Vojko Gorjanc,[1] Simon Krek[1] in Polona Gantar[2] UDK 811.163.6'374:004.6
[1]*Univerza v Ljubljani, Filozofska fakulteta*
[2]*ZRC SAZU, Inštitut za slovenski jezik Frana Ramovša, Ljubljana*
*vojko.gorjanc@guest.arnes.si*
*simon.krek@guest.arnes.si*
*apolonija.gantar@guest.arnes.si*

# SLOVENSKA LEKSIKALNA PODATKOVNA ZBIRKA

V članku predstavimo idejo oblikovanja slovenske leksikalne podatkovne zbirke, pripravljene na podlagi korpusne analize. Predstavljena izhodišča gradnje leksikalne podatkovne zbirke temeljijo na izkušnjah poskusne faze izdelave geslovnika za male dvojezične slovarje DZS s slovenščino kot izhodiščnim jezikom in korpusni analizi v okviru ciljnega raziskovalnega projekta »Zasnova na korpusu temelječih slovarskih in slovničnih opisov slovenskega jezika«. Oblikovanje leksikalne podatkovne zbirke utemeljujemo z dejstvom, da obstoječi slovarji slovenskega jezika ne predstavljajo sodobnega jezika, so tudi metodološko zastareli in nikoli niso dosledno izpeljali jezikovnega opisovanja brez predpisovanja. Predlagana zasnova leksikalne podatkovne zbirke omogoča gradnjo različnih tipov slovarjev; gre namreč za notranje hierarhiziran leksikalni opis sodobnega slovenskega jezika, kakršnega lahko pridobimo na podlagi referenčnega *Korpusa slovenskega jezika FIDA*.

## 1 Izhodišča[1]

Podatki o leksiki slovenskega jezika, kot jih prinašajo obstoječi enojezični slovarji, ne predstavljajo aktualnega stanja v slovenskem jeziku. V primeru *Slovarja slovenskega knjižnega jezika* (*SSKJ*) je že zaradi letnice izida (1970–1991) jasno, da ne more biti več relevanten vir podatkov o sodobnem slovenskem jeziku in normi sodobnega knjižnega jezika, pri *Slovenskem pravopisnem slovarju* (*SPS* 2001) pa analize kažejo, da je glede podatkov o aktualnem stanju v slovenščini celo še manj zanesljiv.[2]

Kljub dejstvu, da *SSKJ* ne more biti nesporen razsodnik o leksikalni normi sodobnega slovenskega jezika, tako v jezikoslovju kot tudi v slovenski družbi deluje kot nesporna avtoriteta. Širše družbeno je to razumljivo, saj je dejstvo, da se določen slovar tudi v okoljih, kjer je na voljo več slovarjev istega tipa, pojavlja v ednini, kot

[1] Del raziskav, ki so osnova za idejno zasnovo slovenske leksikalne podatkovne zbirke, je potekal v okviru ciljnega raziskovalnega projekta »Zasnova na korpusu temelječih slovarskih in slovničnih opisov slovenskega jezika« V6-0122, odgovorni nosilec projekta doc. dr. Vojko Gorjanc.

[2] Prim. npr. drugo številko *Slavistične revije* 2003, ki je v celoti namenjena oceni *SPS* in prinaša tudi polemiko med kritiki *SPS* in njegovim glavnim urednikom Jožetom Toporišičem.

da gre za slovarski unikum (Béjoint 2000: 121–122). Slovar je tako avtoriteta za informacije o tem, kaj v jeziku obstaja s splošno formulo *Besede X ni v slovarju = Besede X ni v jeziku* (Algeo 1990: 32). Ker pa se slovar ob tem v družbi pogosto doživlja, kot da slovarske informacije ne podlegajo času, so večne in nespremenljive, posledično velja tudi obratna formula, četudi raba določenega leksikalnega elementa ne izkazuje več, se je njegov pomen bistveno spremenil ali v razmerju do drugega ne deluje več nevtralno (Béjoint 2000: 122). Nerazumljivo pa je dejstvo, da *SSKJ* v slovenističnem jezikoslovju prav tako pomeni nesporno avtoriteto tako glede norme knjižnega jezika kot slovarske metodologije, kot da se tako eno kot drugo v času od nastanka slovarja sploh ni spremenilo, kar kaže na temeljno nerazumevanje dinamike razvoja jezikovne norme na eni strani in neumeščenost slovenske leksikografije pri slovarski metodologiji v sodobne leksikografske tokove na drugi strani.

Obstoječi leksikalni opisi slovenskega jezika so oblikovani na podlagi jezikovnih podatkov, zbranih na klasičen način z ročnim izpisovanjem na kartotečne listke, ki so glede na sodobne opise, kakršne prinašajo korpusi, tako kvantitativno kot kvalitativno že zdavnaj preseženi, *SPS* je nastal celo brez načrtno zbrane gradivne zbirke, norma v njem pa je določena tudi na podlagi jezikovne intuicije ter ideološkega avtorskega intervencionizma. Izhodišče za sodobne leksikalne opise je analiza velike količine načrtno zbranega avtentičnega gradiva in empirična analiza dejanskih vzorcev jezikovne rabe; šele računalniška tehnologija in oblikovanje metod zbiranja ter gradnje korpusov sta omogočila pridobiti veliko količino relevantnih aktualnih jezikovnih podatkov. Tako jezikovni opisi, ki nastajajo na tej podlagi, temeljijo na empirični analizi zares velike količine avtentičnih načrtno zbranih besedil (Biber et. al. 1998: 5, 9–10). Vse to so značilnosti jezikovnih podatkov, ki jih starejšim klasično zbranim listovnim zbirkam jezikovnih podatkov ne moremo pripisati (Čermák 2002: 265). Bistveno novo kakovost pa daje jezikovnim podatkom tudi oblikovanje meril za zajem besedil v korpuse, ki temeljijo na analizi diskurzivnega prostora. Tako zbrani jezikovni podatki omogočajo v jeziku ločevanje med tipičnim in posebnim oz. individualnim, torej prepoznavanje osrednjih in obrobnih jezikovnih pojavov, hkrati pa tudi opazovanje njihove distribucije glede na posamezni tip besedila. Tovrstni podatki omogočajo res kvalitetno analizo kolokabilnih lastnosti posameznih jezikovnih enot, njihovo tipično ubesediljenje pa razkriva tudi tipične slovnične vzorce analizirane jezikovne enote.

Ker slovarji slovenskega jezika ob jezikovnem opisu nenehno jezik tudi uravnavajo, so vsaj v enem delu predpisovalni in intervencionistični. Tako stanje je v slovenskem prostoru skušal načrtno preseči *SSKJ*:

> Slovenci smo navajeni, morda bolj kakor drugi narodi, da zaradi narodnostne ogroženosti zelo pazimo, da se v knjižni jezik ne vnaša preveč tujega, oz. tega, česar ne izkazuje literarna tradicija. Zdaj bo v slovarju registriranega mnogo več: to, kar je bilo priznano kot dobro, manj dobro in tudi to, kar je veljalo za slabo. Hoteli smo prikazati knjižni jezik v najširšem pomenu besede: živ, poln, z dubletami, notranjimi nasprotji, vzporednimi istočasnimi normami, jezik sredi zagona in razvoja. /.../ Slovar bo registriral dejansko stanje v jeziku, torej osnove njegove norme, s kvalifikatorji in kvalifikatorskimi pojasnili pa bodo vstavljene v ta okvir posebnosti, dvojnosti in izjeme. (Suhadolnik 1968: 221.)

Osredotočenje na jezikovni opis naj bi presegalo vrednotenje besed »pavšalno po tem, ali so pravilne ali nepravilne«, novo izhodišče sodobnega jezikovnega opisa pa pripomoglo »k prenehanju preganjanja izoliranih jezikovnih napak in utrdilo zavest o normalnosti govorjenega oz. pisanega jezika povprečnega izobraženca, istočasno pa poglobilo resnično, široko in poglobljeno zanimanje za slovensko besedo v celoti« (Suhadolnik 1968: 221–222). Celostni opisi slovenskega jezika pa v resnici niso nikoli jasno razmejili jezikovnega opisovanja od predpisovanja in jezikovne intervencije.[3] Čeprav se zavedamo komplementarnosti opisovalnega in predpisovalnega načela v jezikoslovju (Crystal 1997: 2–3), pa je prav tako nesporna potreba po jasnem metodološkem ločevanju enega in drugega. Metodološko neločevanje namreč nemalokrat privede do kvazi jezikovnega opisa, prikritega (namernega ali nenamernega) predpisovanja, ponujenega jezikovni skupnosti kot jezikovni opis.

Namen leksikalne podatkovne zbirke je tako predvsem pridobiti podatke o realnem jeziku, torej aktualnem leksikalnem naboru v slovenščini, o pomenih leksikalnih enot in njihovem tipičnem ubesediljenju. Predlog za oblikovanje leksikalne podatkovne zbirke, ki ga predstavljamo, je nastal na podlagi izkušenj pri pridobivanju leksikalno relevantnih podatkov za slovenski del splošnih malih dvojezičnih slovarjev DZS s slovenščino kot izhodiščnim jezikom in korpusne analize v okviru ciljnega raziskovalnega projekta »Zasnova na korpusu temelječih slovarskih in slovničnih opisov slovenskega jezika«. V poskusni fazi so bile za te namene analizirane različnice črke *b* iz *Korpusa slovenskega jezika FIDA* (http://www.fida.net), za razrešitev nekaterih zahtevnejših vprašanj pa so bile dodatno še problemsko izbrane in analizirane pogostnejše korpusne različnice.[4] Leksikalna podatkovna zbirka ima torej namen popisati stanje v slovenskem jeziku izključno na podlagi podatkov iz referenčnega korpusa slovenskega jezika. Gradnja take zbirke je neodvisna od morebitnih kasnejših slovarskih realizacij, kjer je potrebno upoštevati še npr. tip slovarja, uporabnika, velikost itd.

## 2 Pridobivanje jezikovnih podatkov

Izhodišče za oblikovanje podatkovne zbirke so korpusni podatki. Pri pridobivanju podatkov se ne moremo izogniti veliki količini ročne analize, predvsem takrat, ko želimo pridobiti relevantne podatke o pomenski zgradbi posameznega leksikalnega elementa. Da pa bi delo leksikografom olajšali, so jim bili za vsak analiziran element na voljo že predhodno procesirani korpusni podatki, in sicer naključni izbor do 300 konkordanc za posamezno korpusno različnico in statistična analiza neposrednega ali definiranega besedilnega okolja.

---

[3] Kako močno je (bila) zakoreninjena predpisovalna in intervencionistična tradicija v slovenistiki pri *SSKJ*, lepo kažeta posebna normativna kvalifikatorja *nepravilno* in *neustaljeno*. Kljub programski usmeritvi v jezikovni opis in vzpostavitvi vrednotenja knjižnojezikovne norme s pomočjo širokega nabora raznorodnih kvalifikatorjev, je opis v drobnem, a nikakor ne nepomembnem segmentu, pristal tudi v intervencionističnih vodah.

[4] Za korpusno analizo je del procesiranih korpusnih podatkov pripravilo podjetje Amebis (http://www.amebis.si), tudi sicer projektni partner pri izdelavi referenčnega *Korpusa slovenskega jezika FIDA*. Podjetju Amebis se za pomoč lepo zahvaljujemo.

### 2.1 Lista besed

Lista besed, ki so osnova za analizo in potencialne iztočnice leksikalne podatkovne zbirke, je narejena na podlagi procesiranih podatkov celotnega korpusa *FIDA*. V poskusni fazi gradnje podatkovne zbirke je bilo glede na pogostnost v korpusu izločenih prvih 25.000 lem. Najpogostejša je lema *biti* (7.749.214), zadnja po pogostnosti pa lema *acetat* (108). Črka *b*, ki je bila izbrana za poskusno fazo, zajema 783 lem, od najpogostejše *biti* do *biskvit* (108). Natančno število iztočnic podatkovne zbirke s tem še ni določeno, saj je potrebno preveriti, ali ni lema mogoče precenjena. Ker je bil korpus *FIDA* avtomatsko lematiziran brez razdvoumljanja pri lemah s skupnimi oblikami, je korpusnim pojavnicam lahko pripisanih tudi več lem, kontekstno nedvoumno pa je potrebno šele ugotoviti.

```
<p ID="F0008210.1482"><s ID="F0008210.1482.1">
<w lemma="nabirati">Nabiramo</w>
<w lemma="zdravilen">zdravilne</w>
<w lemma="rastlina">rastline</w>
<c type="PUN">-</c>
<w lemma="hrastov">hrastovo</w>
<w lemma="in">in</w>
<w lemma="vrbov">vrbovo</w>
<w lemma="lubje">lubje</w>
<c type="PUN">,</c>
<w lemma="jeglič">jeglič</w>
<c type="PUN">,</c>
<w lemma="kopriva">koprivo</w>
<c type="PUN">,</c>
<w lemma="ljubica">ljubice</w>
<c type="PUN">,</c>
<w lemma="regrat">regrat</w>
<c type="PUN">,</c>
<w lemma="lapuh">lapuh</w>
<c type="PUN">,</c>
<w lemma="popek">popke</w>
<w lemma="breza">breze</w>
<w lemma="in">in</w>
<w lemma="topol">topola</w>
<c type="PUN">.</c></s></p>
```

**Zgled 1:** *Del nedvoumno lematiziranega besedila iz korpusa* FIDA.

```
/.../ <w lemma="za">Za</w>
<w lemma="priloga">prilogo</w>
<w lemma="h">h</w>
<w lemma="krompir">krompirju</w>
<w lemma="se">se</w>
<w lemma="lepo lep">lepo</w>
<w lemma="podaja podati">podajo</w>
<w lemma="kuhan kuhati">kuhani</w>
<w lemma="list lista">listi</w>
<w lemma="lapuh">lapuha</w> /.../
```

**Zgled 2:** *Del dvoumno lematiziranega besedila iz korpusa* FIDA.

Paziti je torej treba pri tistih iztočnicah, kjer je možna dvojna lema in je druga mož-na lema v korpusu enako pogostna ali pogostnejša. Lahko se zgodi tudi, da sta dve lemi v seznam potencialnih iztočnic prišli z združenimi močmi, saj sami ne bi presegli dogovorjene meje 108 pojavitev. Značilni primeri za to so npr.:

- izpeljava prislova iz pridevnika – v teh primerih gre pogosto za pridevnik v tožilniku, ki mu je pripisana lema prislova: *bakren* (709), *bakreno* (133); *beseden* (2102), *besedno* (358);
- dva možna samostalnika: *beril* (189), *berilo* (344);
- homografi: *bístro* (prislov), *bistró* (samostalnik).

Kot je običajno pri podatkih iz korpusa, prihaja tudi do korpusnega šuma – ponavljajočih se koščkov besedil, ki pridobijo na pogostnosti iz »neupravičenih« razlogov: naslovov rubrik v časopisih, televizijskih programov itd.; npr. zveza *žametna vrtnica* je v korpusu samo zaradi radijske oddaje, pri lemi *odmev* je npr. problematična tako dvojna lema *odmev* : *odmevati*, hkrati pa močan korpusni šum povzročata oddaji *Odmevi* (TV) in *Dogodki in odmevi* (RA).

Kadarkoli gre za iztočnico, ki nenavadno izstopa, mora leksikograf pomisliti na navedene razloge in preveriti, če jo upravičeno obdržimo v podatkovni zbirki. Preverjanje gre tudi v smeri ugotavljanja korpusne razpršenosti – pojavljanje leksikalnega elementa pri različnih avtorjih in v različnih tipih besedil, saj je treba izključiti morebitne leksikalne elemente, ki so značilnost npr. enega avtorja. Pri tem gre v osnovi za subjektivne odločitve, ki pa se jim glede na trenutno označenost korpusa *FIDA* ne moremo izogniti.

### 2.1.1 Zgradba iztočnic

Iztočnice v podatkovni zbirki so eno- ali večbesedne. Tipično so večbesedne iztočnice s povratnoosebnim glagolom, npr. *bati se*; *briti : briti se*, hkrati pa status večbesedne iztočnice pridobijo tudi samostalniške besedne zveze v primeru, če sta obe sestavini pomensko netransparentni in je zveza kot celota dovolj pogostna (z enako ali višjo pogostnostjo kot v primeru enobesednih iztočnic), npr. *bela knjiga*, *bela garda*, *bela pritlikavka*, *beli ovratnik*.

**železen**
vrata, ograja, palica

| **železni** | *pogostnost* | *število pomenov* | *samostojna iztočnica*[5] |
|---|---|---|---|
| zavesa | 315 | 2 | + |
| doba | 138 | 1 | + |
| repertoar | 82 | 1 | – |
| cesta | 41 | 1 | – |

[5] Status samostojne iztočnice je odvisen od obsega leksikalne zbirke. Načelno izhodišče je, da mora biti tudi zveza sama z enako ali višjo pogostnostjo kot posamezne za leksikalno zbirko analizirane leme. Ker gre pri poskusni fazi za zajem lem s pogostnostjo nad 108, je to tudi izhodišče za predstavitev zveze na ravni iztočnice.

| | | | |
|---|---|---|---|
| lady | 41 | 2 | – |
| konjiček | 7 | 1 | – |
| pljuča | 7 | 1 | – |
| srajca **idiom** | 33 | 1 | – |
| križec | 13 | 1 | – |
| kačica | 8 | 1 | – |

**Zgled 3:** *Prikaz kolokacij in stalnih besednih zvez s pridevnikoma* železen *in* železni.

Kot je razvidno iz zgornjega zgleda, pri pridevnikih razlikujemo lastnostne in vrstne. Na podlagi tega predvidevamo tri tipe pridevniških iztočnic:

(a) Če sta realni obe vrsti pridevnika – gre predvsem za razlikovanje med pridevniki na *-en* in *-ni* (Vidovič Muha 2000) – potem kot iztočnici v istem iztočničnem članku navedemo obe obliki, najprej predstavimo lastnostni pridevnik in kot podiztočnico v posebnem delu iztočničnega članka še vrstni pridevnik, npr. *bajen (zaslužek*, *vsota)/bajni (bitje)*, *bajesloven (zaslužek*, *bogastvo)/bajeslovni (podzemlje*, *motiv)*, *bakren (žica*, *pločevina)/bakreni (doba)*, *baročen (razkošnost)/baročni (doba*, *umetnost)*, *bel (barva*, *lisa)/beli (vino*, *moka)*, *briljanten (nastop*, *izvedba)/briljantni (prstan*, *ogrlica)*. Pogostne zveze z vrstnim pridevnikom so tipično obravnavane kot stalne besedne zveze, ki pa jih je mogoče nadalje obravnavati kot pomensko transparentne na ravni kolokacij, npr. *bela (rasa*, *sorta*, *priseljenec)*, ali kot pomensko netransparentne, npr. *bela hiša*, *bela tehnika*.

(b) Če je realen samo lastnostni ali samo vrstni pridevnik, posebnosti ni, npr. *balistični*, *bančni*, *bitni*; *banalen*, *bežen*, *bister*, *bistven*. Pri pridevnikih izberemo tip pridevniške iztočnice izključno glede na **pomen** in nikoli glede na obliko zapisa. To pomeni, da ima lahko tudi vrstni pridevnik (v določeni skladenjski vlogi, npr. za vezjo v vlogi povedkovnika) obliko na *-en*, npr. *barvni*, *bitni*: */.../ epilogu, ki ni barven, ampak črno-bel /.../; /.../ Pravi razlog, zakaj je Cardinal 31-biten in ne 32-biten, je /.../*. S tem ne izgubi statusa vrstnega pridevnika, se pa ta podatek vedno upošteva pri zajemu zgledov rabe.

(c) Dosledno ločevanje lastnostnega in vrstnega pridevnika privede do izpostavitve tudi tistih pridevnikov, kjer gre pri lastnostnem in vrstnem za dva popolnoma pomensko ločena pridevnika, tako v *SSKJ* kot tudi v *SPS* sta zaradi neločevanja med oblikama predstavljena kot homonima, npr. *bučen* 'zelo glasen' *(aplavz, navijanje)*; *bučni* 'o buči' *(olje, seme)*. V teh primerih pridevnika predstavimo kot dve iztočnici s svojima iztočničnima člankoma.

Kot samostojne iztočnice so v podatkovno zbirko lahko sprejete tudi besednovrstno med seboj povezane besede in besedne oblike, če so v korpusu dovolj pogosto izkazane, npr. izpridevniški samostalniki. V primerih, kjer je razmerje med (skladenjsko in pomensko) povezanima besedama mogoče vzpostaviti, je smiselna tudi predstavitev znotraj enega iztočničnega članka s podiztočnico, npr. *brezposelni* (pridevnik) – *brezposelni* (samostalnik).

Pri drugih besednih vrstah se v poskusni fazi projekta posebnosti niso pokazale.

## 2.2 Pomenska analiza

Za vsak element, ki bo postal iztočnica v podatkovni zbirki, se iz korpusa *FIDA* izpišejo konkordance. Ker pa jih je pri posameznih korpusnih različnicah lahko izjemno veliko, pri pogostnejših naredimo naključni filter, s pomočjo katerega število konkordanc zmanjšamo na 300. Izhajamo iz predpostavke, da bomo iz tako izločenega dela konkordanc lahko razbrali pomensko zgradbo leksikalnega elementa, ki ga analiziramo. Tako oblikovan konkordančni niz je osnova za določanje pomenov; pri tem si zaradi lažje analize konkordančni niz poravnavamo levo/desno, kar omogoča na podlagi tipičnih sopojavnic levo/desno lažje razbiranje pomenov.

| *sopojavnica* | *izhodišče analize* | *sopojavnica* | *pomen elementa v izhodišču analize* |
|---|---|---|---|
| belo-, rdeče-, modro- | črn | obleka, avto … | take barve |
| | črn | gradnja, borza | nezakonit |
| | črn | slutnja, misel | neprijeten |
| | črn | točka, petek | tragičen |
| | črn | lista, seznam | nedovoljen, nezaželen |

**Zgled 4:** *Sopojavnice analiziranega elementa in njegovi pomeni.*

V nadaljevanju se število pomenov, razbranih s pomočjo analize konkordanc, lahko tudi primerja s pomensko zgradbo posameznega elementa v drugih slovarjih, predvsem *SSKJ*, vendar je za končno določitev pomenov vedno relevanten le korpus, obstoječi slovarji so pri tem lahko le pomožno sredstvo. Tudi za razporeditev pomenov je vedno relevanten korpus, tako da pomene v podatkovni zbirki nizamo izključno glede na njihovo pogostnost v korpusu.

Pri posameznih **pomenih** je v leksikalni zbirki naveden **pomenski indikator**. Gre za čim krajši pomenski kazalec, ki ima namen le pomene ene iztočnice medsebojno ločiti. T. i. posrednih indikatorjev ne uporabljamo, npr. *glagolnik od*, tudi ne takih, ki govorijo le o besedotvorni možnosti besede; ne navajamo torej pretvorbe tipa *kdor, kar*, saj nič ne povedo o pomenu, ampak pomen le povežejo z drugo iztočnico. Če kratek pomenski indikator po leksikografovem mnenju ne zadostuje, se lahko v opombo zapiše daljša razlaga, pri prenesenih pomenih zadostuje že indikator *figurativno*.

Pri vsakem pomenu iztočnice vnesemo tudi **zglede rabe**, tj. zglede iz konkordanc, ki morajo biti »slovarski«, tj. kratki in čimbolj tipični. Pri izboru zgledov upoštevamo najbolj pogostne kolokacije in/ali najbolj pogostne skladenjske vzorce (ali vsaj del vzorca) analizirane besede, ki je zastopana v iztočnici. Zgledi morajo biti vedno izbrani tako, da kažejo pogosto rabo iztočnice v leksikalni zbirki, ne pa morebitnih frazeoloških enot, v katerih nastopa tudi iztočnica. Zgledi so prvotno namenjeni prikazu kolokabilnosti in skladenjskih vzorcev in ne prikazu družbenih razmer, zato morajo leksikografi pri zbiri paziti na nevtralnost zgledov, da se pri tem čimbolj izognejo ideološkosti skozi preferenčnost pri njihovi izbiri (Béjoint 2000, Gorjanc 2004).

### 2.3 Besedne zveze

Osnova za določanje kolokacij in frazeologije so podatki o besedilni okolici analiziranega elementa z izračuni vrednosti MI³, in sicer v okolici –1, +1, +/–4.[6] Ti podatki služijo kot osnovni namig o tipičnem ubesediljenju analiziranega leksikalnega elementa. Končne odločitve o tem pa se vedno sprejemajo na podlagi nadaljnjega dela s korpusom, predvsem pregledovanja konkordančnih nizov, zvez in njihove okolice. Statistične podatke vrednosti vzajemne povezanosti elementov korpusa MI³ kombiniramo s podatki o absolutni pogostnosti, predvsem zaradi funkcijskih besed, saj se predvsem podatki o predlogih, veznikih in členkih pri vzajemnih vrednostih izgubijo zaradi izjemno visoke pogostnosti v korpusu (Gorjanc in Krek 2001).

čakati - 24443 (čaka(8463), čakajo(3692), čakal(2668), čakali(2277), čakala(1903), čakati(1836), čakam(732), čakamo(648), čakalo(548), čakaj(429), čakata(369), čakale(244), čakaš(180), čakajte(173), čakate(156), čakava(50), čakat(45), čakajmo(16), čakajta(14)

| 1LF | 1LM | 1DF | 1DM | 4XF | 4XM |
|---|---|---|---|---|---|
| biti(2594) | komaj(568) | na(5428) | na(5428) | biti(11178) | na(8215) |
| on(2462) | nestrpno(208) | še(970) | še(970) | na(8215) | biti(11178) |
| in(1511) | potrpežljivo(165) | v(882) | v(882) | on(7560) | on(7560) |
| jaz(1105) | nestrpen(206) | biti(573) | jaz(279) | in(4662) | dati(4332) |
| jesti(1046) | jaz(1105) | on(365) | tudi(318) | dati(4332) | da(4296) |
| še(820) | potrpežljiv(164) | tudi(318) | težek(115) | da(4296) | še(2999) |
| komaj(568) | on(2462) | in(302) | le(191) | v(3960) | jaz(2681) |
| ki(557) | zaman(208) | jaz(279) | ti(198) | jesti(3383) | in(4662) |
| pa(472) | biti(2594) | ti(198) | biti(573) | še(2999) | komaj(640) |
| že(448) | in(1511) | le(191) | nanj(52) | jaz(2681) | v(3960) |
| ti(422) | še(820) | do(168) | pred(155) | ki(2433) | nestrpno(240) |
| ne(385) | dolgo(294) | nov(163) | nov(163) | pa(2099) | jesti(3383) |
| dolgo(294) | jesti(1046) | pred(155) | nova(134) | se(1730) | ki(2433) |
| vedno(291) | dolg(287) | sem(155) | sem(155) | ta(1678) | potrpežljivo(184) |
| dolg(287) | ti(422) | že(150) | do(168) | že(1245) | pa(2099) |
| treba(270) | že(448) | jesti(145) | vas(84) | ne(1198) | nestrpen(245) |
| ta(243) | treba(270) | ta(135) | že(150) | ti(1178) | ti(1178) |
| morati(232) | vedno(291) | nova(134) | on(365) | leto(963) | že(1245) |
| zdaj(231) | vas(231) | pa(118) | in(302) | kaj(960) | dolgo(517) |
| vas(231) | predolgo(57) | težek(115) | presenečenje(37) | za(913) | kaj(960) |
| sem(228) | ki(557) | velik(113) | name(29) | ko(862) | potrpežljiv(189) |
| dati(226) | zdaj(231) | kar(107) | naporen(30) | sem(791) | zaman(252) |
| nestrpno(208) | predolg(57) | več(103) | neljubo(9) | tudi(780) | dolg(667) |
| zaman(208) | pa(472) | sam(98) | križem(18) | ves(756) | ta(1678) |
| da(206) | ne(385) | samo(93) | samo(93) | let(698) | se(1730) |
| nestrpen(206) | morati(232) | pri(88) | zaman(27) | dolg(667) | ne(1198) |
| leto(193) | sem(228) | ves(84) | malo(62) | z(664) | vas(495) |
| sam(191) | morala(163) | vas(84) | več(103) | delo(663) | leto(963) |
| samo(181) | mirno(68) | dva(82) | kar(107) | komaj(640) | sem(791) |
| potrpežljivo(165) | samo(181) | z(71) | velik(113) | nov(588) | ko(862) |
| potrpežljiv(164) | sam(191) | ob(70) | nanjo(26) | deti(583) | čakati(380) |
| morala(163) | miren(69) | veliko(70) | dva(82) | dan(581) | zdaj(542) |

**Zgled 5:** *Statistični podatki o besedilnem okolju za lemo* čakati.

[6] Uporabljene so bile statistične vrednosti, ki jih omogoča spletni konkordančnik ASP32 pri *Korpusu slovenskega jezika FIDA*. Različne statistične analize korpusa, ki se jih najpogosteje uporablja v leksikografiji, so bile preizkušene za slovenski jeziki, na koncu pa izbrane tiste, ki dajejo najboljše rezultate (Gorjanc in Krek 2001).

### 2.3.1 Kolokacije

S svojim izrazito strukturno-pomenskim izhodiščem je bila leksika slovenskega jezika obravnavana predvsem z vidika jezikovnih poimenovalnih enot. Spoznanja o skladiščenju slovarja kot komunikacijskih delov jezika, ki niso le leksemi, ampak večje leksikalne enote, so tudi pri opazovanju in opisovanju jezika sprožila vprašanja o slovarju kot zelo različnih leksikalnih enotah (Hill 2000: 47, Lewis 2000: 8), hkrati pa je razvoj korpusnega jezikoslovja šele zares omogočil kvalitetno analizo pojavov kolokabilnosti, saj je šele velika količina jezkovnih podatkov in njihova avtomatska analiza omogočila pridobivanje relevantnih podatkov o oblikovni in pomenski povezovalni moči posameznih elementov (Sinclair 1991). Opazovanje in opisovanje kolokacij na ravni enega jezika namreč temelji na objektivno merljivem parametru, tj. pogostnosti sopojavljanja. Na podlagi podatkov o pogostnosti sopojavljanja lahko s pomočjo statističnih metod ugotovimo nize besed, ki se pogosteje kot z ostalimi besedami sopojavljajo v besedilih obravnavanega jezika, npr. *rdeč* (*luč, križ, karton*).[7]

Pri **kolokatorjih** v leksikalni podatkovni zbirki vedno navajamo vsaj dva; vnos določimo glede na podatke o vzajemnih vrednostih. Kolokator je lahko tudi lastno ime, vendar nikoli ne osebno. Navajamo tipične nize kolokatorjev za posamezne besedne vrste, kar pa ne pomeni, da v primeru, ko se pojavi drugačen korpusni vzorec, tega ne registriramo.

Pri samostalniku so tipično kolokatorji tako lahko
- pridevniki **[mlad, pozoren, nepoučen] bralec** – pri zgledu rabe v teh primerih skušamo zajeti tudi prislov kot modifikator celotnega dela, če se ta pokaže kot relevanten, npr. *[skrajno, povsem] brezupen (primer)*;
- samostalniki **bralka [revije, časopisa], boj z/s [konkurenco, tekmeci, rakom]**;
- glagoli **[kotirati, trgovati] na borzi**.

Pri pridevniku tipično
- prislovi **[neozdravljivo, duševno, smrtno, kronično] bolan** in
- samostalniki **bolan [otrok, mati, tkivo, pacient]**.

Pri prislovu tipično
- glagoli **boleče [občutiti, odjekniti, zarezati]**,
- pridevniki **bistveno [drugačen, zmanjšan]** in
- prislovi **bistveno [manj, bolj]**.

Pri glagolih kolokatorji zapolnjujejo vezljivostna mesta:

**[veter, burja] brije; bežati pred [vojno, nacizmom, Turki; resničnostjo]; gojiti [ljubezen, upanje, čustvo, zamero]; gojiti [na balkonu, v rastlinjaku, na prostem]** …

oz. glagol modificirajo:

**[panično, brezglavo, množično] bežati**.

---

[7] V slovenskem prostoru je bilo vprašanje kolokabilnosti v glavnem domena anglistike (Jurko 1997 in Gabrovšek 1998), manj tudi slovenistike v okviru frazeološkega razpravljanja (Kržišnik Kolšek 1987); v slovenskem prostoru je šele v zadnjem času postalo del širšega jezikoslovnega zanimanja, slovenističnega v večji meri šele s pojavom koprusov slovenskega jezika (Gantar 2004, Gorjanc in Jurko 2004, Perko 2004).

### 2.3.2 Frazeologija in skladenjski vzorci

Zveze v okviru posameznega pomena ločujemo glede na njihovo pomensko zgradbo in pogostnost, in sicer:

(a) besedne zveze z visoko pogostnostjo, kamor sodijo vsi ponavljajoči se koščki besedil, v katerih nastopa iztočnica kot jedro in jih ne pokrijemo z navajanjem kolokatorjev, npr.: *pahniti (koga) v brezno (česa), bahati se pred (kom)*, *(deskanje, brskanje, naročanje) po internetu*;

(b) pomensko netransparentne zveze, ki imajo lahko tudi nizko pogostnost. Pri vseh zvezah, ki imajo vsaj en element pomensko netransparenten (tj. klasičnih frazemih in idiomih), dodamo pomenski indikator, npr. *barva kože* 'rasna pripadnost'; *zgoraj brez* 'brez zgornjega dela oblačila (kopalk)'. Tako izhodišče omogoča, da v podatkovno zbirko vključujemo besedne zveze v širšem obsegu ne glede na klasično delitev na stalne in nestalne. S tem ko se ne osredotočamo le na stalne besedne zveze, evidentiramo v podatkovni zbirki tudi tipične skladenjske vzorce in njihovo vlogo v besedilu; tako predstavimo vse tiste elemente v korpusu, ki se pojavljajo kot ponavljajoči se korpusni vzorec, npr. *vse/čedalje/vedno bolj (zapleten, pereč; se povečevati)*; *neprimerno/precej bolj (škodovati) kot (koristiti)*; *(biti) bolj ali manj (jasno, uspešno; znan)*.

Namen podatkovne zbirke je pač predstaviti leksikalno zgradbo slovenščine, kot se pojavlja v korpusu. Kot je bilo že rečeno, pa puščamo odprto vprašanje, kako bi se pri realizaciji slovarja odločali o njihovi slovarski predstavitvi. Izbira osnovne enote slovarja je glede na moč besedne povezovalnosti in posledično tvorjenja bolj ali manj trdnih zgradb vedno dogovorne narave.

Kulturološko vezane besede in besedne zveze imajo obvezno opombo z razlago, npr. *bela garda – kolaboracionistična organizacija v Sloveniji med NOB*.

Vse zveze zapisujemo v slovarski obliki, zapolnitev potencialnih udeležencev pa navajamo v oklepaju, npr. *bati se (koga/česa) kot hudič križa*; *imeti (kaj) za bregom*. Na ta način navajamo tudi besedilno okolje zveze, kadar se to v korpusu izkaže kot tipično, npr. *(spraviti, spravljati, pognati, pripeljati) (koga/kaj) na beraško palico*; *(zabrusiti, povedati, vreči) (komu) (kaj) v brk*; *(človek, moški, dečko) na mestu*. Idiome navajamo na koncu iztočničnega članka; idiomu dodamo opombo z razlago pomena. V ta razdelek spadajo tiste zveze, katerih pomen je glede na pomen njihovih sestavnih delov netransparenten in jih glede na pomen iztočnice ni mogoče uvrstiti pod posamičen že obstoječi pomen. Posebno pozorni smo na variantnost; v podatkovno zbirko namreč vnašamo podatke o realnih pojavitvah in ne idealizirane ene oblike, ki iztrgana iz besedila uporabniku slovarja ne more ponuditi funkcionalne informacije. Preučevanje pojava v številnih dejanskih realizacijah, kot nam jih ponuja korpusno okolje, namreč omogoča izločitev najbolj tipičnih in obenem opustitev individualnih rab, ki so slovarsko manj zanimive.

(povedati, razglasiti) brez dlake na jeziku
(biti brez, ne imeti) dlake na jeziku
iskati dlako v jajcu
dlaka gre pokonci (komu)
volk dlako menja, nravi/narave/značaja pa ne/nikoli

**Zgled 6:** *Zajeti podatki o frazeologiji pri iztočnici* dlaka.

## 3 Format podatkovne zbirke

Vsak leksikograf se danes sooča z odločitvijo, v kakšnem računalniškem okolju in v kakšnem formatu bo nastajal njegov slovar, podobno pa velja tudi za leksikalno podatkovno zbirko, ki jo lahko razumemo tudi kot neke vrste slovar. Izhodišča so dokaj jasna: doseči je treba, da je vsebina čim bolj trajno hranljiva, uporabna v čim več različnih računalniških okoljih (programih, operacijskih sistemih) in da je zaradi močne strukturiranosti iztočničnega članka vedno omogočeno prepoznavanje posameznih njegovih delov (Krek 2003).

Zaradi splošne razširjenosti ter prednosti, ki jih prinaša, je bila odločitev za urejanje leksikalne podatkovne zbirke v računalniškem okolju, ki prepoznava in zna shraniti vsebino v formatu XML takorekoč na dlani. Format XML kot naslednik standarda za zapis besedil SGML (*Standard Generalized Markup Language*) ter njegove izvedenke za internet HTML (*Hypertext Markup Language*) izpolnjuje vse zgoraj naštete pogoje, skupaj z veliko razširjenostjo. Za urejanje besedila v tem formatu je bil izbran urejevalnik Corel Word Perfect, ki je bil v času priprav na poskusno fazo projekta ena boljših izbir zaradi lažjega prilagajanja slovarske ekipe na delo, ker ima urejevalnik veliko skupnih funkcij z najbolj razširjenimi klasičnimi urejevalniki besedil, predvsem pa zaradi standardnih funkcij, ki jih omogoča modul za format XML/SGML, kot so validacija zgradbe dokumenta in možnost več različnih izgledov dokumenta ob isti vsebini. V novejšem času je prišlo na tem področju do hitrega razvoja, zato bo v prihodnje specializiran urejevalnik za XML najbrž boljša izbira.

Pri vprašanju slovarskega urejevalnika in končnega formata slovarja je sicer potrebno ločiti med dvema zelo različnima segmentoma. Za leksikografa je pomembno, da ureja slovar oz. leksikalno zbirko v čim udobnejšem okolju, ki je prilagojeno njegovim potrebam pri samem procesu sestavljanja. Klasični urejevalniki besedil pa so za to delo nezadostni, ker je leksikalna zbirka med drugim tudi vrsta baze podatkov, ki je močno notranje strukturirana. Hkrati pa izkazuje lastnosti besedila, zato mora izbrana programska oprema upoštevati elemente splošnih urejevalnikov besedil ter programov za delo z bazami podatkov, z nekaterimi slovarskimi specifikami, kot so hiter dostop do zaključenih nizov (kvalifikatorji), vnaprej nastavljive pogoste sheme geselskih člankov, preverjanje predvidene strukture geselskega članka, hiter dostop do različnih delov zbirke, zahtevna iskanja po različnih kriterijih ipd. Leksikograf mora za delo poznati zasnovo slovarja ali zbirke ter delo z izbrano programsko opremo. Od tega sorazmerno neodvisna, vendar nujna pa je potreba, da izbrana programska oprema zna hraniti in izvoziti slovarske podatke v strukturiranem formatu XML. Šele to namreč omogoča izmenljivost podatkov in lahek prenos v druga računalniška okolja.

<GS>◊◊
<IZ>
<IS>blagoslov</IS>
<I>**blagoslôv**</I>
</IZ>
<ZG>
<BV>*sam.*</BV>
</ZG>
<SM>
<KV>*relig.*</KV>
<IN>**(prošnja *za božjo naklonjenost)***</IN>
<KO>*[papeški, apostolski]*</KO>
<RA>
<ZD>*nekaj tisoč ljudi je čakalo pred gradom na papeški blagoslov*</ZD>
<ZD>*Ob koncu ponižno prosim Vašo svetost za apostolski blagoslov za to škofijo, njenega ponižnega pastirja in romarje*</ZD>
<ZD>*Koledniki naj bi prinašali k hiši blagoslov za letino, zdravje in srečo ljudem ter živini*</ZD>
<ZD>*Božji angeli ga spremljajo na njegovih potih. Blagoslov je z njim*</ZD>
</RA>
<FR>
<ST>
<FI>
<F>**božji blagoslov**</F>
</FI>
<RA>
<ZD>*Z molitvijo kličemo Božji blagoslov na vse ljudi*</ZD>
<ZD>*Naj tudi mi povsod prinašamo božji blagoslov in radi pomagamo ljudem v stiskah*</ZD>
<ZD>*S tako bogato in jedrnato molitvijo hočemo priklicati božji blagoslov na vsa področja človeškega udejstvovanja*</ZD>
</RA>
</ST>
<ST>
<FI>
<F>**[prositi, želeti, biti deležen] božjega blagoslova**</F>
</FI>
<RA>
<ZD>*Ako spolnjujemo božjo voljo, smo deležni božjega blagoslova*</ZD>
<ZD>*vsem ljudem dobre volje želimo ob božičnih praznikih obilo Božjega blagoslova in srečno ter uspešno novo leto*</ZD>
<ZD>*pa je z bolečino v srcu še dolgo prosila Božjega blagoslova za svojega vnuka*</ZD>
</RA>
</ST>
</FR>
</SM>
<SM>
<IN>***(privolitev)***</IN>

<KO>*[državni, vladni, uradni, očetov]*</KO>
<RA>
<ZD>*pripravili so novo uredbo, po kateri bodo investitorji z državnim blagoslovom lahko nadaljevali pogubno poseganje v dragocene vodotoke*</ZD>
<ZD>*Za deset tolarjev pa so cestarji z vladnim blagoslovom podražili tudi smrtno nevarno gradbišče, imenovano Slovenika*</ZD>
<ZD>*po dolgih pogajanjih se je z očetovim blagoslovom in denarjem odpravila v Kalifornijo študirat književnost*</ZD>
<ZD>*darvinistični evolucijski teoriji je pred dvema letoma dal celo svoj uradni blagoslov*</ZD>
</RA>
<FR>
<ST>
<FI>
<F>**brez blagoslova**</F>
</FI>
<RA>
<ZD>*prav od ruskih potez bo v mnogočem odvisna usoda novih posojil, ki jih Moskva brez ameriškega blagoslova nikakor ne bo dobila*</ZD>
<ZD>*bi bil Nato pripravljen posredovati na Kosovu tudi brez blagoslova OZN?*</ZD>
</RA>
</ST>
<ST>
<FI>
<F>**dati blagoslov**</F>
</FI>
<RA>
<ZD>*Javnost je presenečena nad tem, kako se je patriarh pred kamerami državne tv klanjal Miloševiču in njegovi ženi, čeprav je komaj pred nekaj meseci dvakrat sprejel opozicijske voditelje in jim dal blagoslov za str-moglavljenje Miloševičevega režima*</ZD>
<ZD>*Pete Sampras pa je tik preden je odpotoval iz New Yorka dal blagoslov Leveringovi za McEnroejevo imenovanje*</ZD>
</RA>
</ST>
<ST>
<FI>
<F>**dobiti blagoslov**</F>
</FI>
<RA>
<ZD>*Ta prizor so morali posneti devetnajstkrat, preden je dobil blagoslov cenzorjev*</ZD>
<ZD>*Prodajalec da modem XY na atestiranje in dobi blagoslov*</ZD>
<ZD>*Šele ko so na različnih uradih preverili in ugotovili, da je pri nas mir, je le dobil blagoslov za odhod*</ZD>
</RA>
</ST>
</FR>
</SM>

<SM>
<IN>***(sreča, korist)***</IN>
<KO>*[pravi]*</KO>
<RA>
<ZD>*ker ste živahne in brezskrbne narave, ste za svoje domače pravi blagoslov*</ZD>
<ZD>*Toplina, ki jo izžarevate, je pravi blagoslov za tiste, ki nenehno tarnajo in vidijo vse črno*</ZD>
<ZD>*Spoznali boste, da je pravi blagoslov, če imate veliko znancev in prijateljev*</ZD>
<ZD>*O, zdravje, zdravje! Blagoslov bogatih! Bogastvo revnih!*</ZD>
<ZD>*Razglasitev za lepotno kraljico je bila zame hkrati blagoslov in prekletstvo*</ZD>
</RA>
</SM>
<SM>
<IN>***(obred)***</IN>
<KO>*[opraviti]*</KO>
<KO>*[konj, ognja cerkve, prostorov]*</KO>
<RA>
<ZD>*Na Gomilskem želijo, da bi blagoslov konj postal spet tradicija*</ZD>
<ZD>*Na veliko soboto je najpomembnejši blagoslov ognja, vode in jedi*</ZD>
<ZD>*Občinski praznik je bil priložnost za svečano otvoritev in blagoslov novih prostorov občine Šentjernej*</ZD>
<ZD>*Slovesna maša bo ob 11. uri, blagoslov pa dobri dve uri kasneje*</ZD>
<ZD>*Mašno daritev in blagoslov je opravil šenčurski župnik*</ZD>
<ZD>*V Tibetu je v navadi, da vernik po blagoslovu v znak hvaležnosti izroči prostovoljno daritev*</ZD>
<ZD>*vsak blagoslov je hvaljenje Boga in prošnja za doseženje njegovih darov*</ZD>
</RA>
<FR>
<ST>
<FI>
<KO>*[nesti, nositi]*</KO><F>**k blagoslovu**</F>
</FI>
<RA>
<ZD>*na Vipavskem so vsi člani družine nesli k blagoslovu vsak svojo olječno vejico*</ZD>
<ZD>*V vseh slovenskih pokrajinah nosijo k blagoslovu jajca, ki jih imenujejo tudi pisanice*</ZD>
<ZD>*že tisočletje so domorodci v cerkev nosili k blagoslovu poljsko cvetje in zelišča*</ZD>
</RA>
</ST>
</FR>
</SM>
</GS>

**Zgled 7:** *Iztočnica* blagoslov *v leksikalni podatkovni zbirki.*

## 5 Sklep

Predstavljena ideja oblikovanja slovenske leksikalne podatkovne zbirke temelji na izkušnjah pri analizi korpusa *FIDA* za pripravo splošnih malih dvojezičnih slovarjev DZS s slovenskim izhodiščem in metodologiji korpusne analize ter hranjenja korpusno induciranih podatkov v leksiklani podatkovni zbirki, oblikovani v okviru ciljnega raziskovalnega projekta »Zasnova na korpusu temelječih slovarskih in slovničnih opisov slovenskega jezika«. Oblikovanje take podatkovne zbirke utemeljujemo z dejstvom, da obstoječi slovarji slovenskega jezika ne predstavljajo realnega jezika danes, hkrati pa so tudi metodološko zastareli.

Predlog oblikovanja leksikalne podatkovne zbirke za slovenščino temelji na analizi realnega jezika, kot mu lahko sledimo s pomočjo referenčnega pisnega *Korpusa slovenskega jezika FIDA*. Nabor iztočnic je narejen na podlagi korpusne liste besed, ki je osnova za nadaljnjo analizo posameznih leksikalnih enot. Izhodiščno vodilo je v leksikalni zbirki prikazati aktualno stanje slovenščine na leksikalni ravni: obstoj leksikalnih enot, njihovo dejansko obliko in pomen ter tipično ubesediljenje. Poseben poudarek velja registraciji različnih vrst besedne povezovalnosti: kolokacije, skladenjski vzorci, pomensko netransparentne zveze in idiomi. Ob tem pa v podatkovni zbirki z zgledi rabe glede na dejansko življenje leksikalne enote v slovenščini nenehno opozarjamo na leksikalno variantnost.

Pri poskusni izdelavi posameznih iztočničnih člankov podatkovne zbirke so se testirale tudi različne metode korpusne analize za slovenščino. Za hranjenje podatkov se je oblikoval celovit sistem vključevanja leksikalno relevantnih podatkov v podatkovno zbirko, njihove hierarhiziranosti in medsebojne povezanosti. Vse to omogoča odločitev, da podatkovna zbirka nastaja v formatu XML/SGML, ki omogoča tudi trajno hranljivost, uporabnost v različnih okoljih, predvsem pa zaradi velike strukturiranosti podatkov prepoznavanje posameznih segmentov leksikalne zbirke.

Pri nadgrajevanju načel oblikovanja leksikalne podatkovne zbirke bodo v prihodnje uporabljeni novi podatki o slovenskem jeziku, pridobljeni iz korpusa *FidaPLUS* (http://www.fidaplus.net), ob tem pa se bodo preizkušala tudi nova orodja za korpusno analizo, predvsem orodje, ki ga uporabljajo pri analizi češkega in slovaškega korpusa, tj. konkordančnik *Bonitio*. Glede na razpoložljiva finančna sredstva za tovrstne projekte bodo preučene tudi možnosti uporabe komercialnih leksikografskih programov za hranjenje podatkov v podatkovni zbirki, ki so danes že zelo zmogljivi, a za manjše projekte v slovenskem prostoru zaenkrat pomenijo preveliko finančno breme.

### Korpusa

*Korpus slovenskega jezika FIDA*. URL: <http://www.fida.net>.
*Korpus FidaPLUS* (poskusna verzija). URL: <http://www.fidaplus.net>.

### Literatura

Algeo, John, 1990: Dictionaries as seen by the educated public in Great Britain and the USA. Hausmann, F. et al. (ur.): *Wörterbücher. Ein internationales Handbuch zur Lexikographie*. Berlin: de Gruyter. 28–34.

Béjoint, Henri, 2000: *Modern Lexicography. An Introduction*. Oxford: Oxford University Press.

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

Crystal, David, 1997: *The Cambridge Encyclopedia of Language*. 2nd edition. Cambridge: Cambridge University Press.

Čermák, František in Holub, Jan, 1982: *Syntagmatika a paradigmatika českého slova I. Valence a kolokabilita*. Praha: Statní pedagogické nakladatelství.

Čermák, František, Klímová, Jana, Pala, Karel in Petkevič, Vladimír, 2001: The Design of Czech Lexical Database. Rayson, P., Wilson, A., McEnery, T., Hardie, A. in Khoja, S. (ur.): *Proceedings of the Corpus Linguistics 2001 conference*. Lancaster: Lancaster University. 119–125.

Čermák, František, 2002: Today's corpus linguistics. Some open questions. *International journal of corpus linguistics* 7/2. 265–282.

Firth, John Rupert, 1951: Modes of Meaning. *Essays and Studies* 4. Ponatisnjeno v Firth, J. R., 1957: *Papers in Linguistics 1934–51*. London: Oxford University Press.

Fischer, Ute, 1994: Learning Words from Context and Dictionaries: An Experimental Comparison. *Applied Psycholinguistics* 15/4. 551–574.

Fontenelle, Thierry, 1997: *Turning a Bilingual Dictionary into a Lexical-Semantic Database*. Tübingen: Max Niemeyer Verlag.

Gabrovšek, Dušan, 1998: Coping with Stubborn Stains and Persistent Headaches – for What It's Worth: Word Combinability in Action. *Vestnik* 32/1–2. 111–154.

Gantar, Polona, 2003: Stalnost in spremenljivost frazema v slovarju. Vidovič Muha, Ada in Gajda, S. (ur.): *Współczesna polska i słoweńska sytuacja językowa/Sodobni jezikovni položaj na Poljskem in v Sloveniji*. Opole. Uniwersytet Opolski, Instytut Filologii Polskiej in Univerza v Ljubljani, Filozofska fakulteta. 209–224.

Gantar, Polona, 2004: *Frazem in njegovo besedilno okolje. Doktorska disertacija*. Ljubljana: Univerza v Ljubljani, Filozofska fakulteta.

Gorjanc, Vojko in Krek, Simon, 2001: A corpus-based dictionary database as the source for compiling Slovene-X dictionaries. *Proceedings of the COMPLEX 2001 6th Conference on Computational Lexicography and Corpus Research*. Birmingham. 41–47.

Gorjanc, Vojko in Žele, Andreja, 2002: Compound dictionary entries (the case of Slovene noun phrases). Braasch, A. in Povlsen, P. (ur.): *EURALEX 2002: proceedings of the Tenth EURALEX international congress, Copenhagen, Denmark, August 13–17, 2002*. Copenhagen: Center for Sprogteknologi. 607–614.

Gorjanc, Vojko, 2004: Politična korektnost in slovarski opisi slovenščine – zgolj modna muha? Stabej, Marko (ur.): *Moderno v slovenskem jeziku, literaturi in kulturi. 40. seminar slovenskega jezika, literature in kulture*. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete. 153–161.

Gorjanc, Vojko in Jurko, Primož, 2004: Kolokacije in učenje tujega jezika. *Jezik in slovstvo* 49/3–4. 49–62.

Hausmann, Franz Josef, 1989: Le dictionaire de collocations. Hausmann, F. J., Reichmann, O., Wiegand, H. E., Zgusta, L. (ur.): *Wörterbücher* (3 zvezki). Berlin: Walter de Gruyter. 1010–1019.

Hill, Jimmie, 2000: Revising priorities: From grammatical failure to collocational success. Lewis, Michael (ur.): *Teaching Collocation. Further Developments in the Lexical*. Hove: LTP. 47–69.

Jurko, Primož, 1997: Towards a cline of difficulty of lexical collocations: Slovene–English. *Vestnik* 31/1–2. 220–237.

Krek, Simon, 2003: Sodobna dvojezična leksikografija. *Jezik in slovstvo* 49/2. 3–16.

Kržišnik Kolšek, Erika, 1987: Prenovitev kot inovacijski postopek. *Slava*. 49–56.

Lewis, Morgan, 2000: There is nothing as practical as a good theory. Lewis, M. (ur.): *Teaching Collocation. Further Developments in the Lexical Approach*. Hove: LTP. 10–27.

Manning, Christoper in Schütze, Hinrich, 1999: *Foundations of Statistical Natural Language Processing*. Cambridge MA: The MIT Press.

Perko, Gregor, 2004: *Razločevanje prevodnih ustreznic v dvojezičnem uvezovalnem slovarju (predlogi za slovensko-francoski slovar). Doktorska disertacija*. Ljubljana: Univerza v Ljubljani, Filozofska fakulteta.

Sinclair, John, 1991: *Corpus*, *Concordance*, *Collocation*. Oxford: Oxford University Press.

Suhadolnik, Stane, 1968: Koncept novega slovarja slovenskega knjižnega jezika. *Jezik in slovstvo* 13/7. 219–224.

Vidovič Muha, Ada, 2000: *Slovensko leksikalno pomenoslovje. Govorica slovarja*. Ljubljana: Znanstveni inštitut Filozofske fakultete.