



UDK 811.163.6'373:81'374  
*Irena Stramljič Breznik*  
Pedagoška fakulteta v Mariboru

## KVANTITATIVNE LASTNOSTI SLOVENSKEGA TVORJENEGA BESEDJA V POSKUSNEM BESEDNODRUŽINSKEM SLOVARJU ZA ČRKO B

Prispevek predstavlja kvantitativne lastnosti v slovarju zajetega besedja, ki ga omogoča računalniški program SlovarRed 2.0, 2003 Tomaža Seliškara in Sekcije za terminološke slovarje Inštituta za slovenski jezik Frana Ramovša, ZRC SAZU. Tvorbene lastnosti slovenskega besedja izhajajo iz komaj petindvajsetine sicer izčrpno obdelanih besednih družin, zato jih le z določenimi pridržki lahko posplošimo. Prav zaradi tega pa se kaže smiselnost izdelave besednodružinskega slovarja v celoti, saj bi bilo mogoče na zastavljenih temeljih s timskim delom posamezne segmente obdelave izboljšati in končno dobiti dokaj realno in empirično preverljivo sliko o sestavu slovenskega leksikona, ki bi bila dobro izhodišče za kontrastivne jezikoslovne študije in različne leksikografske projekte.

The article presents the quantitative features of the lexicon included in the dictionary compiled with the help of the computer program SlovarRed 2.0, 2003 by Tomaž Seliškar and the Section for terminological dictionaries at the Fran Ramovš Slovene Language Institute at ZRC SAZU. The derivational features of the Slovene lexicon are established on the basis of only one twenty-fifth of (otherwise thoroughly analyzed) word families, therefore they can only be generalized with some reservation. For that reason it would be sensible to compile a complete word-family dictionary, as it would be possible on the given foundation and with team work to improve certain segments of the treatment and in the end acquire a fairly realistic and empirically verifiable picture of the structure of Slovene lexicon, which would be a good source for contrastive linguistic studies and various lexicographic projects.

**Gljučne besede:** besedotvorje, besedna družina, besednodružinski slovar

**Key words:** derivation, word family, word-family dictionary

### 1 Besedna družina in besednodružinski slovar

Besednodružinski slovar slovenskega jezika je zasnovan kot tip enojezičnega specialnega slovarja, ki prinaša podatke o tvorbeni povezanosti besed, kar pomeni, da so ob nemotivirani iztočnici v ustreznih tvorbenih povezavah prikazane vse tvorjenke s skupnim korenem, ki so dosledno členjene na podstavni in obrazilni del.

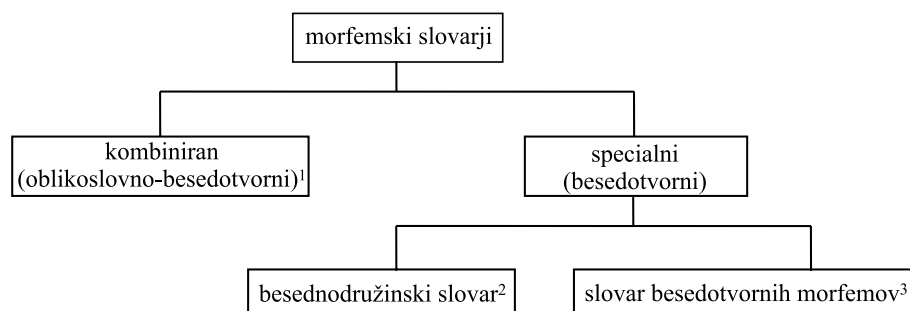
**1.1** Za tako povezane besede velja v slovenščini poimenovanje *besedna družina*, po definiciji pa jo sestavljajo besede istega korena (Toporišič 1992: 7; 1996: 89; 2000: 108, 116 in Vidovič Muha 2000: 107–108). Opozoriti velja, da je v ustrezni ruski, češki, slovaški in bolgarski literaturi mogoče najti poimenovanja *slovoobrazovatel'nyje gnezdo* (Tihonov 1985: 4), *slovotvorné hnízdo* (MČ 1 1986: 207–208), *slovotvorné hniezdo* (Mistrík 1993: 410) in *slovoobrazovatelno gnezdo* (Penčev 1999), medtem ko je v nemškem jezikovnem prostoru ustaljen izraz *die Wortfamilie*, uporabljen v naslovu besednodružinskega slovarja sodobnega nemškega jezika (Augst: 1998).

V slovenistiki je smiselno ostati pri izrazu *besedna družina*, saj najdemo izraz v slovniškem izrazju z začetka 20. stoletja pri Janku Bezjaku (1906: 194): »3. Učenec naj se vadi iz beril in sploh iz pouka znašati besede, med seboj po etimologiji sorodne, ter tako tvarjati *besedne družine*.« Prav tako tudi pri Petru Končniku (1913: 23): »Rastlike, ki se izobrazujejo iz istega korena, delajo *besedno družino ali skupino*.« V novjšem času pa je uporabljen v monografiji *Besedna družina iz korena \*god- v slovenskih jezikih* (Šivic Dular: 1999), kjer gre za besedne družine, zasnovane na diahronih etimoloških merilih.

**1.2** Besednodružinski slovar in slovar besedotvornih morfemov nista bila pogost cilj ne domače ne tuje leksikografske prakse. V vsej dolgi slovenski slovaropisni tradiciji je bil besednodružinski slovar le enkrat deloma uresničen, in to pri Urabanu Jarniku, ki je v naslovu slovar opredelil kot etimološki (Jarnik 1832), nekatere njegove prvine pa so se pojavljale še ob sestavi dvojezičnih slovarjev (Murko 1833), ki so slovensko besedje prikazovali v t. i. besedotvornih gnezdih.

Na tujem so se ti slovarji vsekakor pojavljali, daljšo tradicijo imajo besednodružinski, zametki katerih segajo nekako že v 17. stoletje (Augst 1990: 1145–1152), v sodobnosti pa so izdelani za ruščino (Tihonov 1985: 4), francoščino, nemščino (Augst: 1998), bolgarščino (Penčev 1999) in esperanto. Obstaja pa še primerjalni štirijezični besednodružinski slovar za ruščino, poljščino, češčino in srbohrvaščino (Herman 1975). Slovarji besedotvornih morfemov so kot posebni tipi slovarjev relativno mladi, saj so nastajali praktično v osemdesetih letih 20. stoletja (Link 1990: 1223–1230), in to prav tako le za štiri jezike: nemščino, angleščino, francoščino in ruščino.

**1.3** Besednodružinski slovar je tip morfemskega besedotvornega slovarja, ki jemlje za izhodišče besedno družino, znotraj katere so besede na različnih stopnjah tvorbe vselej prikazane kot sestav podstave in obrazila.



<sup>1</sup> Tipični zgled predstavlja slovaški morfemski slovar avtorjev Sokolova, Moško, Šimon in Benko (1999).

<sup>2</sup> Sestavljalec besednodružinskega slovarja za ruščino je Tihonov (1985), za nemščino Augst (1998).

<sup>3</sup> Kot primer navajamo slovar besedotvornih elementov nemškega jezika, ki ga je sestavila Stepanova (1979).

## 2 Besednodružinski slovar v slovenistiki

Poskusni besednodružinski slovar slovenskega jezika za črko B je rezultat večletnega usmerjenega raziskovalnega dela, katerega potek in delne rezultate prinaša poglavje Sistemskost leksike: Povezanost besed v besedni družini v zborniku *Besedoslovne lastnosti slovenskega jezika in narečij* (Sramljič Breznik 2004: 315–340). Slovar je sestavljen iz treh delov: (1) uvodno-teoretičnega, (2) besednodružinskega in (3) abecednega zaporedja vseh zajetih tvorjenk (213–318).

**2.1 Uvodno-teoretični del (7–67)** uporabnika seznanja s teoretičnimi izhodišči, obsegom in namenom slovarja, nato ga uvaja v zgradbo in uporabo slovarja tako, da predstavi temeljne prvine slovarske strukture, od slovarske enote – besedne družine – do njenih sestavin (iztočnica, tvorjenke) in še posebej splošna načela o notranji organiziranosti in hierarhiji tvorjenk glede na posamezno besedotvorno vrsto.

Temu sledi kratek pregled slovenske besedotvorne teorije s predstavitvijo temeljnih in najpogostejših tvorbenih vzorcev, po katerih se v slovenščini tvorijo posamezne besedne vrste. Tak zgoščen pregled je namenjen zlasti uporabniku, ki ga zanimajo tudi teoretični okviri tvorbe.

**2.2 Besednodružinski slovarski del (71–210)** je trenutno sestavljen za iztočnice na B in obsega 666 besednih družin, v katerih je razvrščenih 11 136 enot vključno z napovedniki. Slovar je pri vključevanju gradiva zavestno in hote izčrpen z namenom, da bi se zajela vsa tvorbena problematika in nakazale njene potencialne rešitve, ki bi lajšale nadaljnjo sestavo. Temeljni gradivni viri so bili: Slovar slovenskega knjižnega jezika (SSKJ), Besedišče slovenskega jezika (BSJ), Slovenski pravopis 2001 (SP) ter po referenčnem korpusu Fida paberkovalno zbrane nekatere besede črke B, manjkajoče v naštetih virih.

Pri iskanju, kako najustrezneje oblikovati besednodružinski slovarski del, v katerem bodo predstavljene besedne družine, sta nastala dva modela, ki sta bila vezana na tehnično ureditev zbranega besedja. Prvi model je izhajal iz t. i. ročnega strukturiranja besednih družin. Zanj je bilo tipično, da je bilo razvrščanje tvorjenk znotraj besedne družine oprto na merila, ki so izhajala iz besedotvornih in besednovrstnih značilnosti tvorjenk. To je hkrati pomenilo tudi, da se je od uporabnika za uspešnejšo uporabo slovarja pričakovala tudi večja teoretična vednost, s čimer se je posredno ožil tudi krog potencialnih uporabnikov. Razvrstitev tvorjenk je namreč urejalo več formalnih načel, ki so bila v naslednjem hierarhičnem odnosu: (1) Besednovrstno merilo: v vsaki besedni družini je veljal ustaljen prikaz tvorjenk po besednovrstnem zaporedju (samostalniške, pridevniške, prislovne, povedkovniške, glagolske, predložne, vezniške, členkovne, medmetne tvorjenke). (2) Besedotvornovrstno merilo: znotraj sleherne besedne vrste je veljalo ustaljeno zaporedje prikaza besedotvornih vrst (izpeljankam so sledile tvorjenke iz predložne zveze, sestavljenke, zloženke medponsko-priponske pred medponskimi, tem so sledili sklopi). (3) Abecedna razvrstitev obrazil, in sicer: odzadnja razvrstitev desnih obrazil, običajna abecedna razvrstitev levih obrazil, običajna abecedna razvrstitev korenskih morfemov pri zloženkah in sklopih. Merila od (1) do (3) so veljala

znotraj vsake, tj. prve, druge, tretje ... stopnje tvorjenosti. Na ta način je bila predstavljena besedna družina besede (Stramljič Breznik 2000: 45–56) in besedne družine predlogov in veznikov (Stramljič Breznik 2001: 197–206).

Drugi modelni prikaz besednih družin je izhajal iz uporabe računalniškega programa SlovarRed 2.0, 2003 Tomaža Seliškarja in Sekcije za terminološke slovarje z Inštituta za slovenski jezik Frana Ramovša, ZRC SAZU, ki je bil za izdelavo besednodružinskega slovarja posebej prilagojen in dograjen. S takim načinom obdelave besedja je postal slovar za uporabnika prijaznejši, saj so se merila razvrščanja, prej temelječa na besedotvornih sistemskih lastnostih leksike, umaknila preprostejšemu abecednemu načelu in zapomnitvi povezave med položajem obrazila ob podstavi ter s tem povezano ustrezno besedotvorno vrsto tvorjenke. Temeljna urejevalna načela besedne družine so tako postala: (a) hierarhičnost tvorjenk, (b) abecedna razvrstitev znotraj iste stopnje, (c) prikaz večkorenskih tvorjenk na koncu besedne družine.

Druga prednost računalniške obdelava je možnost nenehnega posodabljanja in dopolnjevanja gradiva ter njegova takojšnja možnost preureditve iz knjižne v elektronski slovar, ki se lahko uporablja sočasno z drugimi slovenskimi elektronskimi slovarji (SSKJ, Besedišče slovenskega jezika, Odzadnji slovar in SP 2001). Uporaba posebej prilagojenega računalniškega programa je povečala obvladljivost gradiva, saj ga je mogoče urejati po različnih kriterijih. Nekateri izmed teh bodo predstavljeni v nadaljevanju.

**2.2.1** Podatek o deležu besednih vrst v iztočnicah ostaja vsaj za prve tri najpogostejše besedne vrste enak, kot je bil predstavljen v sondažni raziskavi prvotno »ročno« strukturiranih besednih družin z okrog 7720 besedami (Stramljič Breznik 2001/02: 295–301). Trenutna odsotnost predloga, veznika in členska je pogojena zgolj zaradi iztočnic na B, sicer so že bile predstavljene nekatere predložne in vezniške besedne družine (Stramljič Breznik 2001:197–206).

Preglednica 1: *Besednovrstni delež iztočnic.*

Besedna vrsta iztočnic	Število	Delež
Samostalnik	489	73,42 %
Glagol	111	16,66 %
Pridevnik	42	6,30 %
Medmet	19	2,85 %
Prislov	4	0,60 %
Povedkovnik <sup>4</sup>	1	0,15 %

<sup>4</sup> V tem primeru gre za besedo *bot*, ki je v SSKJ označena kot prislov, v SP 2001 pa kot povedkovnik. Novejše jezikoslovne raziskave (Žele: 2004, 33–42) ugotavljajo, da so povedkovniki pomenskoskladenjska in ne že besednovrstna kategorija, zato bi jih bilo v leksikografskem smislu ustrezneje prikazovati kot podiztočnice. Le majhno skupino navadno besednovrstno neopredeljenih besed, pri katerih je povedkovniška vloga primarna, pa bi bilo smiselno predstaviti kot samostojna povedkovniška gesla. Ista avtorica (n. m.) razvršča *bot* v skupino prislovnih povedkovnikov.

**2.2.2** Besednovrstna razvrstitev iztočnic posledično vpliva na besednovrstni delež tvorjenk samo pri povezavi med sorazmerno velikim deležem samostalniških iztočnic (73,42 %) in visokim delež samostalniških tvorjenk (48,93 %), glagol ohranja približno šestinski delež tako iztočnic (16,66 %) kot tvorjenk (15,42 %), medtem ko je pridevnik med iztočnicami zastopan s šestnajstino (6,30 %), pri deležu tvorjenk pa se povzpne skoraj na tretjino (32,22 %). To pomeni, da imajo največjo moč generiranja v slovenščini prav imenske besedne vrste z veliko sposobnostjo tvorbe ne samo iz lastne, ampak tudi iz drugih besednih vrst, npr. samostalnik iz samostalnika, pridevnika, glagola, ali pridevnik iz pridevnika, samostalnika, glagola ipd.

Preglednica 2: *Besednovrstni delež tvorjenk.*

Besedna vrsta	Število tvorjenk	Delež
Samostalnik	4951	48,92 %
Pridevnik in deležnik	2695 + 566 = 3261	26,63 % + 5,59 % = 32,22 %
Glagol	1561	15,42 %
Prislov in deležje	190 + 94 = 284	1,87 % + 0,92 % = 2,80 %
Medmet	38	0,37 %
Členek	21	0,20 %
Predlog	3	0,02 %
Veznik	3	0,02 %

**2.2.3** Tabela o stopenjski razvrstitvi tvorjenk kaže, da je v obdelanih besednih družinah šesta stopnja tvorbe najvišja možna. Na prvi stopnji tvorbe je tako zajetih 41,27 %, na drugi 36,67 %, na tretji 15,61 %, na četrti 5,27 %, na peti 1,04 % in na šesti 0,09 % vseh tvorjenk.

Visoko stopnjo tvorbe dosegajo le predmetnopomenske besedne vrste, medtem ko druge, npr. členek in veznik, le izjemoma dosegajo drugo stopnjo. Tako razvrstitev potrjuje tudi teoretično spoznanje, da tovrstne besedne vrste v glavnem nastajajo s konverzijo ali sklapljanjem neposredno iz drugih besednih vrst. Največji delež tvorjenk prvih petih besednih vrst iz razpredelnice je razporejen od prve do tretje stopnje, na nadaljnjih stopnjah število tvorjenk opazno upada. Pri samostalniku, pridevniku, deležniku in glagolu je opazno zniževanje po stopnjah, v nasprotju s prislovom in deležjem, kjer je tvorjenk na drugi stopnji več kot na prvi. Sicer pa je pri posameznih besednih vrstah opazen kar drastičen upad števila tvorjenk pri višjih stopnjah tvorbe. Za primer vzemimo samostalnik in pridevnik: na prvi stopnji imata dober tretjinski oz. skoraj polovični delež tvorjenk (38,47 % oz. 47,19 %), na drugi stopnji ohranjata prav tako tretjinski delež (36,53 % oz. 31,98 %) na tretji stopnji pa se ta delež zmanjša na dobro oz. slabo šestino vseh tvorjenk (16,52 % oz. 14,47 %), na četrti pa že na petnajstino oz. petindvajsetino vseh tvorjenk svoje besedne vrste (7,0 % oz. 4,82 %).

Ti podatki so lahko pomembni za leksikografijo, ker je s tem postavljeno eno izmed meril o vključevanju tvorjenk v slovar, ki ga seveda mora dopolnjevati še podatek o pogostnosti.

Preglednica 3: *Stopenjska razporeditev tvorjenk po besednih vrstah.*

Besedna vrsta	Št. tvorjenk	Podstava	1. stopnja	2. stopnja	3. stopnja	4. stopnja	5. stopnja	6. stopnja
Samostalnik	4951	489	1905	1809	818	347	67	5
Pridevnik	2695	42	1272	862	390	130	36	5
Deležnik	566		119	259	166	22		
Glagol	1561	111	734	653	152	20	2	
Prislov	190	4	56	85	36	12	1	
Deležje	94		34	38	19	3		
Medmet	38	19	34	4				
Členek	21		19	1				
Predlog	3		3					
Veznik	3		2	1				
Povedkovnik		1						

**2.2.4** Med prvimi štirinštiridesetimi iztočnicami besednih družin, ki obsegajo več kot petdeset tvorjenk, je 19 glagolov (43,18 %), 15 samostalnikov (34,09 %), 8 pridevnikov (18,18 %) ter po en prislov in medmet (2,27 %). V preglednici zajete glagolske besedne družine imajo 21,84 %, samostalniške 16,70 % in pridevniške 8,11 % vseh zajetih tvorjenk. Ugotovimo lahko, da imajo predstavljene glagolske besedne družine v povprečju 116 tvorjenk, samostalniške 113 in pridevniške 103 tvorjenke.

Na podlagi teh podatkov je mogoče z delnim zadržkom glede na obseg v slovar vključenega besedja posplošiti, da so v povprečju med najboljšežnejšimi vendarle glagolske in ne samostalniške besedne družine, čeprav najštevilnejša samostalniška (*bog*) v tem izboru predstavlja 8,34 %, najštevilnejša glagolska (*brati*) pa 6,86 % tvorjenk.

**2.2.5** Zanimive podatke nam lahko ponuja tudi morfemska struktura tvorjenk. Pozornost bomo usmerili le na pripone in priponska obrazila.<sup>5</sup> Računalniški izpis nam glede na število pojavitev hierarhizira 322 pripon. Med njimi so zanimivi zlasti ekstremni primeri, saj lahko ugotovimo, da med njimi nastopa 32 takih, s katerimi je tvorjenih več kot 50 tvorjenk, hkrati pa jih je 131 s samo po eno tvorjenko. Prvih enaintrideset obrazil ima skupno 6069 tvorjenk, kar je 59,83 % delež. To lahko predstavimo tudi drugače: z manj kot desetimi odstotki obrazil (9,93 %) je tvorjenih skoraj šestdeset odstotkov zajetih tvorjenk.

**2.2.6** Nekoliko drugačne podatke o rodnosti obrazil dobimo, če izpostavimo deset najpogostejših za samostalnik, pridevnik in glagol. Deset najpogostejših samostalniških priponskih morfemov tvori več kot polovico (54,25 %) vseh samostalniških tvorjenk oz. dobro četrtino (26,53 %) zajetih tvorjenk sploh.

<sup>5</sup> Pri tem upoštevamo pojavitve vseh priponskih obrazil tako pri izpeljankah, tvorjenkah iz predložne zveze kot pripon pri medponsko-priponskih zloženkah.

Preglednica 4: *Besedne družine z več kot petdesetimi tvorjenkami.*

Besedna družina	Skupaj	1. stopnja	2. stopnja	3. stopnja	4. stopnja	5. stopnja	6. stopnja
1. bóg	406	180	155	52	17	2	
2. bráti	334	25	59	152	82	16	
3. bŕti	298	42	99	78	56	19	4
4. beséda	260	88	101	49	21	1	
5. bél	167	80	46	31	10		
6. bôsti	155	45	78	23	8	1	
7. bráda	139	85	47	7			
8. budíti	136	23	68	31	11	3	
9. braníti	132	54	45	23	7	3	
10. blág	132	36	56	29	11		
11. bʒ	125	62	50	8	5		
12. bók	124	23	32	41	21	7	
13. blodíti	124	20	21	34	33	14	2
14. bléd	120	61	38	13	8		
15. brusíti	115	49	58	7	1		
16. boléti	111	21	43	40	5	2	
17. boríti se	106	45	49	9	2	1	
18. bręsti	100	16	37	33	14		
19. bárva	98	16	41	29	12		
20. brég	95	61	32	2			
21. bíster	95	33	41	16	5		
22. brízgniti	94	3	30	45	16		
23. blízu	91	23	31	13	17	7	
24. brísati	90	23	39	24	4		
25. bába	89	40	37	11	1		
26. brázda	77	29	32	14	2		
27. báker	77	24	44	9			
28. bežáti	66	19	25	10	9	3	
29. bégniti	66	9	33	16	7	1	
30. bręme	65	7	10	15	24	9	
31. bóljši	65	4	9	23	18	9	2
32. bogŕt	64	27	27	7	3		
33. báviti se	64	6	24	26	8		
34. blískniti	60	5	20	29	6		
35. bríti	57	18	22	15	2		
36. blók	54	18	20	16			
37. brčnk	54	2	30	19	3		
38. bʒt	53	13	28	11	1		
39. bolán	53	9	14	20	7	3	
40. bóben	52	8	26	18			
41. bʒniti	52	4	23	22	3		
42. brŕt	51	19	21	5	5	1	
43. bláto	51	13	31	6	1		
44. búriti	51	11	14	16	8	2	



Preglednica 5: *Obrazila z več kot petdesetimi tvorjenkami.*

Zaporedje	Obrazilo	Št. pojavitev
1.	-en	751
2.	-ost	413
3.	-ski	402
4.	-anje	374
5.	-ec	327
6.	-ka	318
7.	-ica	312
8.	-o sam.	310
9.	-n	223
10.	-ev	192
11.	-enje	181
12.	-ik	171
13.	-oč	157
14.	-ast	153
15.	-ina	147
16.	-stvo	133
17.	-ov	132
18.	-ek	127
19.	-ar	126
20.	-o	122
21.	-(š)ki	112
22.	-je	103
23.	-iti	100
24.	-o prid.	91
25.	-aje	87
26.	-ež	83
27.	-in	82
28.	-a	81
29.	-ati	73
30.	-eč	65
31.	-alen	64
32.	-ač	57



Preglednica 6a: *Deset najpogostejših samostalniških priponskih morfemov.*

Zaporedje	Priponski morfem	Št. pojavitev
1.	-ost	413
2.	-anje	374
3.	-ec	327
4.	-ka	318
5.	ica	312
6.	-o sam.	310
7.	-enje	181
8.	-ik	171
9.	-ina	147
10.	-stvo	133

Pri pridevniku prvih deset obrazil tvori 2146 tvorjenk, kar je 65,80 % pridevniških tvorjenk in 21,20 % vseh zajetih tvorjenk.

Preglednica 6b: *Deset najpogostejših pridevniških priponskih morfemov.*

Zaporedje	Priponski morfem	Št. pojavitev
1.	-en	622
2.	-ski	395
3.	-n	217
4.	-ev	192
5.	-oč	155
6.	-ast	153
7.	-ov	132
8.	-(š)ki	107
9.	-o prid.	91
10.	-in	82

Pri glagolu prvih deset obrazil predstavlja slabo petino (19,53 %) vseh glagolskih tvorjenk oz. 3 odstotke vseh tvorjenk sploh.

Preglednica 6c: *Deset najpogostejših glagolskih obrazil.*

Zaporedje	Obrazila	Št. pojavitev
1.	-iti	100
2.	-ati	73
3.	-irati	48
4.	-izirati	24
5.	-ovati	17
6.	-niti	10
7.	-eti	11
8.	-evati	9
9.	-ariti	9
10.	-kati	4

**2.3 Abecedni slovarski del** prinaša abecedno razvrstitev vseh besed iz besednih družin, opremljenih s podatki o besedni vrsti, pogostnosti v korpusu Fida (F) in Nova beseda (NB) ter pripadnosti oštevilčeni besedni družini.

Frekvenčni podatki iz obeh korpusov so pridobljeni s t. i. avtomatskim štetjem. To pomeni, da program ni razlikoval homonimov in homonimnih oblik, tako so npr. pri samostalniku *bil* dejansko podani podatki za deležnik *bil*, kar kaže veliko število pojavitev (F: 204 956; NB: 220 540). Po Fidi, ki ima tudi najbolj izdelane mehanizme iskanja, dostopne slehernemu uporabniku, lahko ob tem primeru opazimo, da samostalnik *bil* ni opredeljen kot lema, marveč je lematizirana le oblika *bilka*, ki se pojavlja 472-krat, pri čemer so vključene tako občno- (*bilka*) kot lastnoimenske pojavitve (*Bilka*). Za Novo besedo velja, da ima vse besede iz Besedišča slovenskega jezika in SSKJ lematizirane, pri Fidi pa je po posredovanih podatkih za črko B lematiziranih 1925 besed, zato je bilo dodatno pregledano še 38 262 nelematiziranih besed, saj je avtomatsko preštevanje znotraj gradiva nelematiziranih besed zaznavalo pojavitve besede *le*, če je ta nastopala v osnovni obliki (npr. *bezljati* (7), ne pa tudi *bezlja* (10), *bezljajo* (8), *bezljajte* (1) *bezljal* (3), *bezljala* (1) *bezljale* (2), *bezljali* (5), *bezljata* (2)). Tako je bilo k prvotnemu podatku 7 prištetih še 32 pojavitev v različnih oblikah, kar znaša skupno frekvenco 39). V nobenem od korpusov pa ni bilo mogoče dobiti razlikovalnih podatkov o pojavitvah glagolov s prostimi morfemi oz. njihovimi vzporednicami brez prostega morfema, ker večinoma niso označeni kot leme. Tako so številski podatki samo za glagol brez prostega morfema, oz. za istokorenski glagol s prostim morfemom, če je *le*-ta lematiziran.

**2.3.1** Glede na pravkar omenjeno kaže na določen odstotek frekvenčnih podatkov v obeh korpusih gledati kritično, saj ne odražajo dejanskega stanja, ker niso bili uporabljeni zahtevnejši iskalni mehanizmi, ki bi podatke avtomatskega štetja ustrezno prečistili.<sup>6</sup> V ta odstotek manj zanesljivih frekvenčnih podatkov se namreč uvrščajo zlasti pojavitve besed, katerih oblike v oblikoslovni paradigmi sovpadajo z oblikami drugih besed, npr. imenovalnik ednine (*barva*) in dajalnik množine (*barvam*) samostalnika *barva* sovpada s tretjo osebo ednine (*barva*) in prvo osebo ednine (*barvam*) pri glagolu *barvati*, tako kot rodilnik množine samostalnika *breza* sovpada s predlogom *brez* ipd. Pogoste so tudi netočnosti, ki izvirajo iz iskanja pojavitve samo oblike besede (celo mimo besednovrstne oznake) in ne celotne leme, npr. *bel* (F: 2223), *belo* prisl. (F: 4555, kjer gre v mnogih primerih sploh za srednji spol pridevnika), *beli* m (F: 5400, kjer je zelo malo verjetno, da gre v tolikšni meri za posamostaljeno pojavitve, ampak za določno obliko pridevnika v besedni zvezi). Do podobnega sovpadanja oblik prihaja tudi npr. pri samostalniku *brana*, ki ima po NB 39696 pojavitev. Dejansko gre najverjetneje za deležnik glagola *brati*, medtem ko je za samostalnik verjetnejši podatek 204 pojavitve. Odras takega načina pridobivanja podatkov je tudi preglednica sedem.

Najpogostejše besede v obeh korpusih niso enako razvrščene, med desetimi so le štiri skupne (*biti*, *breza*, *banka*, *beseda*), vodilnost glagola *biti* v obeh ni presentljiva,

<sup>6</sup> Pogostostni podatki za oba korpusa, objavljeni v slovarju, so v dogovoru z lastniki zaenkrat ostali na ravni avtomatskega štetja, zbiranje in pridobivanje podatkov pa opravljeno tako rekoč prostovoljno in v prostem času Simona Kreka in Primoža Jakopina, ki sta podatke tudi posredovala.

Preglednica 7: Deset najpogostejših besed, zajetih z avtomatskim štetjem po korpusu Fida in Nova beseda.

Fida	x . 10 <sup>3</sup>	Nova beseda	x . 10 <sup>3</sup>
1. biti	738	1. biti	1277
2. breza	83	2. breza	177
3. beseda	61	3. približen	163
4. banka	41	4. banka	124
5. zbor	34	5. beseda	119
6. izbrati	30	6. bolezen	100
7. besedilo	28	7. bližnji	94
8. približen	28	8. britanski	83
9. bog	26	9. bel	71
10. barva	20	10. obramben	63

je pa presenetljiva visoka pojavitvev samostalnika *breza* (dejansko le 380 zadetkov), kar je posledica že omenjenega sovpadanja predloga *brez* z roditeljskim množino *breza*. Prikazano različno razvrstitev besed v obeh korpusih je mogoče pojasniti z dvema pomembnima podatkom: korpusa nista grajena po enakih načelih, kar posledično

Preglednica 8: Dvajset najpogostejših samostalnikov, zajetih v BD z iztočnicami na B po korpusu Fida (F) in Nova beseda (NB).

Samostalnik	F x . 10 <sup>3</sup> februar 2005	F x . 10 <sup>3</sup> avtomatsko	Samostalnik	NB x . 10 <sup>3</sup> februar 2005	NB x . 10 <sup>3</sup> avtomatsko
1. beseda	61	61	1. beseda	46	119
2. banka	41	41	2. banka	39	124
3. bog	26	26	3. zbor	22	27
4. besedilo	22	28	4. odbor	22	26
5. bolezen	19	19	5. boj	19	54
6. barva	17	20	6. bolezen	14	100
7. odbor	17	17	7. bolnišnica	13	46
8. prebivalec	15	15	8. dvoboj	13	32
9. zbor	14	34	9. prebivalec	13	34
10. obramba	13	13	10. obramba	12	39
11. boj	13	20	11. barva	11	30
12. izbira	13	13	12. zbirka	11	14
13. bolnišnica	13	13	13. pobuda	10	29
14. zbirka	12	12	14. besedilo	10	30
15. blago	12	12	15. bolnik	9	23
16. pobuda	11	11	16. prebivalstvo	8	28
17. bralec	10	10	17. izbira	8	9
18. bistvo	10	10	18. bistvo	6	15
19. prebivalstvo	9	9	19. bližina	6	30
20. bližina	9	9	20. bralec	6	16

vpliva na njuno referenčnost, in tudi časovno vključujeta različna besedila, saj ima korpus Fida zajete podatke do leta 2000, Nova beseda pa do leta 2004.

**2.3.2** Zanimiv pregled najpogostejših besed daje razvrstitev po besednih vrstah. Pri tem smo upoštevali glede na število tvorjenk tri najpogostejše besedne vrste in se omejili na samostalnik, pridevnik in glagol, in sicer tako, da smo glede na visok delež tvorjenk pri samostalniku upoštevali razvrstitveno listo prvih dvajsetih, pri pridevniku prvih desetih, pri glagolu in drugih besednih vrstah pa prve tri besede. Kot je že bilo opozorjeno, je avtomatsko štetje lahko zavajajoče zlasti v primerih, kjer prihaja do homonimije različnih paradigmatskih oblik oz. še drugih omenjenih posebnosti. Zato so sledeče pogostostne razvrstitve za samostalnik, pridevnik, glagol in druge besedne vrste narejene na podlagi dodatnih mehanizmov izločanja enakoizraznic in opozoril na nekatere t. i. korpusne šume ob pomoči obeh posredovalcev frekvenčnih podatkov, hkrati pa so dodane tudi pojavitve avtomatskega štetja, kot so objavljene v slovarju.

Pri Fidi prihaja do razlik pri besedah: *besedilo*, ki ima visoko pojavitve zaradi takega Delovega označevanja delov besedila, prav tako se velikokrat pojavlja tudi v transkripcijah sej državnega zbora (skupno 6149 zadetkov), torej je realnejše število  $22 \cdot 10^3$ . Pri besedi *barva* prihaja do že omenjenega prekrivanja oblik pri imenovaniku ednine in dajalniku množine s prvo in tretjo glagolsko osebo ednine, zanesljivo samostalniških zadetkov je  $17 \cdot 10^3$ . Beseda *zbor* je ponovno previsoko na spisku, saj gre spet za transkripcije sej državnega zbora. Če se odštejejo pojavitve v zvezi *državni zbor*, število zadetkov znaša  $14 \cdot 10^3$ . Pri samostalniku *boj* ponovno prihaja do sovpadanja oblik *boj* in *boji* z velelnikom in tretjo osebo ednine glagola *bati se*, zanesljivih je torej le  $13 \cdot 10^3$  zadetkov. Verjetno je vprašljiva tudi uvrstitev besede *izbira*, ki sovpada s tretjo osebo ednine glagola *izbirati*.<sup>7</sup>

Pri Novi besedi so odstopanja od rezultatov avtomatskega štetja precejšnja, saj so prvi podatki nastali na podlagi 148-milijonskega korpusa besed. Novi, kontrolni podatki pa so posredovani iz korpusne različice spomladi 2004, na podlagi katere je bilo ob koncu istega leta poskusno oblikoslovno označenih prvih 100 milijonov besed. Ker je šlo za poskusno označitev, so seveda pri dokončnem postopku možne še razlike.<sup>8</sup>

Podatki iz obeh korpusov kljub različnemu številu pojavitve posameznih samostalnikov in s tem posledično tudi različno uvrstitvijo na lestvici prvih dvajset kažejo na visoko prekrivnost najpogostejših samostalnikov, saj je ta devetdesetodstotna. To pomeni, da se v Fidi med dvajsetimi pojavita *bog* in *blago*, ki nista v Novi besedi, ter v Novi besedi *dvoboj* in *bolnik*, ki nista v Fidi, vendar je to samo posledica omejitve na prvih dvajset samostalnikov. Če bi namreč seznam razširili, se v NB na 21. mestu pojavi *borza* ( $6 \cdot 10^3$ ), na 22. *blago* ( $5 \cdot 10^3$ ) in na 23. *bog* ( $4 \cdot 10^3$ ), v Fidi pa 21. *bolnik* ( $9 \cdot 10^3$ ), 22. *borza* ( $8 \cdot 10^3$ ), 23. *zbornica* ( $8 \cdot 10^3$ ).

Tudi med pridevniki je sedemdestodstona prekrivnost. Na neupravičeno visokem mestu sta se prvotno pojavljala *približen* (F:  $28 \cdot 10^3$ , NB:  $163 \cdot 10^3$ ) in *blag* (F:  $13 \cdot 10^3$ , NB:  $49 \cdot 10^3$ ), ker so k prvemu šteti tudi prislovi oz. členki, pri drugem gre za prekriv-

<sup>7</sup> Podatke za Fido posredoval in komentiral Simon Krek v februarju 2005.

<sup>8</sup> Podatke za Novo besedo posredoval in komentiral Primož Jakopin v februarju 2005.

Preglednica 9: *Deset najpogostejših pridevnikov, zajetih v BD z iztočnicami na B po korpusu Fida(F) in Nova beseda (NB).*

Pridevnik	F x . 10 <sup>3</sup> februar 2005	F x . 10 <sup>3</sup> avtomatsko	Pridevnik	NB x . 10 <sup>3</sup> februar 2005	NB x . 10 <sup>3</sup> avtomatsko
1. bel <sup>9</sup>	27	2	1.bel	18	71
2. bivši	11	11	2. britanski	11	83
3. obramben	11	11	3. bližnji	9	94
4. britanski	11	11	4. obramben	7	63
5. najboljši	10	10	5. bivši	7	61
6. bližnji	10	10	6. boljši	4	8
7. boljši	5	5	7. bodoči	4	37
8. borzen	5	5	8. barven	3	24
9. barven	4	4	9. bolan	3	37
10. dobeseden	4	4	10. borzen	2	28

nost oblik s samostalnikom *blago*. Oba primera sta izločena. Vprašljiva je ponovno visoka pojavitve pridevnika *dobeseden*, saj je velika verjetnost, da gre za vštete tudi pojavitve prislova.

Podatka o pojavitvah glagola *biti* in *izbrati* sta v obeh korpusih primerljiva. Zanimivo pa je, da pri glagolu *bati se* v slovarju ni pojavitve po Fidi, ker je lema le *bati* (17 478), pa še tu je vštete tudi samostalnik *boj*, medtem ko je po Novi besedi pojavitve (22 480). Na tretjem mestu po številu pojavitve v obeh korpusih je torej glagol *bati se*, na četrtem mestu pa mu s približno enako frekvenco sledi *brati* (F: 14 397; NB: 14 402). Frekvenca za glagol *besediti* (23 159) po NB zbuja dvom, verjetno gre za napako.

Preglednica 10: *Trije najpogostejši glagoli, zajeti v BD z iztočnicami na B po korpusu Fida (F) in Nova beseda (NB).*

Glagoli	F x . 10 <sup>3</sup> februar 2005	F x . 10 <sup>3</sup> avtomatsko	Glagoli	NB x . 10 <sup>3</sup> februar 2005	NB x . 10 <sup>3</sup> avtomatsko
1. biti	1000	738	1. biti	591	1277
2. izbrati	30	30	2. izbrati	14	25
3. bati (se)	17	0	3. bati se	11	22

Prvotne pojavitve iz obeh korpusov za *približno*, *najbrž* in *blizu* so po mnenju obeh posredovalcev podatkov dokaj točne. Dejansko pa gre za v veliki meri členkovne izraze, s katerimi najpogosteje modificiramo stavke oz. njihove dele.

<sup>9</sup> V slovarju je pri pridevniku *bel* zabeležen podatek pojavitve po F le za obliko *bel*, ki znaša 2223, medtem ko je za lemo *bel* 27 520 zadetkov.

Preglednica 11: *Prve tri druge besedne vrste, zajete v BD z iztočnicami na B po korpusu Fida (F) in Nova beseda (NB).*

Druge besedne vrste	F x . 10 <sup>3</sup>	Druge besedne vrste	NB x . 10 <sup>3</sup>
1. približno	28	1. približno	32
2. najbrž	13	2. najbrž	15
3. blizu	10	3. blizu	12

### 3 Sklep

Poskusni besednodružinski slovar, izdelan za korenske besede s črko B, v prvi vrsti uzavešča novo leksikografsko možnost urejanja slovenske leksike, saj je besedna družina le ena izmed naravnih sistemskih struktur leksikona slehernega jezika, kar je zaradi abecednega načela slovarske ureditve prepogosto prezrto dejstvo.

Slovar je v slovenskem prostoru prva aplikacija besedotvorne teorije v slovaropisju, temu pa je poskusno dodana nova kvaliteta, saj je gradivo opremljeno s podatki o pojavitvah v dveh besedilnih korpusih, ki so danes postavljeni v izhodišče vseh jezikoslovnih in slovaropisnih projektov. Prav zaradi pionirske narave tega dela je nemogoče pričakovati, da bi prinašalo idelane rešitve slehernega tvorbenega vprašanja oz. ponujalo trenutno zbrane korpusne podatke kot kazalce absolutnega jezikovnega stanja. V resnici slovar odpira večsmerno komunikacijo, ker potencialne uporabnike in zlasti jezikoslovce postavlja pred vprašanja, ki si jih pred tem verjetno še niso zastavljali, in ker neposredno opozarja na zagatnost oblikoslovnega označevanja slovenščine in pomanjkljivosti trenutno razpoložljivih postopkov jezikoslovnih analiz v obeh korpusih. Tovrstni izzivi pa lahko postanejo dobri temelji za morebitni projekt, v katerem bi posamezniki in inštitucije lahko sodelovali glede na svoje strokovne kompetence, ki bi vodile k čim večji objektivizaciji rezultatov takega raziskovalnega dela.

Dolgoročno gledano se zdi izdelava takega slovarja možna in dokaj realna, ker je podprta z ustrezno prilagojenim računalniškim programom, brez katerega v sodobni leksikografiji več ne gre. Po drugi strani pa edino tako zasnovan slovar lahko daje konkretne in preverljive podatke o ustroju slovenskega leksikalnega sistema, tj. o besednovrstnem deležu motiviranih in nemotiviranih besed, o rodnosti posameznih besednih družin, o tvorbeni moči korenskih besed na različnih stopnjah tvorbe, o prevladujočih morfemskih prvinah in s tem posredno o besedotvornovrstni tipologiji tvorjenk. In edino tako zasnovan slovar lahko zelo konkretno odgovarja tudi na najpogosteje zastavljena vprašanja, katere besede in koliko jih predstavlja korensko jedro slovenščine ter kakšne spremembe doživlja slovenska leksika v času in prostoru.

#### LITERATURA

- Janko BEZJAK, 1906: *Posebno ukoslovje slovenskega učnega jezika v ljudski šoli*. Ljubljana: Slovenska Šolska Matica. 194.
- Vojko GORJANC, 2003: Korpusi in jezikoslovje. *Jezik in slovstvo* 48/3–4. 19–28.
- Peter KONČNIK, 1913: *Slovenska slovnica z naukom, kako se pišejo pisma in opravilni sestavki*. Dunaj: V cesarski kraljevi zalogi šolskih knjig. 23.



- Simon KREK, 2003: Jezikovni priročniki in novi mediji. *Jezik in slovnstvo* 48/3–4. 29–46.
- Josef MISTRÍK idr., 1993: *Enciklopedia jazykovedy*. Bratislava: Vydavateľstvo Obzor. 410.
- Mluvnice češtiny I*, 1986. Praha: Academia. 207–208.
- Primož JAKOPIN, 2003: Nekaj zanimivosti iz besedilnega korpusa Nova beseda. *Jezikoslovni zapiski* 9/2. 145–159.
- Irena STRAMLJIČ BREZNIK, 2000: Besedna družina *besede*. *Jezikoslovni zapiski* 6. 45–55.
- 2001: Besedne družine predlogov in veznikov v besedotvornem slovarju. *Jezikoslovni zapiski* 7/1–2. 197–206.
- 2001/2002: Nekaj podatkov iz Besedotvornega slovarja slovenskega jezika za črko B po SSKJ. *Jezik in slovnstvo* 47/7–8. 295–301.
- 2004: O besedi s tvorbenega, izrazijskega in frazeološkega vidika. *Besedoslovne lastnosti slovenskega knjižnega jezika in narečij*. Ur. M. Jesenšek. Zora 28. Maribor: Slavistično društvo Maribor. 315–401.
- Jože TOPORIŠIČ, 1992: *Enciklopedija slovenskega jezika*. Ljubljana: Cankarjeva založba. 7.
- 1996: *Slovenski jezik in sporočanje 2*. Maribor: Založba Obzorja Maribor. 89.
- 2000: *Slovenska slovnica. Četrta, prenovljena in razširjena izdaja*. Maribor: Založba Obzorja Maribor. 108, 116.
- Ada VIDOVIČ MUHA, 2000: *Slovensko leksikalno pomenoslovje. Govorica slovarja*. Ljubljana: Znanstveni inštitut Filozofske fakultete. 107–108.
- Andreja ŽELE, 2004: Povedkovnik v slovenščini. *Jezikoslovni zapiski* 10/1. 33–42.
- Slovarji in korpusni viri
- Gerhard AUGST, 1990: Das Wörterbuch der Wortbildungsmittel. *Wörterbücher/Dictionaries/Dictionnaires*. Berlin, New York: Walter de Gruyter, 1145–1152.
- 1998: *Wortfamilienwörterbuch der deutschen Gegenwartssprache*. Tübingen: Max Niemeyer Verlag.
- Fida: <http://www.fida.net/slo/index.html>
- Jay Louis HERMAN, 1975: *A dictionary of Slavic word families*. New York and London: Columbia University Press.
- Urban JARNIK, 1832: *Versuch eines Etymologikons der slovenischen Mundart in Inner-Oesterreich*. Celovec.
- A. I. KUZNECOVA, T. F. JEFREMOVA, 1986: *Slovar' morfem russkogo jazyka*. Moskva: Ruskij jazik.
- Elisabete LINK, 1990: Das Wörterbuch der Wortbildungsmittel. *Wörterbücher/Dictionaries/Dictionnaires*. Berlin, New York: Walter de Gruyter, 1223–1230.
- Anton MURKO, 1833: *Slovensko-nemški ročni besednik*. Gradec.
- Nova beseda*: <http://bos.zrc-sazu.si>
- Jordan PENČEV, 1999: *Slovoobrazovaten rečnik na s'vremennija b'lgarski knižoven ezik: a–ja*. Sofija.
- Slovenski pravopis*, 2001: Ur. J. Toporišič. Ljubljana: SAZU in ZRC SAZU.
- Slovar slovenskega knjižnega jezika z Odzadnjim slovarjem slovenskega jezika in Besediščem slovenskega jezika z oblikoslovnimi podatki: Elektronska izdaja na plošči CD-ROM*. Ljubljana 1998.
- Miloslava SOKOLOVÁ, Gustáv MOŠKO, František ŠIMON, Vladimír BENKO, 1999: *Morfematically slovník slovenčiny*. Prešov: Náuka Prešov.
- M. D. STEPANOVA, 1979: *Slovar' slovoobrazovatel' nyih elementov nemeckogo jazyka*. Moskva: Ruskij jazik. 164–165.
- Irena STRAMLJIČ BREZNIK, 2004: *Besednodružinski slovar slovenskega jezika. Poskusni zvezek za iztočnice na B*. Zora – Priročniki 1. Maribor: Slavistično društvo Maribor.
- Alenka ŠIVIC - DULAR, 1999: *Besedna družina iz korena \*god- v slovanskih jezikih. Pomenoslovna razčlemba v kulturološkem kontekstu*. Ljubljana: Založba ZRC SAZU.



Aleksander Nikolajevič ТИХОНОВ, 1985: *Slovoobrazovatel'nyj slovar ruskogo jazyka*. Moskva: Ruskij jazik.

#### SUMMARY

The article proceeds from the terminological definition of the basic unit of the dictionary, i.e., the word family, which should be preserved for the sake of continuity in Slovene linguistics despite different terms in other Slavic languages and for the sake of placement of word-family dictionary into the system of morphemic dictionaries.

What has been in older Slovene lexicography only indicated in a form of individual elements, has been in modern lexicography of some Slavic, Germanic, and Romance languages already executed in specialized word-family dictionaries. According to needs and derivational characteristics of Slovene, two word-family lexicographic models were devised. The first one was based on the criteria derived from characteristics concerning derivation and parts of speech of the derivatives. The second model, implemented in the trial issue of the dictionary, is based on the use of a specialized computer program, which is the basis of presentation of quantitative features of the lexicon compiled in the trial dictionary.

An interesting piece of information is the fact that no absolute correlation exists between the part of speech of head words and derivatives, which is indicated by the small share of adjectival head words and several times larger share of adjectival derivatives. Derivatives arranged into word families indicate the sixth degree of derivation as the highest possible degree, but the number of derivatives drastically decreases in all parts of speech after the third degree. Among forty-four word families with more than fifty derivatives in all head words starting in B, there are nineteen verbs, which at this point leads to the conclusion that they have larger families than nominal head words. There is an important difference in the productivity of individual suffixes, as almost sixty percent of all derivatives included in the dictionary are derived with only ten percent of the most productive suffixes.

The corpus data about the number of occurrences based on automatic count are reliable only in those cases where the identical expression of word forms in paradigmatic pattern is excluded. Hence, the trial determination of lexical frequency directly points out the difficulties with morphological marking of Slovene and shows the shortcomings of the methodologies of linguistic analysis in both corpora, which can be extremely helpful for further development of the programs.

The compilation of a complete word-family dictionary of Slovene is supported on the one hand by various possibilities of its use and the variety of data that it will provide. On the other hand, only a dictionary conceived in this manner can provide concrete and verifiable data about the structure of Slovene lexical system and answers the questions what words and how many of them represent the root core of Slovene.