Printed in Göttinger Beiträge zur Sprachwissenschaft, Heft 14, 83-91 (2007)

On diversity of word frequencies and language typology

Ioan-Iovitz Popescu, Bucharest
Gabriel Altmann, Lüdenscheid

Abstract. In the article it will be shown that different morphological structures of languages give rise to different word frequency distributions. The forms of the distributions evaluated by means of repeat rate and entropy show that strongly synthetic and strongly analytic languages are situated on the two poles of a continuous scale.

The fact that words occur with different frequencies in texts is very familiar. If one ranks them according to decreasing frequency or if one orders them according to the number of words occurring x-times (frequency spectrum), one obtains monotone decreasing convex curves (or distributions) that have been modelled in different ways. All this is well known under the collective name "Zipf´s law". Different other presentations are possible yielding respective models (cf. Popescu et al. 2008). While Zipf´s law representing the power or zeta function disseminates in many different sciences on well known grounds, in linguistics it seems to decrease in its importance because it has a very weak and even controversial interpretation.

In this contribution we shall not touch the modelling but look directly at the frequencies and their diversity. Since we shall do it in 20 languages simultaneously, we expect that some conclusions concerning typological properties of languages can be drawn. Of course, this is a pilot study that can be stepwise refined if further languages will be processed.

Consider the set of relative frequencies of words $\{p_1,p_2,\ldots,p_n\}$ taken from a complete text. For the sake of simplicity the index *i* can mean an ordering. There are different indices measuring the diversity of these frequencies and even a chi-square test for homogeneity can be performed. In any case we know that the frequencies are not uniformly distributed, and even if we know that one and the same curve (distribution) can be used as a model, the diversity of frequencies differs from text to text and even from language to language. The latter kind of diversity will be studied here.

We choose only two diversity indices, namely the Shannon entropy (H) which is one of the great number of entropy expressions (cf. Esteban, Morales 1995) and Herfindahl´s concentration measure, used in linguistics since G. Herdan who called it "repeat rate" (RR). The formulas are

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

and

(2) $$RR = \sum_{i=1}^{n} p_i^2$$

where $\sum_i p_i = 1$.

The formulas can be transformed into one another (see Appendix), hence they show the same property in a slightly different manner. The repeat rate moves in interval $RR \in \langle 1/n, 1\rangle$ where

1/n means maximal dispersion (i.e. the case when all frequencies are uniformly distributed) and 1 means maximal concentration (i.e. when all frequencies are concentrated in one value). The entropy lies in the interval $H \in <0, \log_2 n>$, where 0 means maximal concentration, and $\log_2 n$ means maximal dispersion (uniformity). Here n is the inventory size, i.e. the number of classes.

Needless to say, a writer can consciously control his text, especially if it is short enough, but in longer texts a subconscious trend gains the upper hand, the text controls the writer, as Machiavelli used to say, and the text converges to a state somewhere in the given intervals. It can be conjectured that this state depends on some properties of language because word form repetition is associated with the morphological status of language, as has been shown elsewhere (Popescu, Altmann 2008a,b). In strongly synthetic languages the number of forms is greater, thus the number of hapax legomena will be greater, too, and the dispersion increases, the texts get more dispersed, i.e. the entropy increases and the repeat rate decreases. In strongly analytic languages having small number of forms, some (formal) words must be repeated very frequently, the number of hapax legomena decreases, and the text tends to the other extreme: small dispersion, high concentration.

If the above conjecture is correct, the measurement of synthetism/analytism can be performed without caring for the morphological specificities of a language. A word frequency count of several texts must give a measure of this property. In order to test this hypothesis, we used 145 texts from 20 languages (taken from Popescu et al. 2008), prepared a usual word count (rank and spectrum) and computed the above indicators. The detailed results are given in Tables 1 and 3 and eventually gathered in Figure 1 as entropy H in terms of repeat rate RR for both ranks and spectra. This will show that, notwithstanding the different nature of the relative frequencies $\{p_1, p_2, \ldots, p_n\}$, say of ranks and of spectra, there should exist an universal $H = f(RR)$ function.

Table 1

Entropies and repeat rates of ranked word frequencies in 145 texts from 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]

Text	**N**	**V**	**RR ranks**	**H ranks**		**Text**	**N**	**V**	**RR ranks**	**H ranks**
B 01	761	400	0,0092	7,8973		Kn 18	4485	1782	0,0035	9,7515
B 02	352	201	0,0012	7,0994		Kn 19	1787	833	0,0041	8,9712
B 03	515	285	0,0086	7,5827		Kn 20	4556	1755	0,0038	9,6909
B 04	483	286	0,0092	7,598		Kn 21	1455	790	0,0047	8,938
B 05	406	238	0,0112	7,3055		Kn 22	4554	1794	0,0042	9,6289
B 06	687	388	0,0095	7,8501		Kn 23	4685	1738	0,0036	9,6444
B 07	557	324	0,0076	7,7944		Kn 30	4499	2005	0,0032	10,0072
B 08	268	179	0,0105	7,107		Kn 31	4672	1920	0,0028	9,8862
B 09	550	313	0,0093	7,6576		Lk 01	345	174	0,016	6,7685
B10	556	317	0,0113	7,6055		Lk 02	1633	479	0,0181	7,3035
Cz 01	1044	638	0,007	8,6163		Lk 03	809	272	0,0204	6,8508
Cz 02	984	543	0,0078	8,3282		Lk 04	219	116	0,0214	6,2882
Cz 03	2858	1274	0,0086	8,9529		Lt 01	3311	2211	0,0027	10,5032
Cz 04	522	323	0,0076	7,877		Lt 02	4010	2334	0,0038	10,2814
Cz 05	999	556	0,012	8,1959		Lt 03	4931	2703	0,0019	10,5934

Cz 06	1612	840	0,0101	8,6111	Lt 04	4285	1910	0,0034	9,8252
Cz 07	2014	862	0,0101	8,4876	Lt 05	1354	909	0,003	9,3625
Cz 08	677	389	0,008	7,9987	Lt 06	829	609	0,0033	8,4581
Cz 09	460	259	0,0192	7,412	M 01	2062	396	0,0209	6,9856
Cz 10	1156	638	0,0069	8,4876	M 02	1187	281	0,0241	6,7198
E 01	2330	939	0,0099	8,5197	M 03	1436	273	0,0252	6,5851
E 02	2971	1017	0,0098	8,3972	M 04	1409	302	0,0235	6,6909
E 03	3247	1001	0,0137	8,2471	M 05	3635	515	0,0182	7,1346
E 04	4622	1232	0,0139	8,4634	Mq 01	2330	289	0,0244	6,6095
E 05	4760	1495	0,0103	8,7676	Mq 02	451	143	0,0288	6,1063
E 06	4862	1176	0,0172	8,2191	Mq 03	1509	301	0,0379	6,5012
E 07	5004	1597	0,0096	8,8057	Mr 15	4693	1947	0,0032	9,8764
E 08	5083	985	0,0192	7,901	Mr 16	3642	1831	0,0024	10,0120
E 09	5701	1574	0,0102	8,6865	Mr 17	4170	1853	0,0024	9,9799
E 10	6246	1333	0,0159	8,3391	Mr 18	4062	1788	0,0034	9,7898
E 11	8193	1669	0,0129	8,5906	Mr 20	3943	1725	0,0026	9,8472
E 12	9088	1825	0,012	8,5717	Mr 21	3846	1793	0,0022	9,9948
E 13	11625	1659	0,0119	8,4674	Mr 22	4099	1703	0,004	9,6097
G 01	1095	539	0,0117	8,0326	Mr 23	4142	1872	0,0026	9,9538
G 02	845	361	0,0108	7,7006	Mr 24	4255	1731	0,0028	9,8062
G 03	500	281	0,0122	7,4369	Mr 26	4146	2038	0,0025	10,0913
G 04	545	269	0,0123	7,353	Mr 30	5054	2911	0,0018	10,6433
G 05	559	332	0,0103	7,7183	Mr 31	5105	2617	0,002	10,4632
G 06	545	326	0,0087	7,7918	Mr 32	5195	2382	0,0024	10,1882
G 07	263	169	0,0128	6,9781	Mr 33	4339	2217	0,0019	10,3521
G 08	965	509	0,0077	8,2157	Mr 34	3489	1865	0,0019	10,1542
G 09	653	379	0,0085	7,9035	Mr 40	5218	2877	0,0018	10,6589
G 10	480	301	0,0021	7,7245	Mr 43	3356	1962	0,0017	10,2964
G 11	468	297	0,0078	7,7563	R 01	1738	843	0,006	8,7903
G 12	251	169	0,0125	6,9814	R 02	2279	1179	0,0066	9,1346
G 13	460	253	0,0095	7,449	R 03	1264	719	0,0065	8,7035
G 14	184	129	0,0144	6,6629	R 04	1284	729	0,0055	8,7736
G 15	593	378	0,0062	8,081	R 05	1032	567	0,007	8,3954
G 16	518	292	0,0074	7,6923	R 06	695	432	0,0072	8,1436
G 17	225	124	0,0153	6,5269	Rt 01	968	223	0,0338	6,2661
H 01	2044	1079	0,0155	8,838	Rt 02	845	214	0,0256	6,3747
H 02	1288	789	0,0133	8,6954	Rt 03	892	207	0,0216	6,542
H 03	403	291	0,0188	7,5293	Rt 04	625	181	0,0249	6,3644
H 04	936	609	0,0117	8,4426	Rt 05	1059	197	0,0202	6,5085
H 05	413	290	0,013	7,6043	Ru 01	2595	1240	0,0069	9,1104
Hw 01	282	104	0,0243	6,0083	Ru 02	17205	6073	0,0049	10,5714
Hw 02	1829	257	0,0206	6,5548	Ru 03	3853	1792	0,005	9,5531
Hw 03	3507	521	0,0211	7,0628	Ru 04	753	422	0,0079	8,0561
Hw 04	7892	744	0,0218	6,5388	Ru 05	6025	2536	0,0044	9,9181
Hw 05	7620	680	0,0185	7,0618	Sl 01	756	457	0,0088	8,1613
Hw 06	12356	1039	0,0193	7,272	Sl 02	1371	603	0,0078	8,2723

I 01	11760	3667	0,0055	9,8671	Sl 03	1966	907	0,0086	8,7048
I 02	6064	2203	0,0068	9,413	Sl 04	3491	1102	0,0169	8,2855
I 03	854	483	0,0106	8,1008	Sl 05	5588	2223	0,0054	9,6509
I 04	3258	1237	0,0069	8,9123	Sm 01	1487	266	0,0309	6,3481
I 05	1129	512	0,0084	8,0893	Sm 02	1171	219	0,0273	6,3632
In 01	376	221	0,0101	7,2975	Sm 03	617	140	0,0282	5,9515
In 02	373	209	0,0108	7,214	Sm 04	736	153	0,034	5,9275
In 03	347	194	0,01	7,178	Sm 05	447	124	0,0299	5,8972
In 04	343	213	0,0077	7,4299	T 01	1551	611	0,0165	7,6919
In 05	414	188	0,0115	6,9893	T 02	1827	720	0,0167	7,8474
Kn 01	3713	1664	0,0042	9,7114	T 03	2054	645	0,018	7,5103
Kn 02	4508	1738	0,0032	9,7285					

In order to get a more lucid survey, we take the means of individual languages and obtain the concentrated results in Table 2. Of course, the *RR* and *H* could be relativized but we leave them in original state. As can be seen, the order of languages concerning *RR* and *H* is almost the same. Further texts would, perhaps, strengthen the coincidences. If one looks at Table 2, one can see that at one extreme one finds the very analytic Polynesian languages, at the other the well known "synthetic" languages. On the basis of previous studies, this result could be expected. Languages lie on a Humboldtian scale, they can be classified in classes only with smaller or greater force – e.g. by taking some intervals – however, the position of every language in this dimension can be established simply by counting word-forms in texts.

Table 2
Means of repeat rates and entropies of ranking in 20 languages

Language	mean RR ranks	Language	mean H ranks
Marathi	0,0024	Marathi	10,1010
Latin	0,0030	Latin	9,8373
Kannada	0,0037	Kannada	9,5958
Russian	0,0058	Russian	9,4418
Romanian	0,0065	Italian	8,8765
Italian	0,0076	Romanian	8,6568
Bulgarian	0,0088	Slovenian	8,6150
Slovenian	0,0095	English	8,4597
Czech	0,0097	Czech	8,2967
German	0,0100	Hungarian	8,2219
Indonesian	0,0100	Tagalog	7,6832
English	0,0128	Bulgarian	7,5498
Hungarian	0,0145	German	7,5297
Tagalog	0,0171	Indonesian	7,2217
Lakota	0,0190	Maori	6,8232
Hawaiian	0,0209	Lakota	6,8028

Maori	0,0224	Hawaiian	6,7498
Rarotongan	0,0252	Rarotongan	6,4111
Samoan	0,0301	Marquesan	6,4057
Marquesan	0,0304	Samoan	6,0975

Using the spectrum of frequencies representing the numbers of words in individual classes (x = 1 = hapax legomena, x = 2 = dislegomena,…) we see a similar picture. In Table 3 one finds the RR and H values for 145 texts in 20 languages.

Table 3
Repeat rates and entropies for spectra of 145 texts in 20 languages

Text	**N**	**V**	**RR spectra**	**H spectra**		**Text**	**N**	**V**	**RR spectra**	**H spectra**
B 01	761	400	0,5758	1,4725		Kn 18	4483	1782	0,4551	1,9559
B 02	352	201	0,6018	1,3763		Kn 19	1787	833	0,4639	1,8273
B 03	515	285	0,576	1,4807		Kn 20	4556	1755	0,4424	2,0101
B 04	983	286	0,6221	1,3051		Kn 21	1455	790	0,5547	1,4948
B 05	406	238	0,6367	1,2729		Kn 22	4554	1764	0,4731	1,923
B 06	687	388	0,6441	1,2496		Kn 23	4685	1738	0,4536	2,0266
B 07	557	324	0,6014	1,381		Kn 30	4499	2005	0,4916	1,7912
B 08	268	179	0,667	1,1324		Kn 31	4672	1920	0,4944	1,8661
B 09	550	313	0,6137	1,3081		Lk 01	345	174	0,5564	1,5426
B 10	556	317	0,6101	1,3003		Lk 02	1633	479	0,4358	2,0506
Cz 01	1044	638	0,6703	1,1517		Lk 03	809	272	0,4426	2,0157
Cz 02	984	543	0,5971	1,3683		Lk 04	219	116	0,5158	1,592
Cz 03	2858	1274	0,5919	1,4559		Lt 01	3311	2211	0,6259	1,0661
Cz 04	522	323	0,5828	1,3298		Lt 02	4010	2334	0,6608	1,1888
Cz 05	999	556	0,6549	1,2232		Lt 03	4931	2703	0,5937	1,4039
Cz 06	1612	840	0,6349	1,2712		Lt 04	4285	1910	0,5287	1,6817
Cz 07	2014	862	0,5469	1,6084		Lt 05	1354	909	0,6727	1,0968
Cz 08	677	389	0,5671	1,3897		Lt 06	829	609	0,7417	0,8842
Cz 09	460	259	0,5553	1,4539		M 01	2062	396	0,2916	2,8033
Cz 10	1156	638	0,6508	1,2578		M 02	1187	281	0,3198	2,562
E 01	2330	939	0,5243	1,6853		M 03	1436	273	0,268	2,8554
E 02	2971	1017	0,5418	1,7416		M 04	1409	303	0,3483	2,5675
E 03	3247	1001	0,4224	2,0406		M 05	3635	515	0,2618	3,0597
E 04	4622	1232	0,3603	2,3397		Mq 01	2330	289	0,1592	3,5209
E 05	4760	1495	0,4529	1,9698		Mq 02	451	143	0,3686	2,2475
E 06	4862	1176	0,3418	2,4092		Mq 03	1509	301	0,2589	2,7786
E 07	5004	1597	0,4814	1,8753		Mr 15	4693	1947	0,4871	1,8592
E 08	5083	985	0,2824	2,7468		Mr 16	3642	1831	0,5459	1,581
E 09	5701	1574	0,4386	2,0699		Mr 17	4170	1853	0,4761	1,837
E 10	6246	1333	0,3181	2,595		Mr 18	4062	1788	0,5111	1,7482
E 11	8139	1669	0,3176	2,6323		Mr 20	3943	1725	0,4804	1,8362
E 12	9088	1825	0,3735	2,4596		Mr 21	3846	1793	0,4937	1,7674

E 13	11265	1659	0,245	3,0631		Mr 22	4099	1703	0,5124	1,7798
G 01	1095	530	0,5787	1,4542		Mr 23	4142	1872	0,4964	1,7656
G 02	845	361	0,4166	1,9695		Mr 24	4255	1731	0,4396	1,9831
G 03	500	281	0,6301	1,2979		Mr 26	4146	2038	0,5526	1,5766
G 04	545	269	0,5058	1,6561		Mr 30	5054	2911	0,5912	1,4314
G 05	559	332	0,633	1,2309		Mr 31	5105	2617	0,5674	1,4988
G 06	545	326	0,6029	0,6824		Mr 32	5195	2382	0,5445	1,6302
G 07	263	169	0,6322	1,2203		Mr 33	4339	2217	0,5359	1,6015
G 08	965	509	0,5786	1,4472		Mr 34	3489	1865	0,5616	1,5108
G 09	653	379	0,6491	1,2539		Mr 40	5218	2877	0,6072	1,3578
G 10	480	301	0,6398	1,216		Mr 43	3356	1962	0,6085	1,3404
G 11	468	297	0,6302	1,2164		R 01	1738	843	0,5433	1,5685
G 12	251	169	0,7121	1,0321		R 02	2279	1179	0,6125	1,333
G 13	460	253	0,5224	1,0787		R 03	1264	719	0,6383	1,2613
G 14	184	129	0,7084	1,0092		R 04	1284	729	0,6332	1,2975
G 15	593	378	0,6621	1,1787		R 05	1032	567	0,5856	1,3811
G 16	518	292	0,5646	1,4983		R 06	695	432	0,6827	1,1099
G 17	225	124	0,4974	1,6234		Rt 01	968	223	0,3538	2,4921
H 01	2044	1079	0,6279	1,2741		Rt 02	845	214	0,3869	2,3785
H 02	1288	789	0,6698	1,0801		Rt 03	892	207	0,2707	2,743
H 03	403	291	0,8025	0,6781		Rt 04	625	181	0,358	2,3568
H 04	936	609	0,7129	0,9396		Rt 05	1059	197	0,1974	3,1981
H 05	413	290	0,7571	0,8167		Ru 01	2595	1240	0,5992	1,4394
Hw 01	282	104	0,3428	2,256		Ru 02	17205	6073	0,5435	1,6859
Hw 02	1829	257	0,3996	2,1569		Ru 03	3853	1792	0,5975	1,4363
Hw 03	3507	521	0,2752	2,9744		Ru 04	753	422	0,5835	1,3991
Hw 04	7892	744	0,1885	3,3509		Ru 05	6025	2536	0,5522	1,6118
Hw 05	7620	680	0,2338	3,3693		Sl 01	756	457	0,6504	1,1913
Hw 06	12356	1039	0,263	3,2598		Sl 02	1371	603	0,5182	1,7276
I 01	11760	3667	0,4935	1,8826		Sl 03	1966	907	0,54	1,5871
I 02	6064	2203	0,5467	1,7029		Sl 04	3491	1102	0,4397	1,9883
I 03	854	483	0,6415	1,2557		Sl 05	5588	2223	0,5355	1,6775
I 04	3258	1237	0,4963	1,8032		Sm 01	1487	266	0,2565	2,822
I 05	1129	512	0,5093	1,0179		Sm 02	1171	219	0,2318	2,9834
In 01	376	221	0,589	1,3666		Sm 03	617	140	0,321	2,625
In 02	373	209	0,5266	1,5313		Sm 04	736	153	0,2875	2,74
In 03	347	194	0,4892	1,631		Sm 05	447	124	0,3398	2,3792
In 04	343	213	0,508	1,4629		T 01	1551	611	0,5964	1,4762
In 05	414	188	0,4449	1,9195		T 02	1827	720	0,5868	1,4742
Kn 01	3713	1664	0,4995	1,7621		T 03	2054	645	0,5575	1,766
Kn 02	4508	1738	0,4425	2,0255						

Again, the presentation of means would allow us to obtain better lucidity of results. In Table 4 one can see the means.

Table 4
Mean of repeat rates and entropies in 20 languages
based on word-form frequencies

Language	mean RR spectra	Language	mean H spectra
Hungarian	0,7140	Hungarian	0,9577
Latin	0,6373	Latin	1,2203
Romanian	0,6159	German	1,2980
Bulgarian	0,6149	Romanian	1,3252
Czech	0,6052	Bulgarian	1,3279
German	0,5979	Czech	1,3510
Tagalog	0,5802	Russian	1,5145
Russian	0,5752	Italian	1,5325
Italian	0,5375	Tagalog	1,5721
Slovenian	0,5368	Indonesian	1,5823
Marathi	0,5301	Slovenian	1,6344
Indonesian	0,5115	Marathi	1,6532
Lakota	0,4877	Lakota	1,8002
Kannada	0,4771	Kannada	1,8683
English	0,3923	English	2,2791
Rarotongan	0,3134	Rarotongan	2,6337
Maori	0,2979	Samoan	2,7099
Samoan	0,2873	Maori	2,7696
Hawaiian	0,2838	Marquesan	2,8490
Marquesan	0,2622	Hawaiian	2,8946

The difference between ranking and spectra is for some languages very great. Hungarian, Kannada and Marathi exchange their places, so to say. But for the majority of languages this ranking is quite stable.

The situation would be different if we counted lemmas. Such a count would not have any consequences for the typology because the complete morphology would be eliminated. Nevertheless, it could be used for stylistic problems.

Fortunately, the sampling properties of repeat rate and entropy are well known, thus asymptotic significance tests for differences between texts or languages can be performed without any difficulties.

The result shows that word frequencies should be studied more thoroughly and compared with other results of typology. Languages differ not only by having ornot having something but by the weight of the given phenomenon which is represented by its frequency of occurrence.

Though the theoretical relationship between repeat rate and entropy is asymptotic and somewhat raw, as shown in the Appendix, the empirical relationship is quite unambiguous. For both ranking and spectra we obtain the results as presented in Figure 1.

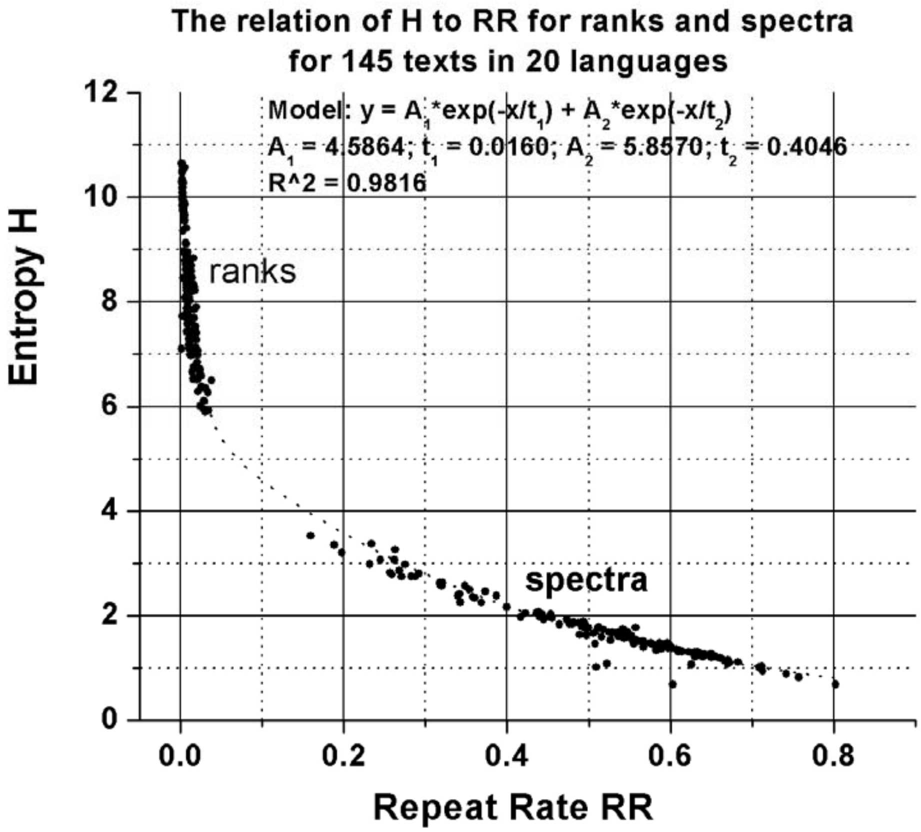

Figure 1. Empirical relationship between repeat rate and entropy in 145 texts of 20 languages

This relationship can be captured by a simple formula such as, for instance, by a two component exponential function which has been proposed as a substitute for Zipf´s law (cf. Popescu, Köhler, Altmann 2008).

References

Esteban, M.D., Morales, D. (1995). A summary of entropy statistics. *Kybernetica 31(4), 337-346.*

Popescu, I.-I., Altmann, **G.** (2008a). Hapax legomena and language typology *(JQL submitted)*

Popescu, I.-I., Altmann, G. (2008b). Zipf´s mean and language typology. *Glottometrics 16, 2008 submitted).*

Popescu, I.-I., Köhler, R., Altmann, G. (2008). Zipf´s law – another view. (submitted)

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*. (submitted).

Appendix

In order to show the asymptotic relation between R (repeat rate) and entropy (H) we consider the expression of chi-square, namely

$$(1)\quad X^2 = \sum_{i=1}^{K} \frac{[f_i - E(f_i)]^2}{E(f_i)}.$$

In case of uniformity $E(f_i) = N/K$ where N is the sample size and K is the number of classes. Expanding (1) we obtain

$$X^2 = \frac{K}{N}\sum_{i=1}^{K} f_i^2 - N\,.$$

Dividing both sides by N we obtain

$$(2)\qquad \frac{X^2}{N} = K\sum_{i=1}^{K}\frac{f_i^2}{N^2} - 1 = KR - 1$$

since the expression under the sum is the definition of repeat rate.

Further, we use the well known result that the information statistics $2I$ is asymptotically chi-square distributed, i.e.

$$(3)\qquad 2I = 2\sum_{i=1}^{K} f_i \ln\frac{f_i}{N/K} \approx X^2\,.$$

Changing the natural logarithms in dyadic ones (ld) by ln x = ld x ln 2 and dividing both sides of (3) by N we obtain

$$\frac{X^2}{N} = 2\ln 2\sum_{i=1}^{K}\frac{f_i}{N}\left(ld\,\frac{f_i}{N} + ld\,K\right)$$

$$= -2\ln 2\,H + 2\ln 2\,ld\,K$$
$$= 2\ln 2(H_0 - H)\,,$$

since ld $K = H_0$. Hence from (2) and (3) we obtain

$$KR - 1 = 2\ln 2(H_0 - H)$$

and finally

$$R = \frac{2\ln 2(H_0 - H) + 1}{K}$$

or vice versa

$$H = H_0 - \frac{KR - 1}{2\ln 2}\,.$$