Primož Jakopin
Ljubljana

UDK 811.163.6:004.5

O NEKATERIH KVANTITATIVNIH KAZALCIH V SLOVENSKEM JEZIKU

V prispevku je predstavljenih nekaj zanimivosti iz treh elektronskih besedilnih zbirk (korpus *Nova beseda* – http://bos.zrc-sazu.si): *Zbranih del Cirila Kosmača* (408.000 besed), *Zbranih del Ivana Cankarja* (2 milijona besed) in časopisa *Delo* 1998–2000 (47 milijonov besed). Prikazane so najpogostejše besedne oblike, samostalniki, povedi, porazdelitve dolžin povedi ter najdaljše povedi.

The article presents several interesting points from three electronic text collections (part of the *Nova beseda* corpus, http://bos.zrc-sazu.si): collected works by Ciril Kosmač (1910–1980, 408.000 words), collected works by Ivan Cankar (1876–1918, two million words), and *Delo* newspaper 1998–2000 (the leading Slovene daily, 47 million words). Top wordforms, nouns and sentences, distribution of sentence lengths, and the longest sentences are given.

1 Uvod

V zadnjih desetletjih so nastali t. i. besedilni korpusi, večje zbirke z besedili, urejene na enoten način in opremljene z internetnim iskalnim vmesnikom predvsem za prikaz konkordanc izbranih besed ali besednih zvez (npr. McEnery –Wilson 2001), najprej za angleški, potem nemški in francoski jezik, v zadnjih nekaj letih pa tudi že za naš jezik. *Nova beseda* (na spletnem naslovu http://bos.zrc-sazu.si) je korpus na Inštitutu za slovenski jezik ZRC SAZU, ki je imel na začetku, leta 1999, 3 milijone besed leposlovnih besedil, bil v letu 2000 (Jakopin 2001) s časopisnimi besedili razširjen na 48 milijonov, v letu 2002 pa je predvidena dopolnitev na 80 milijonov besed. Namenjen je tako za slovaropisno in drugo delo na Inštitutu, poleg tega pa tudi vsem ostalim, ki se ukvarjajo s proučevanjem slovenščine; korpus je prostodostopen in njegove spletne strani so doživele v dobrih dveh letih že več kot 25.000 obiskov.

V prispevku je predstavljenih nekaj zanimivosti iz treh delov korpusa, iz *Zbranih del Cirila Kosmača*, ki so že v korpusu, iz *Zbranih del Ivana Cankarja* (pripravljenih v letu 2001) in besedil dnevnika *Delo* (1998–2000).

Tabela 1: Pregled velikosti dveh leposlovnih opusov in zbirke časopisa *Delo*

	Ciril Kosmač	Ivan Cankar	*Delo* 1998–2000
znakov	2.497.308	11.821.000	307.396.000
besed	407.938	1.991.000	47.219.000
povedi	37.459	157.000	2.005.000

Zbrana dela Cirila Kosmača so že praktično brez napak, tako v besedilu kot pri oznakah, zato števila niso zaokrožena, pri Ivanu Cankarju in časopisu *Delo* pa števila še niso povsem dokončna in so zato zaokrožena na tisočice. Kosmačev opus obsega dobrih 400.000 besed, Cankarjev je približno petkrat večji, v zbranih besedilih časopisa *Delo* pa je letno zaobseženih okoli 16 milijonov besed. Gre za elektronsko izdajo časopisa *Delo*, ki zajema približno tri četrtine dnevne tedenske produkcije, od ponedeljka do sobote, brez reklam, malih oglasov, sporedov javnih prireditev in podobnega.

Tabela 2: Primer besednega iskanja v korpusu *Nova beseda*

nova beseda *iz Slovenije*

Besedno iskanje

obdob (11)

(Št.	**besedna oblika**	pogostost)
1.	obdobij	321
2.	obdobja	1842
3.	obdobje	3577
4.	obdobjece	1
5.	obdobjem	361
6.	obdobjema	3
7.	obdobji	72
8.	obdobjih	406
9.	obdobju	5261
10.	obdobnega	1
11.	obdobnih	2

V tabeli 2 je naveden primer iskanja po slovarju besednih oblik v korpusu *Nova beseda* za niz **obdob**, s krnjenjem spredaj in zadaj. Od skupaj 11 različnih oblik s skupno frekvenco 11.847 se jih 8 nanaša na besedo obdobje (skupaj 11.843). Najmočneje sta zastopana 3. in 5. sklon ednine z dobrimi 44 % vseh pojavitev.

Korpus je mogoče uporabiti tudi za ugibanje besed in v tabeli 3 je naveden primer izpisa konkordanc, dobljen z vrednostjo 1 pri iskalnem določilu *Skrita beseda*. Besedo naj ugane bralec sam, ni težko, še najbolj izdajalska pa je peta konkordanca.

Tabela 3: Primer izpisa konkordanc za neznano besedo

Uganete, kaj je skritega?

Darja Kocbek Z	. . .	sklepi vlade so bili določeni
odstotkov in z dobrimi	. . .	obeti zagotavlja krepko
decembrski zaslužek z	. . .	podražitvami življenjskih
z leto dni starejšimi	. . .	ne toliko, da bi z mirno
razlika med novembrskimi in	. . .	povprečnimi mesečnimi
so se jih tod šli pred	. . .	parlamentarnimi volitvami
letošnjih majskih z lanskimi	. . .	je bilo po računu urada
si je z nekaj uspešnimi	. . .	nastopi v celinskem pokalu
9 do 10 odstotkov nad	. . .	Terme Čatež so večinski
primerjavi z lanskimi	. . .	izplačili, ki so bila
primerjavi z lanskimi	. . .	pa za 4,9 odstotka.
primerjavi z lanskimi	. . .	drobnoprodajnimi cenami
Ljubljana – Med	. . .	predpraznič nimi nakupi
utrditev položajev pred	. . .	spopadi.Afriški in drugi
leti, ko je bil čas pred	. . .	prazniki tisti, ko so
praznujejo, je med tremi	. . .	brati dedek Mraz izgubil
nedeljskih dopoldnevih. Med	. . .	prodajnimi rekorderji
Henryja Sheltona je bilo med	. . .	ameriško-britanskimi zračnimi
primerjavi z lanskimi	. . .	cenami na drobno nižji
Moskva – V Rusiji so pred	. . .	parlamentarnimi volitvami
razorožitvi Start-2 pred	. . .	parlamentarnimi volitvami
avtorjevi režiji pred	. . .	prazniki). S. Pe.
(leva okolica	. . .	desna okolica)

2 Besedne oblike in samostalniki

Pri vsaki kvantitativni obdelavi besedila že v pripravljalnem obdobju, tudi če tega ne bi posebej želeli, ne moremo mimo seznama najpogostejših besednih oblik, če prej ne, potem med preverjanjem. Za vse tri opazovane dele korpusa so najpogostejše besedne oblike navedene v tabeli 4.

Tabela 4: Najpogostejše besedne oblike s pogostnostmi

	Ciril Kosmač		Ivan Cankar		*Delo* 1998–2000	
1.	je	25.798	je	123.281	je	1.570.404
2.	in	18.471	in	78.807	v	1.350.680
3.	se	13.330	se	56.101	in	1.171.370
4.	v	7.809	v	38.931	na	759.128
5.	da	5.412	da	34.552	za	656.814
6.	na	5.124	na	25.983	se	604.642
7.	pa	4.625	ne	24.117	da	583.347
8.	so	4.243	so	22.109	so	561.073
9.	ne	3.695	bi	21.678	ki	510.010
10.	bi	3.221	sem	19.675	pa	472.056
11.	z	3.181	ni	14.182	z	344.917
12.	ga	3.039	pa	13.796	tudi	332.187
13.	ki	2.995	kakor	13.528	bi	314.828
14.	po	2.948	ki	12.741	po	287.911
15.	sem	2.812	bil	12.469	s	281.805
16.	ni	2.636	z	12.107	ne	271.777
17.	s	2.441	še	11.754	bo	251.951
18.	še	2.410	mi	11.116	še	241.063
19.	za	2.327	za	10.719	kot	234.457
20.	tako	2.155	bilo	10.010	ni	190.827

V seznamih po pričakovanju ne najdemo besednih oblik, ki bi izvirale iz besed polnopomenskih besednih vrst, prevladujejo vezniki, predlogi, manj je zaimkov in oblik pomožnega glagola *biti*. Seznami se med seboj že na oko precej ujemajo, še posebej za oba leposlovna vzorca, pri katerih opazimo iste oblike na prvih šestih mestih. Med prvimi dvajsetimi oblikami pri Cirilu Kosmaču in Ivanu Cankarju najdemo 16 enakih, pri Kosmaču in časopisnemu vzorcu 17, še najmanj, le 15, pa po primerjavi najstarejšega in najnovejšega vzorca, Cankarjevega in časopisnega. Oblika pomožnega glagola biti, *je*, je edina nedvomno vedno na svojem odličnem mestu. Povsem verna primerjava z najpogostejšimi besednimi oblikami kakega drugega jezika ni mogoča, enolične preslikave prevodi pač ne morejo biti, je pa vseeno zanimivo pogledati, kaj je skupnega z najpogostejšimi oblikami iz npr. angleškega jezika. V *Britanskem nacionalnem korpusu* najdemo na prvih dvajsetih mestih naslednje besedne oblike: *the*, *of*, *and*, *a*, *in*, *to* (pri nedoločniku), *it*, *is*, *to* (kot predlog), *was*, *I*, *for*, *that*, *you*, *he*, *be*, *with*, *on*, *by* in *at* (Leech – Rayson – Wilson 2001). Struktura je podobna, le zaimkov je več in oblika glagola biti, *is*, je šele na osmem mestu.

Več kot najpogostejše besedne oblike o nekem besedilu povedo najpogostejše polnopomenske besedne vrste. Korpus *Nova beseda* še ni oblikoslovno označen, je le manjši njegov del (ki pa vsebuje *Zbrana dela Cirila Kosmača*) in zato so v prispevku, v tabeli 5, navedeni le najpogostejši samostalniki.

Tabela 5: Najpogostejši samostalniki s pogostnostmi

	Ciril Kosmač		Ivan Cankar		*Delo* 1998–2000	
1.	roka	1.577	oči	6.277	država	89.992
2.	glava	985	srce	5.454	leto	86.151
3.	oči	833	obraz	5.325	čas	65.842
4.	otrok	821	roka	5.216	mesto	61.790
5.	dan	749	človek	4.681	predsednik	55.119
6.	hiša	700	beseda	3.277	zakon	48.240
7.	leto	628	življenje	3.126	odstotek	48.227
8.	vrata	536	dan	2.858	dan	46.679
9.	beseda	513	glava	2.658	konec	45.951
10.	oče	486	ljudje	2.295	ljudje	43.499
11.	človek	453	pot	2.258	tolar	40.618
12.	glas	438	noč	2.235	stranka	37.571
13.	srce	424	gospod	2.216	milijon	37.268
14.	vas	414	čas	2.152	skupina	36.541
15.	obraz	396	lice	2.100	minister	35.129
16.	miza	392	cesta	2.063	podjetje	34.808
17.	noga	384	okno	2.036	vlada	34.550
18.	življenje	373	glas	1.981	primer	32.791
19.	ljudje	369	mati	1.909	vprašanje	32.763
20.	voda	362	miza	1.867	tekma	31.600

Pri Cirilu Kosmaču so frekvence točne, pri obeh ostalih vzorcih pa so ponekod le približne, previsoke. Lema *dan* nastopa pri Ivanu Cankarju npr. s trinajstimi oblikami: *dan* (1363), *dne* (126), *dneh* (141), *dnem* (62), *dneva* (50), *dneve* (11), *dnevi* (167), *dnevih* (8), *dnevom* (5), *dnevoma* (2), *dnevu* (115), *dni* (787) in *dnij* (21). Za vse razen za prvo (in najpogostejšo) lahko z gotovostjo trdimo, da izvira iz te leme, *dan* pa seveda lahko tudi iz glagolske leme *dati* (pri Cirilu Kosmaču oblika *dan* sicer nikoli ne nastopi v glagolski vlogi). Za primerjavo so navedeni še najpogostejši samostalniki iz *Britanskega nacionalnega korpusa* (spet Leech – Rayson – Wilson 2001): *time*, *year*, *people*, *way*, *man*, *day*, *thing*, *child*, *Mr.*, *government*, *work*, *life*, *woman*, *system*, *case*, *part*, *group*, *number*, *world* in *house*. Takoj lahko ugotovimo, da je tudi v BNC veliko časopisnega jezika (7 teh samostalnikov najdemo tudi pri časopisu *Delo* med prvimi dvajsetimi), vendar spet manj kot v *Novi besedi*. Ugotovimo tudi, da besedam, kot so *oči*, *obraz*, *srce*, *roka*, *glava*, mirno lahko rečemo leposlovne, saj jih med prvimi v časopisnem jeziku ne najdemo. Tam prevladujejo politične (*država*, *zakon*, *stranka*, *minister*, *vlada*) in gospodarske (*odstotek*, *tolar*, *milijon*, *podjetje*) teme.

3 Povedi

Če frekvence najpogostejših besed od prve naprej zelo hitro padajo, to seveda še bolj velja za besedne zveze, še bolj kot zanje pa za povedi. Vsi trije vzorci,

predvsem zadnji, so že toliko veliki, da se da nekaj malega pokazati tudi na ravni povedi. V tabeli 6 so navedene najpogostejše povedi iz vseh treh vzorcev.

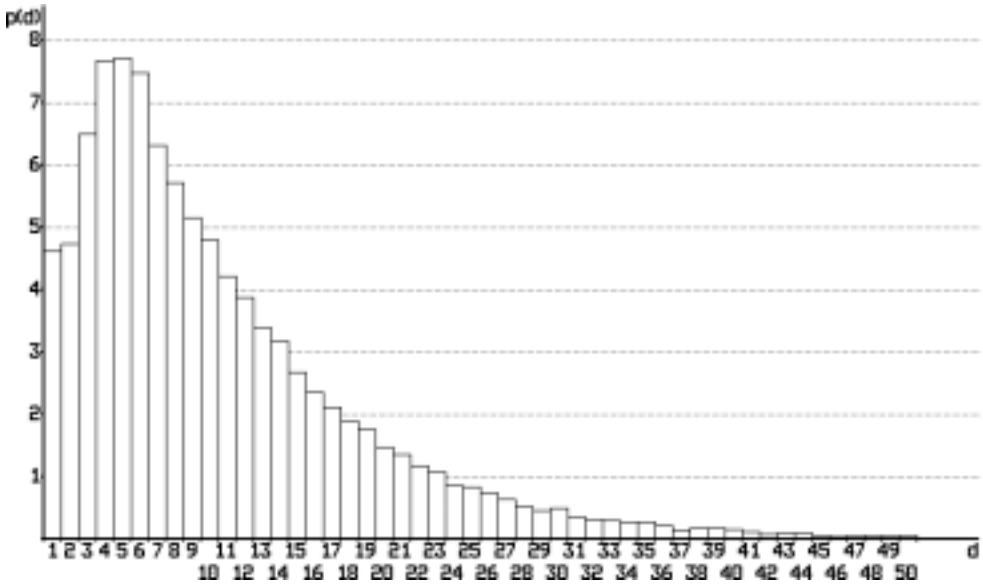
Tabela 6: Najpogostejše povedi s pogostnostmi

	Ciril Kosmač		Ivan Cankar		*Delo* 1998–2000	
1.	»Hm, kajpak,« se je popraskal kmet.	32	Kaj?	111	Ne.	396
2.	Ne!	30	Kam?	59	Da.	198
3.	Tišina.	24	Zbogom!	53	Seveda.	196
4.	Hm, kajpak.	22	Odide.	49	Še več.	131
5.	Kaj?	22	Ne!	47	Ja.	127
6.	Tak!	14	O!	34	Tako je.	93
7.	Aha!	14	Mati!	29	Res je.	91
8.	Hm?	12	Z Bogom!	27	Ne vem.	80
9.	Mhm ...	11	Ah!	25	Kako?	70
10.	Pha!	11	Lahko noč!	25	Zakaj ne?	66
11.	A?	10	Vstane.	21	Pihal bo jugozahodni veter.	64
12.	Seveda.	10	Kako?	19	Vsekakor.	64
13.	»Hm, kajpak,« se je popraskal.	9	Kdo?	19	Sodba še ni pravnomočna.	50
14.	»Prišla bo bridka smrt, moj hramček bo zaprt ...«	9	Kdo si?	18	Kaj to pomeni?	48
15.	Tak!	9	Vida!	17	To je res.	48
16.	Tako je to!	9	Hm!	17	In tako naprej.	43
17.	Zakaj?	8	Jezus!	16	Ne, ne.	42
18.	Drejc!	8	Tako je!	15	Nikakor ne.	39
19.	Tako je!	8	Kaj še!	14	Kako to?	38
20.	Kam?	7	Tako!	14	Ne!	36

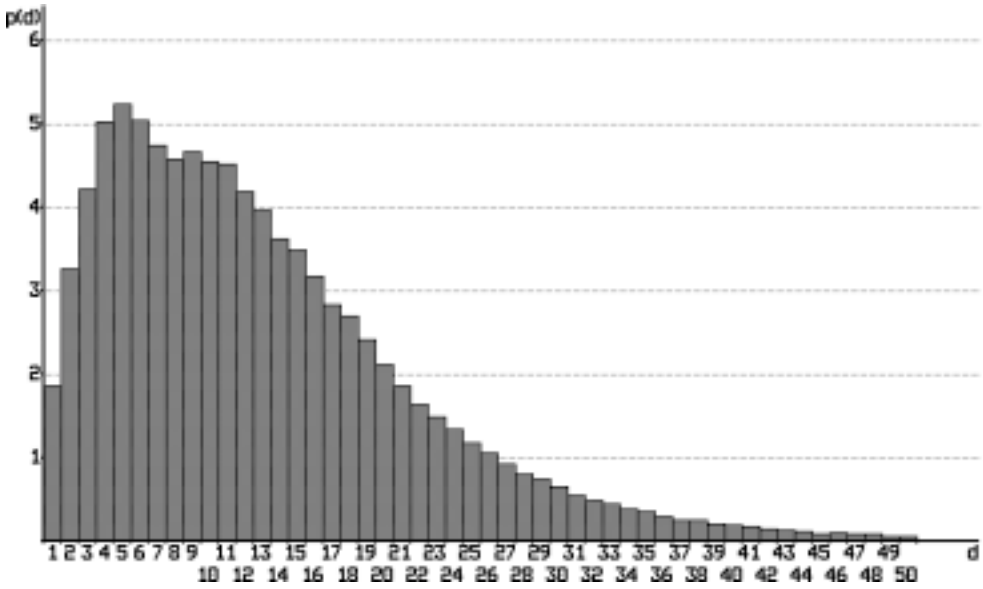
Prevladujejo kratke nikalne, trdilne ali vprašalne povedi, vsak vzorec pa ima pri tem svoje posebnosti. Pri Cirilu Kosmaču 4 najdaljše med najpogostejšimi povedmi hitro lahko povežemo z delom, iz katerega izhajajo (*Balada o trobenti in oblaku*), pri Ivanu Cankarju pa takoj opazimo, da najpogostejše povedi pripadajo dramskim besedilom (*Vstane.*, *Odide.*), včasih tudi konkretnim (*Vida!* iz *Lepe Vide*). Pri besedilih iz časopisa *Delo* sta le dve povedi med najpogostejšimi daljši kot tri besede. Prva: *Pihal bo jugozahodni veter.* je del vremenskih napovedi, druga: *Sodba še ni pravnomočna.* pa je tipična za poročila iz sodnih dvoran.

Na slikah 1, 2 in 3 so v obliki histogramov, pri katerih je na ordinatni osi pogostnost določene dolžine povedi v odstotkih, na abscisni osi pa dolžine povedi.

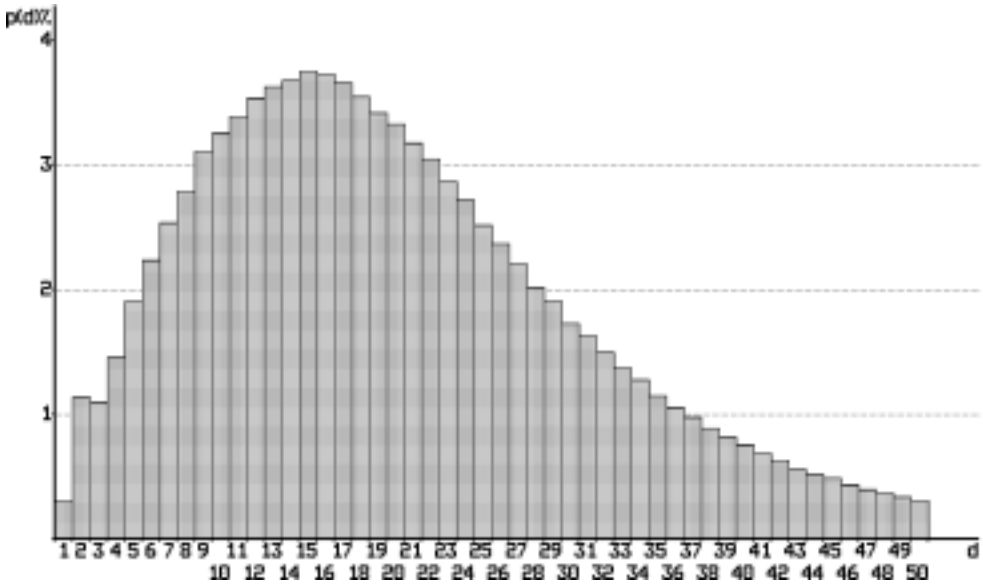
Povedi v delih Cirila Kosmača so večinoma kratke, največ je dolgih 4, 5 in 6 besed, s pogostnostjo približno 7,5 %, krivulja se od dolžine 6 naprej zelo hitro spušča, pri dolžini 24 pogostnost pade pod 1 %, povedi z dolžino nad 40 besed je zelo malo. Velik je tudi delež zelo kratkih povedi, dolgih eno in dve besedi, obojih nad 4,5 %.



Slika 1: Porazdelitev dolžin povedi v *Zbranih delih Cirila Kosmača*



Slika 2: Porazdelitev dolžin povedi v *Zbranih delih Ivana Cankarja*



Slika 3: Porazdelitev dolžin povedi v časopisu *Delo* 1998–2000

Tudi pri Ivanu Cankarju je največ povedi dolgih 4, 5 in 6 besed, le da je njihov delež manjši, rahlo nad 5 %. Padanje od dolžine 6 naprej ni niti hitro niti monotono, saj je povedi z dolžino 9 več kot tistih, ki so dolge 8 besed. Pravo padanje se začne šele pri dolžini 12. Hrbet krivulje je daljši, pod 1 % se krivulja spusti pri dolžini 27, pogostnost se močno približa ničli pri dolžini 50.

Za jezik časopisa *Delo* so značilne precej daljše povedi kot pri obeh leposlovnih vzorcih, krivulja pa ima z izjemo anomalije pri dolžini 2 (dvobesednih povedi je več kot trobesednih) zelo lep in pravilen potek, vidi se, da je bil vzorec za velikostni razred večji od prejšnjih dveh. Največ povedi je dolgih 15 in 16 besed, približno 3,5 %, vrh krivulje je širok in oba, tako vzpon kot padec, lepo položna. Pogostnost pade pod 1 % pri dolžini 37 besed, od ničle pa je pri dolžini 50 še precej daleč.

V nadaljevanju so navedene še najdaljše povedi iz vseh treh vzorcev. Pri vseh treh se vidi, da avtorjem ni šlo za kakšno postavljanje rekordov, najkrajša med njimi je Cankarjeva, najdaljša pa iz časopisa *Delo*.

Najdaljša poved iz *Zbranih del Cirila Kosmača* je dolga 284 besed, vzeta pa je iz novele *Hiša št. 14*:

> Na vse zadnje so se socialno čuteči, po sami rimski in božji previdnosti postavljeni tržaški občinski svetniki do grla nasitili, gledati in poslušati brezposelne, na cesto pognane družinske očete, zapuščene vdove in onemogle starce, ki so slabokrvni in jetični od lakote, garjavi, nadušljivi in polni revmatizma od prenočevanja po zatohlih kavernah in zasmrajenih mestnih podzemskih kanalih, skrhani in sključeni, razjedeni in oglodani od trdega življenja, vsak dan v večjih trumah prihajali na magistrat ter

krevljali po dolgih svetlih hodnikih, puščali na preprogah rjave odtise svojih blatnih, pošvedranih obutev, pobirali čike izza pljuvalnikov, se vsekavali kar s prsti, zamolklo kašljali, grkali, smrkali in metali široke pljunke kamor je naneslo, stresali bolhe, uši in stenice, zaudarjali po kislem smradu raztrganih in gnilih cunj, prijemali s svojimi umazanimi lopatastimi rokami za svetle medene kljuke na vratih, vstopali v sobe in se raztezavali po njih kakor smradljiv, dušeč dim, se obešali za suknje gospodom uradnikom, ki so tako lepo dišali po milu, se odkrivali, držali raztrgana pokrivala z obema rokama nizko na kolenih ter dokazovali, da niso divjaki, temveč ljudje, ustvarjeni po božji podobi in volji, trdili, da so Italijani od pamtiveka, fašisti od vsega začetka, državljani, udani Mussoliniju, kralju in papežu, ter prosili košček strehe svojim prezeblim otrokom in bolnim ženam: saj morajo vendar živeti, dokler jim je usojeno; ali naj se mar lepo zleknejo ob cesti in mirno počakajo na ljubo smrt; ali naj se pobesijo na hlačne jermene; ali naj svoje otroke pobijejo, pokoljejo, pomečejo v morje; ali naj noseče žene usmajajo ponoči po cestah in ponujajo po dve liri svoje kosti – o, saj bi jih, toda nihče se ne zmeni zanje, ker je dovolj bolje ohranjenih žensk; ali naj začnejo krasti, pobijati – ali kaj?

Najdaljša poved iz del Ivana Cankarja je dolga 219 besed in izhaja iz *Knjige za lahkomiselne ljudi*:

Gladile so počasi, z lahnimi, ljubeznivimi, kakor očetovskimi pogledi njene bujne, žarečerdeče lasé, nato so drsali pogledi njegovi navzdol, nekoliko vznemirjeni, nekoliko trepetajoči, dol po nizkem, belem čelu, pobožali so mehke, ozke, lahno obokane obrvi, napotili se počasi na desno stran, dol k drobnemu ušescu, pol zakopanemu pod težkimi, rdečimi kodri, ukvarjali so se tam za trenotek s svetlim smaragdom, ki si je bil izbral svoje mesto na tistem drobnem belem ušescu, in potem so hiteli, oprezno in boječe, a pol že pijani od ljubezni, od poželenja preko mehkih gorkih, belih lic mimo čudovito ustvarjenega, prav majhno, neopazno vzbočenega noska do tiste neizrekljivo umetno zarezane črte, do lahko privzdignjene zgornje ustnice, ozrli so se spotoma z zadoščenjem in radostjo na vrsto krepkih belih zob, ki so gledali skrivaj izza pol odprtih, vročih, napetih ustnic, in vsesali so se naposled v te pol odprte, vroče, napete ustnice, pili so do pijanosti ter zdrknili pijani in polnočno razposajeni dol preko ozke, okrogle, bele bradice – najlepše, beloblešče skalice na svetu –, dol na vrat, na rame, na polne rame, na kraljevski vrat, in pili so ter se igrali, smejali se in se spotikali, in jecali od nečistega veselja ter se izgubili navsezadnje utrujeni od poželenja in nezavedni čez valovito in valujoče belo brdo med dvoje polnih deviških, v tihem spanju se zibajočih prs ...

Najdaljšo poved v časopisu *Delo* v obdobju od leta 1998 do 2000 je napisal dr. sc. med. Janez Rugelj iz Ljubljane kot del članka z naslovom *Replika na zdravljenje neplodnosti*. Dolga je 453 besed, objavljena pa je bila 10. julija 2000:

Podpisnice in podpisniki izjave niso upoštevali naslednjih okoliščin: izhodišče za produktivno starševstvo je po vsem civiliziranem svetu enako: otroke naj zaplodita in rojevata starša, ki sta vsestransko sposobna za zelo zahtevno nalogo, tj. vzgojo otrok za življenje, ki je danes iz dneva v dan bistveno drugačno od življenja staršev – oba starša morata torej biti zreli osebnosti (na ravni genitalne stopnje razvoja), kar pomeni, da se morata spolno zelo privlačiti in da morata imeti tudi ustrezne socialne

in gmotne razmere za produktivno starševstvo; res je, da se visok odstotek partnerstva in starševstva spridi in da so po razidu staršev otroci praviloma prepuščeni »vzgoji« zapuščenih in zagrenjenih mater, ki so se pač izkazale, da niso bile sposobne dobiti in obdržati ustrezne partnerje, ampak to še ne pomeni, da moramo spodbujati celo jalove samske ženske, da rojevajo otroke in jih načrtno prikrajšajo za najpomebnejše življenjsko izkustvo – identifikacijo z očetom; celostna vzgoja otrok je možna izključno pod taktirko vertikalne avtoritete moškega, ki ima v vlogi reprezentanta družbene realitete ustrezne prerogative za prenos številnih znanj iz prejšnjih generacij na naslednjo, medtem ko so matere nepogrešljive asistentke pri vzgojnem procesu, zlasti v pokrivajoči se vlogi (glej na koncu stališče Luigi Zoje o brezpogojni nuji vertikalne avtoritete moških za normalen razvoj otrok); s histeričnim feminizmom ženske kvarijo celotno družbo, saj se moški praviloma izognejo poskusom prevzgajanja s strani žen (ki so čedalje bolj obsedene s spraševanjem, poučevanjem, nadzorovanjem moža ...) tako, da se podajo na pot »odsotnega moža in očeta; odrasel človek, ki ne živi v partnerstvu, je enigma, ker ni nikogar, ki bi ga lahko celostno definiral, saj lahko posameznika celostno definira edino seksualni partner; vsak sposoben moški, ki je dosegel genitalno stopnjo razvoja, ima po 17. letu starosti dekle oziroma partnerko, ali celo eksperimentira z večjim številom kandidatk za morebitno partnerstvo, kar je povsem naravno in celo zaželeno – enako velja za ženske, saj tiste, ki so dosegle genitalno stopnjo razvoja in so vsestransko prebujene ter znajo spuščati Amorjeve puščice (torej se znajo »ponujati«), imajo oboževalce in si lahko med njimi izberejo »žrtve« za »oplojevalce«, če so se pač odločile, da bodo rodile brez vstopa v zakonski pristan, torej da bodo zavestno prikrajšale otroka za doživljanje polnega očetovstva; predvsem nevrotične (torej bolne) ženske niso sposobne (preprosto niso privlačne in so zato v temelju zagrenjene) dobiti moškega za partnerja ali zgolj za »oplojevalca« oziroma so nekatere tako nevrotične, da niso plodne – zavedati se moramo, da so številne aktualno »jalove« ženske takšne zaradi njihove nevrotične situacije, saj v takšnih primerih narava sama poskrbi, da seme pač ne pade na plodna, pač pa na jalova tla – dogaja se, da zaradi nevroze aktualno jalova ženska brez težav zanosi, ko je vsaj za silo pozdravila nevrozo v skupini za pripravo posvojiteljev in dobila otroka v posvojitev.

4 Zaključek

V prispevku je nanizanih nekaj drobnih zanimivosti, ki s kvantitativnega vidika osvetljujejo opusa dveh vidnih slovenskih pisateljev 20. stoletja in zbirko časopisnih besedil iz našega največjega internetnega korpusa s prostim dostopom. Navedeno daje približen občutek, kakšen bo lahko pravi kvantitativni oris slovenskega jezika, ko bodo, predvidoma v ne več tako oddaljeni prihodnosti, besedila iz korpusa tudi oblikoslovno označena.

Literatura

Primož JAKOPIN, 2001: Beseda: a Slovenian text corpus. Ur. M. Fraser – N. Williamson – M. Deegan. *Digital Evidence: selected papers from DRH2000*, *Digital Resources for the Humanities Conference*, *University of Sheffield*, *September 2000* (Office for Humanities Communication publication 14). London: Office for Humanities Communication. 229–241.

Geoffrey LEECH – Paul RAYSON – Andrew WILSON, 2001: *Word Frequencies in Written and Spoken English: based on the British National Corpus*. London: Longman.

Tony MCENERY – Andrew WILSON, 2001: *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

SOME QUANTITATIVE DESCRIPTORS OF SLOVENIAN LANGUAGE

SUMMARY

The paper brings a comparison of three text samples. The first contains the collected works by Ciril Kosmač, one of the leading Slovene fiction authors from the mid-20th century, the second has been made up of works written by Ivan Cankar, the great master of Slovene fiction, and the third brings a major part of texts from the electronic edition of the leading Slovene daily, *Delo*, as made available to the blind.

Table 1: Sample size overview

	Ciril Kosmač	Ivan Cankar	*Delo* 1998–2000
bytes	2.497.308	11.821.000	307.396.000
words	407.938	1.991.000	47.219.000
sentences	37.459	157.000	2.005.000

The texts are part of the *Nova beseda* corpus (http://bos.zrc-sazu.si) compiled at the Fran Ramovš Institute of the Slovene Language, which is part of the Scientific Research Centre of the Slovene Academy of Sciences and Arts (ZRC SAZU). Since 1999, it has been made generally available via the Internet and has been accessed over 25.000 times.

In the first part of the paper the corpus is briefly introduced: besides a concordancer, which can also be used to guess a word or a word sequence, it also offers a very broad wordform search capability.

The second part is devoted to wordforms and nouns. While the table of top twenty wordforms shows good matching, especially in the two fiction samples where we find the same wordforms in the first six positions, the most common nouns, given in Table 2, bring a very different picture. The nouns from five text collections, with frequencies normalized to the ratio of »per million running words« for easier comparison, clearly show a great difference between fiction and newspaper vocabularies (words such as *eyes*, *heart*, *head*, *hand*, *word* are in sharp contrast to *state*, *law*, *president*, *minister*, or *government*). Besides the nouns from the three observed samples, the last two columns represent the nouns from the British National Corpus (100 million words) and the Internet pages in Slovene, collected at the main Slovene web index (http://www.najdi.si) by Noviforum, Ljubljana (460 million words).

Table 2: Top twenty nouns

	Ciril Kosmač		Ivan Cankar		*Delo* 1998–2000		BNC		NAJDI.SI	
1.	hand	3,866	eyes	3,153	state	1,906	time	1,833	article	2,107
2.	head	2,415	heart	2,739	year	1,824	year	1,639	page	1,666
3.	eyes	2,042	face	2,675	time	1,394	people	1,256	day	1,526
4.	child	2,013	hand	2,620	city,place	1,309	way	1,108	year	1,267
5.	day	1,836	man	2,351	president	1,167	man	1,003	work	1,252
6.	house	1,716	word	1,646	law	1,022	day	940	world	1,073
7.	year	1,539	life	1,570	percent	1,021	thing	776	time	827
8.	door	1,314	day	1,435	day	989	child	710	law	799
9.	word	1,258	head	1,335	end	973	Mr.	673	group	790
10.	father	1,191	people	1,153	people	921	government	670	contribution	776
11.	man	1,110	way,path	1,134	tolar	860	work	653	system	773
12.	voice	1,074	night	1,123	party (pol.)	796	life	645	city	717
13.	heart	1,039	Mr.	1,113	million	789	woman	631	connection	690
14.	village	1,015	time	1,081	group	774	system	619	data item	680
15.	face	971	cheek	1,055	minister	744	case	613	school	638
16.	table	961	road	1,036	enterprise	737	part	612	community	608
17.	leg	941	window	1,023	government	732	group	607	right	600
18.	life	914	voice	995	case	694	number	606	use	559
19.	people	905	mother	959	question	694	world	600	court	558
20.	water	887	table	938	sports match	669	house	598	change	556

All the nouns from the Slovene text collections are rough English translation equivalents of the original terms – *človek* in Slovene, translated as *man* in columns 1 and 2, could also be translated as *human*.

The distribution of sentence lengths in the third chapter clearly shows how the curve regularity depends on the sample size. In Figure 1, the distribution of sentence lengths for the newspaper sample (2 million sentences) is shown.

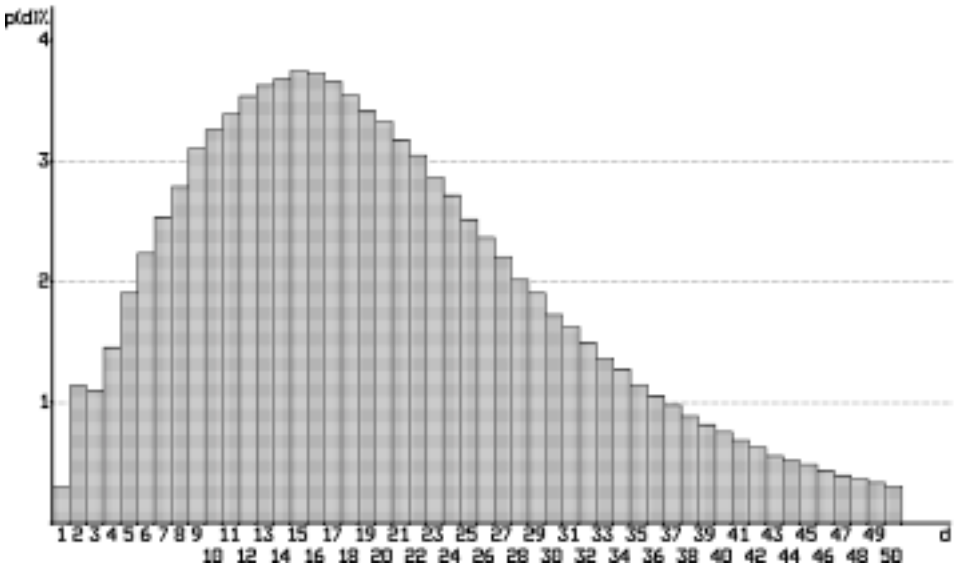


Figure 1: Distribution of sentence lengths, *Delo* 1998–2000