Besedilni korpus *Nova beseda* in geslovnik za *Slovar novejšega besedja slovenskega knjižnega jezika*

Nanika Holz

IZVLEČEK: Jezikovni korpusi se v sodobni leksikografiji uporabljajo kot eden glavnih jezikovnih virov. Prispevek obravnava možnosti izrabe korpusa Nova beseda pri pripravi geslovnika za Slovar novejšega besedja slovenskega knjižnega jezika.

ABSTRACT: Language corpora are used as one of the main language resources in modern lexicography. The article discusses how the Nova beseda corpus could be employed in the formation of the word list for the Dictionary of Newer Standard Slovenian Words.

1 Tradicionalni način sestavljanja slovarjev je temeljil na izpisovanju gradiva, urejenega v (večinoma abecedne) listkovne kartoteke. Iz takih zbirk so slovaropisci zajemali gesla, iskali potrditev pomenov in čim bolj tipične zglede, prepoznavali pogoste oz. stalne zveze, ugotavljali terminološko rabo posameznih besed ali zvez. Podobno velja tudi za črpanje gradiva iz drugih slovarjev in leksikonov. Takšen način ročnega zbiranja in razvrščanja gradiva je zelo zamuden, zbrati in analizirati je mogoče le razmeroma majhno število besedil, presojanje o tem, kaj in v kakšni obliki bo sprejeto v slovar pa je omejeno na subjektivno presojo posameznika oz. skupine. Da bi lahko odpravili dejavnik subjektivnosti in hkrati pridobili čim več objektivnih, empirično preverljivih podatkov, so raziskovalci začeli razvijati elektronske besedilne korpuse (McEnery in Wilson, 2001), ki so že postali standarden vir podatkov za izdelavo slovarjev – sprva predvsem enojezičnih, v zadnjem času pa tudi dvojezičnih (McEnery in Wilson, 2001; Krek, 2003).

Potrebe uporabnikov narekujejo oblikovanje različnih vrst korpusov, ki se razlikujejo po obsegu, zajetju zvrsti, časovnih okvirih, prenosniku, številu upoštevanih jezikov ipd. (Gorjanc, 1999). Najbolj običajni so enojezikovni referenčni korpusi, ki so zaradi uravnoteženih razmerij med vrstami besedil in zajemanja žive, sodobne jezikovne rabe nujni za slovnične, pomenoslovne, besediloslovne, prevodoslovne in druge raziskave in s tem tudi za slovaropisce. Jezikoslovnim raziskavam sta v slovenskem prostoru namenjena Korpus slovenskega jezika *FIDA* (http://www.fida.net – nastaja v sodelovanju založbe DZS, podjetja Amebis, Instituta Jožef Stefan in Filozofske fakultete v Ljubljani) in Besedilni korpus *Nova beseda*

(http://bos.zrc-sazu.si/s_beseda.html – nastaja v Laboratoriju za korpus slovenskega jezika v okviru Inštituta za slovenski jezik Frana Ramovša ZRC SAZU).

2 V Leksikološki sekciji Inštituta za slovenski jezik Frana Ramovša so po izidu *Slovenskega pravopisa* v letu 2001 stekli novi slovarski projekti, med katerimi je tudi *Slovar novejšega besedja slovenskega knjižnega jezika* (dalje SNB).[1] Priprava geslovnika je eden od temeljev za sestavo slovarja, pri čemer je nujna uporaba sodobnih sredstev, konkretno jezikovnih korpusov (poleg tega so potrebni še konkordančniki za obdelavo korpusnih podatkov – seveda pa tudi dovolj zmogljiva računalniška oprema ter zadostno jezikoslovno in računalniško znanje za smiselno uporabo vsega naštetega). Korpusa *FIDA* in *Nova beseda* sta različno zasnovana, zato bi bilo pri sestavljanju geslovnika za SNB zelo koristno kombinirati informacije iz obeh; ker pa dogovori za uporabo korpusa *FIDA* še potekajo, je bil v tem času za pripravo geslovnika na voljo le korpus *Nova beseda*. V literaturi se pojavljajo različni kriteriji, po katerih naj bi se besede iz korpusa uvrščale v geslovnik: lahko so samo frekvenčni, npr. 15 pojavitev določene besede, lahko pa so tudi bolj kompleksni, ko poleg frekvence upoštevajo še npr. korpusni šum ter število in vrsto virov, v katerih se zadetki pojavljajo (Jakopin, 2003; Krek, 2003).

Za potrebe geslovnika SNB so bili v aprilu 2003 pripravljeni spiski nelematiziranega besedja iz korpusa *Nova beseda* za črke **A–K**, in sicer ob predpostavki, da bo med nelematiziranim besedjem mogoče najti glede na *SSKJ* nove besede za vključitev v geslovnik *SNB*. Priprava spiskov je potekala v sedmih korakih:

1. zbiranje vseh enot s frekvenco od 10 do 30 na določeno začetno črko z besednim iskanjem (primer iskalnega pogoja: *be=a* in fr>9 in fr<31*) na internetni strani korpusa *Nova beseda*;[2]

2. kopiranje spiska zadetkov v urejevalnik besedila, npr. Word – zadetki so bili v tabeli, iz katere je bil najprej odstranjen stolpec z zaporednimi številkami,

[1] Avtorica prispevka je bila v projekt izdelave SNB vključena od 1. 11. 2002 do 30. 4. 2003; sprva je za objavo tehnično urejala teoretične in redakcijske prispevke sodelavk SNB (Ljudmila Bokal, mag. Alenka Gložančev, mag. Nataša Jakop, Polona Kostanjevec, Nastja Vojnovič), pozneje pa se je vključila v pripravo geslovnika.

[2] Že v februarju in marcu so bili narejeni prvi poskusi za uporabo korpusa *Nova beseda* kot enega od virov za geslovnik SNB (mag. N. Jakop, N. Holz). Na začetku je bila spodnja meja za število pojavitev posamezne enote pri besednem iskanju postavljena na 6+, vendar je bilo na spisku iz nelematiziranega dela črke **B** skoraj 6.000 enot. Ker zgornja meja števila zadetkov ni bila omejena, bi pregledovanje konkordanc že pri enotah z nad 30 ali 50 zadetki trajalo tako dolgo, da geslovnika ne bi bilo mogoče sestaviti v razumnem času. Zaradi (za leksikografske potrebe) neuravnotežene besedilnovrstne sestave korpusa *Nova beseda* je bil za preizkus tega korpusa kot vira za geslovnik izbran frekvenčni obseg od 10 do 30 pojavitev posamezne enote, kar naj bi tudi sestavljavkam geslovnika omogočilo pregled konkordanc za posamezen zadetek v razumnem času. Nizka zamejitev frekvenčnega obsega je bila postavljena tudi zaradi specifike SNB, saj lahko domnevamo, da imajo novejše besede sorazmerno manjše število pojavitev. – Besedilnovrstna sestava korpusa *Nova beseda* se je deloma izboljšala po razširitvi 12. 4. 2003.

nato pa je bila tabela pretvorjena v besedilo (za ločevanje med enoto in njeno frekvenco je bil uporabljen znak #);

3. kopiranje spiska, pripravljenega v 2. koraku, na internetno stran *Določevanje osnovnih besednih oblik in besednih vrst* (http://bos.zrc-sazu.si/dol_lem.html) in določitev lem;

4. kopiranje rezultatov, dobljenih v 3. koraku, v urejevalnik besedila – dvignjene pike so bile zamenjane z znaki #, oznake za ročni prelom vrstice pa nadomeščene z oznakami za odstavek;

5. spisek iz 4. koraka je bil pretvorjen v tabelo, tabela pa sortirana po stolpcu z lematizacijskimi oznakami – tako je bilo mogoče spisek ločiti na dva dela, tj. na lematiziranega in nelematiziranega;

6. oba dela spiska sta bila pretvorjena nazaj v besedilo, oštevilčena in urejena za tristolpčni (nelematizirani del) oz. dvostolpčni iztis (lematizirani del);

7. iztis nelematiziranega dela spiskov za črke **A–K** (celotne datoteke so sodelavkam SNB na voljo na strežniku v mapi *S:\ISJ\Enozvezkovnik\Geslovnik_NB*).[3]

3 Spiski nelematiziranega besedja za črke **A–K** imajo skupaj 17.532 enot (**A**: 1.962, **B**: 2.677, **C**: 1.597, **Č**: 336, **D**: 1.961, **E**: 1.113, **F**: 1.128, **G**: 1.641, **H**: 1.421, **I**: 967, **J**: 1.042, **K**: 2.687), seveda pa vse enote niso primerne za uvrstitev v geslovnik. Glede na to, da gre za nelematizirani del korpusa *Nova beseda*, se ponavljajo zadetki za isto lemo v različnih sklanjatvenih oz. spregatvenih oblikah, poleg tega pa pri iskanju ni mogoče izločiti lastnoimenskih samostalnikov in iz njih izpeljanih svojilnih pridevnikov. Za okvirno oceno, kolikšen delež zadetkov predstavlja besedje, ki bi ga lahko uvrstili v geslovnik SNB, je bil zbran vzorec tako, da je bil vključen vsak stoti zadetek pri posamezni črki. Vzorec obsega 179 enot (**A**: 19, **B**: 26, **C**: 15, **Č**: 3, **D**: 19, **E**: 11, **F**: 11, **G**: 16, **H**: 14, **I**: 9, **J**: 10, **K**: 26), in sicer: *absorbcijo, adiecto, aganović, aiôna, albana, alijeva, altenmarkt, ameriško-evropski, ancien, aneta, anoreksija, anz, arantxi, aristotelovih, arsović, astorja, atwood, aventin, azerbajdžanski, bahtirija, ballanu, baranji, bartenstein, battisti, beattyja, beit, beltinško, beranek, berner, bettiniju, bilalovič, bishop, blahotova, blum, bogosian, bonaventure, borštnarja, bowiejem, brandom, bregovićem, brika, brotherja, budanova, bulworth, businello, canade, cardin, cassinu, cebita, ceps, chambery, chicco, ciccolini, civilnodružbena, cl50, comenius, convert, coso, cresswell, cujnikova, čepikov, čilencev, črtomirom, damaske, darwinu, deanu, delilca, deprivilegirane, dessau, dezinflacije, didmund, diorju, djukanović, dobrotka, doll, dorica, dragočajna, drk, ds1, duomo, dvojem, džomba, eevropa, eke, elektroinštitutu, emanuelom, enčev, epsonih, esbjerg, etija, evert, evro@delo.si, e-100, fannie, fedjatin, ferlic, fiery, finnu, flensburgu, foo, foxy, frearsa, fromberg, fuzbalu, galino, garms, gcc, georgetown, gfk, ginsburg, glenny, goethejem, gončarenko, gosper, gradaščico, gravitte,*

[3] Za pripravo spiskov nelematiziranega besedja mi je bil na voljo računalnik s 75 MHz procesorjem, 32 MB spomina in 812 MB trdega diska (od tega približno 200 MB prostega), na katerem je obdelava 4.842 zadetkov pri črki **B** trajala več kot tri ure. Pripravo spiskov za črke **A** in **C–K** (skupno 32.016 zadetkov) je bilo mogoče izvesti v 2 delovnih dneh, ko mi je kolega Janez Keber prijazno odstopil v uporabo svoj bistveno zmogljivejši računalnik, dobavljen v l. 2002.

grimščah, grünfeldova, guillemot, gvardiol, hajdošah, handheld, harper's, havlovega, heinzom, herbart, heye, hiroja, hoet, homepage, hotavlje, hrpelj, huj, hypoligo, idz, ilirov, imt, init, interest, ipn, ismn, iversonu, izraelovi, jakubović, janssen, javnofinančnimi, jekaterina, jerneji, jimija, jönköping, joza, juniku, južnoafriškim, kajuhove, kamenjanjem, kapitolu, karija, kastracijski, kazahstanskega, kelmorajnu, keser, kieslowski, kirnu, klavoro, klotz, koblencerja, kojonkoskim, komenski, konradom, kordeža, kosanovič, kotečniku, kračuna, krauthammer, krimovkam, krnskim, kt266, kung, kuvajtsko.

Pregled konkordanc pri zadetkih iz vzorca je trajal približno en delovni dan, saj analiziranje zadetkov, izpisanih na računalniškem zaslonu, terja veliko zbranost in ga ni mogoče izvajati več ur brez premora. Po izločitvi lastnih imen (tudi kratičnih), delov lastnih imen, izlastnoimenskih pridevnikov, elektronskih naslovov, okrajšav in tujejezičnih zadetkov so bili kandidati za vključitev v geslovnik SNB leksemi *absorbcija, anoreksija, civilnodružben, delilec, deprivilegiran, dezinflacija, fuzbal, javnofinančen, kamenjanje, kastracijski* (10 leksemov oz. 5,6 % vzorca).[4] Ker bo kot vir za geslovnik uporabljen tudi občnoimenski fond iz *Slovenskega pravopisa 2001*, je smiselno pregledati, ali so ti leksemi že evidentirani:

Nova beseda	**Slovenski pravopis 2001**
absorbcija	kazalka na *absorpcija*
anoreksija	+
civilnodružben	–
delilec	nezaznamovana vzporednica *delivec*
deprivilegiran	–
dezinflacija	+
fuzbal	–
javnofinančen	–
kamenjanje	+ (*kamenjati*)
kastracijski	–

Obstaja še skupina zadetkov, kjer se mešajo občno- in lastnoimenske enote (frekvenca občnoimenskih enot je lahko tudi manjša od 10): *finn* 'tekmovalna kategorija', *huj* prislov oz. medmet, *kari* 'začimba', *krimovka* 'članica kluba' (4 leksemi oz. 2,2 % vzorca). Posebni primeri so še citatne zveze kot npr. *ancien régime, contradictio in adiecto, cosa nostra*, ki v SNB ne sodijo in bi svoje mesto ob dovolj pogosti rabi našle v slovarju citatno rabljenih besed. V geslovnik bi se lahko uvrstil še *elektroinštitut*, ki v korpusu *Nova beseda* sicer vedno nastopa kot del lastnega imena (*Elektroinštitut Milan Vidmar*).

4 V uvodnem delu so bile že omenjene razlike v zasnovi med korpusoma *Nova beseda* in *FIDA* – tako v besedilnovrstni sestavi kot v možnostih iskanja po korpusu. V korpusu *FIDA* je omogočeno neposredno iskanje po lemi (ne glede

[4] Med nelematiziranim besedjem se je znašla tudi oblika *dvojem* za številnik *dvoj*, ki je že obdelan v *SSKJ*.

na to ali gre za občno- ali lastnoimensko besedje); iz zgornjega vzorca so npr. lematizirane enote *absorbcija, anoreksija, Aristotelov, dvoj, Goethe, Ilir, Jernej/Jerneja, Kajuhov, kamenjanje, Kirn*;[5] že pri t. i. enovrstičnem iskanju je omogočeno iskanje po lematizacijskih oznakah ter kombiniranje pozitivnih in negativnih iskalnih pogojev, s čimer je mogoče izločiti lastnoimenske zadetke, ki v obravnavanem vzorcu iz korpusa *Nova beseda* predstavljajo več kot 80 %. Korpus *FIDA* omogoča tudi kompleksnejša iskanja oz. zamejitve iskanj, npr. po letnici izida ali po drugih podatkih v glavi besedila (zvrst, lektorirano ipd.). Različna sestava korpusov *Nova beseda* in *FIDA* se kaže pri dveh primerih iz obravnavanega vzorca, in sicer pri leksemu *džomba*, kjer *FIDA* izpričuje tudi občnoimenski pomen 'vojak pred koncem služenja vojaškega roka', in pri *huj* pomen 'riba'.

5 Končno oceno uporabnosti korpusa *Nova beseda* pri sestavljanju geslovnika SNB bodo morale podati bolj izkušene geslovničarke in redaktorice, zaenkrat pa se zdi, da bodo spiski nelematiziranega besedja služili kot sekundarni vir, tj. za preverjanje besedja, pridobljenega iz drugih virov. Brez dvoma bi bili sestavljavkam SNB v večjo pomoč frekvenčni seznami z izločenim lastnoimenskim fondom (urejeni po posameznih črkah, morda tudi po najpogostejših lemah), ki bi jih na podlagi korpusa *Nova beseda* verjetno lahko pripravili v Laboratoriju za korpus slovenskega jezika.

Viri in literatura

Besedilni korpus Nova beseda (http://bos.zrc-sazu.si/s_beseda.html.

Določevanje osnovnih besednih oblik in besednih vrst (http://bos.zrc-sazu.si/dol_lem.html.

Gorjanc, Vojko, 1999, Korpusi v jezikoslovju in korpus slovenskega jezika FIDA, *XXXV. seminar slovenskega jezika, literature in kulture*, 47–59.

Gorjanc, Vojko, 2000, Nekatere možnosti jezikoslovne izrabe enojezikovnih korpusov, *XXXVI. seminar slovenskega jezika, literature in kulture*, 335–348.

Jakopin, Primož, 2001, Slovenski nacionalni korpus – idejni osnutek projekta, *Jezikoslovni zapiski* 7, 411–417.

Jakopin, Primož, 2003, Kmalu sto milijonov slovenskih besed, *Delo – Znanost*, 24. 3. 2003, 9.

Korpus slovenskega jezika FIDA http://www.fida.net.

Krek, Simon, 2003, Sodobna dvojezična leksikografija, *Jezik in slovstvo* 48/1, 45–60.

Landau, Sydney I., 1989, *Dictionaries: The Art and Craft of Lexicography,* Cambridge University Press.

McEnery, Tony in Andrew Wilson, 2001, *Corpus Linguistics,* Second Edition, Edinburgh University Press.

Slovar slovenskega knjižnega jezika z Odzadnjim slovarjem slovenskega jezika in

[5] V korpusu *FIDA* zaenkrat ostaja še nerešen problem razdvoumljenja enakopisnih lem.

Besediščem slovenskega jezika z oblikoslovnimi podatki, Elektronska izdaja na plošči CD-ROM, 1998, Ljubljana, DZS.

Slovenski pravopis, 2001, Ljubljana, Založba ZRC.

Stabej, Marko, 1998, Besedilnovrstna sestava korpusa FIDA, *Uporabno jezikoslovje* 6, 96–106.

Weiss, Peter, 2001, Slovenski nacionalni korpus Maks na Inštitutu za slovenski jezik Frana Ramovša ZRC SAZU, utemeljitev, *Jezikoslovni zapiski* 7, 419–428.