Glottometrics 16, 2008, 31-37

Zipf´s mean and language typology

Ioan-Iovitz Popescu, Bucharest
Gabriel Altmann, Lüdenscheid

Abstract. Zipf´s law is not only an expression of the rank-frequency relationship of words but it also enables us to make statements about some morphological features of language, too. In the present study, several indicators are proposed and their mutual relations are studied. The data are taken from 20 languages.

Keywords: Zipf´s law, analytism, synthetism, hapax legomena

In a previous article (Popescu, Altmann 2008) we have shown that in the rank-frequency distributions of word *forms*, hapax legomena (words occurring once) occupy a specific number of ranks, a matter of fact generally known. A function of this number is a characteristic of synthetism/analytism of a language. Zipf´s curve crosses the sequence of hapax legomena (or its prolongation) at a special place depending on the morphological complexity of language. In a strongly synthetic language like Hungarian the empirical hapax legomena are situated for the most part above the Zipfian function, and in a strongly analytic language like Hawaiian, they are situated mostly below it. Thus the fitting of Zipf´s function in the form of a non-linear regression to rank-frequency data reveals not only the validity of this law, but its (say, least square) deviation in the domain of hapax legomena characterizes a language morphologically.

A logical consequence of this finding is the fact that if Zipf´s curve (sequence) runs mostly below the hapax legomena, then its mean must be smaller that the empirical mean

(1) $$M_E = \frac{1}{N}\sum_{r=1}^{V} rf_r\,,$$

where N = text length (number of tokens), V = vocabulary (= number of word form types), r = rank, f_r = frequency at rank r. Similarly, if Zipf´s curve runs mostly above the hapax legomena, its mean must be greater than that of the empirical values in (1). In order to quantitatively express this distance we set up a new indicator B in the form

(2) $$B = \frac{M_E - M_F}{M_E}\,,$$

where M_F denotes the mean of the fitting curve

(3) $f(r) = c/r^a$.

The indicator B has the following properties:

if $B > 0$, then the language tends to contain synthetic phenomena
if $B < 0$, then the language tends to get analytic

if $B = 0$, the language is balanced, containing both types of phenomena.

The greater |B|, the more the language tends to a morphological extreme. As an example consider the frequency count of word forms in the Hawaiian text Hw 05: Moolelo Mokuna III (taken from Popescu et al. 2008, see also Table 1 below). The empirical mean yields M_E = 68.7388. Now, using iterative fitting of (3) we obtain the curve $f(r) = 592.6243r^{0.7267}$. Its mean yields M_F = 170.3493. Inserting these two values in (2) we obtain

B(*Hawaii 05*) = (68.7388 – 170.3493)/68.7388 = -1.4782.

Since the value of B is a direct consequence of the index A denoting the course of Zipf´s curve in its positional relation to hapax legomena and expressed formally as

(4) $$A = \frac{c}{(V - HL/2)^a},$$

where c is the scaling constant of Zipf´s curve, V is the vocabulary of text, $HL/2$ is the half of the range of hapax legomena, and a is the Zipfian exponent; <A, B> must yield a very rigorous relation, especially if one takes the means of all texts written in one language.

Another indicator playing the same role as A is the Zipf curve end frequency, that is, the value of the theoretical Zipf curve in point V, i.e. at the highest rank = V, yielding

(5) $$C = \frac{c}{V^a}$$

which is low in strongly synthetic languages and high in strongly analytic languages.

In Table 1 the results from 100 texts in 20 languages are presented. It can be shown that text length N does not play any role. Since we do not fit a distribution but a curve, the size plays a role only in computing the mean (since $N = \sum f(r)$).

Table 1
Indicators A, B and C from 100 texts in 20 languages
(B = Bulgarian, Cz = Czech, E = English, G = German, H = Hungarian, Hw = Hawaiian, I = Italian, In = Indonesian, Kn = Kannada, Lk = Lakota, Lt = Latin, M = Maori, Mq = Marquesan, Mr = Marathi, R = Romanian, Rt = Rarotongan, Ru = Russian, Sl = Slovenian, Sm = Samoan, T = Tagalog)

ID	**V**	**HL**	**Zipf a**	**Zipf c**	**M_E**	**M_F**	**B**	**A**	**C**
B 01	400	298	0.6850	41.8602	116.4139	109.6275	0.0583	0.9507	0.6909
B 02	201	153	0.5704	17.6950	63.1108	65.6908	-0.0409	1.1292	0.8593
B 03	285	212	0.5550	20.9975	87.5379	93.6461	-0.0698	1.1798	0.9114
B 04	286	222	0.6169	23.6917	91.5569	87.2274	0.0473	0.9790	0.723
B 05	238	187	0.6202	22.0499	75.3153	72.848	0.0328	1.0090	0.7405
Cz 01	638	517	0.7473	54.2844	205.6006	154.7747	0.2472	0.6416	0.4352
Cz 02	543	412	0.7169	51.9648	162.8963	139.7022	0.1424	0.8013	0.5692
Cz 03	1274	964	0.8028	175.4805	311.3947	268.0846	0.1391	0.8261	0.564
Cz 04	323	241	0.6228	23.3822	108.8831	97.3214	0.1062	0.8562	0.6401
Cz 05	556	445	0.8722	77.1944	164.7137	107.4763	0.3475	0.4864	0.3114
E 01	939	662	0.7657	145.9980	216.7004	216.0852	0.0028	1.0783	0.773
E 02	1017	735	0.7434	180.1325	202.6156	242.9598	-0.1991	1.4610	1.0468

E 03	1001	620	0.8179	254.7482	192.996	207.047	-0.0728	1.2123	0.8953
E 04	1232	693	0.8712	385.9532	223.1696	223.4339	-0.0012	1.0449	0.7836
E 05	1495	971	0.8009	319.1386	286.4662	313.164	-0.0932	1.2529	0.9148
E 07	1597	1075	0.7568	300.1258	303.6303	364.9494	-0.2020	1.5416	1.1301
E 13	1659	736	0.8034	811.1689	219.5143	343.8041	-0.5662	2.5688	2.1
G 05	332	250	0.6935	32.8211	105.5599	90.6857	0.1409	0.8129	0.5858
G 09	379	302	0.6523	32.5565	117.9433	109.0793	0.0752	0.9431	0.677
G 10	301	237	0.6053	21.8114	100.7583	92.9696	0.0773	0.9331	0.6893
G 11	297	232	0.5895	19.9677	100.9872	93.5783	0.0734	0.9320	0.696
G 12	169	141	0.6062	14.3627	59.9203	53.4282	0.1083	0.8888	0.6408
G 14	129	107	0.5755	10.8110	47.5543	42.7453	0.1011	0.8977	0.6595
G 17	124	84	0.5515	13.1021	39.8311	42.2041	-0.0596	1.1531	0.9179
H 01	1079	844	1.2268	214.2708	304.7397	69.6929	0.7713	0.0749	0.0407
H 02	789	638	1.1865	122.0057	253.4014	63.2871	0.7502	0.0824	0.0446
H 03	291	259	1.2114	44.9653	107.2308	28.3793	0.7353	0.0950	0.0466
H 04	609	509	0.9549	74.8581	205.1592	97.1793	0.5263	0.2753	0.1642
H 05	290	250	0.8168	30.9795	104.7337	65.8429	0.3713	0.4784	0.3018
Hw 03	521	255	0.7932	329.6012	69.9367	117.9251	-0.6862	2.8821	2.3069
Hw 04	744	347	0.7633	678.1305	75.0495	174.0335	-1.3189	5.3384	4.359
Hw 05	680	302	0.7267	592.6243	68.7388	170.3493	-1.4782	6.2199	5.1825
Hw 06	1039	500	0.7816	1081.7823	91.914	230.7216	-1.5102	5.8855	4.7463
I 01	3667	2514	0.7266	509.5979	677.9826	865.2727	-0.2762	1.7784	1.3109
I 02	2203	1604	0.7488	305.6487	457.5523	505.2243	-0.1042	1.3468	0.9596
I 03	483	382	0.7895	56.8099	146.0597	110.6116	0.2427	0.6427	0.432
I 04	1237	848	0.7014	153.3448	275.2637	315.9784	-0.1479	1.3948	1.0391
I 05	512	355	0.6524	54.5840	134.0469	145.64	-0.0865	1.2306	0.9322
In 01	221	166	0.5809	18.2346	71.4973	71.1092	0.0054	1.0420	0.7926
In 02	209	147	0.5915	19.1717	66.3995	66.5723	-0.0026	1.0509	0.8132
In 03	194	130	0.5417	15.6229	62.7781	65.5138	-0.0436	1.1233	0.9005
In 04	213	145	0.4877	11.9156	74.8338	75.8346	-0.0134	1.0683	0.8721
In 05	188	121	0.5374	19.4218	53.3671	63.8473	-0.1964	1.4347	1.1645
Kn 003	1833	1373	0.6072	66.4545	576.1998	539.1967	0.0642	0.9223	0.6936
Kn 004	720	564	0.5237	22.1001	261.3076	240.2214	0.0807	0.9144	0.7048
Kn 005	2477	1784	0.6621	124.5588	705.5287	664.299	0.0584	0.9480	0.7054
Kn 006	2433	1655	0.5809	95.9573	657.818	740.4287	-0.1256	1.3181	1.0353
Kn 011	2516	1873	0.5786	77.0267	764.0881	767.8495	-0.0049	1.0862	0.8297
Lk 01	174	127	0.6416	23.4838	50.0667	52.6722	-0.0520	1.1474	0.8575
Lk 02	479	302	0.7731	139.2126	89.0171	112.9533	-0.2689	1.5798	1.1788
Lk 03	272	174	0.7512	71.8668	57.7355	68.9918	-0.1950	1.4240	1.066
Lk 04	116	80	0.6792	18.7509	35.3927	34.4326	0.0271	0.9901	0.7429
Lt 01	2211	1792	0.7935	109.3668	771.113	461.7444	0.4012	0.3666	0.2427
Lt 02	2334	1878	0.8047	160.3530	716.6397	474.286	0.3382	0.4729	0.3126
Lt 03	2703	2049	0.6366	109.5291	803.9286	754.7652	0.0612	0.9695	0.7158
Lt 04	1910	1359	0.6505	129.2023	484.4184	525.0506	-0.0839	1.2627	0.9486
Lt 05	909	737	0.5877	34.1056	319.8213	278.5167	0.1291	0.8449	0.6225
Lt 06	609	521	0.5293	19.3370	230.4608	202.4373	0.1216	0.8726	0.6494

M 01	398	202	0.7680	185.4091	63.9248	95.7958	-0.4986	2.3386	1.8677
M 02	277	146	0.8197	123.4636	50.5234	62.835	-0.2437	1.5787	1.2285
M 03	277	133	0.7902	147.8281	46.2162	65.9788	-0.4276	2.1571	1.7364
M 04	326	192	0.8353	137.7184	58.6804	70.9494	-0.2091	1.4664	1.0958
M 05	514	239	0.7484	297.2460	69.4287	125.8978	-0.8133	3.3897	2.7807
Mq 01	289	91	0.8030	240.0615	44.6326	67.1753	-0.5051	2.9102	2.5361
Mq 02	150	86	0.7440	46.4870	33.6324	40.1976	-0.1952	1.4370	1.1177
Mq 03	301	138	0.9795	225.2046	50.6561	50.0045	0.0129	1.0853	0.841
Mr 001	1555	1128	0.6293	78.3965	450.1638	443.8837	0.0140	1.0210	0.769
Mr 018	1788	1249	0.6685	128.5531	454.4077	477.8562	-0.0516	1.1470	0.8606
Mr 026	2038	1486	0.6224	101.6971	559.1975	584.868	-0.0459	1.1758	0.8867
Mr 027	1400	846	0.6166	120.0829	312.5678	408.1721	-0.3059	1.7214	1.3789
Mr 288	2079	1534	0.6304	100.2890	588.117	589.042	-0.0016	1.0857	0.8122
R 01	843	606	0.6720	73.6423	228.9908	228.8815	0.0005	1.0739	0.7961
R 02	1179	908	0.7567	115.8007	328.5853	272.9949	0.1692	0.7930	0.5489
R 03	719	567	0.7175	60.8094	218.4913	182.4494	0.1650	0.7771	0.5423
R 04	729	573	0.6673	52.4236	222.1083	200.3455	0.0980	0.8993	0.6445
R 05	567	424	0.6746	48.1009	169.812	155.514	0.0842	0.9157	0.6677
R 06	432	353	0.6349	30.3691	141.4417	126.7049	0.1042	0.8995	0.6444
Rt 01	223	127	0.8575	123.9533	38.7252	48.4559	-0.2513	1.6008	1.2009
Rt 02	214	128	0.7469	83.2271	39.0686	55.5682	-0.4223	1.9726	1.5128
Rt 03	207	98	0.7208	78.6409	40.7635	55.916	-0.3717	2.0454	1.6835
Rt 04	181	102	0.7359	60.2092	37.5232	48.3329	-0.2881	1.6749	1.3128
Rt 05	197	73	0.6917	87.0541	37.9226	55.5516	-0.4649	2.5959	2.2528
Ru 01	422	316	0.6538	36.1404	129.4329	120.6856	0.0676	0.9437	0.6945
Ru 02	1240	946	0.7713	138.5450	323.625	278.493	0.1395	0.8251	0.5696
Ru 03	1792	1365	0.7106	158.2659	454.9782	445.0264	0.0219	1.0851	0.7719
Ru 04	2536	1850	0.7181	234.3457	598.9348	614.624	-0.0262	1.1661	0.8419
Ru 05	6073	4395	0.7826	775.3826	1215.696	1249.8376	-0.0281	1.2063	0.8488
Sl 01	457	364	0.7467	44.1840	146.7963	113.0045	0.2302	0.6665	0.4561
Sl 02	603	423	0.6846	68.9001	153.3246	162.5056	-0.0599	1.1571	0.8609
Sl 03	907	651	0.7685	115.2402	235.1974	207.9875	0.1157	0.8651	0.6147
Sl 04	1102	701	0.9187	334.8100	213.7368	179.9701	0.1580	0.7633	0.537
Sl 05	2223	1593	0.7232	240.2785	502.7643	535.6631	-0.0654	1.2572	0.9122
Sm 01	267	119	0.8285	177.1858	41.4405	59.8669	-0.4446	2.1315	1.7297
Sm 02	222	96	0.7752	123.5355	38.3578	55.1037	-0.4366	2.2641	1.8745
Sm 03	140	75	0.6858	58.1896	26.4554	40.6778	-0.5376	2.4320	1.9639
Sm 04	153	76	0.7925	89.0771	27.3927	38.2418	-0.3961	2.0738	1.6539
Sm 05	124	66	0.7161	46.3093	25.915	34.9991	-0.3505	1.8312	1.4673
T 01	611	465	0.7624	120.0367	133.617	144.6973	-0.0829	1.2995	0.902
T 02	720	540	0.7803	144.5780	157.1779	163.5462	-0.0405	1.2297	0.8522
T 03	645	447	0.7652	167.7334	119.4537	151.5339	-0.2686	1.6447	1.1877

For the sake of an easier survey we present in Table 2 the means of the above indicators for individual languages. It can easily be seen that the individual languages occupy mostly the same rank with all three indicators, i.e. the indicators are only different expressions of the same property. In order to display the relationships graphically, we use all texts and present the relation <*A*, *B*> in Figure 1 and <*A, C*> in Figure 2. Since the indicators *A* and *C* are both some functions of *V*, they are linked linearly: $C = 0.8408A - 0.0985$. However, *B* and *A* express synthetism/analytism from different points of view, hence their relationship is not quite linear. Nevertheless, we suppose a power curve which must, however, attain also negative values, hence we combine two functions and obtain

$$B = k(A^{-r} - A^{-s}),$$

in our case

$$B = 0.5331(A^{-0.1963} - A^{0.6861})$$

yielding $R^2 = 0.9859$. This curve can be used for typological purposes, too.

Table 2
Means of indicators *A*, *B* and *C* in 20 languages

Language	**mean A**		**Language**	**mean B**		**Language**	**mean C**
Hungarian	0.2012		Hungarian	0.6309		Hungarian	0.1196
Czech	0.7223		Czech	0.1965		Czech	0.5040
Latin	0.7982		Latin	0.1612		Latin	0.5819
Romanian	0.8931		Romanian	0.1035		Romanian	0.6407
German	0.9372		Slovenian	0.0757		Slovenian	0.6762
Slovenian	0.9418		German	0.0738		German	0.6952
Kannada	1.0378		Russian	0.0349		Russian	0.7453
Russian	1.0453		Kannada	0.0146		Bulgarian	0.7850
Bulgarian	1.0495		Bulgarian	0.0055		Kannada	0.7938
Indonesian	1.1438		Indonesian	-0.0501		Indonesian	0.9086
Marathi	1.2302		Italian	-0.0744		Italian	0.9348
Italian	1.2787		Marathi	-0.0782		Marathi	0.9415
Lakota	1.2853		Lakota	-0.1222		Lakota	0.9613
Tagalog	1.3913		Tagalog	-0.1307		Tagalog	0.9806
English	1.4514		English	-0.1617		English	1.0919
Marquesan	1.8108		Marquesan	-0.2291		Marquesan	1.4983
Rarotongan	1.9779		Rarotongan	-0.3597		Rarotongan	1.5926
Samoan	2.1465		Samoan	-0.4331		Samoan	1.7379
Maori	2.1861		Maori	-0.4385		Maori	1.7418
Hawaiian	5.0815		Hawaiian	-1.2484		Hawaiian	4.1487

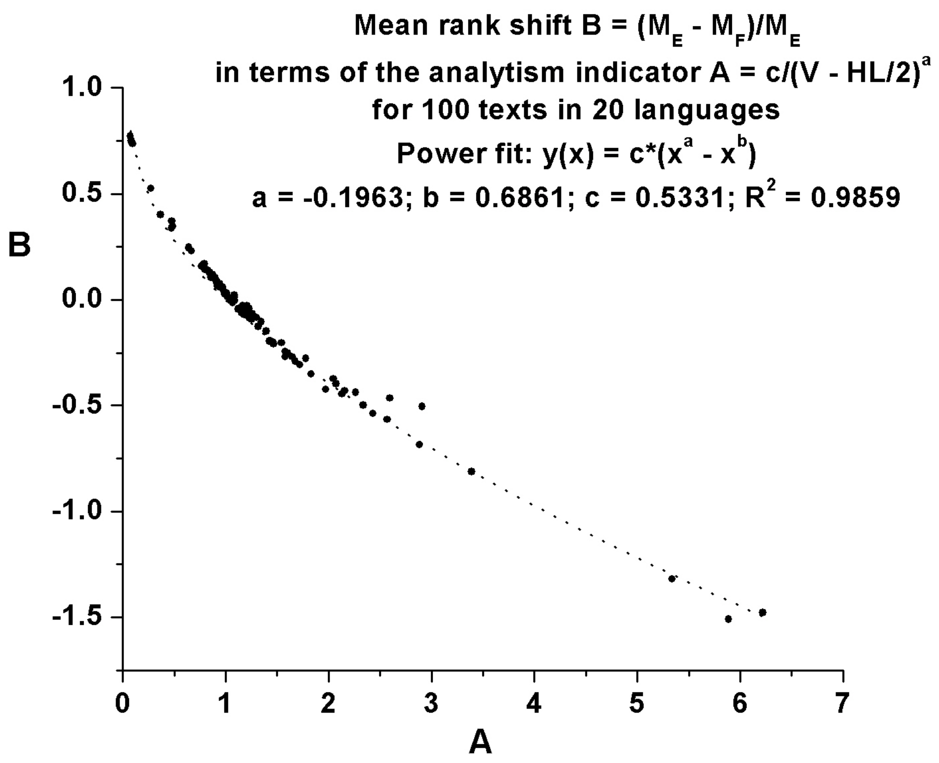

Figure 1. The relationship between indicators A and B

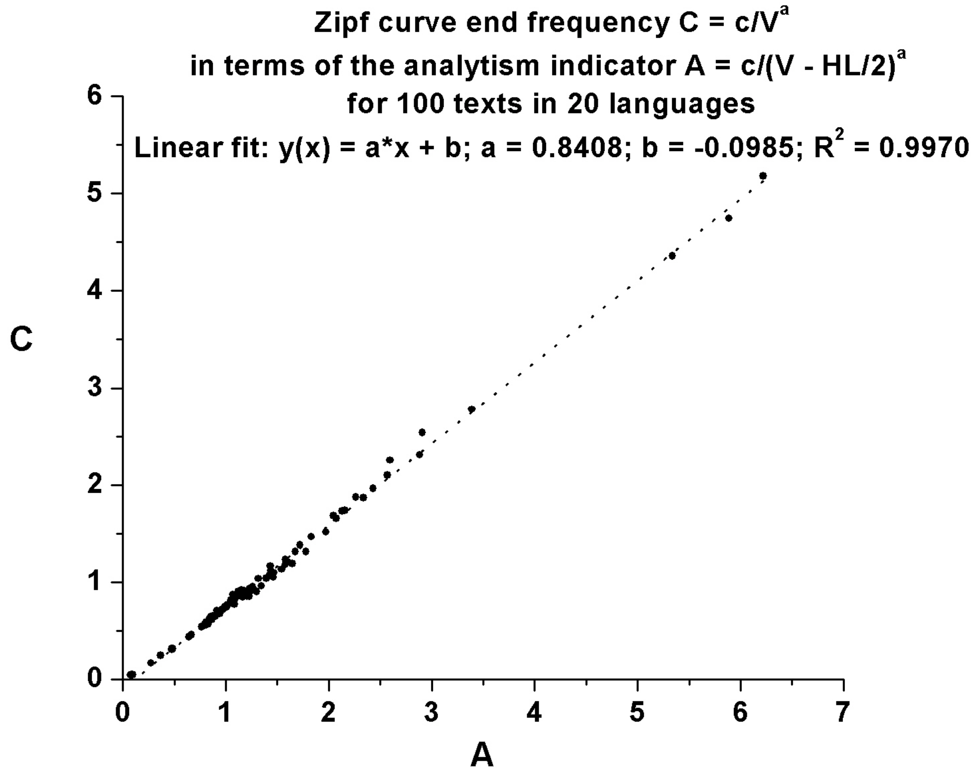

Figure 2. The relationship between indicators A and C

The fact that Zipf´s curve signalizes typological features means that in some cases it may display deviant behaviour when applied to rank-frequency data. Though in the overwhelming majority of fittings of Zipf´s (zeta) distribution to data one obtains very satisfactory results (cf. Popescu et al. 2008), the “best fit” or a fit crossing the hapax legomena exactly in their mean would, perhaps, bring some hint at the modification of Zipf´s curve in this domain. There are the following possibilities: (a) One varies the parameter “a” in order to obtain $M_E = M_F$ or $c/(V\text{-}HL/2)^a = 1$; (b) For B < 0 one uses a modification (e.g. Zipf-Mandelbrot, Lerch, Zipf-Alekseev) and for B > 0 another one. (c) One uses the same modification for both cases but with different parameters. (d) One uses a quite different way of reasoning. Using these possibilities one probably obtains a better fit, but the typological properties of the text (language) must be then inferred from different indicators. In any case we see that Zipf´s law yields deeper insights in language beyond the modelling of rank-frequency distributions.

References

Popescu, I.-I., Altmann, G. (2008) Hapax legomena and language typology (to appear)

Popescu, I.-I., Vidya, M.N., Uhlířová, L., Pustet, R., Mehler, A., Mačutek, J., Krupa, V., Köhler, R., Jayaram, B.D., Grzybek, P., Altmann, G. (2008). *Word frequency studies* (to appear).