UDK 808.63-3(049.3)
Primož Jakopin
Filozofska fakulteta v Ljubljani

NEKAJ ŠTEVILK IZ *SLOVARJA SLOVENSKEGA KNJIŽNEGA JEZIKA*

Gre za kvantitaven opis *Slovarja slovenskega knjižnega jezika* (izdaja 1994): obsega 93.151 gesel, dolg pa je 23.346.100 znakov. Najprej je navedena zastopanost besednih vrst, sledijo podatki o iztočnicah (geslih), porazdelitve črk in dolžin. Poseben poudarek je na informacijski vsebnosti ali entropiji, vrednoteni na n-terčicah do n=8, zaporedjih sosednjih črk v iztočnicah, pri ponazarjalnem gradivu, iztržkih in navedkih v razlagah gesel pa tudi preko besednih mej.
Bistvena novost v prispevku so krivulje rasti n-terčic in možnost napovedovanja ustreznih velikosti vzorca z besedilom, ki sledi iz njih.

A quantitative description of the *Dictionary of the Slovenian Literary Language* (93.151 word entries with a total length of 23.346.100 bytes), obtained during the preparation of its electronic edition, is given. Summary data on word types and on entries are followed by distributions of letters, words and word lengths. Information contents or entropy for character n-tuples, sequences of adjacent letters, up to n=8 has been computed and graphed, for headwords and for text in quotations.
Highly interesting is an estimate of the required size of text samples, derived from the smoothness of n-tuples growth curves, as n increases.

Od jeseni leta 1992 do jeseni 1994 je bil vir za ta prispevek, Slovar slovenskega knjižnega jezika (SSKJ 1994), ki je pred tem v knjižni obliki, v petih knjigah, izhajal od 1970 do 1991, prenesen na računalniški medij. Prenos je bil potreben za izdajo slovarja v eni knjigi konec 1994; izdaja na disketah in CD-ju je v načrtu za drugo polovico letošnjega leta.

Vse gradivo je torej dostopno v obliki, ki jo lahko obdelujemo z računalniki; in če je kaj resnice v trditvi, da ti stroji danes kaj preveč pametnega še ne znajo, je po drugi strani tudi res, da so hitri in neutrudni in da z njihovo pomočjo lahko izbrskamo na dan marsikatero zanimivost, ki bi je sami ne mogli. Navadnim smrtnikom dostopni računalniki so v letu 1994 postali tudi že tako zmogljivi, da z njimi lahko obdelujemo cel slovar v enem kosu. Tako se z večjim veseljem lotimo nalog, ki bi jih morali prej narediti po delih, kar je tudi z računalnikom zamudno opravilo in terja veliko dodatnega preverjanja.

Prešteti in izračunati se da marsikaj, in če gremo zelo v širino, se lahko rezultati proučevanemu delu po obsegu približajo ali ga celo presežejo. Konkordančni slovar Prešernovih poezij je recimo dosti debelejši od njih samih, stvar pa postane temu primerno manj prebavljiva in privlačna. V tem prispevku je bila izbrana srednja pot: izbor tistega, kar je bilo izvedljivo, kar je zanimalo avtorja in kar je

bilo videti zanimivo še za druge.

Poglejmo najprej nekaj splošnih podatkov o slovarju (vsa števila razen pri iztočnicah so zaokrožena na stotice):

Preglednica 1: Število iztočnic, besed in črk

93.151	iztočnic in 13.888 podiztočnic		
362.100	različnih besednih oblik		
3.343.700	vseh besednih oblik		
23.346.100	črk, številk in ločil		
17.517.300	črk, med njimi	271.200	naglašenih ter posebnih črk
5.771.600	ločil, od tega	3.416.900	presledkov,
		1.223.000	sprememb oblike črk in
		57.200	številk

Število iztočnic ne potrebuje posebne razlage. Z besednimi oblikami so označene besede v vseh svojih pregibnih inačicah, npr. *okusen* (2), *okúsen* (2), *okusna* (19), *okusne* (2), *okusnejša* (2), *okúsnejši* (1), *okusni* (6), *okusnih* (2), *okusnim* (1), *okusno* (22); v oklepajih so navedene pogostnosti oblik v ponazarjalnem gradivu slovarja.

Da bi laže pojasnili statistiko črk in ločil, poglejmo najprej nabor znakov, ki je bil uporabljen pri prenosu slovarja v računalniško obliko (pred vrsticami so navedene kode znakov, kot jih je videl urejevalnik):

Preglednica 2: Urejevalniški nabor znakov

```
0-12   : ∘⇧⇩⇨⇦⧗ HI ⧗ ✓◷▲↕ F
13-28  : DOUBLE DOUBLE CONDEN CONDEN EMPHAS EMPHAS STRIKE STRIKE UNDERL UNDERL ITALIC ITALIC SUPERS SUPERS SUBSCR SUBSCR
29-32  : ˘ ‿ ‿
33-64  : !"#$%&'()*+,-./012345́6789:;<=>?@
65-96  : ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`
97-127 : abcdefghijklmnopqrstuvwxyz{|}~△
128-154: ÇüéâäàåçêëèïîìÄÅÉæÆôöòûùÿÖÜ
155-181: ¢£¥ßƒáíóúñÑČčŠšŽžĆćĐđĲĳØøœŒ
182-189: ⁑⁑≍‿◆▌▪©
190-214: ẹọŕâệêîộôûr̂äëïöüāẹ̄ēīọ̄ōūr̄
215-225: αβγδεηϑλπσω
226-252: °-–•◆•●u̯‼«»Íaàāhmḿm̂´śńýÁźżẑžẑm̈m̀řwŕ
258-281: °Óğš̀ṡšŽ‰ı ́^ ̂ ̈ ̑ ̄ ʼ˜̃˜±ˆeoÚÏ⁵⁄₇
```

Vsi znaki na papirju niso bili vidni – tiste s kodami od 0 do 12 in od 182 do 188 je imel urejevalnik za lastno uporabo; znaki od 29 do 32 so različno široki presledki: prvi trije so skrbeli za to, da je bil prazen prostor med besedami pri poravnavi desnega roba čim enakomerneje razporejen, 32 pa je navadni presledek. Posebne pozornosti so vredni znaki od 13 do 28, ki so določali spremembe tiska – iztočnice so bile recimo tiskane krepko, podiztočnice polkrepko, razlage v ležečem tisku, stranske razlage v drobnem ležečem, kvalifikatorji in dodatna slovnična opozorila v drobnem tisku, ponazarjalno gradivo v navadnem tisku (glej tudi SSKJ 94, str. XLIII):

rújen -jna -O prid. (ū) ekspr. **1.** v zvezi z **vino** *dober, plemenit:* kozarec rujnega vina **2.** knjiž. *zlato rumen:* rujna zarja • knjiž. dekle z rujnimi lici *rdečimi*

Gornje besedilo je bilo v računalniku shranjeno nekoliko drugače:

```
rújen -jna -o prid. (ū) ekspr. 1. v zvezi z vino
dober, plemenit: kozarec rujnega vina 2. knjiž. zlato
rumen: rujna zarja • knjiž. dekle z rujnimi lici rdečimi
```

Spremembe vrste tiska, označene s kodami od 13 do 29, so delale pri prenosu slovarja v računalniško obliko tudi največ težav. Bilo jih je veliko, iz preglednice 1 vidimo, da kar 1.223.000 ali skoraj 700 na stran, in prav pri njih so že gotovi programi za strojno razpoznavanje besedila, s katerimi so bili napravljeni začetni poskusi (OmniPage, Lecturus), najbolj šepali. Ni šlo drugače, program za razpoznavanje, po meri slovarja, je bilo treba šele napraviti.

Znaki s kodami od 33 do 127 so velike in male črke angleške abecede, številke ter ločila po ameriškem standardu ASCII. Znaki od 128 do 181 so samoglasniki z naglasi in druge posebne črke, ki so v rabi v abecedah zahodne in severozahodne Evrope (t. i. razporeditev IBM), nekaj valutnih simbolov, proti koncu pa še sičnike in šumevce, ki so v rabi pri nas in pri naših vzhodnih sosedih. Od 182 do 189 najdemo spet nekaj znakov, ki jih potrebuje urejevalnik zase ter simbol za oznako avtorskih pravic. Sledijo znaki za izgovarjavo, od 215 do 225 pa tiste grške črke, ki so se pojavile v SSKJ. Od 226 do 281 najdemo tudi oznake za izgovarjavo, nekaj ločil, npr. veliko in malo izrazijsko (terminološko) in frazeološko gnezdo, dvojni vprašaj, ki je pri strojnem prepoznavanju označeval neznano črko, stopinje Celzija in stopnje grafičnih formatov z zelo podobnim znakom, pa tudi simbola, ki ju je uporabil F. Prešeren v znameniti slovničarski zbadljivki, pri tem, kako se pravilno piše kaša. Kot je videti, si znaki nekako od

kode 190 naprej sledijo brez pravega reda – v nabor znakov, ki prej ni bil točno znan, so prihajali sproti.

Za naborom znakov si oglejmo besedne vrste iztočnic.

Preglednica 3: Besedne vrste gesel

Samostalnik	51448	
m:	21823	m: 21676, m ed. in dv.: 8, m mn.: 49, m neskl.: 84, m ed. neskl: 1, m mn. neskl.: 2, m neskl. in ž neskl.: 3
ž:	21427	ž: 21204, ž im. in tož. ed.: 1, ž mn.: 181, ž neskl.: 37, ž ed. neskl.: 1, ž mn. neskl.: 3
s:	8198	s: 8155, s mn.: 42, s neskl.: 1
Pridevnik	21516	
prid.:	21456,	prid. neskl.: 60
Glagol	16479	
dov.:	9270	dov.: 8569, dov. in nedov.: 701
nedov.:	7209	nedov.: 6537, nedov. in dov.: 672
Zaimek	130	
zaim.:	125	zaim. neskl.: 5
Števnik	122	
štev.:	90	štev. neskl.: 32
Prislov	1325	
Predlog	115	
Veznik	76	
Medmet	615	
Člen in členek	9	
člen:	1	členek: 8
Predpone	406	
		predpona: 19, predpona v im. sestav. in prislovih: 1, predpona v sestavljenkah: 58, prvi del zloženk: 326, prvi del zvez: 2
Druge oznake	910	gl.: 835, prim: 2, neskl. pril.: 67, opisni deležnik od: 1, prihodnji čas od: 1, rod., tož. od: 1, rod., tož. mn. od: 1, tož. od: 1, velelni naklon od: 1

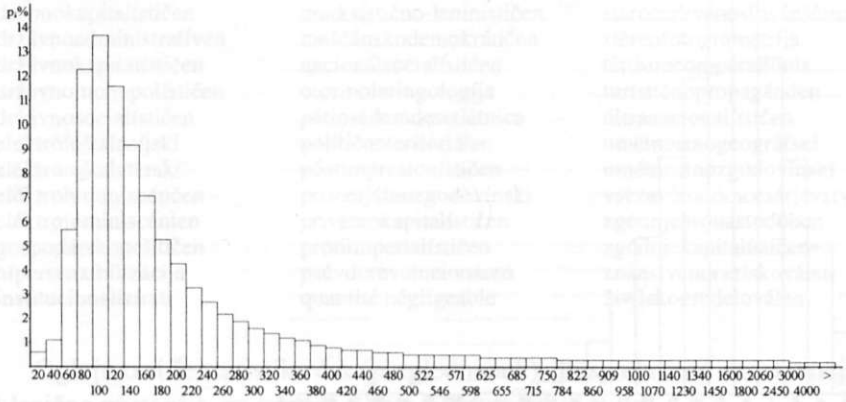
Kakor je bilo pričakovati, med iztočnicami prevladujejo samostalniki, moških je za malenkost več kot ženskih, za njimi pridejo pridevniki in glagoli, drugih besednih vrst pa je bistveno manj. Kot zanimivost velja tu omeniti, da iz 54.522 samostalnikov (iztočnice, dvojnice, podiztočnice) po pregibanju dobimo 468.281 besednih oblik, iz 22.861 pridevnikov pa 277.831; pri tem so oblike, ki so sicer iz istih črk in se razlikujejo samo po naglasu, štete kot različne.

Dolžina razlag: večina iztočnic ima nekaj vrstic dolgo razlago, nekatere prav

kratko, nekaj pa se jih je sestavljalcem slovarja zdelo ali tako težkih ali tako zanimivih, da so jim posvetili kar po nekaj strani. Naslednji podatek in slika 1 sta vsak po svoje zgovorna:

241 črk – povprečna dolžina razlage iztočnic (v črkah)

Slika 1: Porazdelitev dolžin razlag iztočnic



Iz histograma, kjer so pod stolpci napisane zgornje meje razredov, ugotovimo, da je povprečje zelo varljivo. Dolžine razlag so neenakomerno porazdeljene – največ iztočnic, skoraj 14%, ima razlage dolge od 81 do 100 črk, razred s povprečjem, zgornjo mejo 240 črk je pa že precej nizko na strmini. Rep je dolg in razredi na njem zato od 500 naprej tudi niso več enako veliki: v zadnjem so iztočnice, ki imajo razlage daljše od 4000 znakov.

V naslednji preglednici so navedene iztočnice, ki imajo najdaljše razlage – te so spet napisane za vsako iztočnico in pomenijo dolžino razlage v črkah, mednje pa so šteti tudi presledki in ločila:

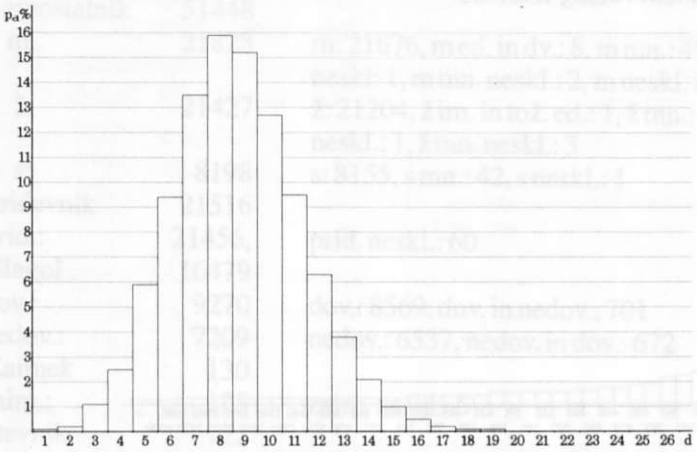
Preglednica 4: 20 iztočnic z najdaljšimi razlagami

príti	19.507	pásti	10.761
íti	17.398	za	10.479
takó	15.958	okó	9.919
vzéti	12.105	rôka	9.776
iméti	12.056	držáti	9.499
rêči	11.684	ujéti	9.389
jêzik	11.448	gláva	9.346
tá	11.375	dóber	9.336
odpréti	10.936	têžek	9.055
dáti	10.801	beséda	9.054

Glagolov je tu kar 10, 5 samostalnikov, samo 2 pridevnika ter po 1 prislov, zaimek in predlog.

Oglejmo si še dolžine iztočnic:

Slika 2: Dolžine iztočnic



Povprečna dolžina znaša 8.7 črke, največ, 14.849 ali skoraj 16% je dolgih 8 črk, od 20 črk naprej jih je pa že zelo malo. Podrobneje si podatke iz histograma lahko pogledamo še v preglednici, kjer pogostnosti niso v odstotkih, ampak absolutne:

Preglednica 5: Dolžine gesel

1	54	6	8790	11	8806	16	465	21	12
2	166	7	12552	12	5872	17	229	22	8
3	1029	8	14849	13	3528	18	114	23	1
4	2313	9	14126	14	1917	19	53	24	3
5	5490	10	11791	15	951	20	31	25	1

Od najkrajših iztočnic so zanimive dve črki dolge, različno pisane so v preglednici 6, pri najdaljših, 57 je dolgih 20 črk ali več, pa v preglednici 7 vidimo, da je kar 41 pridevniških zloženk, le 12 samostalnikov in štirje glagoli:

Preglednica 6: Iztočnice, dolge 2 črki

àd	às	br	čá	do	es	gó	hm	in	ká	kò	li	mm	nì	òh	ór	pí	ré	só	št	tk	TV	úm	XY
àh	áu	bú	če	èh	fá	há	hó	iz	ka	ks	mà	mú	no	òj	ós	pk	Rh	st	tá	tó	úd	úp	za
àj	áv	bù	čì	éj	fc	hé	hò	já	kh	lá	ma	na	nò	ól	òv	pò	-t	sú	tà	ts	ùf	úš	zk
ák	bà	bz	dà	èn	fí	hí	hú	jà	ki	là	mê	ná	nù	om	pa	po	-ž	šc	ta	tt	ùh	vé	zz
ár	bê	cì	da	ép	ga	hì	ìh	jo	ko	le	mh	nà	ob	òn	pá	pš	se	šè	tè	tù	ùj	ví	žé
ás	bi	ck	dó	ês	gá	hk	íl	jó	kó	lè	mí	nè	od	òp	ph	pù	si	šš	tí	tú	úk	vš	

Preglednica 7: Iztočnice, dolge 20 črk ali več

anárhoindividualístičen
anárhosindikalístičen
buržoáznodemokrátičen
buržoáznonacionalístičen
deprofesionalizácija
dialéktičnomaterialístičen
disproporcioníranost
drôbnokapitalístičen
držávnoadministratíven
držávnokapitalístičen
držávnomonopolístičen
držávnosocialístičen
eléktroinštalacíjski
eléktroinštalatêrski
eléktroluminiscéncen
eléktroluminiscénten
gospodárskopolítičen
hípersenzibilizácija
institucionalizírati
internacionalístičen
internacionalizácija
internacionalizírati
kóntrarevolucionáren
krščánskodemokrátski
krščánskosocialístičen
literárnozgodovínski
marksístično-leninístičen
meščánskodemokrátičen
nacionálsocialístičen
otorinolaringologíja
pétinsédemdesetlétnica
políticnoteritoriálen
póstimpresionístičen
primerjálnozgodovínski
privátnokapitalístičen
prótiimperialístičen
psévdorevolucionáren
quantité négligeable
sámozadovoljevátise
skonvencionalizírati
slóvstvenozgodovínski
sociálnodemokrátičen
sociálnorevolucionáren
splôšnoizobraževálen
stárocerkvénoslovánski
stárocerkvénoslovánščina
stêreofotogrametríja
têrtiumcomparatiónis
turístičnopropagánden
últranacionalístičen
umétnostnogeográfski
umétnostnozgodovínski
vsèzavérodómcesárjevstvo
zgódnjebronastodôben
zgódnjekapitalístičen
znánstvenoraziskoválen
živílskopredeloválen

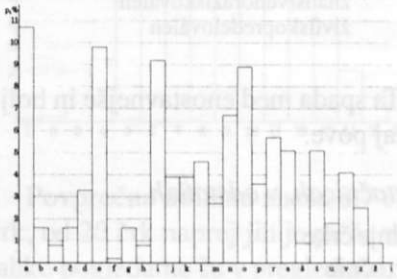
Oglejmo si še statistiko črk (preglednica 8). Ta spada med enostavnejše in bolj klasične postopke, pa vendar o besedilih marsikaj pove.

Preglednica 8: Pogostnosti črk v slovarju in v iztočnicah, v odstotkih

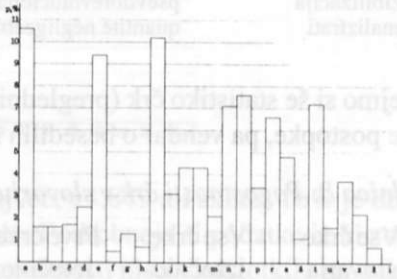
	Vse črke slovarja	Vse črke iztočnic	Prve črke iztočnic	Zadnje črke iztočnic	
a	10.7	10.6	2.8	17.6	a
b	1.6	1.6	4.2	0.2	b
c	0.8	1.6	1.3	3.8	c
č	1.6	1.7	1.3	1.0	č
d	3.3	2.5	5.1	0.8	d
e	9.8	9.4	1.3	.5	e
f	0.2	0.5	1.5	0.2	f
g	1.6	1.3	2.8	0.5	g
h	0.8	0.7	1.8	0.3	h
i	9.2	10.2	3.4	23.3	i
j	3.9	3.1	1.1	0.4	j
k	3.9	4.3	6.5	5.4	k
l	4.6	4.3	2.2	0.9	l
m	2.7	2.1	4.3	1.2	m
n	6.7	7.1	6.2	13.8	n
o	8.9	7.6	5.9	2.8	o
p	3.6	3.4	15.7	0.4	p
r	5.7	6.6	4.9	3.1	r
s	5.1	4.8	8.9	0.9	s
š	0.9	1.1	1.9	0.2	š
t	5.1	7.2	3.8	9.6	t
u	1.8	1.6	2.0	0.1	u
v	4.1	3.7	4.2	3.4	v
z	2.5	2.2	5.5	0.2	z
ž	0.9	0.7	1.0	0.7	ž

Preglednico 8, ki je sicer bolj natančna, zato pa manj pregledna, dopolnjujejo štirje histogrami (slike 3, 4, 5 in 6). Pri črkah v slovarju in v iztočnicah prevladuje črka *a*, kar je tudi sicer značilno za slovenski jezik (v angleškem je npr. najpogostejša črka e). Druga najpogostejša črka pri slovarju je *e*, kar se tudi ujema s slovenskimi besedili, pri iztočnicah pa *i*. To lahko pripišemo predvsem glagolom, ki so v iztočnicah v nedoločniku, ter pridevnikom na *-ski* in *-ški*. Več kot v besedilu je v iztočnicah še črk *t, r, c* in *f*, manj pa *o*-jev in *j*-jev. Pri prvih in zadnjih črkah iztočnic so razlike še dosti izrazitejše: pri prvih je premočno, skoraj s 16%, na prvem mestu *p* (strah in trepet vseh, ki berejo korekture slovarjev), na drugem je *s*. Pri zadnjih črkah so že omenjeni glagoli in pridevniki pomagali *i*-ju do skoraj četrtinskega deleža, samostalniki ž. spola *a*-ju na drugo mesto, tretji je *n* (pridevniki na *-en, -an, -ln*), četrti *t* (samostalniki ž. spola in pridevniki), peti pa *e* (sam. s. spola). Še več o začetkih in koncih iztočnic odkrivata preglednici 9 in 10.

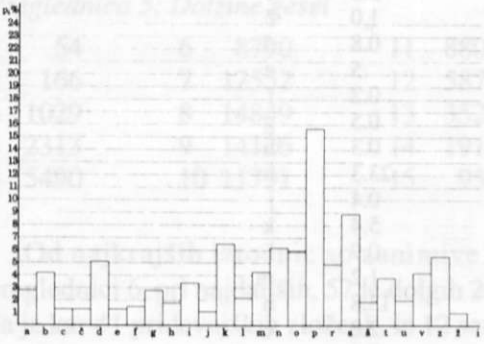
Slika 3: Pogostnosti vseh črk v slovarju



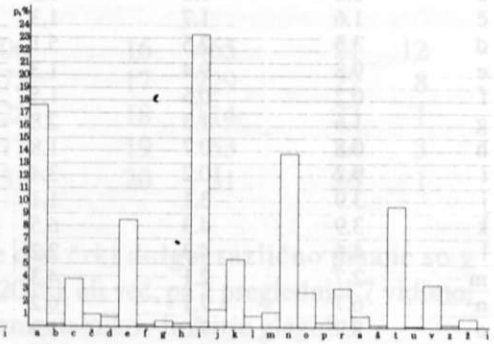
Slika 4: Pogostnosti vseh črk v iztočnicah



Slika 5: Prve črke v iztočnicah



Slika 6: Zadnje črke v iztočnicah



Preglednica 9: Pogostnosti začetnih črk, dvojčic ... sedmerčic pri iztočnicah

1.	p- 14604	pr-	6306	pre-	3206	pred-	483	proti-	207	elektr-	113	elektro-	91
2.	s- 8331	po-	5006	raz-	1836	samo-	443	nepre-	140	široko-	52	četvero-	36
3.	k- 6080	za-	3173	pri-	1685	brez-	408	elekt-	115	kontra-	50	antropo-	26
4.	n- 5785	ne-	2857	pro-	821	pres-	328	inter-	104	kratko-	49	dvanajs-	25
5.	o- 5472	ra-	2409	pod-	753	razp-	294	štiri-	100	četver-	47	general-	21
6.	z- 5166	na-	2283	pol-	645	prep-	283	prest-	99	debelo-	41	pripove-	21
7.	d- 4793	ko-	1933	pos-	573	prot-	278	dolgo-	98	gospod-	39	srednje-	20
8.	r- 4592	iz-	1927	str-	532	nepr-	272	debel-	96	samoza-	37	pripravl-	18
9.	m- 4006	st-	1461	kon-	517	prek-	250	razpo-	90	visoko-	37	harmoni-	17
10.	v- 3901	de-	1270	nep-	499	pris-	245	nepri-	86	prosto-	36	kristal-	17
11.	b- 3896	tr-	1264	bre-	494	prev-	226	neraz-	84	prestr-	35	nasprot-	17
12.	t- 3583	ka-	1233	sam-	482	prem-	212	razpr-	75	sladko-	35	oblikov-	17
13.	i- 3207	do-	1201	zas-	401	pret-	186	preob-	74	veliko-	34	predsta-	17
14.	g- 2636	kr-	1156	pra-	379	nepo-	182	mnogo-	71	zastav-	32	zgodnje-	17
15.	a- 2406	od-	1156	spo-	352	prip-	175	staro-	71	social-	31	elektri-	16
16.	l- 2024	br-	1133	izp-	342	razk-	166	prist-	70	drobno-	29	gospoda-	16
17.	u- 1886	ma-	1118	zap-	340	spre-	166	blago-	69	antrop-	26	prosvet-	16
18.	š- 1800	ob-	1116	dvo-	333	razs-	163	hidro-	67	genera-	26	razkroj-	16
19.	h- 1654	re-	1103	pot-	327	inte-	158	pripo-	67	prista-	26	razstav-	16
20.	e- 1445	sp-	1014	sta-	310	razv-	157	predp-	65	stereo-	26	razsvet-	16
21.	f- 1390	pa-	938	pla-	298	poli-	149	kontr-	62	straho-	26	steklen-	16
22.	č- 1255	me-	913	kol-	297	avto-	144	prepo-	62	svetlo-	26	alkohol-	15
23.	c- 1190	sa-	845	pom-	297	post-	136	prena-	61	svobod-	26	dvajset-	15
24.	j- 1057	so-	791	zak-	296	priv-	135	prepr-	61	dvanaj-	25	preobra-	15
25.	ž- 955	ve-	782	nad-	290	svet-	135	deset-	60	instru-	25	prestre-	15
26.	w- 26	se-	774	spr-	288	zast-	133	preds-	60	predpo-	25	socialn-	15
27.	y- 5	gr-	747	neo-	281	preg-	130	trans-	60	predst-	25	special-	15
28.	q- 4	te-	742	tri-	276	stra-	130	premo-	59	presta-	25	šestero-	15
29.	x- 2	mo-	740	par-	274	neiz-	129	radio-	59	protis-	25	zelenja-	15
30.		pe-	735	pov-	274	pren-	129	preve-	58	politi-	24	dokumen-	14
31.		dr-	730	nas-	273	prim-	129	psiho-	58	srednj-	24	domišlj-	14
32.		pl-	721	kom-	271	anti-	128	razkr-	58	interp-	23	instrum-	14
33.		sk-	700	pop-	261	prik-	127	prekr-	55	magnet-	23	plemeni-	14
34.		go-	694	zav-	260	nedo-	126	prete-	55	poslov-	23	pregled-	14
35.		ba-	688	nap-	257	star-	126	dobro-	54	postav-	23	preskrb-	14
36.		in-	657	nes-	257	preb-	124	gospo-	54	presto-	23	radiote-	14
37.		di-	593	sla-	255	pros-	123	samoo-	54	protip-	23	razstre-	14
38.		bo-	586	tra-	253	izpo-	118	drobn-	53	razstr-	23	sladkos-	14
39.		sl-	566	eks-	250	elek-	116	izpod-	53	knjigo-	22	zavarov-	14
40.		op-	561	kri-	242	preo-	116	prekl-	53	neprem-	22	zdravil-	14
41.		kl-	554	tro-	239	prid-	115	prost-	53	prepre-	22	cvetlič-	13
42.		os-	550	pok-	238	razm-	115	samos-	53	pristo-	22	čeljust-	13
43.		si-	550	gra-	236	prit-	113	sveto-	53	razkro-	22	deseter-	13
44.		ro-	530	sto-	232	kont-	111	širok-	53	razpre-	22	dialekt-	13
45.		le-	524	kra-	231	razb-	111	premi-	51	razred-	22	ekonomi-	13
46.		mi-	497	por-	230	neza-	109	razve-	50	samopo-	22	germani-	13
47.		vr-	492	ste-	224	stre-	109	kratk-	49	sestav-	22	interna-	13
48.		be-	489	obr-	223	dolg-	108	mikro-	49	svetov-	22	izposoj-	13
49.		to-	488	zam-	223	štir-	108	razst-	49	intern-	21	konstru-	13
50.		la-	478	pog-	221	drob-	107	prese-	48	obliko-	21	mravlji-	13
51.		vi-	476	kor-	213	prav-	107	četve-	47	organi-	21	naciona-	13
52.		vz-	452	int-	210	vele-	106	samop-	47	preobr-	21	organiz-	13
53.		pi-	438	dol-	208	prer-	105	zelen-	47	pripov-	21	pridobi-	13
54.		ta-	429	vel-	208	razd-	104	preko-	46	pripra-	21	protina-	13
55.		vo-	420	kar-	203	razt-	104	velik-	46	razdel-	21	razgled-	13
56.		dv-	411	zad-	201	spod-	104	zasta-	46	razpor-	21	razpore-	13
57.		je-	406	ple-	200	pove-	103	brezp-	45	šester-	21	repreze-	13
58.		an-	398	nez-	196	para-	102	predi-	45	zemlje-	21	sporazu-	13
59.		če-	392	neu-	194	nera-	101	razte-	44	bistro-	20	strahop-	13
60.		bi-	388	pon-	194	razg-	100	posta-	43	brezve-	20	svobodo-	13
61.		li-	385	nar-	193	debe-	99	razpl-	43	debelu-	20	univerz-	13
62.		sv-	383	mar-	191	malo-	95	samoz-	43	ekstra-	20	central-	12

Preglednica 10: Pogostnosti končnih črk, dvojčic ... sedmerčic pri iztočnicah

1.	-i	21549	-ti	15385	-ati	9132	-nost	3328	-irati	1744	-ati se	580	-ističen	352
2.	-a	16269	-en	10836	-iti	4855	-anje	3192	-ti se	1226	-iranje	575	-ljivost	340
3.	-n	12779	-st	6488	-ost	4478	-vati	2063	-vanje	999	-iti se	551	-izirati	335
4.	-t	8879	-je	5472	-ski	4282	-rati	2014	-acija	810	-evanje	423	-acijski	248
5.	-e	7895	-ki	5271	-nje	4268	-nica	1906	-evati	759	-stičen	403	-izacija	229
6.	-k	5017	-ca	3685	-ica	3463	-ičen	1460	-tičen	735	-zirati	367	-ologija	176
7.	-c	3486	-ec	3321	-ija	2245	-jati	1249	-enost	686	-cijski	352	-jevanje	166
8.	-v	3112	-ka	3299	-čen	2102	-niti	1246	-ovati	671	-jenost	346	-ljenost	125
9.	-r	2835	-ja	2941	-nik	1885	-i se	1235	-ljati	630	-ovanje	345	-vati se	121
10.	-o	2622	-ik	2207	-len	1769	-nski	1175	-ranje	624	-jivost	345	-njenost	104
11.	-j	1316	-na	1912	-ten	1597	-alen	1147	-ijski	591	-arstvo	304	-niti se	103
12.	-m	1142	-ek	1817	-ina	1439	-enje	1037	-avati	506	-tirati	267	-tiranje	102
13.	-č	914	-vo	1452	-tvo	1385	-cija	1034	-janje	495	-zacija	240	-ziranje	97
14.	-l	862	-ar	1349	-ast	1340	-stvo	1027	-arski	442	-jevati	232	-avljati	94
15.	-s	822	-se	1252	- se	1236	-rski	927	-jenje	424	-ičnost	221	-čevanje	93
16.	-d	704	-iv	1035	-ven	1131	-kati	891	-rstvo	409	-evalen	204	-jati se	92
17.	-ž	612	-ev	1018	-eti	995	-jski	864	-niški	395	-ljanje	203	-ljevati	88
18.	-g	437	-an	830	-ški	947	-ljiv	856	-čnost	378	-ništvo	189	-iranost	88
19.	-.	413	-er	790	-nec	886	-itev	705	-ivost	375	-alnica	184	-jevalen	81
20.	-p	367	-at	778	-jiv	866	-alec	598	-anost	365	-logija	182	-valnica	80
21.	-h	274	-em	730	-lec	837	-izem	587	-valen	361	-ljenje	176	-tivnost	80
22.	-z	211	-ta	643	-vec	796	-vost	570	-inski	343	-evalec	173	-tičnost	78
23.	-š	195	-lo	611	-ren	763	-čiti	558	-anski	341	-alnost	169	-grafija	77
24.	-b	152	-ov	583	-tev	744	-riti	541	-valec	334	-nirati	164	-vljanje	77
25.	-f	150	-ak	563	-ček	734	-tati	517	-lnica	326	-avanje	163	-icirati	73
26.	-u	81	-ra	552	-den	734	-iški	494	-tnost	326	-čevati	163	-njevati	73
27.	-y	44	-če	516	-zem	602	-išče	454	-lnost	280	-kovati	150	-jevalec	72
28.	-x	5	-ež	513	-nka	585	-ilen	450	-ariti	258	-ovalec	138	-onirati	71
29.	-w	4	-or	505	-ist	538	-sten	437	-vnost	251	-rirati	138	-ikacija	70
30.	-	2	-av	461	-jen	522	-liti	404	-ovski	247	-lirati	126	-kovanje	70
31.	-q	1	-in	440	-ika	511	-enec	390	-kniti	242	-ranost	126	-arjenje	70
32.			-on	433	-vka	465	-vski	378	-tnica	238	-ološki	123	-kati se	69
33.			-it	418	-šče	459	-titi	374	-onski	235	-ovalen	122	-čiti se	60
34.			-..	413	-lka	415	-anec	360	-arica	234	-čenost	119	-grafski	58
35.			-ač	411	-...	413	-štvo	357	-nstvo	225	-ativen	116	-matičen	58
36.			-la	410	-sen	393	-lnik	350	-tiven	223	-anstvo	116	-titi se	57
37.			-lj	407	-žen	385	-rica	344	-kanje	222	-cirati	115	-čevalec	56
38.			-da	382	-elj	359	-diti	339	-enski	211	-ališče	113	-čevalen	56
39.			-va	379	-zen	357	-nina	331	-dnost	211	-vljati	113	-tati se	55
40.			-ba	377	-než	348	-iven	314	-jenec	203	-nosten	110	-liti se	55
41.			-nt	359	-rka	336	-čati	310	-ištvo	203	-rjenje	109	-riti se	55
42.			-et	357	-sti	315	-eten	309	-čenje	201	-ivnost	109	-atorski	55
43.			-el	353	-tka	303	-ovec	306	-ovina	199	-atičen	108	-karstvo	55
44.			-ma	348	-jak	302	-vina	303	-ščina	197	-ajanje	107	-inirati	53
45.			-aj	343	-nja	296	-alka	299	-ničen	196	-ilnica	103	-ulirati	52
46.			-ji	315	-ter	295	-tski	283	-alnik	193	-njenje	101	-ntirati	52
47.			-za	290	-ben	290	-tnik	278	-čnica	192	-racija	99	-oslovje	49
48.			-ič	279	-tor	286	-rija	266	-rnost	192	-torski	98	-ionalen	49
49.			-ij	274	-ilo	281	-čina	266	-ogija	184	-dirati	94	-teljica	48
50.			-ča	241	-vje	273	-o...	260	-avost	183	-tacija	91	-niranje	48
51.			-al	231	-tje	261	-lski	260	-rnica	178	-alizem	91	-rati se	48
52.			-og	224	-nat	240	-gati	252	-ajati	178	-kacija	90	-ulacija	46
53.			-ga	218	-alo	229	-avec	251	-njati	175	-ljenec	89	-dovanje	46
54.			-ce	203	-kar	224	-pati	251	-etati	175	-bljati	83	-vališče	43
55.			-či	189	-ava	223	-viti	246	-osten	169	-lacija	82	-riranje	43
56.			-ok	189	-men	221	-anka	232	-orski	166	-kljati	82	-rjenost	43
57.			-us	180	-šen	215	-aren	230	-erski	145	-pljati	80	-ševanje	42
58.			-is	169	-čar	212	-tika	229	-ikati	144	-kavati	80	-diti se	42
59.			-ed	167	-eta	209	-aven	227	-nitev	143	-rafija	79	-liranje	41
60.			-ot	160	-rat	195	-čnik	218	-vnica	142	-istika	77	-kcijski	41
61.			-op	151	-tek	187	-enka	217	-lišče	141	-tavati	77	-narstvo	41
62.			-od	149	-ran	180	-gija	216	-dnica	139	-etnica	76	-ciranje	40

V preglednici 9 vidimo, kako zelo se poveča razgibanost besed po tretji črki – če bi se najpogostejša dvojčica in trojčica še uvrstili na tretje mesto med enojčice oziroma dvojčice, bi se prva četverčica pri trojčicah komaj na dvanajsto in peterčica med četverčice na trinajsto. Zanimivo je tudi gibanje najpogostejše sedmerčice, *elektro-*, katere prvi predhodnik, *elek-*, se pojavi šele med četverčicami, na 39. mestu, med peterčicami je *elekt-* že tretja, med šesterčicami pa *elektr-* prva. V nasprotju s tem so predhodniki prve končne sedmerčice (*-ističen* iz preglednice 10) ves čas pri vrhu. Tudi sicer med najpogostejšimi začetnimi in končnimi n-terčicami po pričakovanju najdemo predvsem predpone in pripone, oziroma natančneje (npr. po (Toporišič 1994)) predponska in priponska obrazila. Hitrost povečevanja števila n-terčic na začetkih in koncih iztočnic kaže tudi preglednica 11.

Preglednica 11: Število različnih začetkov in koncev na n-črk v iztočnicah

1	29	31	10	86.216	85.395	19	90.661	90.659
2	445	397	11	88.525	88.238	20	90.662	90.662
3	3.548	2.839	12	89.789	89.541	21	90.662	90.662
4	14.725	10.180	13	90.272	89.564	22	90.662	90.662
5	32.258	22.220	14	90.502	90.165	23	90.662	90.662
6	50.992	37.965	15	90.596	90.430	24	90.662	90.662
7	65.572	55.593	16	90.640	90.564	25	90.662	90.662
8	75.519	70.295	17	90.653	90.634	26	90.662	90.662
9	82.170	80.047	18	90.659	90.652			

(Končno število je manjše kot 93.151, ker se preglednica nanaša na iztočnice brez naglasov). Da se po drugi črki od začetka iztočnice zelo razgibajo, smo iz preglednic 9 in 11 tako na oko že ugotovili; lahko bi rekli, da se količina obvestila (informacije) s tretjo črko močno poveča. Domnevo potrdimo z orodjem iz teorije informacij.

Količino obvestila, lahko bi ji rekli obvestilnost, merimo z entropijo – večja ko je neurejenost, večja je entropija kakega sistema, in večjo obvestilnost potrebujemo, da ga opišemo. Entropijo v jeziku je prvi, za angleško besedilo, določil Claude E. Shannon (Shannon 1948), za slovensko besedilo pa je o njej pred kratkim izšel prispevek študentov in učiteljev Fakultete za elektrotehniko in računalništvo (Kristan et al. 1994). Entropija zaporedja dogodkov je v splošnem podana z zvezo:

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i \qquad (1)$$

kjer je n število vseh različnih dogodkov, p_i verjetnost pojavitve i-tega dogodka,

$\log_2$ pa dvojiški logaritem. Entropija, enota zanjo je bit (od angl. *binary digit*, po slovensko bi rekli dvojiška števka), je največja, če so vsi dogodki enako verjetni.

V našem primeru je najbolj zanimiva pogojna entropija n-te črke v iztočnici, če prejšnjih n-1 črk že poznamo. Označujemo jo z F_n, dobimo pa jo iz povezave:

$$F_n = H_n - H_{n-1} \tag{2}$$

in s privzetkom, da je $F_1 = H_1$. Pri tem je H_n nepogojna entropija vseh n-teric z začetka (ali konca) iztočnice in zanjo, po (Jakopin 1981), zlahka izpeljemo izraz:

$$H_n = \log_2 N - \frac{1}{N}\sum_{i=1}^{m} f_{in} \log_2 f_{in} \tag{3}$$

kjer je N število vseh n-teric (v našem primeru 93.151), m število vseh različnih n-teric, f_{in} pa frekvenca i-te izmed njih. Povedano preprosteje bi pogojni entropiji F_n rekli mera za verjetnost, da bomo zadeli n-to črko v iztočnici, če prejšnjih n-1 črk že poznamo.

Oglejmo si vrednosti F_n za začetne in končne n-terice pri iztočnicah.

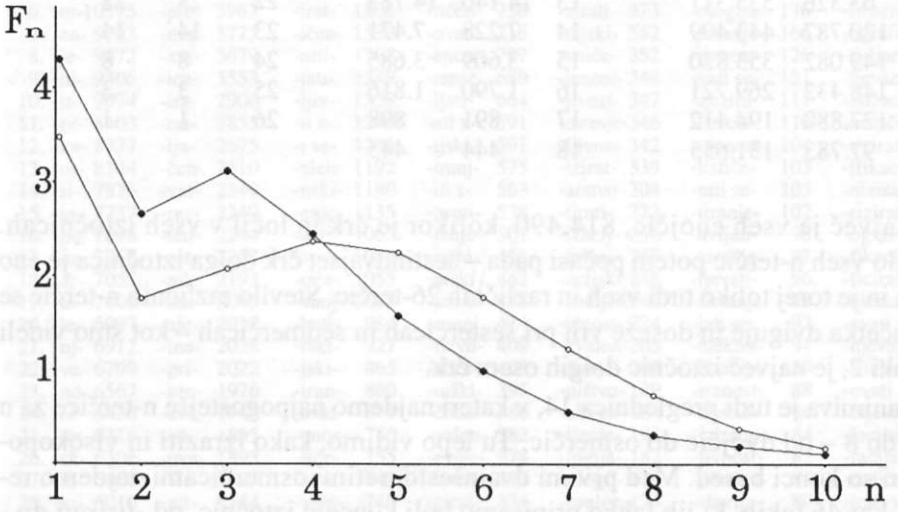
Preglednica 12: Pogojne entropije začetnih in končnih n-teric pri iztočnicah

1	4.31	3.48
2	2.67	1.77
3	3.13	2.09
4	2.46	2.38
5	1.59	2.27
6	1.00	1.79
7	0.56	1.24
8	0.32	0.74
9	0.19	0.38
10	0.11	0.18

Grafična upodobitev je sicer spet manj natančna, zato pa toliko bolj nazorna (polne pike označujejo vrednosti za začetne n-terice, votle pa za končne). Iz slike vidimo, da se obe pogojni entropiji močno razlikujeta od pogojne entropije v črkah besedila, ki monotono pada – če poznamo več črk besedila, je možnost, da bomo pravilno uganili naslednjo črko, večja kot če jih poznamo manj, entropija n-te črke in z njo količina obvestila pa manjša. Pri začetnih in končnih n-tericah iztočnic začne entropija monotono padati šele od trojčic oziroma četverčic naprej, od n = 2 na n = 3 pa pri obojih beležimo močan vzpon. Verjetnost, da bomo zadeli tretjo črko, če poznamo prvi dve, je torej dosti manjša, kot pa da bomo zadeli drugo, če poznamo prvo. Pri slovenskih besedilih (Kristan et al. 1994) pogojna entropija vseskozi monotono pada, pri imenih in priimkih (Jako-

pin 1981) pa: pri imenih tudi, pri priimkih pa, podobno kot pri naših iztočnicah, šele od trojčic naprej.

Slika 7: Pogojne entropije začetnih in končnih n-teric pri iztočnicah



Pri začetkih je prva pogojna entropija 4.31 bita, kar pomeni, da imamo možnost približno 1 : 20 ($2^{4.31}$ = 19.84), da bomo uganili prvo črko naključno izbranega gesla. Pri koncih je možnost večja, že 1 : 11, ker je prva entropija nižja, le 3.48 bita ($2^{3.48}$ = 11.16). To tudi ni čudno, če pogledamo sliko 6 – črke *i, a in n* odnesejo skupaj že več kot polovico iztočniških koncev. Če bi bile črke zastopane enakomerno, bi, s privzetkom, da imamo opravka samo z malimi črkami, ki jih je 25, bili ti možnosti obe 1 : 25. Med začetki pade pogojna entropija na 1 bit (enaka možnost, da bomo zadeli kot tudi, da bomo zgrešili) pri šesterčicah, med konci pa šele pri osmerčicah.

V vsakdanjem življenju pridejo bolj kot začetni deli besed v poštev njihovi konci. V vseh pregibnih oblikah jih močno potrebujejo pesniki pri iskanju rim. Slovenci smo tak seznam, Prvi slovenski pesniški priročnik, s podnaslovom Slovar odzadnjih zlogov slovenskih besed (SBD 1993) dobili pred dvema letoma. Izkaže pa se, kot bomo malo kasneje videli, da tudi začetne n-terice niso brez uporabne vrednosti.

Oglejmo si zdaj še seznam vseh n-teric, to je takih, ki jih dobimo, če jih ne jemljemo samo od začetka ali s konca temveč tudi iz sredine iztočnic. Teh zaporedij črk je seveda še dosti več; koliko, nam pove preglednica 13.

Preglednica 13: Število različnih in vseh n-teric v iztočnicah

1	82	814.490	10	67.510	83.746	19	221	221
2	1.471	721.339	11	42.629	49.936	20	111	111
3	15.060	628.242	12	24.740	27.971	21	54	54
4	63.326	535.311	13	14.140	14.785	22	28	28
5	120.787	443.409	14	7.228	7.471	23	14	14
6	149.082	353.820	15	3.609	3.685	24	8	8
7	148.432	269.721	16	1.790	1.816	25	3	3
8	127.880	194.412	17	891	898	26	1	1
9	97.783	131.655	18	444	445			

Največ je vseh enojčic, 814.490, kolikor je črk in ločil v vseh iztočnicah. Število vseh n-terčic potem počasi pada – šestindvajset črk dolga iztočnica je eno sama in je torej toliko tudi vseh in različnih 26-terčic. Število različnih n-terčic se od začetka dviguje in doseže vrh pri šesterčicah in sedmerčicah – kot smo videli na sliki 2, je največ iztočnic dolgih osem črk.

Zanimiva je tudi preglednica 14, v kateri najdemo najpogostejše n-terčice za n od 2 do 8 – od dvojčic do osmerčic. Tu lepo vidimo, kako izraziti in visokopogostni so konci besed. Med prvimi dvainšestdesetimi osmerčicami najdemo recimo kar 46 takih, ki jih lahko pripišemo bolj koncem iztočnic, od *-lizirati* do *-iologija*, ter samo pet takih, ki so jasno prepoznavni začetki: *nacional-, gospodar-, zgodovin-, oblikova- in demokrat-*. Podobno s primerjavo preglednic 10 in 14 takoj ugotovimo, da so v obeh na prvih treh mestih, od dvojčic do sedmerčic iste n-terčice. Edina izjema je *-rati*, ki je pri vseh n-terčicah tretji, pri končnih pa šele četrti.

Kakor na hitro morda ni videti kake splošno koristne uporabe n-terčic, se izkaže, da je z njihovimi pogostnostmi ter z iztočnicami slovarja mogoče sestaviti razmeroma učinkovit algoritem za deljenje besed, eno od šibkejših točk sodobnega slovenskega trenutka. Osnovna ideja je v tem, da vzamemo najmanjšo skupino črk, ki sega do prvega samoglasnika ali »samoglasniškega« *r* in potem ta, recimo mu kar zlog, podaljšujemo tako dolgo, dokler najmanjši naslednji zlog nima manjše entropije (se pravi večje pogostnosti).

Preglednica 14: Pogostnosti prvih n-terčic – dvojčic do osmerčic pri iztočnicah

1.	-ti-	22498	-ati-	10664	-nost-	3506	-irati-	1786	-ati se-	580	-ističen-	352	-lizirati-	110
2.	-en-	18163	-iti-	5803	-anje-	3314	-ti se-	1226	-iranje-	575	-ljivost-	342	-lističen-	103
3.	-st-	14916	-ost-	5611	-rati-	2233	-vanje-	1001	-iti se-	551	-izirati-	339	-alizirat-	86
4.	-at-	13822	-nje-	4985	-vati-	2225	-acija-	811	-evanje-	423	-acijski-	248	-alističe-	85
5.	-ra-	12863	-ski-	4345	-nica-	1915	-evati-	805	-stičen-	403	-izacija-	229	-iziranje-	84
6.	-re-	10375	-pre-	3963	-irat-	1806	-tičen-	736	-zirati-	373	-ologija-	176	-lizacija-	76
7.	-an-	9703	-nos-	3775	-ičen-	1532	-ovati-	716	-cijski-	352	-jevanje-	166	-ografija-	69
8.	-je-	9572	-anj-	3679	-niti-	1368	-enost-	707	-ističe-	352	-ljenost-	126	-avljanje-	68
9.	-ni-	9306	-ica-	3553	-jati-	1349	-ranje-	680	-jenost-	348	-vati se-	121	-ljevanje-	66
10.	-it-	9074	-ira-	2906	-ljiv-	1328	-ljati-	664	-jivost-	347	-alizira-	111	-alizacij-	63
11.	-pr-	8403	-rat-	2855	-ti s-	1249	-ati s-	591	-ovanje-	346	-lizirat-	110	-ništičen-	62
12.	-ka-	8377	-ija-	2675	-i se-	1235	-ijski-	591	-ljivos-	342	-njenost-	104	-tizirati-	59
13.	-os-	8144	-čen-	2610	-alen-	1192	-iranj-	575	-izirat-	339	-lističe-	103	-ifikacij-	58
14.	-ri-	7876	-vat-	2349	-nski-	1180	-iti s-	563	-arstvo-	304	-niti se-	103	-nizirati-	56
15.	-ja-	7737	-raz-	2249	-enje-	1135	-avati-	538	-tirati-	272	-tiranje-	102	-rizirati-	54
16.	-po-	7698	-len-	2244	-acij-	1065	-janje-	501	-izacij-	256	-avljati-	98	-ografski-	51
17.	-na-	7268	-ast-	2207	-stvo-	1058	-izira-	451	-jevati-	249	-ziranje-	97	-fikacija-	49
18.	-ar-	7053	-nik-	2193	-cija-	1055	-arski-	442	-acijsk-	248	-ljevati-	96	-ficirati-	45
19.	-te-	7000	-ten-	2043	-vanj-	1024	-jenje-	433	-zacija-	240	-čevanje-	93	-nizacija-	45
20.	-le-	6997	-ale-	2037	-kati-	984	-evanj-	424	-ičnost-	224	-jati se-	92	-evati se-	44
21.	-nj-	6912	-ina-	2028	-rski-	927	-rstvo-	409	-evalen-	205	-elektro-	91	-njevanje-	42
22.	-va-	6799	-pri-	2022	-jski-	865	-stiče-	405	-ljanje-	203	-lizacij-	89	-ificirat-	41
23.	-no-	6567	-nic-	1976	-iran-	860	-niški-	395	-ništvo-	189	-iranost-	88	-ovati se-	40
24.	-al-	6456	-eva-	1940	-evat-	814	-čnost-	386	-alnica-	184	-alistič-	86	-eljevati-	39
25.	-ov-	6316	-ist-	1895	-enos-	756	-valen-	382	-ljenje-	183	-iziranj-	84	-lacijski-	39
26.	-sk-	6306	-ova-	1893	-tiče-	755	-evale-	378	-logija-	182	-tivnost-	82	-racijski-	39
27.	-in-	6060	-iče-	1827	-ovat-	741	-ivost-	377	-ologij-	179	-jevalen-	81	-rističen-	39
28.	-ki-	6010	-nit-	1744	-vale-	740	-zirat-	374	-evalec-	173	-stavlja-	80	-irati se-	38
29.	-ro-	5863	-ran-	1742	-ranj-	719	-anost-	368	-alnost-	171	-valnica-	80	-iteljica-	38
30.	-lj-	5782	-lja-	1723	-itev-	715	-istič-	364	-čevati-	170	-tičnost-	78	-ovalnica-	37
31.	-ne-	5704	-jen-	1711	-riti-	696	-ovanj-	355	-nirati-	169	-grafija-	77	-rizacija-	37
32.	-la-	5678	-sti-	1674	-ljat-	689	-cijsk-	352	-jevanj-	166	-vljanje-	77	-tacijski-	37
33.	-ko-	5663	-eti-	1654	-ljen-	654	-jenos-	351	-avanje-	163	-icirati-	76	-tologija-	37
34.	-et-	5491	-lji-	1589	-tati-	643	-jivos-	347	-kovati-	157	-njevati-	76	-dljivost-	36
35.	-li-	5475	-tvo-	1529	-čiti-	621	-anski-	344	-jevale-	153	-onirati-	73	-ativnost-	35
36.	-ve-	5394	-str-	1509	-jeva-	620	-inski-	343	-lizira-	148	-jevalec-	72	-rljivost-	35
37.	-če-	5370	-ven-	1506	-alec-	598	-ljivo-	343	-rirati-	139	-arjenje-	70	-entirati-	34
38.	-av-	5232	-enj-	1454	-izem-	595	-tnost-	342	-ovalec-	138	-ikacija-	70	-nacional-	34
39.	-ta-	5020	-cij-	1441	-ati -	591	-valec-	334	-elektr-	133	-kovanje-	70	-povedova-	33
40.	-er-	4830	-jat-	1421	-ijsk-	591	-lnica-	326	-lirati-	129	-kati se-	69	-tavljati-	33
41.	-ol-	4810	-ite-	1392	-vost-	591	-arstv-	319	-ovalen-	127	-ografij-	69	-cionalen-	32
42.	-lo-	4807	-van-	1389	-pred-	590	-lnost-	287	-ljenos-	126	-avljanj-	68	-zacijski-	32
43.	-za-	4796	-ren-	1383	-iti -	563	-tirat-	272	-ranost-	126	-ljevanj-	66	-gospodar-	31
44.	-ev-	4773	-val-	1365	-sten-	563	-zacij-	272	-vati s-	124	-alizaci-	63	-nljivost-	31
45.	-od-	4743	-jiv-	1354	-išče-	562	-vnost-	268	-ološki-	123	-nističe-	62	-ostavlja-	31
46.	-ik-	4690	-ika-	1346	-liti-	553	-ariti-	267	-ališče-	120	-čiti se-	60	-tljivost-	31
47.	-or-	4624	-stv-	1280	-avat-	548	-ovale-	265	-cirati-	119	-fikacij-	59	-zgodovin-	31
48.	-tr-	4576	-ali-	1265	-isti-	546	-izaci-	256	-čenost-	119	-tizirat-	59	-atizirat-	30
49.	-vo-	4448	-i s-	1259	-eval-	544	-jevat-	252	-vljati-	118	-grafski-	58	-avati se-	30
50.	-ir-	4360	-kat-	1251	-janj-	510	-kniti-	250	-ivnost-	117	-ifikaci-	58	-eljevanj-	30
51.	-el-	4305	-ti -	1249	-zira-	509	-acijs-	248	-anstvo-	116	-matičen-	58	-evalnica-	30
52.	-ek-	4256	-ari-	1246	-iški-	495	-avlja-	248	-ativen-	116	-titi se-	57	-ljati se-	30
53.	-iz-	4215	-nsk-	1244	-ovan-	488	-ovski-	247	-stavlj-	115	-čevalec-	56	-tizacija-	30
54.	-ic-	4200	- se-	1235	-evan-	478	-tnica-	238	-čevale-	112	-čevalen-	56	-bljenost-	29
55.	-se-	4187	-aci-	1218	-samo-	464	-arica-	237	-iziran-	112	-nizirat-	56	-oblikova-	29
56.	-ca-	4172	-nja-	1215	-ilen-	463	-dnost-	235	-alizir-	111	-acional-	55	-acionali-	28
57.	-to-	4097	-eno-	1192	-tira-	462	-onski-	235	-nosten-	110	-atorski-	55	-demokrat-	28
58.	-as-	3996	-sta-	1176	-izir-	454	-ovina-	234	-rjenje-	109	-karstvo-	55	-edovanje-	28
59.	-on-	3927	-rit-	1173	-jenj-	452	-kanje-	228	-atičen-	108	-liti se-	55	-istovski-	28
60.	-de-	3888	-ter-	1151	-oval-	447	-nstvo-	225	-ističe-	108	-riti se-	55	-onalnost-	28
61.	-ij-	3855	-red-	1150	-stič-	447	-ičnos-	224	-ajanje-	107	-tati se-	55	-stavljat-	28
62.	-me-	3746	-den-	1143	-arsk-	445	-tiven-	223	-niti s-	105	-ificira-	54	-iologija-	27

Vzemimo za primer, kako bi pri iztočnici *predčutje* poiskali prvo delilno mesto. Prvi najmanjši zlog je *pre*, *p* ni samoglasnik, *r* tudi ni samoglasniški in se zato ustavimo pri *e*. Pogostnost te trojčice je 3964, pogostnost naslednjega najmanjšega zloga, *dču* pa samo 3, zato gremo naprej. Dodamo še eno črko: *pred* ima pogostnost 590, kar je še vedno več kot *ču* s pogostnostjo 383. Pri *predč* se pa zadeva ustavi: pogostnost je samo 4, medtem ko ima *u*, naslednji najmanjši zlog, pogostnost 13306. Ugotovimo, da je zadnji kandidat za prvi zlog, ki je imel pogostnost še večjo od naslednjega, *pred* in za njim delimo. Algoritem ponovimo na nadaljnem zlogu – najmanjši naslednji je *ču*. Smer je prava, izkaže pa se, da je postopek, če naj bi bili z njim kolikor toliko zadovoljni, potrebno še dopolniti. Okvir in prostor tega prispevka žal ne dopuščata, da bi se spustili malo bolj v širino in v podrobnosti, s katerimi metoda šele dobi pravo težo.

Kakorkoli že, postopek (Jakopin 1995) ni ravno zapleten in s pravili v Slovenskem pravopisu (SP 94) se tepe dosti manj kot algoritmi za deljenje angleških, nemških ali italijanskih besed, ki so že vgrajeni v tuje urejevalnike besedil. Za ponazoritev si oglejmo še dobrih sto iztočnic na *pred-*, deljenih s to metodo. Vzorec je bil med drugim izbran tudi zato, ker je med temi besedami veliko zloženk: pri njih se pojavijo težave pri strojnem ločevanju delov (npr. *predel_ovalen*, *pre_dahniti* proti *pred_ahniti* ali *pred_vojen* proti *pre_dvojen*). V preglednici 15 je z znakom podčrtano (_) označeno mesto, kjer je algoritem našel mejo iz več besed sestavljenega gesla, s pomišljajem (-) pa mesta, ugotovljena z entropijo n-terčic.

Preglednica 15: Prvih 105 iztočnic na pred-, deljenih z entropijo n-terčic

préd	pre_da-ní-ti	prèd-dôb-nost	pre_de-lí-tev	prèd_fílm
pred	pre_dá-nost	prèd_dr-žá-ven	pre_de-lí-ti	pred_gór-je
pred...	prèd_a-príl-ski	pred-dúr-je	pre_de-lo-vá-lec	pred_gó-vor
pre-dàh	pre_dá-ti	pred-dvér-je	pre-del_o-vá-len	prèd_go-vór-nik
pre_dah-ní-ti	pre-da-vá-len	prèd_dvòr	pre-de-lo-vál-ni-ca	prèd_gré-ti
pre-dá-ja	pre-da-vál-ni-ca	pred-dvór-je	pre_de-lo-ván-je	prèd_grét-je
pre_dá-jan-je	pre-dá-van-je	pre_de-ba-tí-ra-ti	pre_de-lo-vá-ti	prèd_gré-va-ti
pre_dá-ja-ti	pre-da-vá-telj	pre_dê-bel	pré-den	pred_hí-šen
pre-dá-jen	pre-da-vá-te-lji-ca	pré-dec	pré-de-nast	pred-híš-je
pre-dál	pre-da-vá-tel-jski	pre_de-já-ti	pré_den-ce	pred_hó-den
pre-dá-last	pre-dá-va-ti	pre-dél	pré-den_čast	pred-hód-ni-ca
pre-dál-ce	pre_dá-ven	pre-dé-la	préd_e-ni-ca	pred_hód-nik
pre-dál_čast	pred-ba-cí-va-ti	pre-dé-la-nost	pre-dên-je	pred-hód-niš-tvo
pre-dál-ček	prèd_bo-ží-čen	pre_dé-la-ti	pré-de-no	pred-hód-nost
pre-dál-čen	pred_čá-sen	pre-de-lá-va	pre-de-róč	pre-dí-ca
pre-dál-čje	pred_čás-nost	pre-de-lá-van-je	pre_de-sti-lí-ra-ti	pre-di-dóč
pre-dál-čnik	prèd_člo-véš-ki	pre-de-lá-va-ti	pre-de-sti-ná-ci-ja	prèd_í-gra
pre_dá-leč	pred_čút-je	pre_dél-ček	pre-de-sti-ní-ra-nost	pre-díh
pre-dá-len	prèd_dé-la-vec	pre-dé-lek	pre-de-sti-ní-ra-ti	pre_dí-ha-ti
pre-dál-nik	prèd_dé-lav-ka	pre_dé-len	pre_dé-ti	pre_díh-ni-ti
pred_álp-ski	prèd_dé-lo	pre_de-li-ká-ten	pre_dé-va-ti	pre-di-kánt

Ponazarjalno gradivo

Namenimo nekaj pozornosti še najpomembnejšemu delu razlag gesel, ponazarjalnemu gradivu. Sem sodijo iztržki in navedki iz razlag, ki so od vsega slovarskega gradiva še najbližji običajnemu besedilu. Vzemimo dva primera: najprej najpomembnejše geslo iz, po Matjažu Kmeclu, edinega slovenskega korena na *f-*:

frfotáti -ám tudi -óčem nedov. (á â, ọ́) **1.** *hitro, slišno mahati s perutmi:* obglavljena kokoš je še nekaj časa frfotala; ptica v njegovi roki je silovito frfotala // *frfotaje letati:* metulji frfotajo; splašena ptica je frfotala nad gnezdom / listje je frfotalo na vse strani; ekspr. gospa je kar frfotala po sobi *lahkotno, urno tekala* **2.** *plapolati, vihrati:* v vetru so mu frfotali lasje; zastave frfotajo; šal ji je frfotal okrog ram **frfotáje:** ptiči so se frfotaje razbežali **frfotajóč** -a -e: jata frfotajočih ptic; v vetru frfotajoča ženska krila

In še eno krajše, pri katerem je naveden tudi avtor navedka:

otávnica -e ž (â) nar. *(posušena) trava četrte košnje:* Senožet je treba samo malo prekopati .. pa bo arnice, otave in otavnika pa še otavnice, kolikor bo kdo hotel (C. Kosmač)

Geslo *frfotati*, kakor ni prav dolgo, vsebuje štiri krajše ponazarjalne enote, iztržke:

metulji frfotajo, zastave frfotajo, jata frfotajočih ptic
v vetru frfotajoča ženska krila

in osem daljših, navedkov:

obglavljena kokoš je še nekaj časa frfotala
ptica v njegovi roki je silovito frfotala
splašena ptica je frfotala nad gnezdom listje je frfotalo na vse strani
gospa je kar frfotala po sobi, v vetru so mu frfotali lasje
šal ji je frfotal okrog ram, ptiči so se frfotaje razbežali

V preglednici 16 si oglejmo najprej nekaj številk o ponazarjalnem gradivu na splošno:

Preglednica 16: Ponazarjalno gradivo v SSKJ

399.181	iztržkov in navedkov,
1.616.200	besednih oblik, od tega 222.176 različnih,
10.303.648	znakov

Gradivo je približno trikrat večje od vzorca, obdelanega v (Kristan et al. 1994) (3.806.201 znakov), ki obsega predvsem Sveto pismo nove zaveze, sodobno časopisno besedilo (Moj mikro, Dnevnik) in prozo Josipa Jurčiča (Deseti brat), Ivana Cankarja (Moje življenje), Mire Mihelič (Ure mojih dni, Bertl), Miloša Mikelna (Veliki voz) ter Iva Zormana (Deklica iz Mihovega mlina). Ker pa se slovar nanaša na tako rekoč vsa področja našega življenja, smemo upati, da bo seznam najpogostejših besed iz ponazarjalnega gradiva zelo zanimiv (preglednici 17 in 18).

Žal je ustrezno strojno orodje za lematizacijo besedila, postopek, pri katerem vse besedne oblike v besedilu spremenimo v osnovne, kakršne so npr. iztočnice v slovarju, še kako leto pred nami in v obeh seznamih zato namesto besed nastopajo njihove oblike.

Na začetku prve preglednice vidimo, da je malta, ki povezuje nosilce sporočila v besedilu, samostalnike, glagole in pridevnike, veliko manj raznorodna od teh gradnikov in zato njene oblike nastopajo z zelo visokimi pogostnostmi. Prvi nedvoumni samostalnik (v obliki *med* je prav gotovo veliko več predlogov kot samostalnikov), *človek*, je šele na 39. mestu.

Omenjeni pojav še dosti lepše vidimo iz krivulje rasti, ki kaže, kako hitro se polni kak seznam enot s pogostnostmi. Podana je z zvezo:

$$S_i = \frac{1}{N} \sum_{j=1}^{i} f_j, \quad i = 1, 2, \ldots, n \qquad (4)$$

kjer je v našem primeru S_i skupna pogostnost do i-te besedne oblike, n število vseh različnih besednih oblik (222.176), N število vseh besednih oblik (1.616.200), f_j pa pogostnost j-te oblike v po pogostnostih padajoče urejenem seznamu, kakršen je recimo tisti iz preglednice 17.

Preglednica 17: Najpogostejše 504 besedne oblike iz ponazarjalnega gradiva s pogostnostmi

je	81145	te	1478	otroka	714	pogled	514	čisto	403	nima	336	peči	294	vreme	263
se	34464	če	1450	misli	713	kdo	510	obleko	403	podjetje	336	smeh	294	govoril	262
v	34067	samo	1412	denar	712	stroj	510	otroke	402	dosti	335	jesti	292	piti	262
so	21898	otrok	1410	govoriti	712	noč	507	pet	400	mati	335	kraj	292	skupina	262
in	21679	veliko	1396	imel	700	hiša	502	vsako	398	njegovih	334	naredil	292	stara	262
na	20819	komu	1388	jezik	699	narediti	502	svojega	397	zmeraj	334	svetloba	292	vedenje	262
za	14255	ko	1385	bom	691	kmalu	501	tak	396	stran	333	bile	290	luč	261
z	13227	življenje	1385	razvoj	690	mogel	498	zrak	396	vodi	332	eno	290	mizi	261
s	10661	imeti	1259	veter	690	noge	498	misel	394	lase	329	red	289	nova	261
ga	8974	mi	1239	hiše	673	komaj	495	moč	392	njena	329	močno	286	solze	260
po	8935	oči	1238	ker	672	vas	493	vprašanje	390	vina	329	svojim	286	velike	260
ne	7774	tega	1235	iti	671	začel	493	hoditi	388	gozd	328	takoj	286	družbe	259
mu	7738	obraz	1213	fant	655	knjiga	490	kom	388	more	328	obrazu	285	nov	259
da	7385	brez	1184	vodo	650	ljubezen	489	konja	386	sebi	328	vlak	285	stoji	259
iz	6828	dan	1147	precej	649	hotel	487	film	383	gledati	326	zato	285	drug	257
ni	6230	njegovo	1134	človeka	639	mesta	486	kruh	383	njeno	326	umetnost	284	konec	257
to	5672	nad	1100	doma	639	nove	486	domov	381	a	325	poslušati	283	skupaj	257
si	4956	svojo	1098	preveč	639	stvari	486	države	379	barve	325	prava	283	vprašanja	257
že	4749	besede	1083	delu	636	let	480	leto	379	šola	325	ravnanje	283	prsti	256
pri	4448	njegova	1050	nas	636	velika	479	nebo	378	celo	324	oblika	282	čevlji	254
od	4429	ves	1036	bodo	634	teh	478	papir	375	ime	324	sobi	279	prišlo	254
še	4153	ali	1020	pesmi	631	živali	477	avtomobil	374	pismo	324	razmere	278	žena	254
bil	3939	ljudi	1019	njim	628	saj	475	denarja	374	videl	324	bodi	277	okna	253
bi	3621	mesto	1017	hišo	626	pravi	473	svoji	374	zemlje	323	družba	277	dogodek	252
bo	3478	kako	1013	blago	624	dve	470	nam	373	mleko	322	mizo	277	roman	252
kot	3418	jim	1009	srce	624	žival	467	skoraj	372	pol	322	naše	277	mimo	251
ta	3356	pot	1003	življenja	622	vode	466	zna	372	vzeti	320	postopek	277	sredstva	251
vse	3353	lahko	999	časa	620	prostor	463	rok	369	bolnik	319	popolnoma	276	umetnosti	251
pa	3287	čez	991	glas	617	druge	461	vasi	369	kamen	319	pride	276	določiti	250
kaj	3105	ki	990	svoj	609	niti	461	načrt	368	strah	319	ceste	275	letalo	250
do	3037	skozi	983	velik	607	bolezni	454	sobo	366	telo	319	plačati	275	pravice	250
med	2794	naj	972	lepo	603	vseh	452	voz	365	barva	318	vrat	275	rastlina	250
ob	2697	ljudje	963	roko	603	danes	448	toliko	363	knjigo	318	ženske	275	pesnik	249
ima	2655	nič	960	sonce	603	poti	448	bomo	362	število	318	cesto	274	država	248
o	2459	kakor	948	obleka	599	tri	446	drevo	362	zanj	318	hodi	274	krilo	248
tako	2371	hitro	918	vsak	595	strani	445	zemlja	362	beseda	317	jezika	274	postati	248
biti	2298	otroci	893	tla	589	dober	441	igrati	358	dovolj	317	odšel	274	reka	248
bilo	2274	dolgo	879	okrog	586	snov	440	priti	358	ura	317	sredi	274	stene	248
človek	2229	njegov	858	sneg	582	bolezen	437	besedo	357	list	316	palico	273	vojaki	248
bila	2140	on	854	ženska	580	slabo	437	sistem	357	rastline	315	pomoč	273	žalost	248
koga	1977	spet	853	dobiti	571	mestu	436	snovi	356	vsem	315	vam	273	doba	247
pred	1964	dati	832	gre	571	cesti	435	mora	355	smrt	314	deset	272	drevesa	247
zaradi	1932	rad	827	ipd.	565	kupiti	435	dal	351	imeli	312	drugega	272	srca	247
delo	1922	bolj	825	pes	565	oče	435	morje	351	njen	312	kje	272	ostal	246
več	1904	me	820	prišel	565	stvar	435	veje	350	mogoče	310	take	271	drevje	245
ji	1900	roke	805	dekle	564	položaj	433	moral	349	začela	310	način	270	kolo	245
zelo	1882	njegove	804	dva	563	zemljo	429	besed	348	sebe	309	mir	269	tisoč	245
nekaj	1856	niso	794	les	562	res	427	sveta	347	dni	308	pravo	269	ideje	244
kar	1827	vsi	791	sam	561	vsa	425	mož	345	drugo	308	stanovanje	268	njene	244
jo	1777	malo	789	drugi	558	leta	422	nisem	344	gibanje	308	jaz	267	službo	244
sta	1755	voda	782	novo	553	njegovi	419	ladja	343	krompir	308	napisati	267	oblast	243
k	1691	čas	775	prav	551	živeti	419	lasje	343	svojih	308	zakon	267	slika	243
dela	1670	smo	775	vso	541	kri	414	cesta	342	takrat	308	odpreti	266	politika	242
ti	1628	treba	774	naprej	538	boj	413	seboj	342	blaga	306	pogovor	266	ampak	241
tem	1615	delati	768	del	537	okoli	412	hiši	341	glavi	304	začeli	266	delavci	241
dobro	1567	težko	766	pesem	534	šel	412	njega	341	imela	304	hlače	265	kam	241
pod	1565	bili	757	tu	531	dobil	411	ogenj	341	dež	302	gojiti	264	ure	241
jih	1556	zdaj	752	počasi	529	konj	409	njegovega	340	skrbi	301	igra	264	dobra	240
tudi	1547	vedno	741	vino	528	prvi	408	očmi	340	nikoli	300	kaže	264	globoko	240
sem	1513	svet	736	tam	525	večkrat	407	postaviti	339	času	299	sadje	264	lesa	240
svoje	1508	glavo	733	tej	520	meso	405	postal	338	daleč	299	visoko	264	uspeh	239
le	1488	vrata	727	knjige	518	imajo	404	okno	337	vsega	298	nihče	263	nanj	238
proti	1480	boš	717	njem	518	pisatelj	404	dejanje	336	plašč	295	taka	263	ustnice	238

Preglednica 18: Abecedni seznam najpogostejših 504-ih besednih oblik iz ponazarjalnega gradiva

a	325	dni	308	in	21679	mati	335	noč	507	prav	551	srce	624	večkrat	407
ali	1020	do	3037	ipd.	565	me	820	noge	498	prava	283	sredi	274	vedenje	262
ampak	241	doba	247	iti	671	med	2794	nov	259	pravi	473	sredstva	251	vedno	741
avtomobil	374	dober	441	iz	6828	meso	405	nova	261	pravice	250	sta	1755	veje	350
barva	318	dobil	411	jaz	267	mesta	486	nove	486	pravo	269	stanovanje	268	velik	607
barve	325	dobiti	571	je	81145	mesto	1017	novo	553	precej	649	stara	262	velika	479
besed	348	dobra	240	jesti	292	mestu	436	o	2459	pred	1964	stene	248	velike	260
beseda	317	dobro	1567	jezik	699	mi	1239	ob	2697	preveč	639	stoji	259	veliko	1396
besede	1083	dogodek	252	jezika	274	mimo	251	oblast	243	pri	4448	strah	319	ves	1036
besedo	357	dolgo	879	ji	1900	mir	269	obleka	599	pride	276	stran	333	veter	690
bi	3621	določiti	250	jih	1556	misel	394	obleko	403	prišel	565	strani	445	videl	324
bil	3939	doma	639	jim	1009	misli	713	oblika	282	prišlo	254	stroj	510	vina	329
bila	2140	domov	381	jo	1777	mizi	261	obraz	1213	priti	358	stvar	435	vino	528
bile	290	dosti	335	k	1691	mizo	277	obrazu	285	prostor	463	stvari	486	visoko	264
bili	757	dovolj	317	kaj	3105	mleko	322	oče	435	proti	1480	svet	736	vlak	285
bilo	2274	drevesa	247	kako	1013	moč	392	oči	1238	prsti	256	sveta	347	voda	782
biti	2298	drevje	245	kakor	948	močno	286	očmi	340	prvi	408	svetloba	292	vode	466
blaga	306	drevo	362	kam	241	mogel	498	od	4429	rad	827	svoj	609	vodi	332
blago	624	drug	257	kamen	319	mogoče	310	odpreti	266	rastlina	250	svoje	1508	vodo	650
bo	3478	druge	461	kar	1827	mora	355	odšel	274	rastline	315	svojega	397	vojaki	248
bodi	277	drugega	272	kaže	264	moral	349	ogenj	341	ravnanje	283	svoji	374	voz	365
bodo	634	drugi	558	kdo	510	more	328	okna	253	razmere	278	svojih	308	vprašanja	257
boj	413	drugo	308	ker	672	morje	351	okno	337	razvoj	690	svojim	286	vprašanje	390
bolezen	437	družba	277	ki	990	mož	345	okoli	412	red	289	svojo	1098	vrat	275
bolezni	454	družbe	259	kje	272	mu	7738	okrog	586	reka	248	še	4153	vrata	727
bolj	825	država	248	kmalu	501	na	20819	on	854	res	427	šel	412	vreme	263
bolnik	319	države	379	knjiga	490	način	270	ostal	246	rok	369	šola	325	vsa	425
bom	691	dva	563	knjige	518	načrt	368	otroci	893	roke	805	število	318	vsak	595
bomo	362	dve	470	knjigo	318	nad	1100	otrok	1410	roko	603	ta	3356	vsako	398
boš	717	eno	290	ko	1385	naj	972	otroka	714	roman	252	tak	396	vse	3353
brez	1184	fant	655	koga	1977	nam	373	otroke	402	s	10661	taka	263	vsega	298
celo	324	film	383	kolo	245	nanj	238	pa	3287	sadje	264	take	271	vseh	452
cesta	342	ga	8974	kom	388	napisati	267	palico	273	saj	475	tako	2371	vsem	315
ceste	275	gibanje	308	komaj	495	naprej	538	papir	375	sam	561	takoj	286	vsi	791
cesti	435	glas	617	komu	1388	naredil	292	peči	294	samo	1412	takrat	308	vso	541
cesto	274	glavi	304	konec	257	narediti	502	pes	565	se	34464	tam	525	vzeti	320
čas	775	glavo	733	konj	409	nas	636	pesem	534	sebe	309	te	1478	z	13227
časa	620	gledati	326	konja	386	naše	277	pesmi	631	sebi	328	tega	1235	za	14255
času	299	globoko	240	kot	3418	ne	7774	pesnik	249	seboj	342	teh	478	začel	493
če	1450	gojiti	264	kraj	292	nebo	378	pet	400	sem	1513	tej	520	začela	310
čevlji	254	govoril	262	kri	414	nekaj	1856	pisatelj	404	si	4956	telo	319	začeli	266
čez	991	govoriti	712	krilo	248	ni	6230	pismo	324	sistem	357	tem	1615	zakon	267
čisto	403	gozd	328	krompir	308	nič	960	piti	262	skoraj	372	težko	766	zanj	318
človek	2229	gre	571	kruh	383	nihče	263	plačati	275	skozi	983	ti	1628	zaradi	1932
človeka	639	hiša	502	kupiti	435	nikoli	300	plašč	295	skrbi	301	tisoč	245	zato	285
da	7385	hiše	673	ladja	343	nima	336	po	8935	skupaj	257	tla	589	zdaj	752
dal	351	hiši	341	lahko	999	nisem	344	počasi	529	skupina	262	to	5672	zelo	1882
daleč	299	hišo	626	lase	329	niso	794	pod	1565	slabo	437	toliko	363	zemlja	362
dan	1147	hitro	918	lasje	343	niti	461	podjetje	336	slika	243	treba	774	zemlje	323
danes	448	hlače	265	le	1488	njega	341	pogled	514	službo	244	tri	446	zemljo	429
dati	832	hodi	274	lepo	603	njegov	858	pogovor	266	smeh	294	tu	531	zmeraj	334
dejanje	336	hoditi	388	les	562	njegova	1050	pol	322	smo	775	tudi	1547	zna	372
dekle	564	hotel	487	lesa	240	njegove	804	politika	242	smrt	314	umetnost	284	zrak	396
del	537	ideje	244	let	480	njegovega	340	položaj	433	sneg	582	umetnosti	251	žalost	248
dela	1670	igra	264	leta	422	njegovi	419	pomoč	273	snov	440	ura	317	že	4749
delati	768	igrati	358	letalo	250	njegovih	334	popolnoma	276	snovi	356	ure	241	žena	254
delavci	241	ima	2655	leto	379	njegovo	1134	poslušati	283	so	21898	uspeh	239	ženska	580
delo	1922	imajo	404	list	316	njem	518	postal	338	sobi	279	ustnice	238	ženske	275
delu	636	ime	324	ljubezen	489	njen	312	postati	248	sobo	366	v	34067	žival	467
denar	712	imel	700	ljudi	1019	njena	329	postaviti	339	solze	260	vam	273	živali	477
denarja	374	imela	304	ljudje	963	njene	244	postopek	277	sonce	603	vas	493	živeti	419
deset	272	imeli	312	luč	261	njeno	326	pot	1003	spet	853	vasi	369	življenja	622
dež	302	imeti	1259	malo	789	njim	628	poti	448	srca	247	več	1904	življenje	1385

Slika 8: Krivulja rasti za besedne oblike v ponazarjalnem gradivu

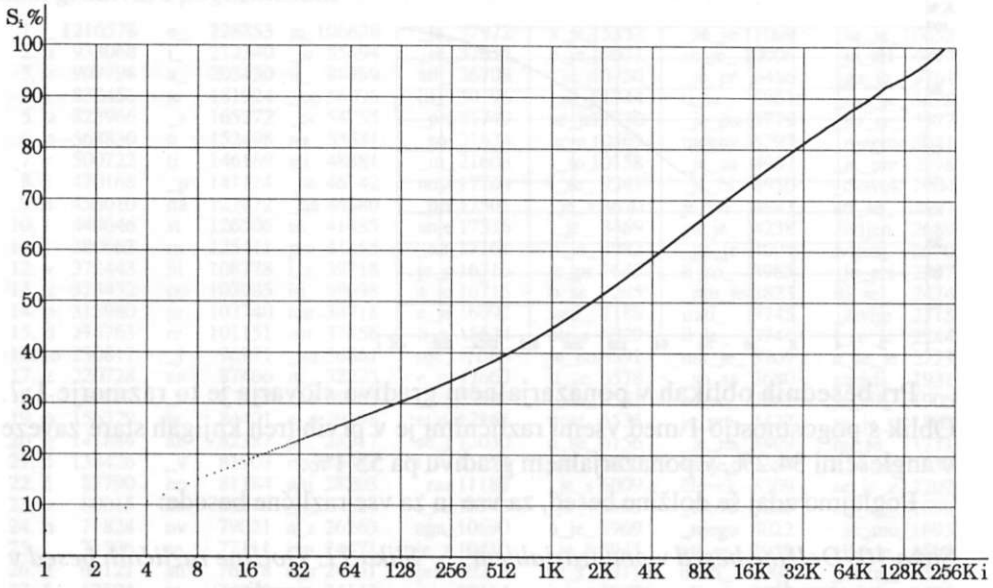


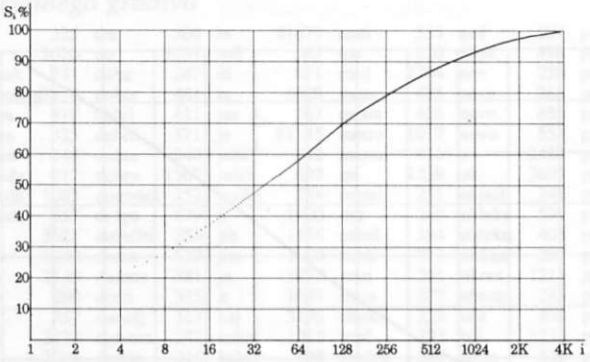
Diagram je pollogaritemski – ordinatna os je linearna in predstavlja S_i iz zveze (4) v odstotkih, abscisa predstavlja pa dvojiški logaritem od i: $\log_i$; K na abscisi pomeni 1.024, zaradi krajše pisave.

Poskusimo zdaj pike na krivulji razložiti še malo manj zavito. Prva predstavlja najpogostejšo besedno obliko, *je*, z deležem 5.02% med vsemi (pogostnost 81.145, vseh oblik je pa 1.616.200). Ko mu dodamo še *se* (2.13% delež), smo že na naslednji piki, ki ima za x 2, za y pa 7.15%. Dodamo še *v* (2.11% vseh oblik) in *so* (1.35%), pa smo pri četrti piki (x 4, y pa slabih 11%). Poglejmo še pomembnejše mejnike: s prvimi 49. besednimi oblikami, t. j. z

je, se, v, so, in, na, za, z, s, ga, po, ne, mu, da, iz, ni, to, si, že, pri, od, še, bil, bi, bo, kot, ta, vse, pa, kaj, do, med, ob, ima, o, tako, biti, bilo, človek, bila, koga, pred, zaradi, delo, več, ji, zelo, nekaj, kar

pokrijemo že četrtino vsega ponazarjalnega gradiva, če jih vzamemo 1.724, dobimo točno polovico gradiva, 17.515 besednih oblik potrebujemo za tri četrtine, 80.185 pa za 90%. Opazimo tudi, da potek krivulje ni regularen – približno pri 62.000 zaniha, nekako pri 90.000-i besedni obliki se pa dokončno zlomi. Pojav kaže na to, da naš vzorec ni zaokrožen, saj se pri večjih besedilih, še posebej v jezikih, ki ne poznajo toliko pregibanja kot slovenski, krivulja asimptotično približuje zgornji vodoravni črti, premici y = 100%. Tak primer so npr. prve tri knjige stare zaveze Svetega pisma, kjer je na 96.474 besednih oblik samo 3.867 različnih (Mejak, Holz 1995) in je razmerje različne : vse oblike kar 1 : 25:

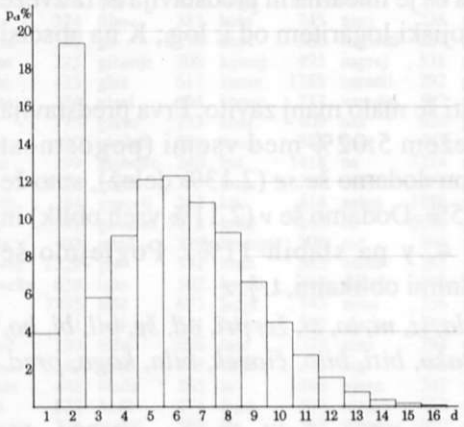
Slika 9: Krivulja rasti besednih oblik v delu »The Bible (The Old Testament I-III)«



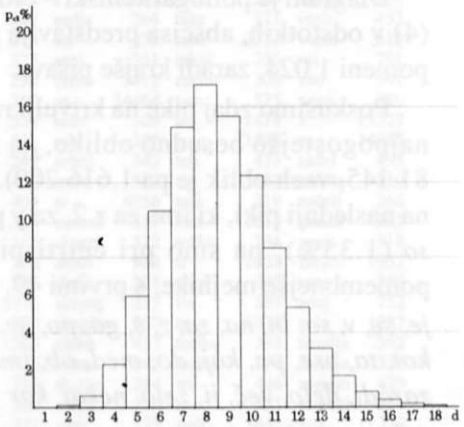
Pri besednih oblikah v ponazarjalnem gradivu slovarja je to razmerje 1:7. Oblik s pogostnostjo 1 med vsemi različnimi je v prvih treh knjigah stare zaveze v angleščini 34.2%, v ponazarjalnem gradivu pa 55.1%.

Poglejmo zdaj še dolžine besed, za vse in za vse različne besede:

Slika 10: Dolžine besed v ponazarjalnem gradivu



Slika 11: Dolžine različnih besed v gradivu



Podatki na sliki 11 se s tistimi s slike 2, kjer je navedena porazdelitev dolžin iztočnic, precej dobro ujemajo. Ustreznih rezultatov v (Kristan et al. 1994) žal ne najdemo, so pa podatki za dolžine vseh besed. Precej dobro se ujemajo s vrednostmi iz slike 10 – le prvi vrh, tudi pri dolžini 2 (besedice *je, se, so, in, na, za, ...*) je tam še višji, kar 27%, drugi vrh je pri dolžini 5 (v našem primeru 6) in rep histograma je pri nas močnejši. Razlike so, glede na naravo iztržkov, ki so navadno brez pomožnega glagola, v skladu s pričakovanji.

Preglednica 19: Najpogostejše črke, dvojčice, trojčice ... sedmerice v ponazarjalnem gradivu, s pogostnostmi

1.	_	1216578	e_	228853	je_	106628	_je_	77972	a_je_	15352	_se_je	11069	_se_je_	10653
2.	a	933068	i_	212340	_je	85494	_se_	37615	e_je_	15071	se_je_	10906	_je_bil	4877
3.	e	907798	a_	203430	ti_	84759	ati_	36708	_je_p	13750	_je_pr	6456	_ga_je_	3761
4.	i	832456	je	181924	_po	56096	iti_	30196	_se_j	11744	ti_se_	5961	_mu_je_	3628
5.	o	823966	_s	165272	_pr	54755	_pre	21749	se_je	11330	_je_po	5779	_so_se_	3527
6.	n	568830	o_	152498	na_	53531	_so_	21634	anje_	10565	njegov	5297	_njegov	3021
7.	r	500722	ti	146169	ati	48881	_in_	21608	i_so_	10158	_je_za	4967	_je_pre	2998
8.	t	470168	_p	141124	_se	46142	nost	17704	i_se_	9341	_je_bi	4930	_človek	2904
9.	s	458010	na	127872	_na	44380	_pri	17501	_je_z	8670	je_bil	4883	ati_se_	2697
10.	l	449646	st	126506	se_	41485	anje	17336	l_je_	8469	ga_je_	4238	življen	2689
11.	j	380663	ra	125411	pre	41455	_na_	17161	o_je_	8083	_ga_je	4004	ivljenj	2670
12.	v	372443	ni	108778	i_s	39718	je_p	16716	je_pr	7645	li_so_	3985	_je_pri	2537
13.	k	323432	po	103985	iti	39698	a_je	16715	o_se_	7285	_mu_je	3823	iti_se_	2474
14.	p	315980	pr	103740	nje	38718	e_je	16092	vati_	7188	irati_	3745	_življe	2315
15.	d	274763	re	101151	ost	37456	ti_s	15624	ati_s	6929	il_je_	3741	l_se_je	2244
16.	m	230817	_j	96971	_za	36861	nje_	15366	je_po	6891	mu_je_	3709	a_se_je	2223
17.	z	220728	en	87466	ni_	32223	e_pr	12663	ti_se	6578	_so_se	3680	zaradi_	1931
18.	u	166752	_n	87457	_so	29863	_za_	12605	,_da_	6541	so_se_	3562	je_bil_	1906
19.	b	156329	la	84401	e_p	29198	e_po	12162	nost_	6535	je_pre	3427	se_je_p	1809
20.	g	141884	al	82387	pri	28517	se_j	12030	a_se_	6156	_je_na	3409	bil_je_	1771
21.	č	138426	_v	81903	no_	28433	i_so	11924	jati_	6030	vanje_	3368	i_so_se	1730
22.	š	87790	no	81184	anj	28203	_raz	11188	_je_s	6009	človek	3309	se_je_z	1707
23.	c	80015	ko	80870	_v_	27524	i_se	10943	_je_v	5983	_človе	3236	ali_so_	1699
24.	h	71824	ov	79021	a_s	26263	ega_	10690	u_je_	5969	_njego	3022	_se_mu_	1693
25.	,	71506	se	77111	i_p	24471	je_z	10410	_je_b	5943	ati_se	2959	vljenje	1669
26.	ž	69121	an	76178	e_s	24361	prav	10253	iti_s	5911	ovati_	2901	_nekaj_	1555
27.	f	17504	at	74048	_in	24142	je_s	10154	i_je_	5887	vljenj	2886	je_bilo	1550
28.	é	2456	li	73123	sta	23829	ti_p	9638	_je_n	5789	al_je_	2878	ovanje_	1475
29.	á	1938	_z	72336	sti	23512	e_za	9568	je_za	5787	je_pri	2867	e_bilo_	1472
30.	.	1763	ri	72197	so_	23315	i_pr	9561	e_pre	5537	a_je_p	2794	_so_ga_	1459
31.	ó	1635	ka	72133	ne_	23044	e_na	9529	_pred	5453	_velik	2778	ti_koga	1429
32.	í	1449	ne	71989	raz	22916	i_po	9413	njego	5302	ivljen	2744	gospoda	1406
33.	ê	1029	,_	70659	_ko	22783	l_je	9327	jegov	5297	iti_se	2715	njegove	1394
34.	è	864	nj	70031	in_	22751	ost_	9024	vanje	5281	življe	2690	veliko_	1386
35.	S	668	el	69923	ga_	22646	ski_	8949	_je_o	5268	ovanje	2659	ospodar	1375
36.	?	644	ja	68407	a_p	21860	o_je	8828	nosti	5234	_besed	2635	o_se_je	1374
37.	P	533	il	67238	o_s	21530	i_na	8793	rati_	5043	e_je_p	2630	je_bila	1365
38.	ô	510	za	65009	_st	20997	a_po	8699	je_bi	5032	govori	2538	e_bila_	1352
39.	!	465	va	64363	nos	20454	osti	8673	govor	4962	ljati_	2482	_proti_	1346
40.	:	447	ve	64265	la_	20442	_ga_	8651	e_bil	4956	evati_	2449	_je_raz	1334
41.	-	423	_d	64205	a_j	18661	pred	8578	_svoj	4878	ti_na_	2436	i_koga_	1332
42.	K	382	te	64129	ova	18627	vati	8525	_so_s	4835	ljenje	2433	_svoje_	1320
43.	A	375	in	63837	_ne	18621	o_se	8235	niti_	4698	aviti_	2382	_veliko	1320
44.	ú	375	lo	63742	o_p	18516	o_po	8166	ga_je	4621	_življ	2319	ili_so_	1304
45.	J	323	od	62708	li_	18307	ska_	8141	i_in_	4586	a_se_j	2316	a_je_pr	1289
46.	T	322	os	62645	rav	18266	je_v	8139	je_na	4562	l_se_j	2292	e_je_po	1259
47.	M	319	le	62574	e_z	18259	ljen	8109	e_pri	4551	la_je_	2281	_da_je_	1258
48.	N	314	it	62537	ka_	18126	_del	7730	pravi	4516	_otrok	2248	_gospod	1238
49.	ŕ	288	ro	60543	e_j	17717	_pos	7729	ati_p	4509	i_so_s	2245	ljenje_	1234
50.	B	275	_k	60250	ko_	17624	je_n	7726	irati	4277	e_je_z	2095	_da_bi_	1230
51.	L	271	_o	59575	ali	17600	_pro	7725	li_so	4277	to_je_	2047	,_da_je	1222
52.	à	236	ta	59075	_do	17533	ti_k	7632	ti_ko	4215	_svoje	2017	e_je_pr	1206
53.	ì	231	av	57307	jen	17526	vanj	7614	e_in_	4191	aradi_	1932	ti_komu	1196
54.	V	220	lj	53360	_ra	17490	ali_	7613	e_so_	4134	zaradi	1932	pravlja	1190
55.	D	219	ar	52157	e_n	17404	_mu_	7591	i_na_	4078	e_bil_	1923	je,_da_	1188
56.	1	202	_m	51704	ki_	17284	a_pr	7587	il_je	4061	_bilo_	1909	evanje_	1187
57.	I	192	sk	51110	e_v	17283	a_se	7585	_ga_j	4009	nekaj_	1838	a_je_po	1186
58.	C	181	vo	51013	del	17264	ila_	7552	ti_na	3985	_je_v_	1835	e_mu_je	1170
59.	G	176	et	50732	red	17262	enje	7395	ovanj	3961	ati_ko	1830	i_komu_	1159
60.	R	169	_i	49927	_iz	16717	_bil	7273	mu_je	3927	jenje_	1827	raviti_	1153
61.	Z	169	me	48404	nik	16421	o_pr	7191	ljenj	3907	ali_so	1815	njegovo	1134
62.	0	166	to	47861	il_	16217	_da_	7168	človе	3895	_obraz	1805	jegovo_	1132

Povprečna dolžina besedne oblike, kjer so upoštevane vse oblike, ne samo vse različne, znaša 5.2 črke, dobre pol črke več kot 4.6, vrednost, izmerjena v že omenjeni raziskavi slovenskih besedil.

Zelo zanimive so tudi najpogostejše n-terice, ki so, od enojčic do sedmerčic, prikazane v preglednici 19 (n-terice tu niso razpete samo znotraj besed, ampak tudi čez besedne meje). Zelo pomembne so za določanje obvestilnosti oziroma entropije kakega besedila ali jezika na splošno.

V preglednici 19 so presledki označeni z znakom podčrtano. Takoj opazimo, da se kot rdeča nit med najpogostejšimi n-terčicami vlečeta *je* glagola biti in povratni osebni zaimek *se*; druge kratke besedne oblike pa za njima tudi ne zaostajajo dosti.

Število različnih n-terčic se od n na n-1 najprej skokovito, potem pa vedno počasneje povečuje, kar lepo kaže preglednica 20.

Preglednica 20: Število različnih in vseh n-terčkov v ponazarjalnem gradivu

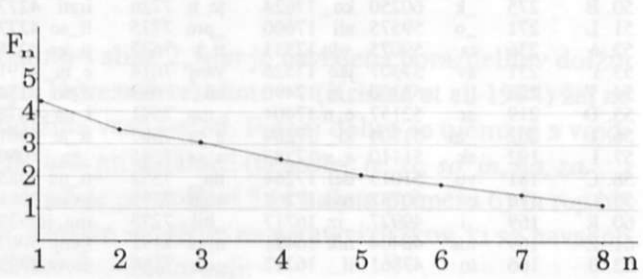
enojčice	106	10.303.648
dvojčice	1.067	9.904.230
trojčice	10.218	9.505.104
četverčice	66.289	9.106.062
peterčice	274.062	8.707.414
šesterčice	777.721	8.309.736
sedmerčice	1.600.735	7.913.616
osmerčice	2.569.124	7.519.853

Ustrezne številke za različne n-terčke v preglednici 5 raziskave (Kristan et al. 1994) so najbližje pri trojčicah (tam 9.477), od tod naprej pa se razlikujejo vedno bolj (osmerčic je bilo 1.670.456). Razliko lahko pripišemo predvsem velikosti vzorca. Pri pogojnih entropijah F_n (glej zvezo (2)), ki povedo, koliko bitov informacije potrebujemo, da bi določili n-ti znak v besedilu, če prejšnjih n-1 znakov že poznamo, ugotovimo, da so v ponazarjalnem gradivu višje kot v približno trikrat manjšem vzorcu omenjene raziskave; razlika je najpomembnejša pri n=7 in n=8:

Preglednica 21:

n	F_n	razlika
1	4.324	-0.081
2	3.468	-0.012
3	3.067	-0.068
4	2.554	-0.061
5	2.095	-0.005
6	1.765	-0.000
7	1.467	-0.051
8	1.151	-0.090

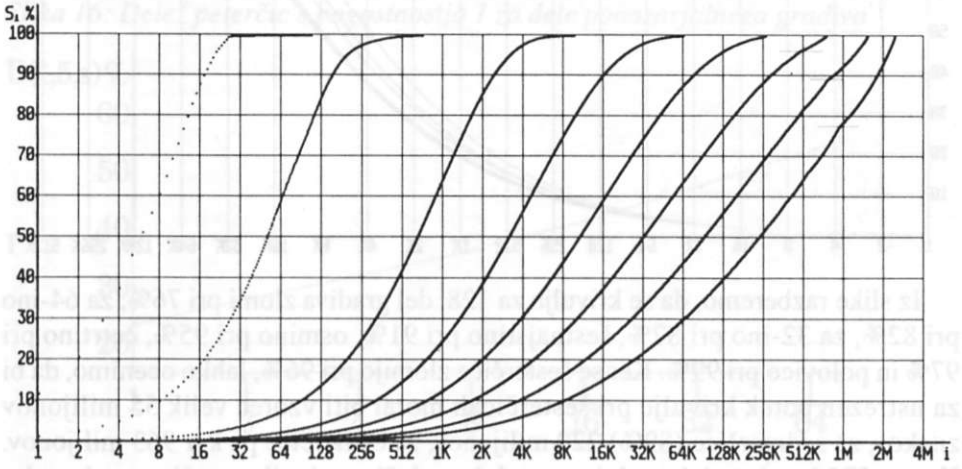
Slika 12: Pogojne entropije F_n



Da bo ponazarjalno gradivo entropijsko bogatejše od leposlovnih besedil, ni presenetljivo, saj mora na kratkem prostoru učinkovito dopolniti iztočnice; malo gre pa višja entropija, še posebej pri n=8, v prid le na prvi pogled presenetljivi ugotovitvi štirih raziskovalcev, da višje pogojne entropije z naraščanjem vzorca rastejo, ne pa padajo. Res je pa tudi, da je entropija vzorca lahko le približek za entropijo v dejanskem jeziku. Najmanjša zgornja meja za angleški jezik, ki so jo z ustreznim modelom, poiskali IBM-ovi raziskovalci (Brown et al. 1992), vzorec je bil dolg 583 milijonov besed, znaša recimo okrogla dva bita za poljubno dolge nize besedila. Ali, preprosteje povedano, verjetnost da bomo v angleškem besedilu pravilno zadeli naslednji znak, znaša eno četrtino.

V ostrejši luči se pokažejo n-terčice še s krivuljo rasti:

Slika 13: Krivulja rasti za črke, dvojice, trojčice, četverčice, peterčice, šesterčice, sedmerčice in osmerčice v ponazarjalnem gradivu



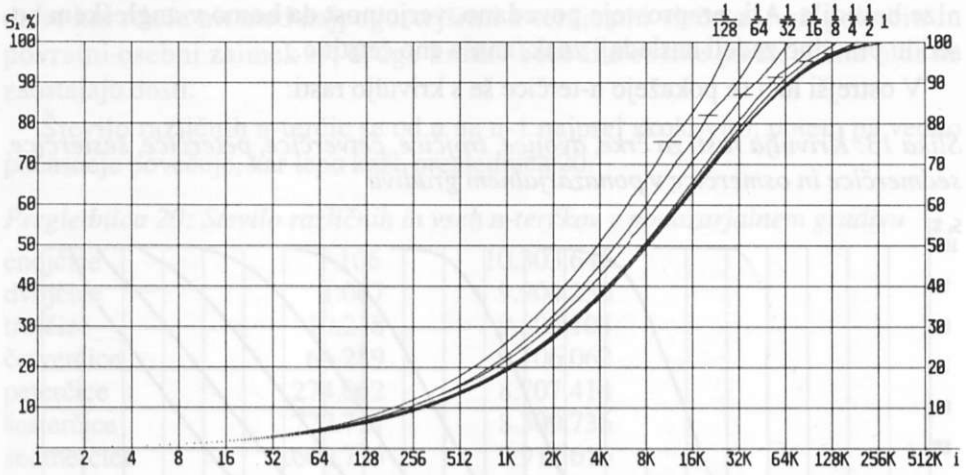
Potek krivulje pri enojčicah, ki se strmo dvignejo od presledka z 12% vsega gradiva (če mu dodamo še črke *a, e* in *i*, zaobsežemo že 37% tega besedila, z osmimi črkami 60%, s šestnajstimi pa že skoraj 90%), je tako pravilen, da je že grd. Tudi dvojčici rastejo zelo hitro, pri trojčicah je potek krivulje na oko nekako najlepši, še sprejemljiv pri četverčicah in peterčicah, pri šesterčicah, sedmerčicah in osmerčicah pa čedalje slabši.

Ponazarjalno gradivo je zanje očitno prekratek vzorec. Krivulje šesterčic, sedmerčic in osmerčic proti vrhu slike najprej zanihajo, potem se pa, mesto je označeno s črtkano črto, dokončno zlomijo. Šesterčice pri 96%, sedmerčice pri 89%, osmerčice pa že pri 77.5%.

Samo od sebe se vsiljuje vprašanje, kako velik bi moral biti vzorec besedila, da bi tudi zadnje tri n-terčice mehko zmogle pot do vrha.

Napravimo v ta namen majhen poskus, s peterčicami, katerih krivulja je bila zadnja še kolikor toliko sprejemljiva. Vzemimo najprej eno stoosemindvajsetino celotnega ponazarjalnega gradiva (to je še vedno 2992 iztržkov in navedkov) in napravimo krivuljo rasti na njene peterčke. Nadaljujmo z eno štiriinšestdesetino, dvaintridesetino, šestnajstino, osmino, četrtino, polovico in na koncu potegnimo krivuljo še za celo gradivo. Dobimo naslednjo sliko:

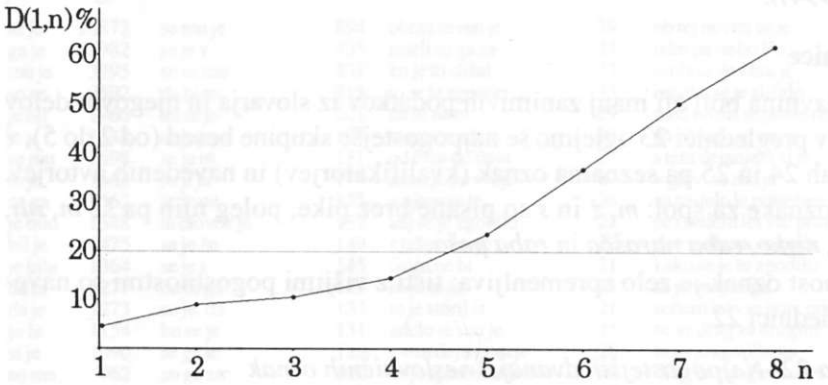
Slika 14: Krivulje rasti peterčkov za dele ponazarjalnega gradiva



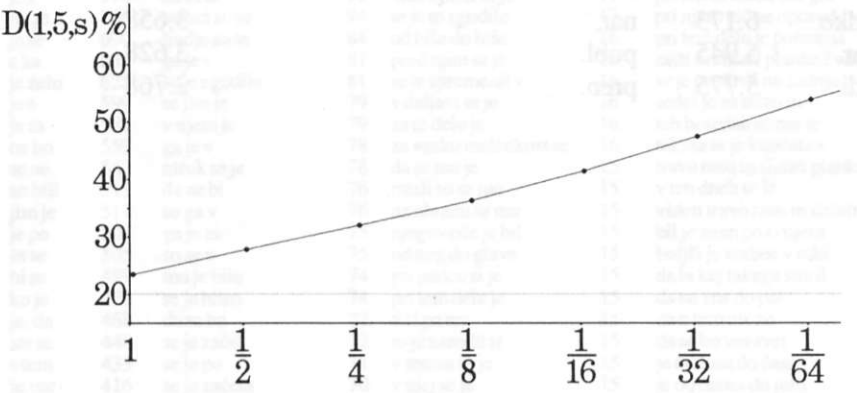
Iz slike razberemo, da se krivulja za 128. del gradiva zlomi pri 76%, za 64-ino pri 82%, za 32-ino pri 87%, šestnajstino pri 91%, osmino pri 95%, četrtino pri 97% in polovico pri 99%. Ker se šesterčice zlomijo pri 96%, lahko ocenimo, da bi za ustrezen potek krivulje pri šesterčicah moral biti vzorec velik 55 milijonov znakov, za sedmerčice (89%) 220 milijonov, za osmerčice pa kar 963 milijonov. Ker so IBM-ovi raziskovalci za model angleškega jezika vzeli v markovsko verigo drugega reda sestavljene trojčice – tri zaporedne trojčice, povezane med seboj (kar ustreza eni deveterčici), velikost njihovega vzorca, 583 milijonov besed ali približno 3.2 milijarde znakov, ne zveni več niti megalomansko niti čisto na pamet.

Poleg ugotavljanja, kdaj se krivulja zlomi, si lahko pomagamo še z enim kazalcem, ki je z njo v primeru n-terčic v precej pravilni zvezi. Gre za delež tistih n-terčic med vsemi različnimi, ki imajo pogostnost 1 – takih torej, ki se pojavijo samo enkrat. V preglednici 15 vidimo te deleže za n-terčice celega gradiva, v 16 pa za peterčice na posameznih delih:

Slika 15: Delež n-terčic s pogostnostjo 1 med vsemi različnimi



Slika 16: Delež peterčic s pogostnostjo 1 za dele ponazarjalnega gradiva



Na podlagi slike 15 in 13 lahko sklepamo na ugoden potek krivulje rasti pri n-terčkih, če med njimi takih s pogostnostjo 1 ni več kot petina.

Preden se poslovimo od ponazarjalnega gradiva, si oglejmo še seznam najpogostejših parov besed. Na naslednji strani, v preglednici 23, so navedene najpogostejše skupine po dveh, treh, štirih in petih besed. V tem seznamu že brez krivulje rasti takoj ugotovimo, da bi moral biti vzorec z lepo zaokroženim seznamom že, recimo, skupin po pet besed, izredno velik. Zadnji par v prvem stolpcu, *da bo*, ki ima pogostnost 311, bi se v drugem stolpcu uvrstil na tretje mesto, zadnja iz drugega in tretjega stolpca, bi se uvrstila v naslednji stolpec pa že kar na prvo mesto. V vsej množici besed ponazarjalnega gradiva imata najpogostejši skupini po pet besed, *ob tej novici se je* in *tako pa ne bo šlo*, komaj pogostnost 10. Tudi med pogostimi skupinami besed, kot že pri n-terčicah, močno prevladujejo kratke besedne oblike, predvsem dve črki dolge. Zelo malo najdemo

klasičnih stavčnih gradnikov, recimo povedkovih zvez (npr. *videti travo*, po (Orešnik 1994)).

Drobtinice

Ko je glavnina bolj ali manj zanimivih podatkov iz slovarja in njegovih delov za nami, si v preglednici 23 oglejmo še najpogostejše skupine besed (od 2 do 5), v preglednicah 24 in 25 pa seznama oznak (kvalifikatorjev) in navedenih avtorjev. Slovnične oznake za spol: *m*, *ž* in *s* so pisane brez pike, poleg njih pa še *in*, *ali*, *tudi*, *redko*, *nizko*, *raba narašča* in *raba peša*.

Pogostnost oznak je zelo spremenljiva; tisti z višjimi pogostnostmi so navedeni v preglednici 22.

Preglednica 22: Najpogostejših dvanajst neslovničnih oznak

ekspr.	29.970	pog.	5.276
knjiž.	13.742	zastar.	4.355
in	6.910	nav.	4.153
redko	6.173	nar.	3.658
star.	5.945	publ.	3.628
tudi	5.775	pren.	2.768

Preglednica 23: Pogoste skupine po 2, 3, 4 in 5 besed s pogostnostmi v ponazarjalnem gradivu

se je	10872	se mu je	894	obraz se mu je	39	ob tej novici se je	10
ga je	3982	se je v	435	imeli so ga za	33	tako pa ne bo šlo	10
mu je	3795	so se mu	231	ko je to slišal	33	od časa do časa je	9
so se	3592	da bi se	215	to se je zgodilo	33	iz sobe se je slišalo	8
je bil	1966	da se je	201	da se ne bi	29	med njima je prišlo do	8
je v	1843	se je z	194	se mu je zdelo	28	na obrazu se mu je	8
se mu	1698	se je in	181	od časa do časa	27	s tem dejanjem si je	8
to je	1663	se ji je	179	znan je po svoji	27	v glavi se mu je	8
so ga	1562	se je na	155	v njem se je	26	za to delo je potrebna	8
je bilo	1548	ta človek je	152	kaj se je zgodilo	22	čez sedem let vse prav	7
bil je	1475	se je že	149	s težavo se je	22	dve muhi na en mah	7
je bila	1364	se je s	145	da ga ne bi	21	kako se je to zgodilo	7
da bi	1277	da bi ga	135	se je in se	21	se je vanj in ga	7
da je	1273	se je, da	133	to je storil iz	21	sedem let vse prav pride	7
je že	1154	ko se je	131	zdelo se mu je	21	so se drug za drugim	7
si je	1090	se ga je	123	s tem dejanjem je	20	to se je zgodilo na	7
so mu	962	so ga za	120	se je spremenila v	20	drug za drugim so se	6
ji je	937	ga je z	113	se mu je, da	19	enem ušesu mu gre noter	6
da se	911	se je za	106	v njenih očeh je	19	iz sobe je bilo slišati	6
je na	898	na vse strani	102	drug za drugim so	18	ne zna do pet šteti	6
jo je	874	so se ji	101	se mu je v	18	pa naj bo po tvojem	6
je z	774	da se ne	99	med njima se je	17	pri enem ušesu mu gre	6
jih je	706	otroci so se	94	se je to zgodilo	17	pri njem nič ne opraviš	6
je še	690	ljudje so se	84	od hiše do hiše	16	pri tem delu je potrebna	6
e za	637	ga je s	81	pred njim se je	16	rasti in slišati planke žvižgati	6
je zelo	628	se je zgodilo	81	se je spremenil v	16	se je postavil na zadnje	6
je s	599	se jim je	79	v daljavi se je	16	sedel je za mizo in	6
je in	581	v njem je	79	za to delo je	16	teh besedah se mu je	6
ne bo	559	ga je v	78	za vsako malenkost se	16	tekma se je končala s	6
se ne	542	otrok se je	78	da se mu je	15	travo rasti in slišati planke	6
so bili	519	da ne bi	76	misli so se mu	15	v teh dneh se je	6
jim je	517	so ga v	76	na obrazu se mu	15	videti travo rasti in slišati	6
je po	507	ga je za	75	njegov oče je bil	15	bil je znan po svojem	5
in se	505	so se v	75	od nog do glave	15	boljši je vrabec v roki	5
bi se	489	mu je bilo	74	pri padcu si je	15	da bi kaj takega storil	5
ko je	479	se je hitro	74	pri tem delu je	15	da ne zna do pet	5
je, da	468	da se bo	72	ti si pa res	15	da o tem nič ne	5
sta se	448	se je začel	72	to je naredil iz	15	da se bo ves svet	5
s tem	433	se je po	71	v mestu se je	15	je od časa do časa	5
je vse	416	se je začela	70	v njej se je	15	je od danes do jutri	5
on je	414	to je bilo	70	znašel se je v	15	je od nog do glave	5
se ji	406	se je zelo	68	bil je znan kot	14	je vrabec v roki kakor	5
je treba	397	je bil v	67	da se je kar	14	kaj takega pa še ne	5
bila je	388	slišati je bilo	67	iz sobe se je	14	kraja ga pa nič ni	5
ne bi	384	njegove besede so	66	oči so se ji	14	le s težavo se je	5
v tem	360	pa se je	66	odločil se je za	14	misliš, da se bo ves	5
otroci so	359	to je bil	66	pazi, da se ne	14	misliš, da ti bodo tam	5
o tem	342	ga je po	65	polastila se ga je	14	na levo ne na desno	5
se ni	342	je in se	65	ta človek je pravi	14	na stara leta se je	5
je kot	341	se je vrnil	65	glas se mu je	13	ne na levo ne na	5
je nekaj	341	da ga je	64	kaj bo iz tega	13	njegova življenjska pot ni bila	5
ni mogel	338	drug za drugim	64	kdo bi si mislil	13	o tem ni da bi	5
se bo	335	mu je v	64	njegove besede so bile	13	ob teh besedah ga je	5
je imel	334	da so se	62	ob teh besedah je	13	ob teh besedah se mu	5
da ne	330	je prišlo do	62	se mu je zdela	13	okrog kraja ga pa nič	5
so že	328	bil je velik	61	v roki je držal	13	pojesti vso modrost z veliko	5
mi je	325	da ne bo	61	z veje na vejo	13	pred očmi se mu je	5
da so	324	očitali so mu	60	iz dneva v dan	12	pri padcu si je zlomil	5
da ga	319	so se na	60	iz rok v roke	12	roki kakor golob na strehi	5
je le	318	v sobi je	60	je bilo tudi nekaj	12	se je v zadnjem času	5
z njim	318	ne da bi	59	je to slišal, je	12	se mu je zataknila v	5
da bo	311	se je kot	59	njene besede so ga	12	spoštovali so ga zaradi njegove	5

Preglednica 24: Oznake v SSKJ

adm.	166	fin.	309	med.	1812	petr.	169	šalj.	354
aer.	119	fiz.	959	medm.	617	pisar.	59	šol.	528
agr.	1301	fot.	213	mest.	120	pog.	5276	šport.	1011
ali	638	friz.	19	metal.	299	polit.	172	št.	10
alp.	263	gastr.	445	meteor.	199	poljud.	139	štev.	123
anat.	718	geod.	39	min.	336	pooseb.	23	teh.	1540
ant.	95	geogr.	408	mitol.	23	predl.	122	tekst.	402
antr.	70	geol.	210	mn.	2181	preg.	508	tisk.	280
arheol.	179	geom.	452	mont.	129	preh.	246	tož.	77
arhit.	199	gl.	859	muz.	821	pren.	2768	trg.	68
astr.	244	gled.	211	nam.	177	prid.	20594	tudi	5775
avt.	248	gost.	40	nar.	3658	pril.	424	tur.	65
bibl.	76	gozd.	262	nav.	4153	prim.	814	um.	456
biblio.	69	grad.	404	nav.mn.	1	prisl.	4916	urb.	59
biol.	751	igr.	130	navt.	400	psih.	176	usnj.	87
bot.	1651	im.	65	nedov.	7995	psiht.	84	var.	12
brezoseb.	718	in	6910	neprav.	31	publ.	3629	vet.	436
čeb.	214	ipd.	602	nepreh.	91	raba narašča	16	vez.	86
čl.	1	iron.	229	neskl.	752	raba peša	336	voj.	604
daj.	86	itd.	7	neskl.pril.	1	rad.	137	vrtn.	351
dov.	10052	jur.	1440	nestrok.	54	redko	6173	vulg.	165
dv.	54	kem.	1013	neustalj.	96	rel.	1024	vznes.	298
ed.	288	knjiž.	13742	nizko	262	rib.	122	zaim.	131
ekon.	637	kor.	39	num.	58	rod.	77	zal.	66
ekspr.	29970	kozm.	51	obl.	75	s	8247	zastar.	4355
elektr.	749	l. r.	1	obrt.	391	sam.	926	zgod.	645
elipt.	368	les.	338	or.	41	slabš.	2182	zool.	1510
etn.	440	lingv.	1248	os.	271	soc.	137	ž	21511
evfem.	341	lit.	584	otr.	119	star.	5945	žarg.	1249
farm.	140	ljubk.	109	pal.	55	stil.	755	žel.	209
filat.	39	lov.	358	papir.	93	strojn.	417		
film.	150	m	21943	ped.	153	strok.	105		
filoz.	469	mat.	592	pesn.	289	šah.	249		

Preglednica 25: V slovarju navedeni avtorji, s pogostnostmi večjimi od ena

L. Andrejev		Homer		A. Novačan	
- J. Vidmar		- A. Sovre	2	V. Ocvirk	
A. Aškerc	4	A. Ingolič	4	B. Pahor	7
H. Balzac		J. Jalen	11	J. Pahor	
- O. Župančič		M. Jarc	3	R. Polič	
F. Bevk	32	S. Jenko		I. Potrč	12
M. Bor		J. Jurčič	3	I. Pregelj	9
H. Bratož		Kajuh	3	F. Prešeren	26
I. Cankar	25	E. Kardelj		Prežihov	48
E. Cevc		V. Kavčič		M. Pugelj	
J. Conrad		B. Kidrič	2	A. Rebula	3
- O. Župančič		M. Klopčič	2	E. Rostand	
D. Debič		E. Kocbek		- O. Župančič	
Delo		I. Koprivec	11	W. Shakespeare	
F. Detela		C. Kosmač	15	- M. Bor	
Ch. Dickens		S. Kosovel	3	W. Shakespeare	
- O. Župančič		F. Kozak		- O. Župančič	4
D. Druškovič		A. Kraigher	2	G. Strniša	
J. Dular	3	L. Kraigher		T. Svetina	
F. Erjavec	2	M. Kranjec	32	D. Šega	
L. Fatur		B. Kreft	2	L. Škerjanc	
F. Finžgar	18	F. Levstik	8	I. Šorli	
N. Gaborovič		D. Lokar	5	I. Tavčar	10
J. Galsworthy		A. Luther		Tito – M. Močnik	
- O. Župančič	2	- M. Klopčič		J. Trdina	4
J. Glazer		A. Medved	2	J. Vidmar	
F. Godina	17	J. Mencinger		V. Vodnik	
P. Golia		K. Meško	2	S. Vuga	
K. Grabeljšek	4	M. Mihelič		S. Vuk	
A. Gradnik	2	Molière		P. Zidar	3
S. Gregorčič	7	- J. Vidmar		B. Ziherl	
I. Gruden		J. Murn	2	I. Zorec	9
M. Hus		R. Murnik		O. Župančič	25

V zadnji preglednici, nosi številko 26, si v razlagah oglejmo še nekaj letnic: nekatere so bolj znamenite, druge spet manj.

Preglednica 26: Letnice v slovarju in njihove pogostnosti:

1830	9	1929	16	1962	10
1848	39	1930	9	1963	7
1861	7	1941	24	1965	15
1900	7	1943	11	1968	7
1917	11	1945	131	1970	7
1918	27	1946	10	1980	20
1919	17	1950	7	1989	8
1920	7	1954	7	1990	11

Sklep

Številk je bilo veliko, lahko bi bila kakšna manj, kaka zanimivost je pa gotovo še ušla. Vse navedeno vsekakor utrjuje prepričanje, da je jezik zelo bogat in raznoter in da je tudi na področju, ki se ga loteva prispevek, še dosti raziskovalnega prostora. Predvsem podatki, dobljeni za n-terice in povezane skupine besed, kažejo, da bo treba v bodoče segati po vsaj še za velikostni razred večjih vzorcih besedila. Novi, bistveno hitrejši in zmogljivejši stroji, ki se obetajo že nekaj časa, bodo imeli veliko dela in ne bodo niti prehitri niti preveliki.

Naj se ob koncu zahvalim še prof. dr. Jožetu Toporišiču, ki mu ni bilo žal časa in truda, da je prispevek jezikovno popolnejši, bolj gladek in bogatejši za marsikatero besedo, ki se je pisec teh vrstic sam gotovo ne bi mogel spomniti.

Navedenke:

Slovar slovenskega knjižnega jezika, 1994. Inštitut za slovenski jezik Frana Ramovša ZRC SAZU, Državna založba Slovenije, Ljubljana.

P.F. BROWN, V.J. Della PIETRA, S.A. Della PIETRA, J.C. LAI, R.L. MERCER, 1992. An Estimate of an Upper Bound for the Entropy in English. Computational Linguistics, Vol. 18, No. 1, 31–40.

P. JAKOPIN, 1981. *Entropija imen in priimkov v Sloveniji*. Magistrsko delo. Zagreb: Sveučilište u Zagrebu.

P. JAKOPIN: *Deljenje besed s pomočjo entropije n-terčkov*. Rokopis, Ljubljana 1995.

B. KRISTAN, M. JUG, S. KOVAČIČ, L. GYERGYÉK: *Entropija slovenskih besedil*. Elektrotehniški vestnik, Vol. 61, št. 4, 171–179, 1994.

T. MEJAK, N. HOLZ: *Statistični prikaz dela The Bible (The Old Testament I-III)*. Seminarska naloga pri predmetu »Besedilo in računalnik«, Ljubljana: Filozofska fakulteta, 1995.

J. OREŠNIK: *Slovenski glagolski vid in univerzalna slovnica*. Ljubljana: Slovenska akademija znanosti in umetnosti, 1994.

SHANNON C.E.: *A Mathematical Theory of Communication*. Bell Systems Technical Journal, Vol. 27, 379–423, 623–656, 1948.

Slovensko bibliofilsko društvo: *Slovenski pesniški priročnik*. Ljubljana/Kamnik: SBD (spremna beseda Janez Menart), 1993.

Slovenski pravopis; Pravila. Ljubljana: Državna založba Slovenije, 1994.

J. TOPORIŠIČ: *Slovenski jezik in sporočanje*. Maribor: Obzorja, 1994.

Summary

The Dictionary of the Slovenian Literary Language (SSKJ), produced at the Institute for the Slovenian Language on the Academy of Sciences and Arts, has been published from 1970 to 1991, in five volumes. From 1992 till 1994 it has been, mostly

through OCR at the Institute, transferred into electronic form, to be available on CD late in 1995. The author, who has also written the text editor/OCR program used in the project, presents in this paper some quantitative data about the dictionary.

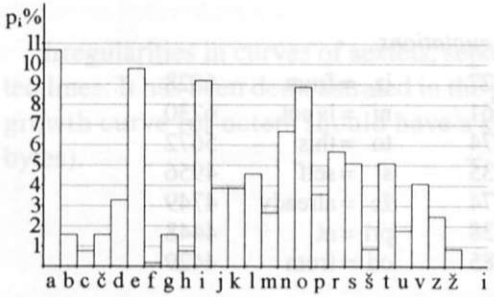
The dictionary includes 93,151 word entries with a total of 3,343,700 word forms of which 362,100 are different, all together 23,346,100 bytes long. The character set consists of 281 letters, numbers and special characters. The majority of the 93,151 headwords are nouns (51,448), followed by adjectives (21,516) and verbs (16,479). The twenty longest entries, with their English equivalents, are given below:

Table 1: Top 20 words with the lengths of their entries in bytes

príti	- to come	19,507	pásti	- to fall	10,761
íti	- to go	17,398	za	- for	10,479
takó	- so	15,958	okó	- eye	9,919
vzéti	- to take	12,105	rôka	- hand	9,776
iméti	- to have	12,056	držáti	- to hold	9,499
rêči	- to say	11,684	ujéti	- to catch	9,389
jêzik	- language	11,448	gláva	- head	9,346
tá	- this	11,375	dóber	- good	9,336
odpréti	- to open	10,936	têžek	- heavy	9,055
dáti	- to give	10,801	beséda	- word	9,054

Letter statistics, taken across the whole dictionary, reveal a somewhat different picture from its English counterpart – the most frequent letter is A and F is the least common (frequencies of q, w, x and y are negligible):

Figure 1: Distribution of letters in SSKJ



The distribution of headword letters is not much different, yet the histograms for distributions of first and last headword letters (Figures 2 and 3) bring a very different picture. For the initial letter, **p** is outstanding (po-, pod-, pri-, pred-), while for the final, verbs which in the infinitive all end with -i (such as *videti*, to see) have helped **i** to the first place, feminine nouns have brought **a** to second place (most end with -a, e.g., *raca*, the duck) and adjectives **n** to the third place (such as *strahoten*, terrible).

Figure 2: First keyword letters

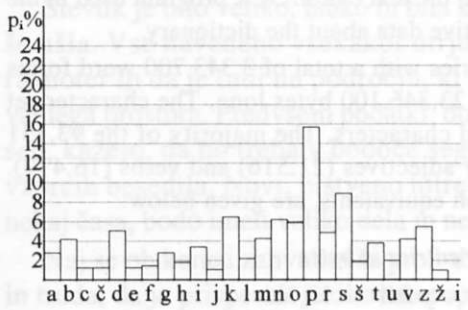
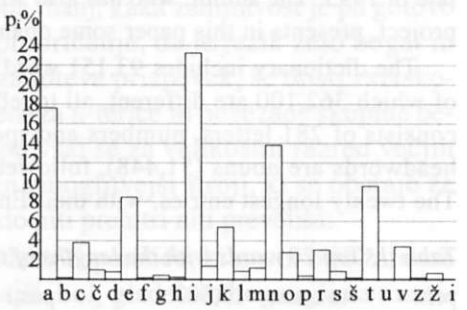


Figure 3: Last keyword letters



Basic statistics on headword letters, lengths and entry lengths are followed by lists of the most common n-tuples, strings of n characters, taken from headwords, their beginnings and endings. Conditional entropies for beginning and ending n-tuples up to n=10 are given, where it is of interest that after the initial fall from values at n=1 to values at n=2 both conditional entropies rise at n=3. Use of such n-tuples entropy is demonstrated by a simple, yet fairly effective algorithm for the division of words.

In the second half of the paper quotations from headword explanations (1,616,200 words and 10,303,648 bytes) are examined. The 504 words with the highest frequencies are given; the first 21, with approximate English translations, are shown in Table 2.

Table 2: Top 21 word forms from headword quotations

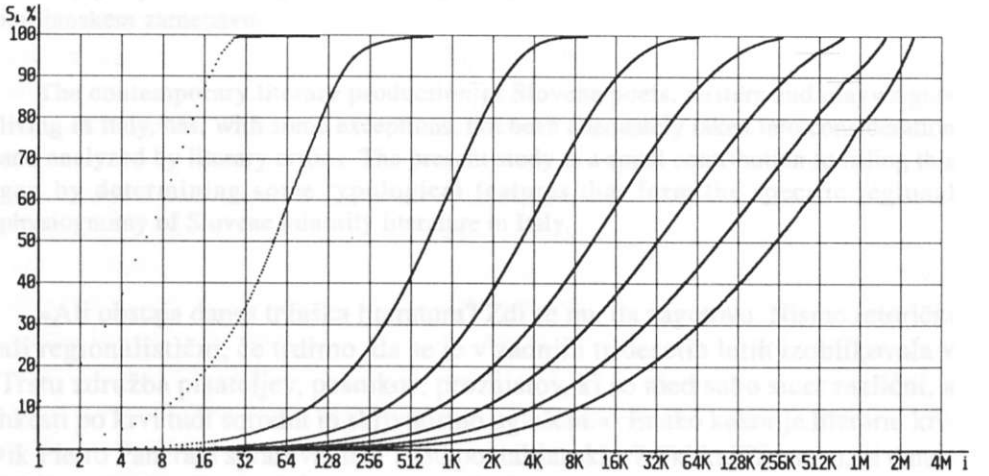
je	=is	81145	z	=with	13227	iz	=from	6828
se	=self	34464	s	=with	10661	ni	=is not	6230
v	=in	34067	ga	=him	8974	to	=this	5672
so	=are	21898	po	=after	8935	si	=self	4956
in	=and	21679	ne	=not	7774	že	=already	4749
na	=on	20819	mu	=him	7738	pri	=at	4448
za	=for	14255	da	=that	7385	od	=from	4429

N-tuples from the quotations part of the SSKJ, taken across boundaries (including punctuation and the word space), and their measure of information content, are discussed in detail. Growth curves for n-tuples from n=1 to n=8 (Figure 4) are shown and evaluated. Such curves, drawn for a list of elements (words, n-tuples) with their frequencies, sorted in descending order, are defined by the expression:

$$S_i = \frac{1}{N} \sum_{j=1}^{i} f_j, \quad i = 1, 2, ..., n \qquad (1)$$

where (in the case of Fig. 4) S_i stands for the sum of frequencies up to n-tuple i, n is the number of all different n-tuples, N the number of all n-tuples and f_j the frequency of n-tuple j.

Figure 4: Growth curves for the n-tuples from n=1 to n=8 (quotations):



Irregularities in curves of sextets, septets and octets are shown with small dotted lines. It has been demonstrated in the paper that the text sample with a smooth growth curve for octets should have a size in the vicinity of 1 GB (963 megabytes).