr,Sommr" lemmas="indeks"
msds="Somdr,Sommr">indeksov</w>
<w>SBI</w>
<w lemma="in" msd="Vpe" lemmas="in" msds="Vpe">in</w>
<w>Bio</w>
<w lemma="zrasel zrasti" msd="Pkomdi,Pkomdt,Pkosmi,Pkosmt,Pkozei
Gpdrxdmtxxxxxxd,Gpdrxeztxxxxxxd,Gpdrxmstxxxxxxd" lemmas="zrasel
zrasti" msds="Pkomdi,Pkomdt,Pkosmi,Pkosmt,Pkozei
Gpdrxdmtxxxxxxd,Gpdrxeztxxxxxxd,Gpdrxmstxxxxxxd">zrasla</w>
<c type="PUN">,</c>
<w lemma="indeks" msd="Somei,Sometxxn" lemmas="indeks"
msds="Somei,Sometxxn">indeks</w>
<w>Pix</w>
<w lemma="pa" msd="L,Vpe" lemmas="pa" msds="L,Vpe">pa</w>
<w lemma="se" msd="Zpxxxxxxd,Zpxxxrxxdos,Zpxxxtxxdos"
lemmas="se" msds="Zpxxxxxxd,Zpxxxrxxdos,Zpxxxtxxdos">se</w>
<w lemma="biti jesti on" msd="Gvps3exxn Gpps3exxnxxxxxn
Zo3zerxxdxs" lemmas="biti jesti on" msds="Gvps3exxn
Gpps3exxnxxxxxn Zo3zerxxdxs">je</w>
<w lemma="znižati" msd="Gpdrxemtxxxxxxd" lemmas="znižati"
msds="Gpdrxemtxxxxxxd">znižal</w>
```

Zgled 1: Del standardno označenega besedila iz Korpusa slovenskega jezika FIDA

Kot je iz zgornjega zgleda razvidno, je pri avtomatskem oblikoslovnem označevanju variantnost MSD-jev še vedno tako velika, da bo za jezikoslovno izrabo tovrstnih podatkov v naslednjih korakih nujna disambiguacija. Iskanje po MSD-jih namreč zdaj daje tako razpršene rezultate, da je iz njih težko izbrati relevantne zglede.

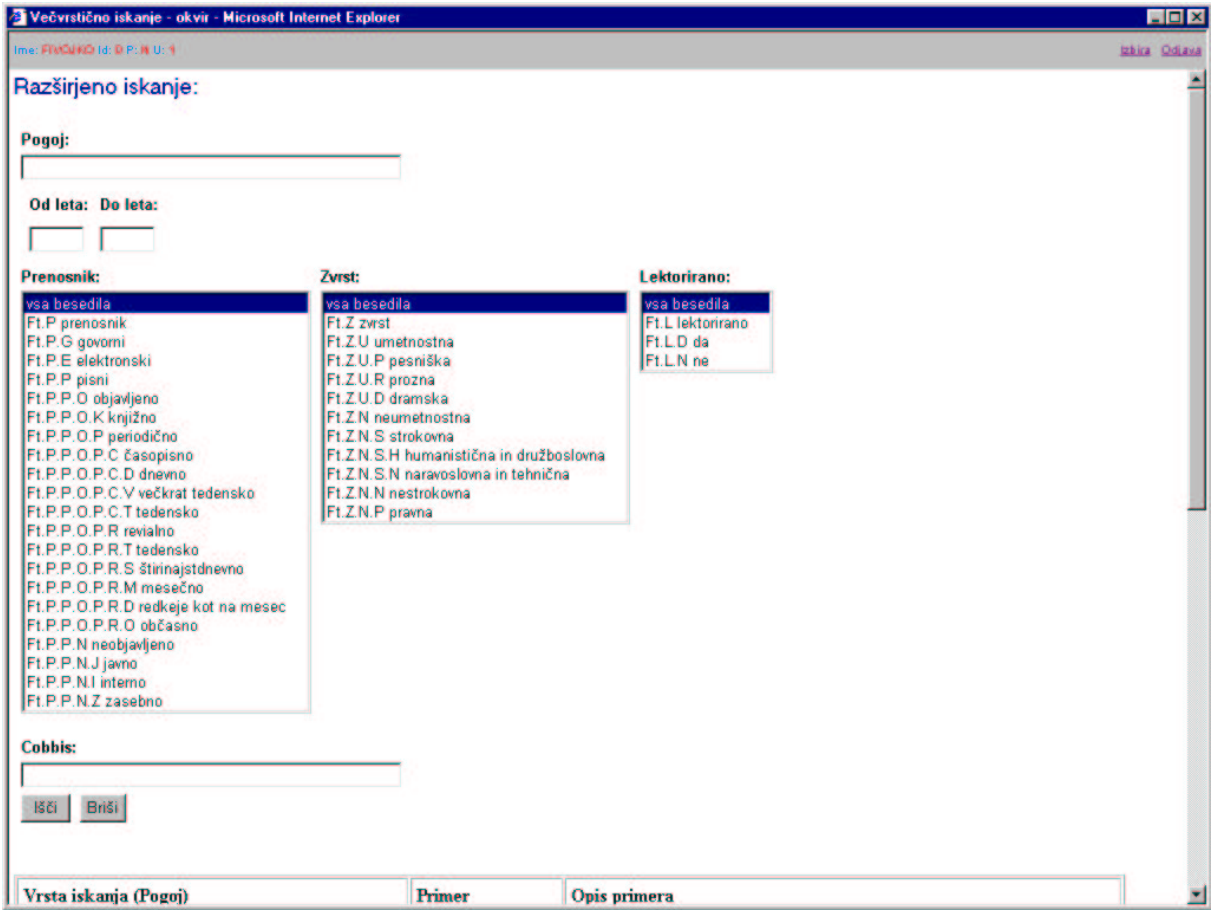
Glede na iskano besedo ali iskalni niz se nam kot rezultat iskanja izpišejo konkordance iskane besede oz. iskalnega niza, tj. seznama vseh pojavitev iskanega niza v korpusu s svojim minimalnim besedilnim okoljem. Konkordance so torej neke vrste osnovni tip korpusne analize; velikokrat je za nadaljnjo analizo v veliko pomoč že urejanje konkordanc po besedah pred ali za zadetkom; na tak način dobimo spisek konkordanc, strukturiranih po bližnjih besedah: ob poravnavi pred zadetkom se nam v slovenščini tako npr. združijo pojavitve z levim prilastkom ipd.



Slika 2: Zgled dela konkordanc iskalnega polja *#1borza*. Desno so šifre s povezavami (a) v levem stolpcu na podatke iz glave besedila zadetka, (b) v desnem stolpcu na razširjeno besedilno okolje obsega odstavka

3.2. Razširjeno iskanje

Pri enostavnem iskanju so bili vsi iskalni pogoji vezani na besedila korpusa FIDA; razširjeno iskanje pa nam omogoča tudi izbiro iskalnega pogoja glede na podatke v glavi posameznega besedila, in sicer iskanje po določenih besedilih glede na (a) taksonomije korpusa FIDA (zvrst, prenosnik, lektorirano) in (b) bibliografske podatke iz osrednjega slovenskega bibliotečnega sistema COBISS v formatu Comarc; posebej je omogočeno iskanje glede na (c) letnico izida besedila.



Slika 3: Maska razširjenega iskanja z ASP32

S pomočjo razširjenega iskanja lahko preverimo prisotnost določene besede[2] v različnih zvrsteh; tako lahko mimogrede preverimo, kako je z besedo *borza* v sodobni slovenščini.

umetnostno	strokovno	nestrokovno
0,08	6,62	93,30

Zgled 2: Odstotkovna zastopanost besede *borza* po besedilih različnih zvrsti

Še pred kratkim beseda, s katero se nismo srečevali prav pogosto, je postala del našega vsakdana, absolutno prevladujoča v nestrokovnih besedilih, torej tistih, s katerimi se povprečni uporabnik slovenščine danes največ srečuje. Znotraj nestrokovnih pa lahko nadalje preverimo, v katerem tipu prenosnika se z borzo danes največkrat srečujemo. Tudi ti podatki potrjujejo zastopanost besede v pisnih medijih, s katerimi se srečujemo največkrat.

knjiga	dnevni časopis	revija
0,70	62,44	36,86

Zgled 3 : Odstotkovna zastopanost besede *borza* po različnih pisnih prenosnikih

[2]Beseda je v korpusni analizi le izrazno definirana. Seveda to ni relevanten podatek o zastopanosti borze kot leksema.

3.3. Obdelava rezultatov

Pogostnost je sicer pomemben podatek, vendar nam ne daje prave slike o medsebojni odvisnosti jedrne besede in njenih kolokatov, torej kolikokrat se ob jedrni besedi pojavi določena beseda glede na njeno pogostnost v korpusu sploh. Zanima nas torej, kako se npr. beseda *borza* povezuje z drugimi besedami v večbesedni leksem. Konkordančnik ASP32 pri statistični analizi poleg abecednega seznama besed ob jedrni besedi ali nizu besed in seznama besed po pogostnosti omogoča tudi izpis rezultatov MI in MI[3].

Eden bolj uporabljanih statističnih podatkov pri analizi korpusov je rezultat MI (*mutual information*, vzajemnost).[3] Rezultat je vezan na področje informacijskih znanosti, izračunan je glede na par besed ali katerikoli par iskalnega niza v korpusu, poda pa verjetnost, da se bosta dve besedi pojavili v korpusu skupaj ali narazen oz. da se bosta elementa pojavila v določenem iskalnem polju (Biber, Conrad, Reppen 1998: 265). Visoka pogostnost seveda ne pomeni nujno tudi visoke vzajemnosti; visoko vrednost ima MI npr. v primeru, ko se določeni besedi pojavljata v korpusu izključno samo ena ob drugi, izračunan je po formuli

$$MI = \log_2 \frac{P(x'y')}{P(x')P(y')}$$

kjer P pomeni število pojavitev, x in y pa sta elementa korpusa.

1	=velikoprodajne[4]	1	1[5]
2	=frankurtska	1	1
3	=frankfurtstki	2	2
4	=koresopondenci	1	1
5	=ljubljnaska	1	1
6	=tinskoameriške	1	1
7	=njuyorške	1	1
8	=turističlni	1	1
9	=vsevropske	1	1
10	=svetvnih	1	1
11	=taipejske	1	1
12	=ljubljanjske	1	1
13	=produktna	1	1
14	=neinstitucionalizirane	1	1
15	=večerovo	57	95
16	=neobdavčevanje	1	2
17	=kopenhagenski	1	2
18	=večerova	25	63
19	=3sat	19	56
20	=ljubljnski	1	3

Zgled 4: Rezultati MI –1 jedrne besede *borza* v Korpusu slovenskega jezika FIDA

Prvih dvajset kolokatov besede *borza* glede na rezultat MI nam izpostavi predvsem v korpusu enkratno, v našem primeru so to velikokrat tipkarske napake, kar hkrati pomeni, da izpostavi slabosti korpusa oz. kaže na kvaliteto elektronski besedilnih virov. Rezultat MI pač poleg relevantnih rezultatov glede vzajemnosti izpostavi tudi vse enkratno; če je enkratno vezano na nekvaliteten vir, potem so rezultati, kot je to v našem primeru, za jezikoslovno obdelavo nerelevantni.

Uporaba rezultata MI je bila v zadnjem času večkrat kritizirana. Gre za dejstvo, da v veliko primerih prav zaradi neupoštevanja pogostnosti pojavitve korpusnega elementa v korpusu sploh ni ustrezen za merjenje vzajemne odvisnosti dveh korpusnih elementov, samo enkratna pojavnost korpusnih elementov enega ob drugem rezultat popači, tako da rezultat MI ni primeren, ko gre za nizko pogostnost pojavitve določenega elementa v koprusu (Manning, Schütze 1999).[6]

Če pri omenjenem rezultatu MI –1 ob jedrnem elementu *borza* upoštevamo pogostnost in se v konkretnem primeru odločimo za pogostnost pojavitve v korpusu 100- ali večkrat, so rezultati takoj relevantnejši:

tokijski	33	246
frankfurtski	26	401
newyorški	98	1728
varšavski	29	538
ljubljanski	1396	26910
budimpeštanski	13	257
londonski	124	2498
terminski	22	476
valuten[7]	11	384
=honkgkonški	4	147
blagoven	78	3531
luksemburški	5	234
vseevropski	3	159
donavski	4	216
zagrebški	45	2457
podjetniški	27	1658
=delove	2	146
milanski	10	790
sat	5	396
srednjeevropski	19	1681

Zgled 5: Seznam kolokatov –1 jedrne besede *borza* v korpusu FIDA, in sicer prvih dvajset glede na rezultat MI, ki se v celotnem korpusu pojavijo vsaj 100-krat

Seveda je tako početje dokaj zamudno in preveč subjektivno; zahteva namreč pri vsakem korpusnem elementu glede na njegovo pogostnost oceno relevantne pogostnosti, hkrati pa tudi ročno izbiranje relevantnih rezultatov.

[3]Poleg rezultata MI se variantno za iste namene uporablja tudi rezultat Z, za primerjavo dveh različnih elementov ob jedrnem je zanimiv rezultat T (Biber, Conrad, Reppen 1998).

[4]Enačaj pred besedo pomeni, da beseda v korpusu nima svoje leme, podatek se tako nanaša samo na to besedno obliko.

[5]V desnem stolpcu je podatek o absolunem številu pojavitev besede (ali besedne oblike pri nelematiziranih) v korpusu, v levem je število pojavitev iste besede ob jedrni besedi.

[6]Manning in Schütze (1999) pri angleško-francoskem paralelnim korpusu primerjata rezultate MI in testa Π^2 in pri slednjem ugotavljata boljše rezultate glede relevantnosti kolokacij.

[7]Tako vrstni pridevniki z obrazilom *-ni* kot kakovostni z obrazilom *-en* imajo v korpusu FIDA enotno lemo *-en*, saj trenutno ni mogoče avtomatično prepoznati pomena pridevnika.

Kot neke vrste korekcija rezultata MI se pojavlja rezultat MI^3, izračunan po naslednji formuli

$$MI^3 = \log_2 \frac{a^3}{(a+b)(a+c)}$$

kjer pomenijo *a* število sopojavitev določene sekvence, *b* število pojavitev prvega elementa brez drugega in *c* število pojavitev drugega elementa brez prvega.

1	ljubljanski	1396	26910
3	=večerovo	57	95
4	londonski	124	2498
5	newyorški	98	1728
6	posloven	210	22081
8	=večerova	25	63
9	svetoven	213	46449
11	tokijski	33	246
12	blagoven	78	3531
15	ljubljansko	60	2150
16	črn	109	26836
17	varšavski	29	538
18	frankfurtski	26	401
19	turističen	91	17722
20	zagrebški	45	2457
21	terminski	22	476
22	evropski	85	54888
23	podjetniški	27	1958
24	budimpeštanski	13	257
26	azijski	22	2372
27	pariški	22	2496
28	srednjeevropski	19	1681
30	valuten	11	384

Zgled 6: Pridevniki kot kolokati neposredno pred besedo *borza* glede na rezultat MI^3 v korpusu FIDA

Za besedo pred zadetkom tako dobimo za slovenščino relevantne podatke, saj so to večinoma vrstni pridevniki, ki z jedrno besedo tvorijo stalno besedno zvezo. Če izvzamemo izlastnoimenske vrstne pridevnike na *-ski* in jih primerjamo s podatki o tovrstnih stalnih besednih zvezah v Slovarju slovenskega knjižnega jezika (SSKJ) ne glede na to, kje v slovarju se besedna zveza pojavi, lahko ugotovimo, da se je glede na spremembe v slovenski družbi in posledično slovenskem jeziku pomensko polje zelo razširilo:

borza	
Korpus FIDA	*SSKJ*
poslovna	
blagovna	blagovna
črna	črna
turistična	
terminska	
podjetniška	
valutna	
	efektna
	produktna
+ izlastnoimenski vrstni pridevniki	

Zgled 7. Primerjava besednih zvez z jedrno besedo *borza* z vrstnim pridevnikom v SSKJ in najpogostnejših vrstnih pridevnikov ob besedi borza v korpusu FIDA

Avtomatsko zajemanje stalnih besednih zvez z desnim dopolnilom in vseh več kot dvobesednih stalnih besednih zvez je bistveno bolj zahtevno; desne predložne so avtomatično več kot dvobesedne in prav te predstavljajo že večji problem. Danes je pri zajemanju večbesednih stalnih zvez vse bolj aktualno kombiniranje statističnih rezultatov s skladenjskimi vzorci; pri tovrstnih tehnikah je rezultat avtomatskega zajetja stalnih besednih zvez v korpusu bistveno boljši (Dias 1999; Vintar 2000).

4. Dostopnost

Korpus FIDA s konkordančnikom ASP32 bo dostopen za delavce, sodelavce in študente projektnih partnerjev, tj. Filozofske fakultete Univerze v Ljubljani, Instituta Jožef Stefan, DZS d. d. in podjetja Amebis d. o. o. Dostop bo možen prek interneta: na spletni strani http://www.fida.net bodo vsi zainteresirani lahko zaprosili za individualno geslo. Vsi drugi, ki bi želeli uporabljati korpus v raziskovalne ali pedagoške namene, se bodo za dostop do korpusa lahko dogovarjali z lastnikoma avtorskih pravic korpusa, tj. založbo DZS d. d. in podjetjem Amebis d. o. o.; projektnima partnerjema, ki sta zagotovila celotno finančno konstrukcijo projekta.

5. Sklep

V članku smo prikazali nekatere možnosti uporabe Korpusa slovenskega jezika FIDA s pomočjo spletnega iskalnika ASP32. Opozorili smo na problematiko v zvezi z oblikoslovnim označevanjem slovenščine, ki razvoju tovrstnih jezikovnih tehnologij še vedno pomeni posebne vrste izziv.

Z veliko količino jezikovnih podatkov v korpusih za slovenski jezik se postopoma srečujemo šele v zadnjem času. Uporaba statističnih metod ter drugih metodoloških postopkov v jezikoslovni korpusni analizi je bolj ali manj na samem začetku; šele srečevanje s posameznimi problemi bo postopoma pripeljalo do uporabe ustreznih metodoloških pristopov in oblikovalo morebiti tudi nove, primerne jezikovnim virom za slovenščino.

6. Literatura

Aston, Guy, Burnard, Lou, 1998. The BNC Handbook. Exploring the British National Corpus wth SARA. Edinburgh University Press.

Barlow, Michael, 1999. MonoConc 1.5 and ParaConc. International Journal of Corpus Linguistics IV/1. 319-327.

Biber, Douglas, Conrad, Susan, Reppen, Randi, 1998. Corpus Linguistics. Investigating Language Structure in Use. Cambridge University Press.

Christ, Oliver. 1993. The Xkwic User Manual. Institut fuer maschinelle Sprachverarbeitung, Universitaet Stuttgart.

Dias, Gaël et al, 1999. Multilingual Aspects of Multiword Lexical Units. Language Technologies – Multilingual Aspects. Proceedings of the workshop within the framework of the 32th Annual Meeting of the Societa Linguistica Europea, 8-11 July 1999, Ljubljana. Ur. Špela Vintar. Filozofska fakulteta v Ljubljani, Odelelk za prevajanje in tolmačenje: 11–21.

Džeroski, Sašo, Erjavec, Tomaž, 1998. Inductive Learning of Multilingual Morphology. *Electrotechnical Review* 65 (6), 296–302.

Erjavec, Tomaž, Gorjanc, Vojko, Stabej, Marko, 1998. Korpus FIDA. Jezikovne tehnologija za slovenski jezik/Language Technologies for the Slovene Language. Ur. Tomaž Erjavec in Jerneja Gros. Institut Jožef Stefan Ljubljana: 124–127. URL: http://www.fida.net

Gorjanc, Vojko, 1999. Korpusi v jezikoslovju in korpus slovenskega jezika FIDA. 35. seminar slovenskega jezika, literature in kulture: 47–59. URL: http://www.fida.net

Kennedy, Graeme, 1998. An Introduction to Corpus Linguistic. Longman.

McEnery, Tony, Langé, Jean-Marc, Oakes, Michael, Véronis, Jean, 1997. The Exploitation of Multilingual Annotated Corpora for Term Extraction. Corpus Annotation. Linguistic Information from Computer Text Corpora. Longman: 220–230.

Manning, Christopher, D., Schütze, Hinrich, 1999. Foundations of Statistical Natural Language Processing. The MIT Press Cambridge MA.

Mason, Oliver, 1997. CUE - A Software System for Corpus Analysis. Proceedings of the Second TELRI Seminar, Kaunas, 85-90.

Scott, Mike, 2000. Wordsmith Tools. http://www.liv.ac.uk/~ms2928/

TACT – Text Analysis Computing Tools. URL: http://www.chass.utoronto.ca/cch/tact.html

Vintar, Špela, 2000. Računalniško podprto iskanje terminologije v slovensko-angleškem korpusu. Uporabno jezikoslovje 7-8, 156-169.

Zupan, Jure, 1999. Problemi in nekaj rešitev računalniških obdelav slovenskih besedil. Slavistična revija 47 (3), 277–296. http://www.ff.uni-lj.si/sr/index.html