Nekaj zanimivosti iz besedilnega korpusa *Nova beseda*

Primož Jakopin

IZVLEČEK: V članku je nekaj zanimivih podatkov o distribuciji črk, besed in stavkov v besedilnem korpusu v Laboratoriju za korpus slovenskega jezika pri Inštitutu za slovenski jezik Frana Ramovša ZRC SAZU. Korpus Nova beseda je glavni prek spleta prosto dostopni vir za kvantitativno raziskovanje slovenskega jezika (http://bos.zrc-sazu.si) in zdaj obsega 100 milionov besed, pretežno iz časopisnih tekstov in leposlovja.

***Some Details from the Text Corpus* Nova beseda**
ABSTRACT: In the paper some interesting data about the distribution of letters, words and sentences from the text corpus at the Corpus Laboratory of the Fran Ramovš Institute of the Slovenian Language are revealed. The corpus, Nova beseda, is the main freely accessible online source for the quantitative research of Slovenian language (http://bos.zrc-sazu.si) and currently consists of 100 million running words, mainly newspaper texts and fiction.

Nova beseda je besedilni korpus, ki nastaja v Laboratoriju za korpus slovenskega jezika pri Inštitutu za slovenski jezik Frana Ramovša ZRC SAZU. Namenjen je predvsem za slovaropisne namene Inštituta, lahko pa po njem s konkordančnim iskalnikom in iskalnikom v slovarju besednih oblik preko svetovnega spleta (na naslovu **http://bos.zrc-sazu.si**) prosto poizvedujejo tudi vsi drugi uporabniki, iz domovine in tujine. Dnevno je na spletne strani korpusa približno 600 dostopov, od začetkov leta 1999 s 3 milijoni besed pa je bila zbirka do julija 2003 povečana na 100 milijonov besed. Zvrstna struktura korpusa je razvidna iz tabele 1.

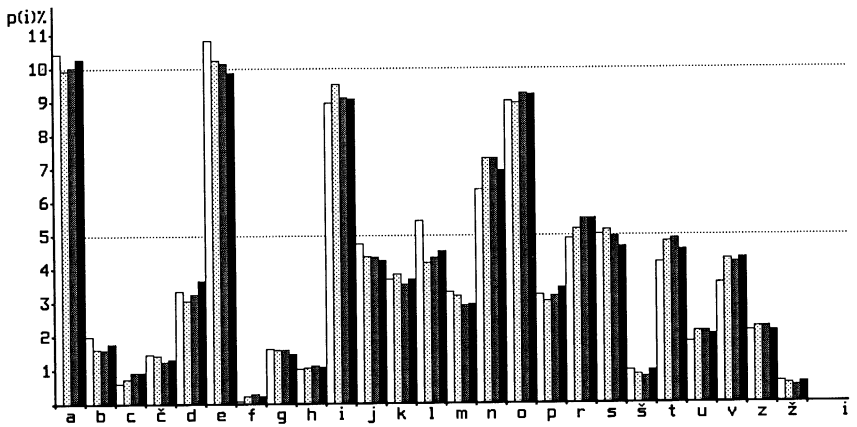
Tabela 1: Število besedil in besed po posameznih delih korpusa

	besedil	besed
leposlovje (A)	557	5.536.160
polleposlovje (B)	227	1.008.645
monografije (C)	18	1.640.447
periodika (P)	1.829	92.191.047
Skupaj	2.631	100.376.299

Med leposlovnimi besedili je največ proze (523 del), sledita dramatika (25 del) in poezija (9 del). V polleposlovju je 165 esejističnih besedil, 58 besedil korespondence, po dvoje pa je potopisov in spominov. Z besedo monografije so v tabeli označene znanstvene in strokovne publikacije: 17 družboslovnih in eno naravoslovno delo. Periodične publikacije obsegajo predvsem izvode časopisa DELO od januarja 1998 do junija 2003, poleg njih pa še 33 izvodov računalniške revije Monitor (letniki 1999, 2000 in 2001) ter 8 izvodov revije za zdravo življenje Viva (od oktobra 2002 do maja 2003). V prvih treh delih korpusa (A, B in C) je 777 izvirnih del in 25 prevodov. Zastopanih je 70 avtorjev, med njimi s celotnimi opusi Ciril Kosmač (52 del), Drago Jančar (107 del) in Ivan Cankar (537 del), več kot eno delo pa imajo še Drago Bajt (2), Lewis Carroll (2), Gitica Jakopin (24), Josip Jurčič (2), Janko Kersnik (3), Tomo Križnar (2), Fran Levstik (3), Florjan Lipuš (2), Platon (2), Ivan Pregelj (2), Fran Saleški Finžgar (2), Ignac Sivec (2), Ivan Tavčar (3) in Prežihov Voranc (2).

Podrobneje bo korpus s kvantitativnega vidika predstavljen v okviru posebne publikacije, v tem prispevku pa je navedenih nekaj zanimivosti, ki so se pojavile ob njegovi obdelavi. Na sliki 1 je prikazana porazdelitev pogostejših črk, ki so obenem tudi črke slovenske abecede, v vseh štirih delih korpusa.

Slika 1: Porazdelitev pogostejših črk v leposlovnih (belo), polleposlovnih (svetlosivo), monografskih (temnosivo) in periodičnih publikacijah (črno)



Porazdelitve so si razmeroma blizu, še najbolj izstopa leposlovje, kjer je npr. delež najpogostejše črke *e* znatno večji kot delež črke *a*, ki je sicer najpogostejša v časopisnih oz. revijalnih besedilih. Zanimivo je tudi pri najredkejši črki, *f*, ki je je v leposlovju le 0,11 %, v drugih besedilih z večjim deležem tujk pa 0,26 oz. 0,34 oz. spet 0,26 %, več kot dvakrat toliko.

Besede so, kot tudi sicer v jezikoslovju, najzanimivejši predmet obdelave, so tudi osnovne enote besedilnih korpusov. Pri obdelavi gradiva se pojavijo mnoga vprašanja, ki jih na prvi pogled ne bi pričakovali. Tako npr. odločitev, kaj je beseda

in kaj ne, nikakor ni enostavna – napačen pristop lahko zelo poveča (in nasmeti) slovar besednih oblik. V tabeli 2 je naveden seznam pri obdelavi korpusnih besedil uporabljenih besednih enot (angl. *token*), ki so bile uporabljene pri obdelavi korpusnih besedil, s pogostnostmi.

Tabela 2: Pregled besednih enot v korpusu

	primer	različnih	vseh
enostavne besede	jezik	1.085.659	98.167.579
sestavljene besede	(pre)večkrat	61.373	272.919
enostavna števila	2003	42.946	1.451.863
časi športnih rezultatov	4;07:02	31.824	97.497
številski pridevniki	9-članski	9.776	81.526
spletni naslovi	www.daruj.com	7.757	13.193
elektronski naslovi	pisma@delo.si	508	1.479
datumi	6.febr.1902	3.166	15.839
denarne vsote	1.071.390,00	3.020	3.992
računalniške oznake	set.add("četrtek")	2.507	3.948
avtomobilske reg. št.	LJ 47-83P	1.924	2.013
oznake ISBN, ISSN, ISO	ISBN 86-11-15086-4	933	1.385
telefonske številke	031/589-863	827	1.201
UDK klasifikatorji	821.163.6-32	417	730
drugo	Čeljabinsk-70	26.947	261.135

Klasičnih besednih oblik, ki so sestavljene samo iz črk, je v celoti več kot 98 %, v seznamu različnih enot pa seveda nekoliko manj. Zanimiv je tudi razmeroma velik delež zloženih besed, med katerimi prevladujejo z vezajem povezana imena.

Za jezikoslovno rabo so najpomembnejše oblike enostavnih besed, pisanih z malo začetnico, kakršnih je 546.539, nekaj nad polovico celote. Ker besedila v korpusu še niso oblikoslovno označena, je mogoče razmeroma hitro poiskati le leme tistih besednih oblik, ki imajo samo eno lemo. Besedna oblika *boj* npr. ni taka, saj lahko izvira iz dveh lem, iz glagola *bati se* ali iz samostalnika *boj*. V tabeli 3 so navedene pogostejše leme pri tistih samostalnikih, pri katerih so imele ustrezne besedne oblike eno samo lemo.

Tabela 3: Najpogostejše samostalniške leme enolično določljivih besednih oblik po delih korpusa s frekvencami

	leposlovje		*polleposlovje*		*monografije*		*periodika*	
1.	roka	12088	dan	2172	organizacija	4051	država	179792
2.	oči	11575	čas	2062	telo	3115	leto	167640
3.	obraz	9059	življenje	1791	čas	2177	dan	139584
4.	človek	8291	človek	1619	oblika	1934	čas	130489
5.	glava	7966	ljudje	1435	leto	1806	mesto	119210

6.	srce	7786	stvar	1429	stoletje	1795	predsednik	103815
7.	čas	7380	knjiga	1279	skupina	1756	odstotek	98355
8.	beseda	7125	beseda	1178	primer	1710	zakon	90988
9.	ljudje	6781	leto	998	način	1687	konec	89851
10.	življenje	5905	pesem	906	subjekt	1639	tolar	85505
11.	hiša	5857	konec	847	cilj	1531	ljudje	82513
12.	glas	5595	narod	847	življenje	1525	milijon	76446
13.	pot	5242	pismo	769	mesto	1506	podjetje	76088
14.	otrok	5082	literatura	766	vprašanje	1486	skupina	68680
15.	noč	5011	srce	739	vrsta	1426	program	68480
16.	gospod	4961	mesto	732	proces	1359	primer	68295
17.	dan	4950	umetnost	722	ljudje	1358	minister	65389
18.	miza	4696	pot	705	podjetje	1351	družba	65055
19.	oče	4532	svet	629	površina	1303	vprašanje	64093
20.	okno	4340	roka	624	sistem	1293	vlada	63070

Seznami se precej razlikujejo med seboj, le dve besedi, *čas* in *ljudje* najdemo v vseh štirih stolpcih, in samo štiri v treh: *dan*, *leto*, *mesto* in *življenje*. Besedam iz prvega stolpca dobesedno lahko rečemo leposlovne, v stolpcu časopisnih in revijalnih besedil pa so v ospredju izrazi političnega in gospodarskega besednjaka. Revija *Monitor* je na vidnejše mesto pomagala predvsem besedi *program*.

V naslednji tabeli (4) so prikazane leme najdaljših besednih oblik samostalnikov, pridevnikov in glagolov vseh štirih delov korpusa. Upoštevane so le enostavne besede, brez vezajev in drugih posebnih znakov, pa tudi brez števk.

Tabela 4: Leme najdaljših enostavnih besednih oblik polnopomenskih besednih vrst po delih korpusa

leposlovje	polleposlovje	monografije	periodika
jugovzhodnovzhoden	institucionalizirati	abstraktnoekspresionističen	dermatokozmetologija
konstruktivističen	klaustrofiloksenofobija	evangeličanskoluteranski	desizopropilatrazin
kontrarevolucionar	klavstrofilofoboksenofilofobija	forenzičnostomatološki	elektrodistributer
kontrarevolucionaren	klavstrofiloksenofobičen	funkcionalnostrukturalen	elektroluminescenten
nedolžnoodkritosrčen	kolonialnoarestantski	kadrovskoinformacijski	hiperholesterolemija
prestolonaslednikov	kompozicijskotehničen	krščanskosocialističen	hiperlipoproteinemija
protiimperialističen	nacionalboljševističen	literarnoizdajateljski	hiposenzibiliziranje
protiracionalističen	nacionalsocialističen	nacionalnošovinističen	nacionalsocialističen
socialnoreformatorski	neinstitucionaliziran	narodnostnoohranjevalen	nevrofizioterapevtski
splošnoizobraževalen	podatkovnoinformacijski	organizacijskoračunalniški	ovolaktovegetarijanski
starojugoslovanski	socialističnokatoliški	političnoorganizacijski	transakcijskopsihoanalitičen
sviiiiiiiiiinjaaaaaaaaaa	socialističnorealističen	političnoraznarodovalen	videoelektroencefalografija
trebušnomišičnokrepilen	umetnostnozgodovinski	poslovnoorganizacijski	visokospecializiran
zgodnjepomladanski	znanstvenofantastičen	slovstvenofolkloristični	znanstvenofantastičen
zgodnjespomladanski	znanstvenoraziskovalen	znanstvenoraziskovalen	znanstvenoraziskovalen

Prevladujejo zloženi pridevniki, glagol je en sam, *institucionalizirati* pri polleposlovju, samostalnikov pa tudi ni veliko – v časopisnih in revijalnih besedilih slaba tretjina, 6, pri leposlovju in polleposlovju po 2, pri znanstvenih in strokovnih

monografijah pa ni celo nobenega. Opazimo tudi, da gre povsod pretežno za strokovne izraze, razen morda dveh pridevnikov in enega samostalnika pri leposlovju.

Seveda je veliko enostavnih besednih oblik, predvsem številkov, tudi daljših. Najdaljše besede, navedene so v tabeli 5, pa so pretežno zložene.

Tabela 5: Leme najdaljših besednih oblik v korpusu

belgijsko-francosko-slovensko-italijansko-britanski
bio-psiho-socialno-kulturno-zgodovinsko-ekonomsko-filozofski
boemsko-anarhistično-narodnjaško-boljševiško-individualističen
filmsko-kulturno-izobraževalno-festivalsko-dokumentaren
francosko-ameriško-slovensko-judovsko-madžarski
francosko-belgijsko-italijansko-angleško-slovenski
francosko-belgijsko-slovensko-angleško-italijanski
francosko-slovensko-italijansko-belgijsko-britanski
geografsko-zgodovinsko-filozofsko-kulturno-literaren
glicidilmetalkrilaten-etilen-dimetakrilaten
gorenjsko-ljubljansko-dolenjsko-notranjsko-belokranjski
kinotečno-dokumentarno-gejevsko-lezbično-filozofsko-muzikalen
pneumonoultramicroscopicsilicovolcanoconiosis
poveljniško-nadzorno-računalniško-komunikacijsko-obveščevalen
slovensko-angleško-nemško-francosko-italijanski
šestmilijontidvestotriindvajsettisočtristodvaintrideseti
športno-poslovno-loterijsko-medijsko-oglaševalski
turistično-trgovsko-garažno-olimpijsko-športen
2-(benziloksikarbonil)amino-3-dimetilaminopropenoat
2-(2-acetil-2-etoksikarbonil-1-etenil)amino-3-dimetilaminopropenoat

13 oblik med naštetimi dvajsetimi je sestavljenih iz 5 delov, ena celo iz desetih, le dve pa sta enostavni – številnik *šestmilijontidvestotriindvajsettisočtristodvaintrideseti* in pa medicinska diagnoza *pneumonoultramicroscopicsilicovolcanoconiosis*, ki ni bila uvrščena v tabelo 4 zaradi tega, ker ni poslovenjena. Najdaljša besedna oblika (izvzete so oblike, kjer je avtor zaradi posebnega poudarka daljšo zvezo ali celo celo poved napisal brez presledkov med besedami, npr. *kartugolčanjepijkleppovsodzdlgmorzusničjovšeklazno*, ki ga je *V Filisteji* zapisal Drago Jančar) je po pričakovanju kemična spojina, *2-(2-acetil-2-etoksikarbonil-1-etenil)amino-3-dimetilaminopropenoat*. Zanimivo je tudi, da so skoraj vse oblike iz tabele vzete iz časopisnega besedila – edina izjema je *geografsko-zgodovinsko-filozofsko-kulturno-literaren*, ki je vzet iz polleposlovnega besedila, iz eseja *V burji besed*, ki ga je napisal Drago Bajt.

Tabela 6: Leme na *jezič-* in *jezik-* s pogostnostmi v različnih delih korpusa

jezičast 4: A,P3
jeziček 455: A12, C3,P440
jezičen 97: A32, B2,C,P62
jezičiti P
jezičnež P
jezičnica 2: A
jezičnik 3: A2, P
jezičnodohtarski P
jezičnost 5: A, P4
jezik 24325: A1449, B512,C1259,P21105
jezikač 2: B, P
jezikanje 17: A2, B,P14
jezikast P
jezikati 12: A6,P6
jezikav 37: A7,P30
jezikavka 13: P
jezikavost 7: A, P6
jezikavt P
jezikljast A
jezikohitrec P
jezikolomen P
jezikomrznica P
jezikoslovček P
jezikoslovčev 2: P
jezikoslovec 464: A5, B8,C21,P430
jezikosloven 201: A3, B2,C26,P170
jeziko(slo)ven P
jezikoslovje 229: A, B3,C38,P187
jezikoslovka 12: P
jezikoslovno-literaren C
jezikoslovnostističen P
jezikoslovstvo B
jezikoslužen P
jezikotvornost P
jezikov 4: A,C2,P
jezikoven 3379: A31, B83,C285,P2980
jezikov(en) C
jezikovno-dramaturški P
jezikovnoestetski P
jezikovno-estetski P
jezikovno-historičen C
jezikovno-izrazen P
jezikovno-književen C
jezikovno-kompozicijski B
jezikovno-komunikacijski P
jezikovnokotičkarski P
jezikovnokulturen 14: P
jezikovno-kulturen 6: P
jezikovno-literaren 2: P
jezikovno-logičen P
jezikovnooblikoven 2: P
jezikovnoohranjevalen P
jezikovnopolitičen 8: P
jezikovno-političen P
jezikovnorazvojen P
jezikovno-semantičen C
jezikovnosintaktičen C
jezikovnoskladenjski C
jezikovnoslogoven P
jezikovno-stilen 3: B,P2
jezikovnostilističen 2: P
jezikovnostilski P
jezikovnostrateški P
jezikovnostrukturen P
jezikovno-tematski C
jezikovno-varnosten P
jezikovno-vsebinski 2: P
jezikovnozakonski P
jezikovnozemljepisen P
jezikovnozgodovinski 3: P
jezikovno-zgodovinski P

Zanimivi so tudi obsežni grozdi satelitov, ki jih za seboj potegnejo pogoste podstave. Primer je naveden v tabeli 6, kjer so naštete vse leme (71 = 62 + 9) na *jezik-* in *jezič-*, z absolutnimi pogostnostmi in pogostnostmi po posameznih zvrsteh. Pogostnost ni navedena, kadar znaša 1; če je večja, pa se je besedna oblika pojavila samo v enem delu korpusa, je pogostnost zapisana samo prvič. Kar 41 lem (58 %) iz seznama je bilo v korpusu zapisanih samo enkrat.

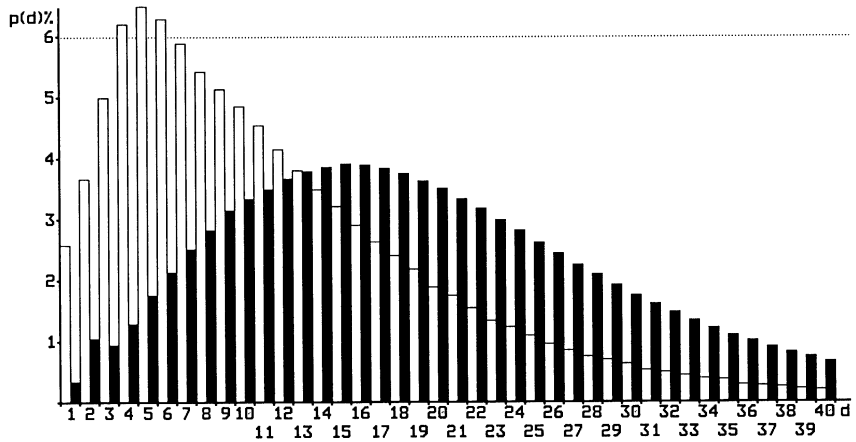
Jezik je izredno gibek in raznolik sistem, že pri besedah je večina, kot je npr. videti iz prejšnjega primera, bolj redko uporabljenih. Če se od besed premaknemo k besednim zvezam, začnejo pogostnosti še hitreje padati, pri povedih, katerih podatki so tudi zelo zanimivi, pa je že za bolj preproste kvantitativne analize potreben zelo velik vzorec. Iz tabele 7 vidimo, da je vseh povedi v korpusu slabih 6 milijonov, od tega dobrih 5 milijonov različnih – se pravi, da je velika večina enkratnic (angl. *hapax legomena*).

Tabela 7: Število različnih in vseh povedi po posameznih delih korpusa

	različnih povedi	vseh povedi
leposlovje	417.154	432.263
polleposlovje	60.139	63.302
monografije	85.952	89.787
periodika	4.680.225	5.169.663
Skupaj	5.192.713	5.701.224

Povprečna dolžina povedi v korpusu je torej nekaj manj kot 20 besed (17.5), se pa vrednosti po posameznih zvrsteh precej razlikujejo. Na sliki 2 so prikazane dolžine povedi v leposlovnih (beli stolpci) in časopisnih oz. revijalnih besedilih (črni stolpci).

Slika 2: Porazdelitev dolžin povedi v leposlovnih (belo) in časopisnih (črno) besedilih



Leposlovna besedila imajo, tudi zaradi večjega deleža dialogov, veliko krajše povedi od novinarskih prispevkov. Pri leposlovnih besedilih se krivulja od visoke začetne vrednosti 2,5 % za povedi z eno samo besedo hitro dviguje do vrha pri dolžini 5 besed z dobrimi 6 % povedi, potem pa se skoraj linearno spušča in pade pod 1 % že pri dolžini 25 besed na poved. Pri periodičnih besedilih je potek veliko pravilnejši, vrh dosežen pri 15 besed dolgih povedih (slabe 4 %), padanje pa naprej zelo počasno in enakomerno.

V zadnji tabeli so prikazane najpogostejše povedi iz prvega in zadnjega dela besedil - polleposlovnih besedil in besedil iz monografij za tak prikaz trenutno še ni dovolj. V glavnem gre za zelo kratke pritrdilne, nikalne in vprašalne povedi, v prvem stolpcu hitro prepoznamo tudi nekaj dramskih povedi (*Tišina.*, *Odide.*) in eno, ki jo takoj lahko pripišemo njenemu avtorju (*Hm, kajpak, se je popraskal kmet.*). V drugem stolpcu so štiri vsebinsko zelo zgovorne povedi: *Parlament. Vlada. Pihal bo jugozahodni veter. Sodba še ni pravnomočna.* Prvi dve imata opraviti z dnevno politiko, tretja z vremenskimi napovedmi, zadnja pa je iz sodnih klopi.

Tabela 8: Najpogostejše povedi iz leposlovnih in časopisnih besedil s frekvencami

	Leposlovje		**Periodika**	
1.	Da.	277	Zakaj?	1308
2.	Tako je.	200	Ne.	1009

3.	Ne.	174	Seveda.	448
4.	Kaj?	158	Da.	437
5.	Seveda.	150	Še več.	250
6.	Zakaj?	105	Ja.	247
7.	Ne!	94	Parlament.	243
8.	Vsekakor.	94	Tako je.	209
9.	Kako?	91	Ne vem.	206
10.	Dobro.	70	Res je.	205
11.	Ja.	69	Nasprotno.	201
12.	Tišina.	69	Dobili smo:	189
13.	Ne vem.	68	Vlada.	172
14.	Gotovo.	64	Vsekakor.	155
15.	Nič.	64	Kako?	151
16.	Odide.	53	Zakaj ne?	150
17.	Nikakor ne.	47	Kaj to pomeni?	125
18.	Prav.	40	Morda.	118
19.	Ali ne?	37	To je res.	115
20.	Kdo?	36	Pihal bo jugozahodni veter.	114
21.	Prav gotovo.	36	Nič.	107
22.	sem rekel.	36	Drži.	102
23.	Kam?	35	Sodba še ni pravnomočna.	101
24.	Res je.	34	Nikakor.	98
25.	Vem.	34	Ne, ne.	94
26.	Nujno.	33	Ne!	91
27.	sem nadaljeval,	33	Nikakor ne.	91
28.	Hm, kajpak, se je popraskal kmet.	32	Kako to?	87
29.	Lahko noč!	31	Mislim, da ne.	84
30.	Tako je!	29	In tako naprej.	80