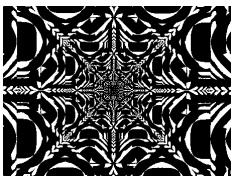


ZAKON O VELIČINI VOKABULARA TEKSTA
Heapsov zakon i određivanje veličine vokabulara tekstova na hrvatskom jeziku

Miroslav TUĐMAN
Filozofski fakultet, Zagreb

UDK: 811.163.42'32
004.9:81
Izvorni znanstveni rad
Primljeno: 29. 3. 2004.

Postojeća formula Heapsova zakona o veličini vokabulara teksta nije univerzalna te zakon treba redefinirati, kako bi se mogao rabiti za analizu korpusa na raznim jezicima. Analiza korpusa tekstova na hrvatskom jeziku potvrđuje hipotezu da je broj funkcionalnih pojavnica u tekstu konstantan te iznosi 21% veličine teksta. Autor dokazuje da se postotak funkcionalnih pojavnica u tekstu može uzimati kao vrijednost za parametar K te da je parametar K konstantna vrijednost za svaki jezični korpus. Empirijska istraživanja potvrđuju autorovu tezu da se broj funkcionalnih pojavnica u tekstu može izračunati po formuli $F = nK/100$, a da za veličinu najfrekventnije pojavnice (MF) vrijedi formula $MF = n\,(K/100)^2$. Vrijednost drugoga parametra Heapsova zakona također se može precizno odrediti i zato autor predlaže novi oblik zakona o veličini vokabulara teksta. Istraživanja potvrđuju da je vrlo visoka korelacija između izračunanih i stvarnih vrijednosti veličine vokabulara, odnosno između stvarnih i izračunanih vrijednosti jednokratnih riječi u tekstu. Ovako interpretiran i definiran zakon o veličini vokabulara teksta omogućuje izračun veličine vokabulara teksta na svakom jeziku, kada se zna postotak funkcionalnih riječi koji je konstantan za taj jezik. No ova interpretacija zakona omogućuje, osim izračuna veličine vokabulara teksta, i određivanje broja funkcionalnih pojavnica u tekstu, veličine najfrekventnije riječi u tekstu te broja jednokratnih pojavnica koje tvore vokabular teksta.

✉ Miroslav Tuđman, Filozofski fakultet Sveučilišta u Zagrebu, Odsjek za informacijske znanosti, I. Lučića 3, 10000 Zagreb, Hrvatska.
E-mail: miroslav.tudjman@zg.htnet.hr

Lotkin zakon o produktivnosti autora, Bradfordov zakon o pravilnosti razdiobe članaka po časopisima te Zipfov zakon o razdiobi riječi u tekstu – tri su temeljna zakona na kojima počivaju empirijska istraživanja u informacijskoj znanosti. Ovi zakoni doživjeli su svoje modifikacije i razne interpretacije (V. Oluić-Vuković, 1999.), a postoje i pokušaji da se sva tri zakona izvedu iz jedne jednadžbe ili svedu na jednu zajedničku logiku.

Mnogo je manje poznat u teoriji, a manje se primjenjuje i u praksi, Heapsov zakon.

Heapsov zakon (Heaps, 1978.) omogućuje izračunavanje fonda riječi, tj. veličine vokabulara, koji se rabi u nekom dokumentu ili nizu dokumenata. Zakon je formuliran na sljedeći način:

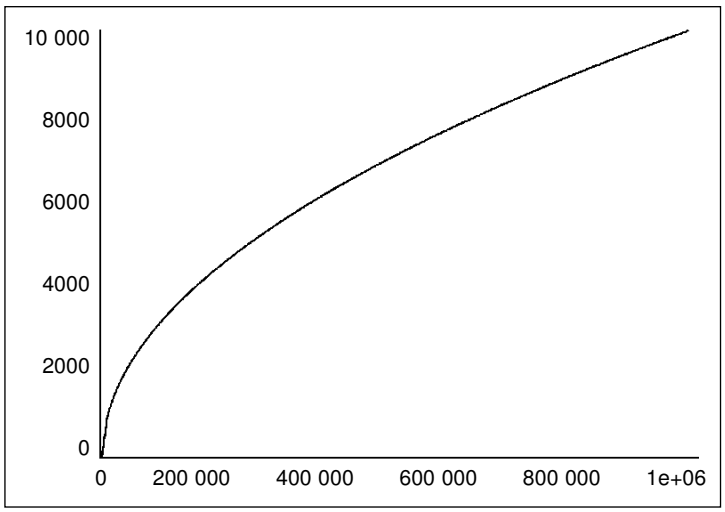
$$(1) \qquad V_R(n) = Kn^{\beta}$$

V_R je oznaka za dio vokabulara ($V_R \subseteq V$) koji se rabi u tekstu veličine n. K i β su parametri koji se određuju empirijski. Za korpus tekstova na engleskom jeziku vrijednost K je između 10 i 100, a za β iznosi:

$$0{,}4 \leq \beta \leq 0{,}6$$

Tipične razdiobe vokabulara prema Heapsovu zakonu prikazane su na sljedećem grafikonu.

➲ Grafikon 1
Razdiobe vokabulara prema Heapsovu zakonu*



* Na osi x prikazane su veličine teksta, a na osi y veličina vokabulara, tj. fonda riječi koji se rabi u tekstu određene veličine.
Izvor: http://www.PlanetMathOrg.PlanetMathHeaps'Law.htm

Heapsov zakon u teoriji informacijske znanosti nije nepoznat (Gelbukh, Sidorov, 2001.; Turner, 2001.), ali se o njemu manje raspravlja i manje se rabi nego ostali bibliometrijski zakoni: Bradfordov, Lotkin i Zipfov zakon. Tome su vjerojatno tri razloga.

Prvo, Heapsova formula za izračunavanje veličine vokabulara (tj. fonda riječi) nekoga teksta po svojoj formi odgovara metodi najmanjih kvadrata i iste rezultate možemo dobiti koristeći se standardnim statističkim programima (poput SPSS).

Drugo, Heapsova formula nije univerzalna jer ne vrijedi za sve jezike (Gelbukh, Sidorov, 2001.). Odnos fonda riječi, tj. upotrijebljenoga vokabulara, i veličine teksta različit je od jezika do jezika. Kao dokaz za ovu tezu dostatan je i primjer knjige Antoinea de Saint-Exuperyja "Mali princ" na osam jezika:

Tekst	Veličina teksta	Vokabular Vr	Funkcionalne riječi Fs
Mali princ – engleski prijevod	16884	2203	4953
Mali princ – njemački prijevod	14354	2664	3951
Mali princ – španjolski prijevod	13612	2865	4462
Mali princ – češki prijevod	13500	3141	2370
Mali princ – talijanski prijevod	13213	2558	4934
Mali princ – hrvatski prijevod	13029	3250	2595
Mali princ – poljski prijevod	12810	3364	1850
Mali princ – srpski prijevod	12411	3332	2335

TABLICA 1
"Mali princ" na osam jezika; osam tekstova i vokabulara različitih veličina

Iz ovoga se primjera vidi da je jedna te ista poruka ("Mali princ") na različitim jezicima iskazana različitom dužinom teksta. (Engleska verzija ima 36% više riječi od srpske verzije.) Odnos između veličine vokabulara i veličine teksta kreće se od 1:4 do 1:8. Ukupan broj funkcionalnih riječi također je različit: od 1/3 do 1/7 ukupnoga broja riječi u tekstu.

Treće, činjenica je da bi se prema Heapsovu zakonu vrijednosti parametara i K i β trebale određivati empirijski. Dosadašnja pak istraživanja potvrdila su njihove vrijednosti u prvom redu za engleski jezik. Da bi se Heapsov zakon mogao primjenjivati u istraživanju vokabulara tekstova na jezicima koji nisu engleski, trebalo bi izračunati vrijednosti parametara K i β za svaki jezik posebno. Takva istraživanja očito nisu brojna, a primjena Heapsova zakona u istraživanju korpusa tekstova na hrvatskom jeziku na samom je početku (M. Tuđman, 2003.; M. Tuđman i dr., 2003.).

PROBLEM

Problem kojim se bavimo u ovom istraživanju jest primjena Heapsova zakona u istraživanju vokabulara tekstova na hrvatskom jeziku. Zato je potrebno odrediti vrijednosti parametara K i β koje bi vrijedile za obradbu korpusa hrvatskih tekstova. Potrebno je odrediti empirijske vrijednosti parametara K i β ali i istražiti (međusobnu) uvjetovanost veličina tih parametara. Zato su pitanja na koja tražimo odgovore: u kakvim su odnosima vrijednosti parametara K i β s drugim odrednicama veličine vokabulara tekstova na hrvatskom jezi-

ku? Želimo istražiti može li Heapsov zakon naći svoju primjenu u izračunavanju veličine vokabulara tekstova na hrvatskom jeziku, ali i za izračunavanje broja funkcionalnih riječi, jednokratnih riječi te najfrekventnije riječi u pojedinim tekstovima ili korpusu tekstova.

O TERMINOLOGIJI

Terminologija koja se rabi u analizi jezičnih korpusa nije ustaljena. Jednako tako termini koji se rabe u primjeni Heapsova zakona nisu standardizirani ili se pak rabe uvjetno, s određenim ograničenjima. Zato ćemo izložiti kako smo upotrijebili neke osnovne termine vezane za tumačenje Heapsova zakona.

– *Veličina teksta* (*text size*) određena je ukupnim brojem riječi koje se pojavljuju u tekstu. No i *broj pojavnica* (*token*) uvriježen je naziv za ukupan broj riječi teksta, za sve riječi koje tvore neki tekst. Zato se veličina teksta i broj pojavnica teksta rabe kao termini koji označuju isti sadržaj.

– *Veličina vokabulara teksta* određena je fondom riječi koje se rabe u nekom tekstu. Vokabular teksta jest rječnik teksta, preciznije: vokabular je fond različnica (*types*) jednoga teksta. Veličina vokabulara i broj različnica teksta rabe se kao termini koji označuju isti sadržaj: broj različitih riječi u tekstu.

Različnice u istraživanju korpusa (engleskih i hrvatskih) tekstova nisu lematizirane (tj. svedene na osnovni oblik – nominativ ili infinitiv). Zato se (semantički) ista riječ može pojaviti više puta kao različnica u vokabularu teksta, ako se javlja u promijenjenu obliku (u različitim padežima ili vremenima).

Posljedica i ograničenje ovakva određivanja veličine vokabulara teksta jest u tome da flektivni jezici s razgrananom morfologijom imaju veći vokabular od jezika čija morfologija nije jednako razgranana.

– *Funkcionalne riječi* sinonim su za "gramatičke riječi", "prazne riječi" ili "stop words". Funkcionalne riječi tvore malen i konačan razred riječi kao što su prijedlozi (*u, na, s, od, do, pri* itd.), veznici (*i, ili, ni* itd.), odnosno članovi u nekim jezicima (npr. *a, the* u engleskom). Funkcionalne riječi tvore *zatvoreni razred* riječi (R. Carter, 1998.), one su malobrojne i same za sebe ne nose semantičke poruke, nego služe za gramatičku tvorbu teksta. Zato su to najfrekventnije riječi u tekstu.

– *Hapax legomena* (grč. u značenju "jednom izgovoreno") upotrebljava se kao termin u analizi veličine tekstova, kao oznaka za one riječi koje se javljaju jednokratno, samo jedanput, ili, preciznije, čija je frekvencija pojavljivanja jedan. Da bismo označili riječi koje se javljaju jednokratno u tekstu, služili smo se kao sinonimima terminima *hapax legomena*, *jednokratne riječi*, odnosno *jednokratnice*.

– *Višekratne riječi* ili *višekratnice* rabimo kao termin da bismo opisali one riječi u tekstu čija je frekvencija pojavljivanja veća

od jedan. Jednokratnice i višekratnice tvore vokabular teksta, s tim da se jednokratnice javljaju samo jedanput u cijelom tekstu.

– *Maksimalna frekvencija* jest oznaka vrijednosti za onu riječ koja se najviše puta javlja u nekom tekstu.

O METODI

Analizu vokabulara tekstova na hrvatskom jeziku provodili smo na korpusu koji se sastoji od 111 hrvatskih tekstova, ili ukupno 5.343.624 pojavnica. Tekstovi su preuzeti u digitalnom obliku i nisu posebno prilagođivani za ovu analizu. Pri odabiru tekstova nastojali smo uvrstiti samo prozne tekstove, tako da ne bude velike razlike u žanrovima. (Korpus od 111 tekstova nastao je od nova 42 teksta koja su pridodana korpusu od 69 tekstova što smo ih analizirali u radu M. Tuđman i dr., 2003.)

Kao kontrolnu grupu uzeli smo korpus od 35 tekstova na engleskom jeziku, korpus veličine 4.536.115 pojavnica. I ovaj korpus sastoji se pretežno od književnih tekstova, jer pretpostavljamo da na razlike u veličini vokabulara tekstova mogu utjecati i pojedini žanrovi sa svojim stilskim specifičnostima.

Svaki je tekst iz korpusa odabranih tekstova podvrgnut obradbi kako bi se dobili osnovni podaci o tekstu: veličina teksta (broj pojavnica), veličina vokabulara (broj različnica), broj funkcionalnih riječi, broj jednokratnih i broj višekratnih riječi, pregled različnica po frekvenciji (izdvojili smo maksimalnu frekvenciju u svakom tekstu). (Obradbom smo došli do niza podataka koji nisu predmetom ove analize: broj znakova, broj slova, broj paragrafa, broj rečenica teksta.) Softverski program za ovu obradbu izradio je prof. dr. Damir Boras u okviru projekta "Modeli znanja i obrada prirodnog jezika".

Nismo radili lematizaciju, tako da se rezultati analize veličine vokabulara tekstova mogu upotrijebiti samo za kvantitativne, a ne i za kvalitativne, prosudbe.

Statistička analiza podataka rađena je u Excelu, a grafički prikaz u Microsoft Wordu.

Rezultati statističke obradbe podataka korpusa tekstova na hrvatskom jeziku i korpusa tekstova na engleskom jeziku prikazani su u prilogu: M. Tuđman, D. Boras, N. Mikelić "Heapsov zakon i određivanje veličine vokabulara tekstova na hrvatskom jeziku. Dokumentacija" (Filozofski fakultet, Zagreb, 2004.).

Bez spoznaja do kojih smo došli analizirajući podatke prikazane u spomenutoj "Dokumentaciji" nije moguća analiza Heapsova zakona u ovom radu. Nažalost, zbog opsega dokumentacije prikupljene i obrađene analizom vokabulara tekstova na hrvatskom jeziku nismo mogli cijelu građu uvrstiti u ovaj članak, nego se možemo samo na nju pozivati. Međutim, osnovni podaci ovoga empirijskog istraživanja prikazani su u Tablici 2. Tako se rezultati istraživanja na koja upućuju sljedeći grafikoni i formule mogu provjeriti priloženim podacima u Tablici 2.

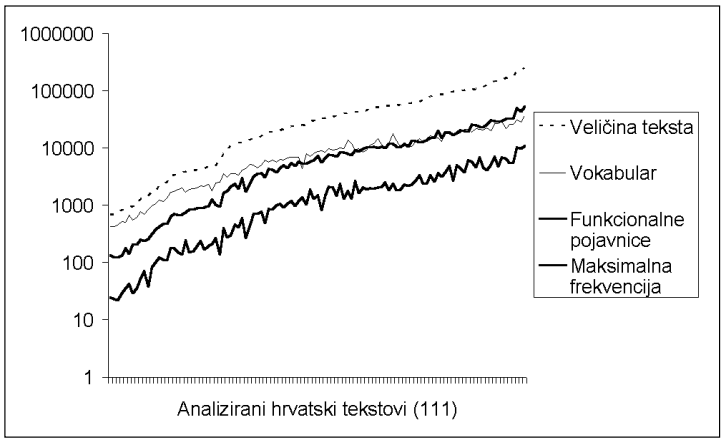
REZULTATI (1): KONSTANTE U HEAPSOVU ZAKONU

Broj funkcionalnih riječi je konstantan

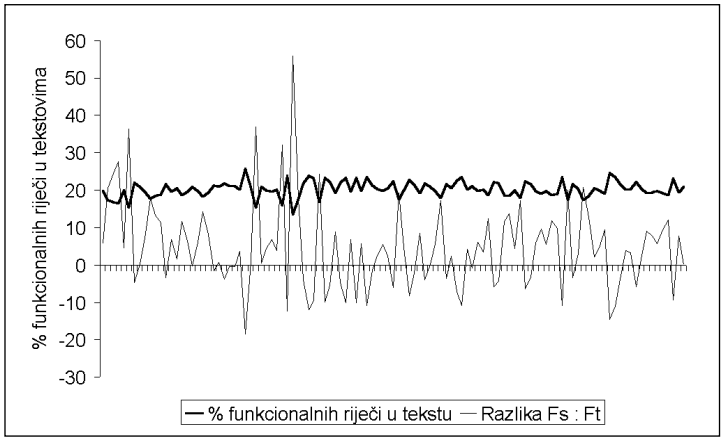
U Heapsovu zakonu K i β su parametri koji se određuju empirijski. No što utječe na njihovo određivanje i mogu li ovi parametri – ili jedan od njih – imati stalnu vrijednost za pojedine jezike?

Na Grafikonu 2 prikazan je rast vokabulara (V(n)), broja funkcionalnih pojavnica (Fs) i maksimalne frekvencije (MF) ovisno o veličini teksta (n). S veličinom teksta rastu i svi ostali pokazatelji i imaju isti trend rasta, svi osim funkcionalnih riječi. Naime, kod tekstova većih od 60.000 riječi broj funkcionalnih pojavnica postaje veći od veličine vokabulara. Međutim, pravi pokazatelj o odnosu broja funkcionalnih pojavnica i veličine teksta može se vidjeti na Grafikonu 3.

➲ GRAFIKON 2
Logaritamski prikaz osnovnih pokazatelja analiziranih tekstova



➲ GRAFIKON 3
Postotak funkcionalnih riječi i razlike između Fs i Ft



Analizom podataka o broju funkcionalnih pojavnica u tekstu možemo zaključiti da je u prosjeku 21% funkcionalnih riječi u svakom tekstu. Štoviše, čak i ukupan broj riječi cijeloga korpusa /N=5.343.624 podijeljene s ukupnim brojem funkcionalnih riječi Fr=1.093.554/ upućuje na isti zaključak jer daje sljedeći rezultat: 20,46% (Tablica 2).

➲ TABLICA 2
Osnovni podaci za analizu vokabulara hrvatskih tekstova

	Oznaka teksta	Veličina teksta: broj pojavnica	Vokabular: različnice	Izračunani vokabular VRt=(Nk)$^{2/3}$	Fs: broj funkcionalnih riječi	% funkcionalnih riječi	Izračun funkcionalnih riječi Ft=0,2In	Hapax legome na HLs: frekven.=1	HLt=((Nk)/2)$^{2/3}$	MFs: maksimalna frekvencija	MFt=N(0,21)$^{2/3}$
1	Miloš: Petar Pan	681	425	608	135	19,82	143	336	382	25	30
2	Miloš: Jezero	706	422	623	123	17,42	148	330	392	23	31
3	Miloš: Povratak	723	453	633	122	16,87	152	363	398	22	32
4	Miloš: Gibraltar	815	515	686	134	16,44	171	422	431	29	36
5	Miloš: Ex picador	886	509	726	178	20,09	186	395	456	36	39
6	Kovačina: Bakrena svila	922	681	745	142	15,40	194	592	468	43	41
7	Miloš: Penelope	949	544	760	209	22,02	199	416	477	29	42
8	Lokotar: Eseji o moru	999	617	786	209	20,92	210	508	494	36	44
9	Kovačina: Paučina	1281	779	929	251	19,59	269	641	584	53	56
10	Kovačina: Ljepota	1318	699	947	235	17,83	277	535	595	71	58
11	Miloš: Afrodizijak	1415	851	993	262	18,52	297	681	624	38	62
12	Nekoć i sad - pogovor	1702	985	1124	321	18,86	357	796	706	86	75
13	Školarac - pogovor	1758	1061	1148	382	21,73	369	866	722	100	78
14	Tonio Kroger - pogovor	2149	1218	1314	422	19,64	451	978	826	124	95
15	Ljubej: Priča	2271	1157	1363	469	20,65	477	865	857	113	100
16	Branisavljević: Mosq	2510	1324	1458	472	18,80	527	1044	916	111	111
17	Žudnja za ljubavlju - pogovor	3261	1699	1737	643	19,72	685	1312	1092	179	144
18	O tragičnom - pogovor	3336	1740	1764	702	21,04	701	1340	1109	181	147
19	Ljudska sudbina - pogovor	3403	1893	1787	674	19,81	715	1484	1123	152	150
20	Berlin Alexander Platz - pogovor	3723	2002	1898	684	18,37	782	1556	1193	139	164
21	Branisavljević: Vjeverica	3819	1666	1931	742	19,43	802	1242	1214	246	168
22	Vremenski stroj - pogovor	3927	1903	1967	837	21,31	825	1460	1237	150	173
23	Ivanhoe - pogovor	4055	1961	2010	846	20,86	852	1489	1263	161	179
24	Put do Indije - pogovor	4130	2014	2035	901	21,82	867	1539	1279	195	182
25	Goli i mrtvi - pogovor	4201	2162	2058	885	21,07	882	1713	1294	235	185
26	1984 - pogovor	4361	2131	2111	919	21,07	916	1608	1327	171	192
27	Kontrapunkt - pogovor	4789	2315	2247	970	20,25	1006	1746	1412	198	211

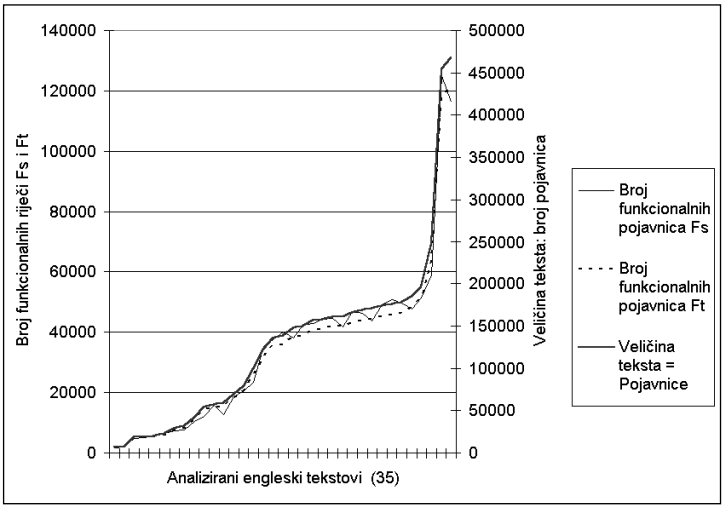
	Oznaka teksta	Veličina teksta: broj pojavnica	Vokabular: različnice	Izračunani vokabular VRt=(Nk)$^{2/3}$	Fs: broj funkcionalnih riječi	% funkcionalnih riječi	Izračun funkcionalnih riječi Ft=0,21n	Hapax legomena HLs: frekven.=1	HLt=((Nk)/2)$^{2/3}$	MFs: maksimalna frekvencija	MFt=N(0,21)$^{2/3}$
28	Kiš: Ženik	4862	1795	2270	1250	25,71	1021	1260	1427	212	214
29	Plodovi gnjeva - pogovor	4898	2466	2281	1019	20,80	1029	1907	1434	264	216
30	Mažuranić1	6134	2492	2653	941	15,34	1288	1692	1667	140	271
31	Blaga je noć - pogovor	7828	3581	3123	1634	20,87	1644	2730	1963	402	345
32	Tomić: Opančareva kći	8994	3091	3428	1808	20,10	1889	2125	2154	273	397
33	Pasagić	10927	3571	3906	2148	19,66	2295	2375	2455	300	482
34	Donadini: Bauk	11772	3556	4105	2380	20,22	2472	2478	2580	466	519
35	Mali princ (hrvatski)	13029	3250	4394	2595	19,92	2736	2011	2657	420	542
36	Aralica1	12480	4132	4269	2989	23,95	2621	2898	2683	597	550
37	Šenoa1	13001	4347	4388	1752	13,48	2730	2712	2758	264	573
38	Šenoa: Karanfil	13775	5183	4561	2455	17,82	2893	3696	2867	445	607
39	Kiš: Obijalo	14650	4968	4753	3203	21,86	3077	3572	2988	705	646
40	Aralica4	14813	4531	4789	3533	23,85	3111	3099	3010	712	653
41	Aralica3	15808	5074	5002	3672	23,23	3320	3448	3144	772	697
42	Šenoa: Prijan Lovro	18148	6227	5487	3064	16,88	3811	4202	3448	484	800
43	Aralica5	18647	5654	5587	4344	23,30	3916	3930	3512	873	822
44	Aralica2	18930	6123	5644	4201	22,19	3975	4318	3547	836	835
45	Donadini: Novele	19609	5630	5779	3782	19,29	4118	3930	3632	826	865
46	Tonio Kroger	20058	6321	5867	4421	22,04	4212	4386	3687	1074	885
47	Aralica9	21360	6205	6120	4982	23,32	4486	4184	3846	921	942
48	Novak1	22807	6607	6394	4487	19,67	4789	4442	4019	1050	1006
49	Aralica6	23494	6823	6523	5484	23,34	4934	4645	4100	1182	1036
50	Lisac: Savršeni krug	24512	6838	6711	4870	19,87	5148	4645	4218	981	1081
51	Aralica7	24917	6881	6785	5866	23,54	5233	4602	4264	1176	1099
52	Hlapić	24930	4402	6787	5302	21,27	5235	2367	4266	1388	1099
53	Jorgovanić	25702	7905	6927	5261	20,47	5397	5448	4354	1038	1133
54	Donadini: Kroz šibe	28727	7142	7463	5715	19,89	6033	4788	4691	1881	1267
55	Irena Vrkljan	30019	8011	7687	6160	20,52	6304	5182	4831	1311	1324
56	Školarac	32259	8568	8066	7211	22,35	6774	5530	5070	1512	1423
57	Šenoa: Čuvaj se	32269	9010	8068	5680	17,60	6776	5584	5071	814	1423
58	Wiesner-Livadić: Novele	32923	8437	8177	6593	20,03	6914	5586	5140	1450	1452
59	Kozarac: Đuka	33453	9620	8265	7648	22,86	7025	6768	5195	2109	1475
60	Kamov: Novele	34526	9056	8442	7396	21,42	7250	6075	5306	2065	1523
61	Leskovar: Propali dvori	36674	9384	8790	7100	19,36	7702	6173	5525	1444	1617
62	Gjalski: Dolazak Hrvata	38234	10006	9039	8355	21,85	8029	6338	5681	2352	1686
63	Leskovar: Novele	39616	9780	9257	8312	20,98	8319	6489	5818	1300	1747
64	Kovačić I. G.: Pripovijetke	39814	13821	9288	7906	19,86	8361	9693	5837	1764	1756
65	Kovačić: Pripovijesti	42054	11969	9635	7534	17,92	8831	8039	6055	1302	1855
66	Ministar Obrane	42271	9219	9668	9211	21,79	8877	5873	6076	2677	1864
67	Kozarac1	42575	9831	9715	8741	20,53	8941	6493	6106	1629	1878
68	Aralica8	43572	8379	9866	9846	22,60	9150	3015	6201	2036	1922
69	Priče iz davnine	44027	9225	9935	10367	23,55	9246	5360	6244	1900	1942

	Oznaka teksta	Veličina teksta: broj pojavnica	Vokabular: različnice	Izračunani vokabular VRt=(Nk)$^{2/3}$	Fs: broj funkcionalnih riječi	% funkcionalnih riječi	Izračun funkcionalnih riječi Ft=0,21n	Hapax legomena HLs: frekven.=1	HLt=((Nk)/2)$^{2/3}$	MFs: maksimalna frekvencija	MFt=N(0,21)$^{2/3}$
70	Kozarac2	50361	11277	10872	10151	20,16	10576	7541	6833	2008	2221
71	Gjalski: Borislavić Janko	50521	11876	10895	10700	21,18	10609	7693	6847	1957	2228
72	Brlić: Jaša Dalmatin	50961	14535	10958	10093	19,81	10702	9360	6887	2071	2247
73	Tomić 2	52217	11560	11138	10587	20,28	10966	6940	7001	2122	2303
74	Naši i vaši	52966	9737	11245	9889	18,67	11123	5794	7068	2498	2336
75	Škola plivanja	54021	12007	11395	12051	22,31	11344	7563	7162	1858	2382
76	Matoš: Pripovijetke	54933	17652	11523	12042	21,92	11536	12278	7242	2395	2423
77	Kovačić: Fiškal	55761	13526	11639	10494	18,82	11710	8667	7315	1846	2459
78	Šenoa: Prosjak Luka	55985	10432	11671	10343	18,47	11757	6055	7335	1882	2469
79	Tajna jednog videozapisa	58938	12193	12080	11844	20,10	12377	7186	7592	2300	2599
80	Kumičić 2	59687	10250	12182	10697	17,92	12534	5819	7657	2178	2632
81	Usmene narodne priče	60253	10517	12259	13502	22,41	12653	4499	7705	2336	2657
82	Nekoć i sad	62149	12462	12517	13487	21,70	13051	7615	7867	2653	2741
83	Žudnja za ljubavlju	63115	14398	12647	12521	19,84	13254	9433	7949	3358	2783
84	Kumičić: Začuđeni svatovi	67831	12742	13272	12990	19,15	14245	7420	8342	2377	2991
85	Novak2	69282	14660	13462	13779	19,89	14549	9198	8461	2382	3055
86	Ljudska sudbina	79428	16156	14753	14916	18,78	16680	9798	9272	3374	3503
87	Šenoa3	81500	15493	15009	15597	19,14	17115	8893	9433	2567	3594
88	O tragičnom osjećanju života	84097	15202	15328	19802	23,55	17660	9280	9634	3689	3709
89	Kumičić1	85276	12960	15472	14930	17,51	17908	7166	9724	3169	3761
90	Vremenski stroj - Rat svjetova	87971	17751	15798	19108	21,72	18474	11103	9929	3824	3880
91	1984	92242	17819	16308	18781	20,36	19371	10860	10249	4917	4068
92	Šenoa2	95526	17548	16694	16605	17,38	20060	9900	10492	2661	4213
93	Tomić1	97411	17534	16914	18129	18,61	20456	10509	10631	4958	4296
94	Put do Indije	99062	18331	17106	20340	20,53	20803	11232	10751	4445	4369
95	Gjalski1	100702	21243	17295	20206	20,07	21147	13274	10870	3816	4441
96	Blaga je noć	101775	21454	17418	19525	19,18	21373	13861	10948	6183	4488
97	Šimunović: Pripovijetke	103691	19039	17637	25483	24,58	21775	7767	11085	5842	4573
98	Drugo lice marketinga	107788	21689	18101	25420	23,58	22635	13457	11377	4598	4753
99	Kamov: Isušena kaljuža	108598	21452	18192	23673	21,80	22806	13127	11434	6823	4789
100	Jagma	114720	20438	18873	23177	20,20	24091	12013	11862	4548	5059
101	Gjalski2	126203	22229	20119	25653	20,33	26503	13312	12645	4093	5566
102	Berlin Alexander Platz	137531	19835	21312	30635	22,27	28882	11538	13394	5162	6065
103	Kovačić1	144724	26603	22052	29612	20,46	30392	16209	13860	7156	6382
104	Ivanhoe	149596	26951	22547	28824	19,27	31415	15439	14171	4709	6597
105	Kontrapunkt	154148	29265	23004	30035	19,48	32371	17950	14458	6838	6798
106	Plodovi gnjeva	164244	21502	24003	32613	19,86	34491	11845	15086	6526	7243
107	Kumičić: Urota	170792	25622	24640	32796	19,20	35866	14228	15486	5462	7532
108	Kumičić3	175021	25630	25047	32805	18,74	36754	14228	15742	5463	7718
109	Aralica1-9	217515	30432	28974	50401	23,17	45678	14284	18210	10287	9592
110	Goli i mrtvi	221609	27788	29338	43193	19,49	46538	14802	18439	9761	9773
111	Begović: Giga Baićeva	251013	36021	31892	52678	20,99	52713	21500	20044	10966	11070

Isti stalni omjer veličine teksta i broja funkcionalnih riječi u tekstu pokazuju podaci iz kontrolne grupe – za tekstove na engleskom jeziku (Grafikon 4). Samo u korpusu tekstova na engleskom udio je funkcionalnih pojavnica u veličini teksta 26%.

Grafikon 4 koristi se dvjema ordinatama s različitim skalama vrijednosti za prikaz veličine teksta, odnosno broj funkcionalnih riječi. To omogućuje da se zorno pokaže kako su obje krivulje jednake, odnosno da se omjer između ovih veličina ne mijenja s veličinom teksta.

➲ GRAFIKON 4
Veličina teksta i broj funkcionalnih riječi Fs i Ft



Zato možemo zaključiti da je postotak funkcionalnih riječi u tekstu konstantan. On za tekstove na hrvatskom jeziku iznosi 21%, a za one na engleskom 26%. Možda bi mogle postojati varijacije između žanrova, no to je već predmet novih analiza.

Broj stvarnih funkcionalnih riječi (Fs) dobili smo empirijski. Teorijski izračun postotka funkcionalnih riječi (Ft), kada je K = 21, pokazuje minimalna odstupanja.

$$Ft = n(K/100) \quad (2)$$

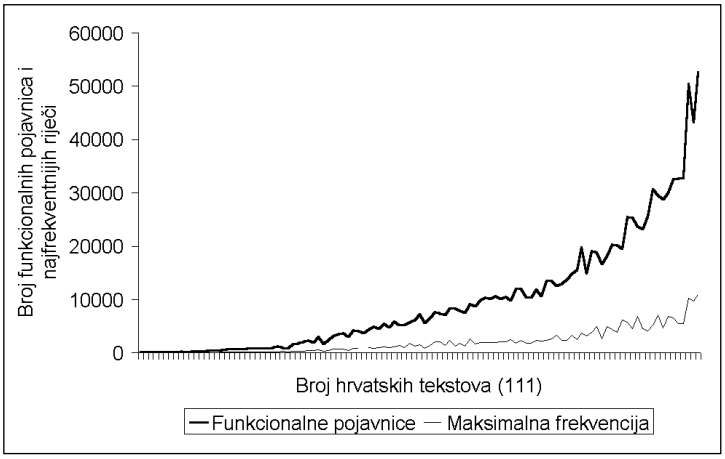
Grafički prikaz razlika između Fs i Ft potvrđuje tezu da pokazatelj postotka funkcionalnih riječi u tekstu možemo uzeti kao konstantu za pojedini jezik. Funkcionalne su riječi prazne riječi, služe za gramatičku gradnju rečenica, a ne za oblikovanje obavijesti. Udio broja funkcionalnih pojavnica u tekstu ne mijenja se s veličinom teksta nego ostaje isti. "Problem" jest kod kratkih tekstova do 1000 riječi, gdje je postotak funkcionalnih pojavnica nešto manji (17-20%) i gdje je najveći odmak između stvarnoga (Fs) i procijenjenoga broja (Ft) svih funkcionalnih riječi (Tablica 2). To je isti problem s kojim su suočeni teoretici bibliometrijskih zakona, jer je na početku logaritamske krivulje koja pokazuje bibliometrijske razdiobe najviše odstupanja (V. Olujić-Vuković, 1999.; R. Rousseau, 1990.).

Analiza obrađenoga korpusa tekstova na hrvatskom i engleskom jeziku vodi k zaključku da je udio funkcionalnih pojavnica u tim korpusima konstantan i da za hrvatske tekstove K = 21, a za engleske K = 26.

Maksimalne su frekvencije konstanta

Ako usporedimo odnose između ukupnoga broja funkcionalnih riječi (Fs) i najfrekventnije riječi u tekstu (MFs), možemo zaključiti da je riječ o razdiobama koje su pravilne (Tablica 2). To nam pokazuje i Grafikon 5.

➲ GRAFIKON 5
Odnos broja funkcionalnih riječi i maksimalna frekvencija



Štoviše, statistička analiza odnosa vodi nas k hipotezi da je odnos veličine teksta prema broju funkcionalnih pojavnica jednak omjeru broja funkcionalnih pojavnica prema maksimalnoj frekvenciji (MF). Zato taj odnos možemo prikazati i ovako:

(3) $MF = Ft(K/100)$

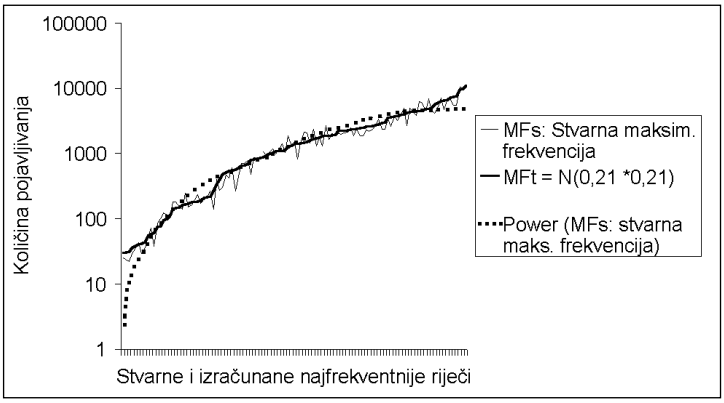
odnosno:

(4) $MF = n(K/100)^2$

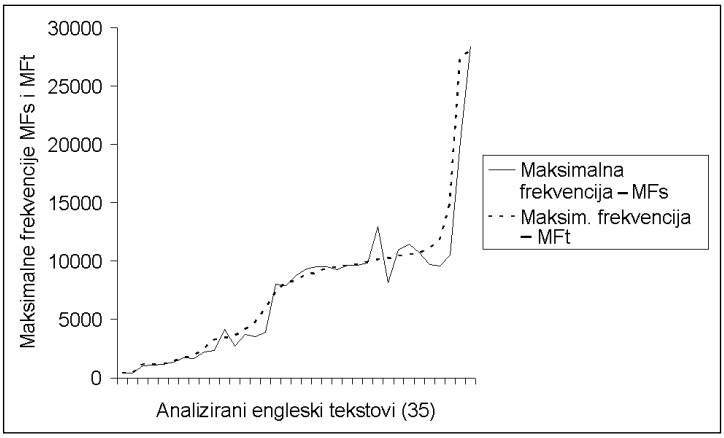
Na Grafikonu 6 prikazani su odnosi između stvarnih (MFs) i izračunanih vrijednosti (MFt), prema formuli (4) za najfrekventnije riječi u analiziranim tekstovima. Odstupanja između stvarnih i izračunanih vrijednosti mnogo su manja nego u slučajevima nekih drugih bibliometrijskih pojava i procesa; u gotovo 40% slučajeva odstupanja su čak manja od ± 10% (Tablica 2). Korelacija između stvarnih i izračunanih vrijednosti najfrekventnijih riječi jest 0,976.

Na Grafikonu 7 prikazane su maksimalne frekvencije za tekstove na engleskom jeziku, stvarne i izračunane:

I na temelju ovoga grafikona može se zaključiti da su razlike između stvarnih i izračunanih maksimalnih frekvencija u statistički dopuštenim granicama.



➲ GRAFIKON 6
Odnos stvarnih i izračunanih najfrekventnijih riječi (MFs i MFt)



➲ GRAFIKON 7
Maksimalne frekvencije riječi u tekstovima – MFs i MFt

Konstante u Heapsovu zakonu (komentar)

Do sada smo uočili dvije konstante u analizi teksta:

– Prvo, udio funkcionalnih pojavnica u tekstu, neovisno o duljini teksta, jest konstantan.

Komentar: funkcionalne riječi ne grade poruku teksta, nego strukturiraju tekst; *K* se može mijenjati prema jezicima i žanrovima, ali je konstantan za pojedini jezik i pojedini žanr.

– Drugo, maksimalna frekvencija (MF) u tekstu jest konstanta, jer je određuju gradbeni elementi teksta, a ne poruke.

MFt izračunana je po formuli:

$$MFt = F(K/100)$$

ili

$$MFt = n(K/100)^2$$

Komentar: maksimalna frekvencija (MF) uvijek je konstanta i iznosi 4,41% od ukupnoga broja pojavnica; odnosno MFt je 4,41% od veličine teksta na hrvatskom jeziku, ili 6,76% od veličine teksta na engleskom jeziku (prema $K^2/100$).

– Treće, možemo zaključiti da postoji i pravilnost u gradnji teksta, odnosno odnosima koji postoje između veličine tek-

sta (n), broja svih funkcionalnih pojavnica (F) i najfrekventnije riječi (MF). Ako je naša analiza točna, onda su ti odnosi ovakvi:

$$n : F = F : MF$$

Odnosno: broj pojavnica ili dužina teksta (n) prema broju funkcionalnih pojavnica (F) u istom je omjeru kao broj funkcionalnih pojavnica prema maksimalnoj frekvenciji (MF).

Ako je tako, onda vrijede i sljedeći izvodi:

$$n = F^2/MF$$
$$MF = F^2/n$$
$$F = (nMF)^{1/2}$$

– Četvrto, nema razloga da se za parametar *K* u Heapsovu zakonu ne uzme pokazatelj o broju svih funkcionalnih riječi u tekstu. To je empirijska veličina koja se izračunava za broj pojavljivanja svih funkcionalnih riječi u tekstovima na različitim jezicima. Vrijednost ovoga parametra u granicama je onoga što je određeno Heapsovim zakonom (1), ali je prednost u tome što se mogu izračunati i drugi pokazatelji relevantni za veličinu teksta.

REZULTATI (2): VARIJABLE U HEAPSOVU ZAKONU

Vrijednost parametra β je konstantna?

Podsjetimo da je Heapsov zakon formuliran na sljedeći način:

(1) $$V_R(n) = Kn^{\beta},$$

a da su *K* i β parametri koji se prema Heapsu određuju empirijski. Za korpus tekstova na engleskom jeziku vrijednost za β iznosi (Heaps'Law.PlanetMath.Org.):

$$0{,}4 \leq \beta \leq 0{,}6$$

Mogu li se ove vrijednosti teorijski utemeljiti ili samo izvući iz empirije? Vrijednost parametra β za izračunavanje vokabulara tekstova na hrvatskom jeziku iznosi 0,67, ako je K = 21 (vidi Tablica 2; "Dokumentacija"), a prema drugim istraživanjima može iznositi i 0,74 (vidi M. Tuđman i dr., 2003.). No nas zanima može li se odrediti po nekoj drugoj osnovi vrijednost parametra β, kako bi bio općevrijedan i kako se ne bi morao empirijski određivati u svakom istraživanju posebno.

Podsjetili smo na to da analitičari vokabulara tekstova na engleskom jeziku tvrde kako eksponent β ima vrijednost između 2/5 i 3/5. Naše istraživanje (Tablica 2) pokazuje da je vrijednost β za hrvatske tekstove 0,67 ili 2/3. Nije teško zaključiti da su ove vrijednosti zapravo logaritamske vrijednosti od K/100.

Naime:

Logaritam broja 0,21 = – 0,67778
Logaritam broja 0,22 = – 0,65757
Logaritam broja 0,23 = – 0,63827

Logaritam broja 0,24 = – 0,61978
Logaritam broja 0,25 = – 0,60205
Logaritam broja 0,26 = – 0,58502

Znamo da je logaritam broj kojim treba potencirati bazu *a* da se dobije broj *x*, tj. iz $a^y = x$, slijedi $y = log\ a^x$ (*y* je logaritam od *x* po bazi *a*).

Ali što je priroda logaritma i kakav je odnos između logaritamskih i aritmetičkih progresija? Već opći rječnici tumače logaritam (*logos* – govor, odnos, račun, računanje + *arithmos* – broj) kao matematički broj uzet u jednoj aritmetičkoj progresiji, koji odgovara broju uzetom u geometrijskoj progresiji, pri čemu su obje progresije prilagođene određenim uvjetima (B. Klaić, 1978.).

Teorijsko je pitanje možemo li istu paralelu povući i u našem slučaju i zagovarati tezu da je eksponent β u formuli (1) vrijednost logaritma od K/100? Zašto bismo mogli zastupati ovakvu hipotezu? Zato što je vrijednost *K* u formulama (2) i (3) empirijska veličina, parametar kojim utvrđujemo omjer veličine teksta i broja funkcionalnih pojavnica (2), odnosno veličine teksta i najfrekventnije pojavnice (3). Ali ako tim istim parametrom *K* želimo izračunati veličinu vokabulara teksta, onda to znači da odnos funkcionalnih riječi i veličinu vokabulara teksta stavljamo u novu progresiju, koju određuje eksponent β. Zato se i vrijednost *K* mijenja u malo *k*, tj. $k = K^{\beta}$.

To ima za posljedicu da formulu (1) $V_R(n) = Kn^{\beta}$ moramo prikazati na sljedeći način:

(5) $$V_R(n) = K^{\beta}n^{\beta},$$

ili $$V_R(n) = (kn)^{\beta}.$$

Parametri *K* i *k* imaju istu početnu vrijednost koja se mijenja kada se rabi za izračunavanje odnosa u različitim progresijama.

Empirijskih potvrda za ovakvu vrstu tumačenja možemo naći već u činjenici da pojedini elementi teksta rastu u različitim omjerima: veličina teksta raste eksponencijalno u odnosu na vokabular teksta, koji raste linearno; ali vokabular teksta raste linearno u odnosu na broj funkcionalnih pojavnica, koji raste eksponencijalno; ili broj jednokratnih riječi raste linearno u odnosu na broj najfrekventnijih riječi koji raste eksponencijalno itd. (vidi Tablica 2, "Dokumentacija"). Očito je da pojedini elementi teksta rastu različitim progresijama i da imaju različite protežnosti u tekstu i korpusu tekstova. Neki autori s pravom upozoravaju na to da "priroda ovih zakona (Zipfova i Heapsova, op. M.T.) nije poznata" (Gelbukh, Sidorov, 2001.). To nas upućuje na spoznajno-teorijski problem multidimenzionalnosti teksta i potrebu epistemološkoga tumačenja pojedinih dimenzija teksta. Samo bismo na taj način mogli teorijski obrazložiti tezu zašto vrijednost parametra β može biti $\beta = \log K/100$.

Kako se u ovom radu bavimo empirijskom provjerom zakona o veličini vokabulara teksta, tako ćemo samo upozoriti na mogućnost koju treba potvrditi novim istraživanjima, ali i epistemološkim tumačenjima, da vrijednost parametra β može biti logaritam od K/100. Što bi značilo i da vrijednost parametra β može biti konstantna za tekstove na istom jeziku.

Izračun veličine vokabulara tekstova na hrvatskom jeziku

Utvrdili smo da je broj funkcionalnih pojavnica u tekstu konstantan. Postotak funkcionalnih riječi uzeli smo kao vrijednost parametra *K*. Empirijska istraživanja pokazuju da parametar β može biti logaritamska vrijednost od K/100. Zato za formulu (1) za izračunavanje veličine vokabulara tekstova na hrvatskom jeziku vrijede parametri K = 21, i β = log K/100, pod uvjetom da izvornu formulaciju Heapsova zakona (1) redefiniramo:

(5) $$V_R(n) = (Kn)^{\beta}$$

Za ovakvu formulaciju zakona o veličini vokabulara teksta nalazimo potvrdu u empirijskim istraživanjima. Teorijsko tumačenje trebalo bi slijediti iz razumijevanja multidimenzionalnosti teksta. Operativno smo u formulu (5) uvrstili parametre do kojih smo došli empirijski:

(6) $$V_R(n) = (21n)^{2/3},$$

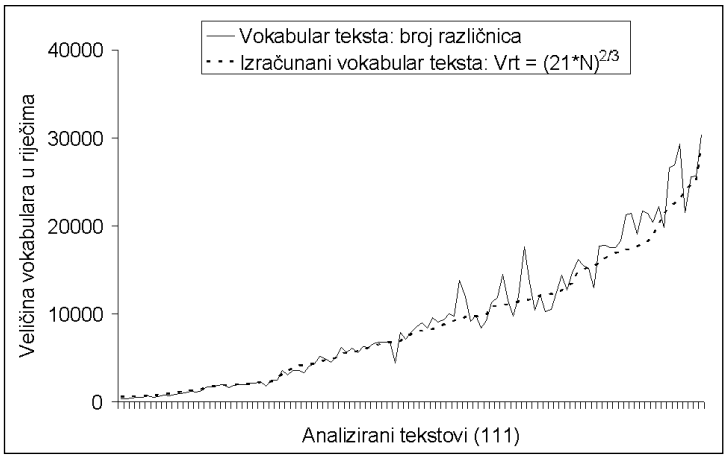
odnosno:

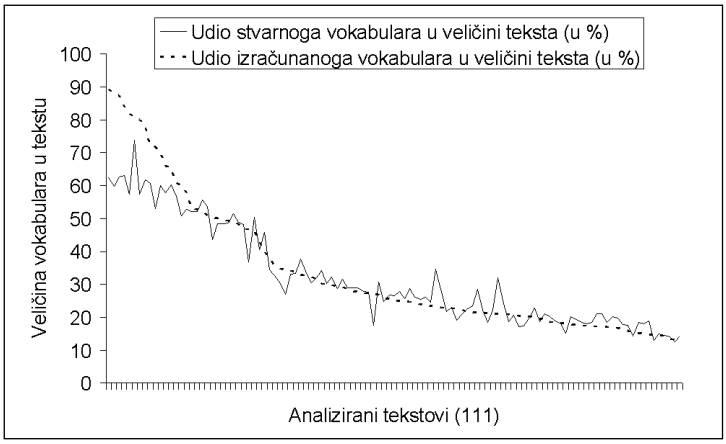
(6b) $$V_R(n) = (21n)^{0,67}.$$

Koristeći se ovom formulom, dobili smo sljedeće rezultate (Vrt) koji su prikazani na Grafikonima 8 i 9 te u Tablici 2.

Na oba prethodna grafikona, kao i u Tablici 2, prikazane su stvarne i izračunane vrijednosti vokabulara tekstova na hrvatskom jeziku. Iz grafikona i Tablice 2 možemo zaključiti da je formula (6, 6b) pouzdano sredstvo za procjenu vokabulara tekstova, jer je korelacija između dviju varijabli (stvarne i izračunane veličine vokabulara) 0,984.

➲ GRAFIKON 8
Prikaz stvarnih i izračunanih vrijednosti vokabulara





➲ GRAFIKON 9
Udio vokabulara
u veličini teksta (u %)

Pogreška prognoze veličine vokabulara najveća je na početku krivulje, tj. kod tekstova do 1000 riječi. Kod tekstova do 1000 riječi udio vokabulara u veličini teksta iznosi otprilike 60%, a procjene se kreću od 90% do 78%. Kod tekstova od 100.000 do 250.000 riječi stvarni udio vokabulara u odnosu na veličinu teksta pada od 21% prema 14% (odnosno, izračunana vrijednost vokabulara u odnosu na veličinu teksta pada od 17% prema 13%).

Zato moramo postaviti metodološko pitanje: koja je najmanja veličina teksta koji možemo smatrati tekstom, odnosno predmetom kvantitativnih analiza? Ili, tekst koje veličine ima značajke što se mogu kvantitativno analizirati? Prije ili kasnije morat ćemo naći odgovor na ovo pitanje, jer o njemu ovisi i odgovor na pitanje o prihvatljivosti ove ili one formule za izračunavanje veličine vokabulara teksta ili korpusa tekstova.

Broj jednokratnih riječi u vokabularu teksta jest konstantan

Na Grafikonu 10 prikazan je eksponencijalni rast i veličine vokabulara teksta, ali i jednokratnih riječi (*hapax legomena*) u tekstu. Očito je da ove dvije varijable veličine teksta rastu eksponencijalno i da su stalno u istom omjeru. Veličinu vokabulara teksta izračunavamo po formuli (5). Omjer veličine vokabulara i jednokratnih riječi možemo izračunati ako taj omjer uključimo u formulu (5), jer je očito da je posrijedi opis empirijskih odnosa.

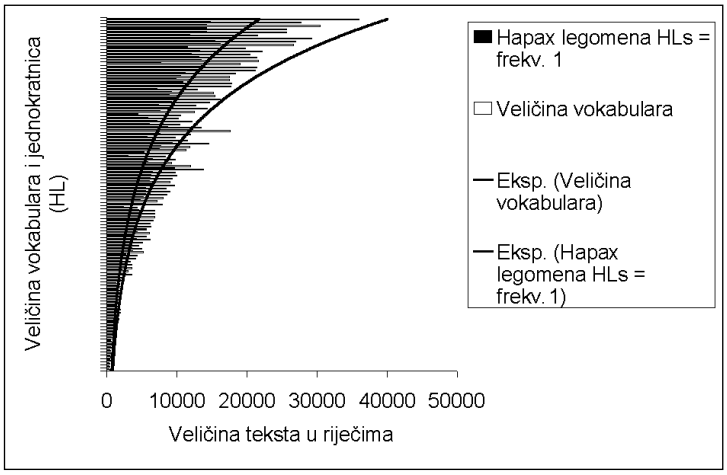
Zato se broj jednokratnih riječi (HL) u tekstu može izračunati po sljedećoj formuli:

(7) $HL = ((Kn)/2)^{\beta}$

(8) $HL = ((21n)/2)^{2/3}$,

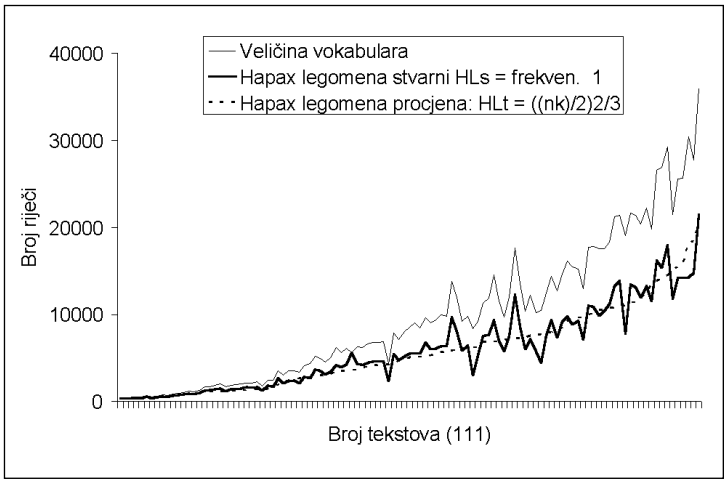
odnosno:

(9) $HL = ((21n)/2)^{0,67}$.



➲ GRAFIKON 10
Prikaz eksponencijalnoga rasta vokabulara i *hapax legomena*

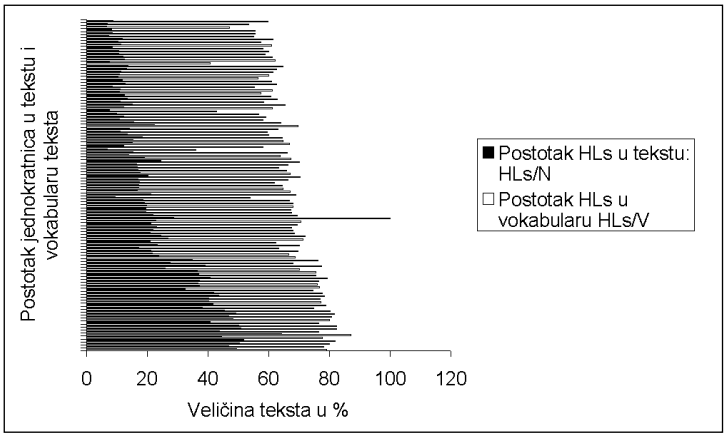
Grafički prikaz odnosa veličine vokabulara, stvarnih (HLs) i izračunanih (HLt) jednokratnih riječi (formula (9)), dan je u Grafikonu 11 i u Tablici 2.



➲ GRAFIKON 11
Prikaz veličine vokabulara i veličine *hapax legomena*

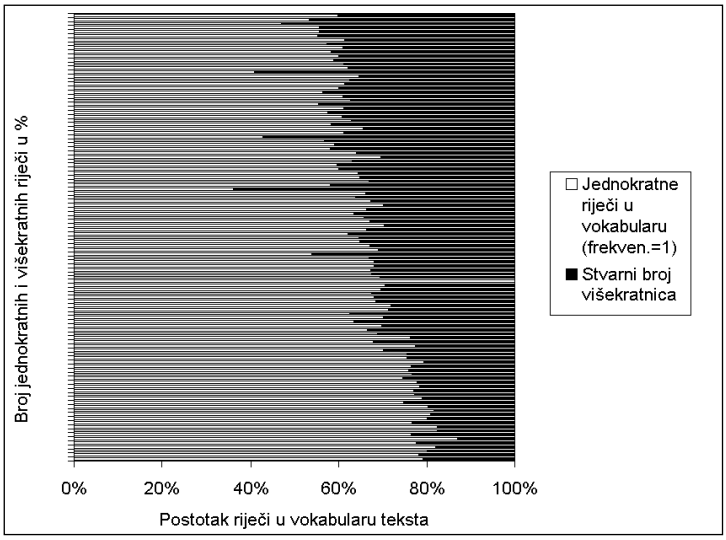
Korelacija između stvarne i izračunane vrijednosti jednokratnih riječi vrlo je visoka: 0,956. Statističke razlike između stvarnih i izračunanih vrijednosti i u ovom su slučaju prihvatljive. Osim toga, omjeri koji postoje između broja jednokratnica i teksta te broja jednokratnica i vokabulara teksta pokazuju pravilnosti koje se mogu očitati i u Grafikonu 12.

Već smo upozorili da postoje samo značajnija odstupanja u izračunu vokabulara, ali i jednokratnih riječi (HL), kod kratkih tekstova (do 1000 riječi). Ta se odstupanja možda mogu tumačiti i kao nepreciznosti, jer je parametar *K* izračunan kao prosjek na temelju velikog uzorka. Realni *K* pak varira, jer kod malih tekstova (do 1000 riječi) iznosi čak 17%. Ako se izračunava Vr (vokabular) i HL s realnim *K*, onda su odstupanja u granicama dopuštenog.



➲ GRAFIKON 12
Postotak jednokratnih riječi u tekstu i vokabularu teksta

Broj višekratnih riječi u tekstu lako se može izračunati kada se zna veličina vokabulara i broj jednokratnih riječi. Kako višekratnice nisu predmet naše analize, nego zakon o veličini vokabulara teksta, u Grafikonu 13 prikazujemo samo omjere između jednokratnih i višekratnih riječi u analiziranom korpusu tekstova.



➲ GRAFIKON 13
Broj jednokratnih i višekratnih riječi u vokabularu teksta

ODREĐIVANJE VELIČINE VOKABULARA TEKSTOVA NA HRVATSKOM JEZIKU

Izračunavanje veličine vokabulara teksta bit će to preciznije što se parametri koji određuju veličinu teksta mogu preciznije odrediti. U ovom smo pristupu pošli od empirijskoga podatka da je broj svih funkcionalnih pojavnica onaj parametar koji se dosta precizno može odrediti za svaki korpus tekstova na nekom jeziku. Analizirajući podatke o korpusu tekstova koje smo istraživali, ustanovili smo da je *K* za korpus teksto-

va na hrvatskom 21, a za korpus tekstova na engleskom jeziku K = 26. Štoviše, ustanovili smo i omjere koji postoje između broja funkcionalnih pojavnica i najfrekventnije riječi. Time smo dobili empirijsko polazište za istraživanje odnosa između dva parametra u Heapsovu zakonu. Utvrdili smo da drugi parametar (β) može biti logaritam od K/100. Slijedom toga naše je istraživanje potvrdilo sljedeće odnose u korpusima tekstova koje smo istraživali, a koji se mogu opisati formulama:

– za broj funkcionalnih pojavnica (F) u tekstu ili korpusu tekstova

$$F = n(K/100) \quad (2)$$

– za maksimalnu frekvenciju (MF) riječi u tekstu

$$MF = n(K/100)^2 \quad (4)$$

– za veličinu vokabulara (Vr) teksta

$$V_R(n) = (Kn)^{\beta} \quad (5)$$

– za broj jednokratnih riječi (HL) u (vokabularu) tekstu

$$HL = ((Kn)/2)^{\beta}. \quad (7)$$

Vrijednost je parametara K i β za korpus tekstova (n) na hrvatskom jeziku 21, odnosno 0,67.

Prema ovim formulama izračunane vrijednosti veličine vokabulara teksta, broja funkcionalnih i najfrekventnijih pojavnica te jednokratnih riječi u tekstu prikazane su u Tablici 3. Za usporedbu, izračunana je veličina vokabulara (Vr) i prema izvornoj Heapsovoj formuli (1).

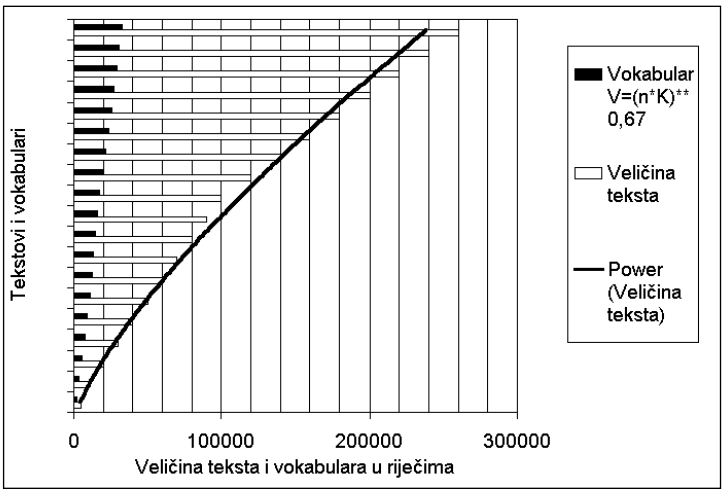
TABLICA 3
Prikaz teorijskih veličina vokabulara teksta, broja funkcionalnih i najfrekventnijih pojavnica te jednokratnih riječi u tekstu

Veličina teksta n	Vokabular $Vr=(nk)^{0,67}$	Funkcionalne $F=nk/100$	$MF=N(K/100)^2$	Jednokratnice $HL=((nk)/2)^{0,67}$	Heapsova formula $Vr=21n^{0,67}$
5000	2313	1050	221	1454	6317
10000	3680	2100	441	2313	10051
20000	5856	4200	882	3680	15992
30000	7683	6300	1323	4829	20984
40000	9317	8400	1764	5856	25445
50000	10819	10500	2205	6800	29548
60000	12225	12600	2646	7683	33387
70000	13555	14700	3087	8519	37020
80000	14824	16800	3528	9317	40485
90000	16041	18900	3969	10082	43809
100000	17214	21000	4410	10819	47013
120000	19451	25200	5292	12225	53122
140000	21567	29400	6174	13555	58901
160000	23586	33600	7056	14824	64414
180000	25522	37800	7938	16041	69703
200000	27389	42000	8820	17214	74801
220000	29195	46200	9702	18349	79734
240000	30948	50400	10584	19451	84520
260000	32653	54600	11466	20522	89177

Usporedba podataka sa stvarnim vrijednostima iz Tablice 2, s izračunanim podacima iz Tablice 3, upućuje na zaključak o prihvatljivosti predloženih formula. No isto tako možemo razabrati da je izvorna Heapsova formula neuporabljiva, ako želimo kao vrijednost parametra *K* rabiti postotak funkcionalnih riječi u tekstu.

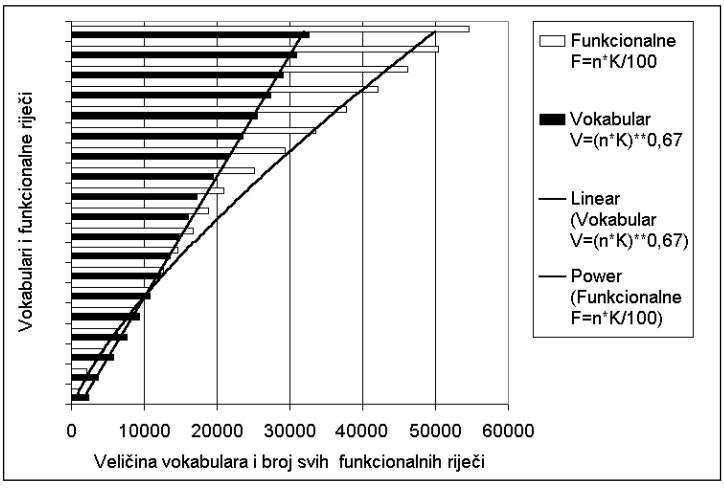
Iako se to može razabrati iz podataka u Tablici 3, na sljedećem je grafikonu (14) još očitije da tekstovi i korpusi teksova rastu eksponencijalno, a da veličine vokabulara tekstova rastu po drugoj progresiji.

➲ GRAFIKON 14
Teorijske veličine teksta i vokabulara



Grafički prikaz veličina vokabulara tekstova i svih funkcionalnih riječi prikazan je u Grafikonu 15. I ovdje ukupan broj funkcionalnih pojavnica raste eksponencijalno, a veličina vokabulara linearno. No mnogo je važnije podsjetiti na to da je postotak funkcionalnih pojavnica u tekstovima stalan, pa zato tu vrijednost možemo rabiti kao parametar u formuli za izračunavanje veličine vokabulara teksta.

➲ GRAFIKON 15
Teorijske veličine vokabulara teksta i funkcionalnih riječi u tekstu



Predložene formule za izračunavanje veličine vokabulara tekstova, broja svih funkcionalnih pojavnica u tekstovima, broja jednokratnih riječi i najfrekventnije riječi imaju empirijsku potvrdu u našim istraživanjima. Naša formula temelji se pak na Heapsovu zakonu i izvedena je iz njega. Kako je ipak Heapsova formulacija zakona o veličini vokabulara teksta ponešto promijenjena, potrebno je da iduća istraživanja potvrde uporabljivost, predloženih formuli te vrijednost parametara za tekstove na hrvatskom jeziku.

MOGUĆE PRIMJENE ZAKONA O VELIČINI VOKABULARA TEKSTA

Heapsov se zakon nije ustalio kao istraživačka metoda, a zašto nije naveli smo u uvodnom dijelu. Zato je još uvijek sam zakon predmet istraživanja (G. R. Turner, 2001.; A. C. Fang, 1997.) i pokušaja da se definiraju parametri presudni za njegovo razumijevanje (C. M. Urzua, 2000.; A. Gelbukh, G. Sidorov, 2001.).

Heapsov zakon smatra se izvedenicom iz Zipfova zakona i zato je povezan s istraživanjem jezika. Međutim, sve više nalazi primjenu u istraživanju kvantitativnih odrednica tekstova, korpusa tekstova te umreženih tekstova na internetu.

Na to upućuju pokušaji primjene toga zakona na raznim područjima. Naznačimo samo neka istraživanja koja se provode u nabrojenim područjima, iako se čini da istraživači međusobno nisu upućeni jedni na druge. Ova istraživanja navodimo kao primjere moguće primjene Heapsova zakona, a ne kao temeljna istraživanja za ta područja.

U lingvistici Heapsov zakon ima svoju primjenu u istraživanju veličine vokabulara teksta (G. R. Turner, 2001.; P. Batke), leksičke gustoće i segmenata diskursa (J. Ure, 1971.; L. Y. L. Cheung i dr., 2001.; P. Batke) te topologije teksta (M. M. A. Juillard, N. X. Luong, 1996.). Svoju primjenu čini se da će naći i u istraživanju podataka na web stranicama (J. C. French, 2002.) te istraživanju veličine servera i broja HTML dokumenata (S. Sanguanpong, S. Warangrit, 2000.).

Nije teško pretpostaviti i druga područja primjene ovoga zakona: u kriptologiji i zaštiti podataka. Ali i u rekonstrukciji dokumenata i korpusa dokumenata na raznim područjima: od arheologije i povijesti do kvantitativnih metoda za analizu stila i vokabulara djela u književnosti i lingvistici. Analiza vokabulara pojedinih struka ili socijalnih grupa može biti zanimljiva sociologiji, leksikologiji i kulturnoj politici. No isto tako raščlamba i prepoznavanje osnovnoga vokabulara jezika socijalnih grupa može biti od koristi za sve one discipline koje komuniciraju tekstom sa svim tim grupama – bilo u medijima ili u formalnim i neformalnim komunikacijama. Poznavanje razvoja i rasta vokabulara teksta te segmenata od kojih se tekst ili korpus tekstova sastoji može biti od praktične koristi i za razvoj novih metoda i tehnika učenja (stranih) jezika.

Ovaj zakon zacijelo će imati mnogo primjena, ali prethodno valja razumjeti njegovu prirodu i precizno odrediti parametre bez kojih se ne može izračunati veličina vokabulara teksta ili korpusa tekstova.

LITERATURA

Batke, P. The Philology and Philosophy Project. Methodology. http://www.princeton. edu/čbatke/phph/meth/meth_wm1.htm

Carter, R. (1998.), *Vocabulary. Applied Linguistic Perspectives*. London: Routledge.

Cheung, L. Y. L., Lai, T. B. Y., Tsou, B. K., Chik, F. C. Y., Luk, R. W. P., Kwong, O. Y. (2001.), *A preliminary Study of Lexical Density for the Development of XML-based Discourse Structure Tagger*. 1st NLP and XML Workshop Tokyo, Nov. 2001.

Fang, A. C. (1997.), *STRATA 4.0*. Department of Phinetics and Linguistics, University College London. www.phon.ucl.ac.uk/home/alex/project/strata/strata.htm. (Last updated: 9 September 1997.)

French, J. C. (2002.), *Modeling Web Data*. Department of Computer Science University of Virginia Charlotttesville, VA. Proceedings of the second ACM/IEEE-CS joint conference on Digital libraries, July 14-18, 2002, Portland, Oregon, USA.

Gelbukh, A., Sidorov, G. (2001.), *Zipf and Heaps' Laws Coefficients Depends on Language*. http://www.cic.ipn.mx/~gelbukh/CV/Publications/2001/CICLing-2001-Zipf.htm [03/19/2003]

Heaps H. S. (1978.), *Information Retrieval: Computational and Theoretical Aspects*. New York: Academic Press.

Heaps' Law. PlanetMath.Org. http://www.PlanetMathOrg\PlanetMathHeaps'law.htm

Juillard, M. M. A., Luong, N. X. (1996.), New maps of text: a new way to account for the distribution of lexems in texts. In *ALLC-ACH'96*: 157-158.

Klaić, B. (1978.), *Rječnik stranih riječi*. Zagreb: Matica Hrvatska.

Oluić-Vuković, V. (1999.), *Vremenska komponenta u informetrijskim razdiobama*. Zagreb: Sveučilište u Zagrebu, Filozofski fakultet, doktorska disertacija.

Rousseau, R. (1990.), Relations between continuous versions of bibliometric laws. *Journal of the American Society for Information Science*, 41 (3): 197-203.

Sanguanpong, S., Warangrit, S. (2000.), *Facts about the Thai Web*. Department of Computer Engineering, Kasetsart University Bangkok, Thailand. Proceedings of National Computer Science and Engineering Conference (NCSEC-2000, Bangkok), pp. 102-103.

Tuđman, M. (ur.) (2003.), *Modeli znanja i obrada prirodnog jezika*. Zagreb: Zavod za informacijske studije.

Tuđman, M., Nives, M., Boras, D. (2003.), *Vocabulary size prediction of Croatian texts*. Proceedings of the 25th Int. Conf. Information Technology Interfaces ITI 2003, June 16-19, Cavtat, Croatia.

Tuđman, M., Boras, D., Mikelić, N. (2004.), *Heapsov zakon i određivanje veličine vokabulara tekstova na hrvatskom jeziku*. Dokumentacija. Filozofski fakultet, Odsjek za informacijske znanosti, Zagreb.

Turner, G. R. (2001.), *Relationship Between Vocabulary, Text Length and Zipf's Law*; http://www.btinternet.com/~g.r.turner/ZipfDoc.htm [02/20/2003]

Ure, J. (1971.), Lexical density and variety differentiation. U: G. Perren, J. Trim (ur.), *Applications of Linguistics*. Papers From the 2nd AILA Congress. Cambridge: Cambridge University Press; str. 443-52.

Urzua, C. M. (2000.), A simple and efficient test for Zipf's law. *Economics Letters*, 66: 257-260.

The Text Vocabulary Size Law. Heaps' Law and Determining Text Vocabulary Size in Croatian Language

Miroslav TUĐMAN
Faculty of Philosophy, Zagreb

The existing formula / $Vr(n)=Kn^{\beta}$ / of Heaps' Law regarding the size of a text's vocabulary is not universal, thus the law needs to be redefined, in order to be used for analysis of a different language corpus. The analysis of a corpus of texts in the Croatian language confirms the hypothesis that the number of functional items (F) in a text is constant and amounts to 21% of the size of the text n (there are 26% of functional items in English texts). The author proves that the percentage of functional items in a text can be used as the value for the parameter K, and that the parameter K presents a constant value for every language corpus. Empirical research has confirmed the author's thesis that the number of functional items in a text can be calculated according to the formula F=nK/100, and that for the value of the most frequent item (MF) the formula $MF=n(K/100)^2$ can be applied. The value of the other parameter of Heaps' Law can also be accurately determined: $\beta=\log K/100$. The author therefore suggests a new form of the text vocabulary size law: $Vr(n)=(Kn)^{\beta}$. The number of words appearing only once (HL) in the text can be calculated according to the formula: $HL= ((Kn)/2)^{\beta}$. Research confirms that there is a very high correlation between the calculated and real values of the vocabulary size, i.e. between the real and calculated values of single words in the text. Interpreted and defined in such a way, the law of the text vocabulary size enables the calculation of the text's vocabulary size in every language, if the percentage of constant functional words for this language is known. However, this interpretation of the law enables, apart from determining the size of the text's vocabulary, also the calculation of the number of functional items in the text, the size of the most frequent word in the text, and the number of single items comprising the text's vocabulary.

DRUŠ. ISTRAŽ. ZAGREB
GOD. 14 (2005),
BR. 1-2 (75-76),
STR. 227-250

Gesetz zur Bestimmung des Wortschatzumfangs von Texten. Das Heaps'sche Gesetz und die Bestimmung der Wortschatzgröße in kroatischen Texten

Miroslav TU MAN
Philosophische Fakultät, Zagreb

Die bestehende Formel / $Vr(n)=Kn^{\beta}$ / des Heaps'schen Gesetzes zur Bestimmung des Wortschatzumfangs von Texten hat keine universale Gültigkeit, sodass das Gesetz, soll es zur Textkorpusanalyse in verschiedenen Sprachen angewandt werden, redefiniert werden muss. Die Analyse von Textkorpora in kroatischer Sprache bestätigt die Hypothese, dass die Zahl funktionaler Wörter (F) in einem Text konstant ist und 21% der Größe eines Textes n ausmacht (in englischen Texten beträgt die Zahl funktionaler Wörter 26%). Der Verfasser weist nach, dass der in einem Text vertretene Prozentsatz funktionaler Wörter als Wertangabe für den Parameter K benutzt werden kann und dass der Parameter K einen gleichbleibenden Wert für jedes sprachliche Korpus darstellt. Empirische Forschungen bestätigen die These des Verfassers, dass die Zahl funktionaler Wörter in einem Text mit der Formel $F = nK/100$ errechnet werden kann, dass wiederum für die Größe der häufigsten Wörter (MF) die Formel $MF = n(K/100)^2$ gilt. Der zweite Parameter des Heaps'schen Gesetzes kann ebenfalls genau bestimmt werden: ß = log K/100. Der Verfasser schlägt daher vor, das Heaps'sche Gesetz in neuer Form zu bestimmen: $Vr(n) = (Kn)^{\beta}$. Die Zahl der nur einmal im Text vorkommenden Wörter (HL) kann anhand der folgenden Formel errechnet werden: $HL = ((Kn)/2)^{\beta}$. Forschungen haben bestätigt, dass die errechneten und die wirklichen Werte des Vokabularumfanges, bzw. dass die wirklichen und die errechneten Werte von einmalig vorkommenden Wörtern in einem Text in hohem Maße miteinander korrelieren. Ein solchermaßen interpretiertes und definiertes Gesetz zur Bestimmung des Wortschatzumfangs ermöglicht uns, den Wortschatzumfang eines Textes in jeglicher Sprache auszurechnen, hat man erst einmal den Prozentsatz funktionaler Wörter, der für die betreffende Sprache gleichbleibend ist, erstellt. Des Weiteren ermöglicht diese Interpretation des Heaps'schen Gesetzes, die Zahl der funktionalen Wörter, den Umfang der am häufigsten vertretenen Wörter sowie die Zahl der einmalig vorkommenden Wörter in einem Text zu bestimmen.