Göttinger Beiträge zur Sprachwissenschaft 13, 2006, 87-98

Some geometric properties of word frequency distributions

Ioan-Iovitz Popescu, Bucharest
Gabriel Altmann, Lüdenscheid

Abstract: The present article shows two complementary methods for estimating the technique of word exploitation (repetition) with a given vocabulary. For the sake of simplicity word forms are counted. The methods are based on the geometric properties of the rank-frequency distribution and the frequency spectrum.

Keywords: Indices, word frequency, rank-frequency

1. Introduction

In a previous article (Popescu, Altmann 2006) we defined some crucial points (*h, k, m, n*) for word frequency distributions that can be used – cum grano salis – for the characterization of vocabulary richness and for mechanical distinguishing of auxiliary words from content words. We are aware of the fact that for all these points confidence intervals must be set up and that the membership in these classes is fuzzy. Further, some of the points are appropriate only for long-tailed monotonously decreasing distributions, other ones can be used generally. Nevertheless, all of them can serve as starting points for analysing the geometry of word frequency distributions.

In this article we restrict ourselves to the *h*- and *k*-points and some geometric properties of the word frequency distributions and propose some indices which show how the writer managed to find a balance between *N* (= text length in word forms), *V* (= vocabulary = number of different words = highest rank) and the exploitation of individual words. It must be noted that our proposals hold for units whose inventories are not very small, i.e. rather for words and morphemes whose inventories can be potentially infinite but not for phonemes, letters and in some languages not even for syllables. In those cases the argumentation and the interpretation must be modified.

Consider first some basic definitions:

The *h*-point of a rank-frequency distribution (of words) is usually the point at which $r \approx f(r)$, i.e. the rank equals the frequency at this rank. For other monotonously decreasing distributions it is always possible to find a point whose distance to [0,0] is minimal, namely $h = \min\sqrt{x^2 + f(x)^2}$. Another possibility is to join the points *V* and f(1) with a straight line and seek a point *h* on the frequency sequence yielding a triangle with the two mentioned points with maximal area (see below). The *h*-point need not be an integer but we adhere here to such a determination.

The *k*-point is an analogy to the *h*-point but used with frequency spectra of words. For the sake of differentiation in relation to the word frequency f(x) we shall call the frequencies of the spectrum *g*(*x*), hence *g*(1) is the number of words occurring exactly once. Here the total sum of g(x) equals the vocabulary V just as, similarly, the total sum of f(x) equals the text length N. Also W is the total number of different non-zero frequency classes just as, similarly, V is the total number of different words.

A typical case of situating the *h*- and the *k*-point can be seen in Figure 1 and 2 respectively.

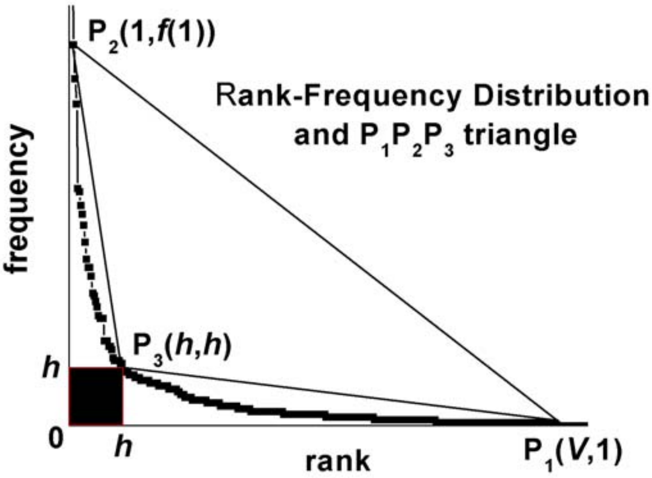

Figure 1. Typical word rank-frequency distribution where h is the unique point at which rank equals frequency, $h = f(h)$; V is the vocabulary (the maximum word rank); and $f(1)$ is the maximum occurrence frequency (of the word of rank one). These three remarkable points define a characteristic $P_1P_2P_3$ triangle. Notice that the sum of all occurrence frequencies f (that is the total area covered by the distribution curve) is equal to the total word count (text length or text size) N.

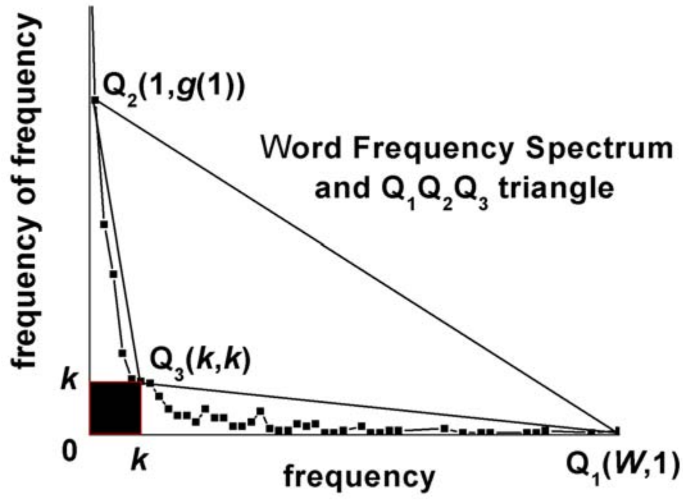

Figure 2. Typical word frequency spectrum where k is the unique point at which the frequency equals the frequency of frequency, $k = g(k)$; W is the number of non-zero frequency classes; and $g(1)$ is the occurrence of the words having the frequency equal to unity (the maximum frequency of frequencies). These three remarkable points define a characteristic $Q_1Q_2Q_3$ triangle. Notice that the sum of all occurrence frequencies g (that is the total area covered by the frequency spectrum) is equal to the text vocabulary V.

For all computations we used the counter that can be found on Internet (http://www.georgetown.edu/faculty/ballc/webtools/web_freqs.html) which works well for English but texts in other languages need slight cosmetic improvements, e.g. in Italian it counted "anch´io" (also I), "all´orecchio" (to the ear), "l´età" (the age) as one word respectively. But since we were interested only in the method, we left cases like that in all languages unchanged. However, if conclusions should be drawn about individual texts, the texts must be pre-processed.

2. The rank-frequency distribution

Consider first the rank-frequency distribution of words, Fig. 1. The author "plans" (cf. Orlov, Boroda, Nadarejšvili 1982) to write a text of a certain length N and convey a certain (epistemic) information. In order to do it, he needs a certain vocabulary V which can be smaller or greater according to the aim of the text. Didactic or scientific texts usually use a smaller number of words repeating them more frequently, literary texts contain relatively more words. The author adapts his technique to the aims of the texts. The result of this adaptation can be measured and characterized. The basic points of the text are given empirically: we have $f(1)$, i.e. the frequency of the most frequent word whose relative frequency is almost a constant not depending on N and beginning the distribution on the left hand side. Further we have V, the highest rank whose value is usually $f(V) = 1$. And finally, we have the h-point on the frequency sequence which is the nearest to the origin. One could take as a special characteristic the area between the straight line connecting $P_1(V,1)$ and $P_2(1,f(1))$ and the frequency sequence along the distribution arc length. but in that case we would be forced to compute and add V-1 trapezoids. For the sake of simplicity we approximate the real fulfilment of writer´s aim by a triangle taking into account the remarkable h-point $P_3(h,h)$ by computing the area A_h of the corresponding $P_1P_2P_3$ triangle of Figure 1 as follows

(1) $$A_h = (1/2)[Vf(1) + 2h - h(V + f(1)) - 1]$$

The triangle $P_1P_2P_3$ becomes both maximal and rectangular for $P_3(1,1)$, that is for $h = 1$, with an area

(2) $$A_{max} = (1/2)(V - 1)(f(1) - 1)$$

so that we can define a new normalized indicator

(3) $$A = A_h/A_{max}.$$

The ratio $A = A_h/A_{max}$ shows the exploitation of the given vocabulary for the given aim. It can be seen that the nearer the curve (the real frequency sequence) to the upper line ($\overline{P_1P_2}$), the stronger is the exploitation of some few words, i.e. the smaller is vocabulary richness. On the contrary, the greater the area A_h, the smaller the exploitation of individual words (only some few words are strongly exploited) and the more words must be used in the text. Thus $f(1)$, h and V are sufficient to give a first characteristic of the word exploitation in a text.

In a similar way, this time with reference to Figure 2, we can consider the $Q_1Q_2Q_3$ triangle, with the area B_k, the maximal area B_{max}, and the corresponding normalized indicator $B = B_k/B_{max}$ given respectively by

(4) $B_k = (1/2)[Wg(1) + 2k - k(W + g(1)) - 1]$

(5) $B_{max} = (1/2)(W - 1)(g(1) - 1]$

and

(6) $B = B_k/B_{max}$

In Table 1 and Table 2 respectively we show some results concerning the indicators *A* and *B*. For the sake of simplicity, we give integer *h*-values, defined as the closest first integer rank to a frequency. The same convention is maintained for integer *k*-values. Let us illustrate the simple computation using the Nobel lecture of Frederick G. Banting from Table 1a. Using formula (1) we obtain

A_h(Banting) = 1/2[1669(622) + 2(32) – 32(1669+622) – 1] = 482435

A_{max}(Banting) = 1/2[(1669 –1)(622 – 1)] = 517914

A = 482435/517914 = 0.9315.

It is to be noted that *A* is simply a proportion that can be treated further statistically, i.e. it allows tests for difference between writers to be made. Consider the quantities A_1 and A_2 of two texts. We build a mean of both setting

$$\bar{A} = (A_{h1} + A_{h2})/(A_{max1} + A_{max2})$$

and insert it in the criterion

(7) $$z = \frac{A_1 - A_2}{\sqrt{\bar{A}(1-\bar{A})\left(\frac{1}{A_{\max 1}} + \frac{1}{A_{\max 2}}\right)}}$$

which is asymptotically normally distributed. Let us illustrate the procedure comparing two English texts with very similar *A*s, namely Banting (*A* = 0.9315) and Mcleod (*A* = 0.9303). The weighted mean $\bar{A}$ = (482435 + 250872)/(517914 + 269663) = 0.9311, hence

$$z = \frac{0.9315 - 0.9303}{\sqrt{0.9311(1-0.9311)\left(\frac{1}{517914} + \frac{1}{269663}\right)}} = 1.995$$

being significant at the 0.05 level. It must be noted that greater differences are all significant because of very great A_{max}, hence an ordering of writers according to *A* is at the same time an estimation of their technique of *word exploitation* using the given vocabulary and writing a text of a certain length. We do not speak about vocabulary richness but rather about stylistic differences concerning word repetition.

Table 1a
English texts (Nobel lectures)

Text	*N*	*V*	*f*(1)	*h*	*A*
Frederick G. Banting, Med 1925	8193	1669	622	32	0,9315
John Macleod, Med 1925	4862	1176	460	24	0,9303
Linus Pauling, Peace 1963	6246	1333	546	28	0,9302
Richard P. Feynman, Phys 1965	11265	1659	780	41	0,9245
J.M. Buchanan Jr., Econ 1986	4622	1232	366	23	0,9219
Ernest Rutherford, Chem 1908	5083	985	466	26	0,9208
Pearl Buck, Lit 1938	9088	1825	617	39	0,9175
George C. Marshall, Peace 1953	3247	1001	229	19	0,9031
Bertrand Russell, Lit 1950	5701	1574	342	29	0,9001
Saul Bellow, Lit 1976	4760	1495	297	26	0,8988
Sinclair Lewis, Lit 1930	5004	1597	237	25	0,8833

Table 1b
German texts

Author	Text	N	V	f(1)	h	A
Schiller, F.v.	Der Taucher	1095	530	83	12	0,8451
Anonym	Fabel - Zaunbär	845	361	48	9	0,8076
Krummacher, F.A.	Das Krokodil	500	281	33	8	0,7563
Anonym	Fabel - Mäuschen	545	269	32	8	0,7482
Goethe, J.W.v.	Der Gott und die Bajadere	559	332	30	8	0,7375
Sachs, H.	Das Kamel	545	326	30	8	0,7371
Heine, H.	Belsazar	263	169	17	5	0,7262
Droste-Hülshoff, A.	Der Geierpfiff	965	509	39	11	0,7172
Goethe, J.W.v	Elegie 19	653	379	30	9	0,7030
Goethe, J.W.v	Elegie 13	480	301	18	7	0,6271
Goethe, J.W.v	Elegie 15	468	297	18	7	0,6268
Goethe, J.W.v	Elegie 2	251	169	14	6	0,5861
Fontane, Th.	Gorm Grymme	460	253	19	8	0,5833
Goethe, J.W.v	Elegie 5	184	129	10	5	0,5243
Moericke, E.	Peregrina	593	378	16	8	0,5149
Lichtwer, M.G.	Die Rehe	518	292	16	8	0,5094

Table 1c
Romanian Text

Author	Text	N	V	f(1)	h	A
Eminescu, M.	Scrisoarea III - Satire III	2279	1179	110	16	0.8497
Eminescu, M.	Scrisoarea IV - Satire IV	1264	719	65	12	0.8128
Eminescu, M.	Scrisoarea I - Satire I	1284	729	49	10	0.8001
Eminescu, M.	Luceafarul - Lucifer	1738	843	62	14	0.7714
Eminescu, M.	Scrisoarea V - Satire V	1032	567	46	11	0.7601
Eminescu, M.	Scrisoarea II - Satire II	695	432	30	10	0.6688

Table 1d
Indonesian (online) newspaper texts

Text	*N*	*V*	*f*(1)	*h*	*A*
Assagaf-Ali Baba	346	221	16	6	0,6442
BRI Siap Cetak	373	209	18	7	0,6182
Pengurus	347	194	14	6	0,5896
Pemerintah	343	213	11	5	0,5811
Pelni Jamin Tiket Tidak Habis	414	188	16	8	0,4961

Table 1e
Hungarian (online) newspaper texts

Text	*N*	*V*	*f*(1)	*h*	*A*
Orbán Viktor beszéde	2044	1079	225	12	0,9407
A nominalizmus forradalma	1288	789	130	8	0,9369
Népszavazás	403	291	48	4	0,9259
Egyre több	936	609	76	7	0,9101
Kunczekolbász	413	290	32	6	0,8214

Table 1f
Italian texts

Author	Text	*N*	*V*	*f*(1)	*h*	*A*
Silvio Pellico	Le mie prigioni	11760	3667	388	37	0,8972
Alessandro Manzoni	I promessi sposi	6064	2203	257	25	0,8954
Giacomo Leopardi	Canti	854	483	64	10	0,8385
Grazia Deledda	Canne al vento	3258	1237	118	21	0,8129
Edmondo de Amicis	from Il cuore	1129	512	42	12	0,7102

Table 1g
Latin texts

Author	Text	*N*	*V*	*f(1)*	*h*	*A*
Vergil	Georgicon liber primus	3311	2211	133	12	0,9117
Apuleius	Fables, Book 1	4010	2334	190	18	0,9028
Ovidius	Ars amatoria, liber primus	4931	2703	103	19	0,8169
Cicero	Post reditum in senatu oratio	4285	1910	99	20	0,7962
Martialis	Epigrammata	1354	909	33	8	0,7735
Horatius	Sermones.Liber 1, Sermo 1	829	609	19	7	0,6568

As can be seen in Tables 1a to 1g, the *A* does not depend either on *N* or on *V*. In Figure 3 one finds the relation to *N*. The variability at low *N* is enormous – perhaps because we have many texts of this size - but does not change with increasing *N*. Even if one would suppose a dependence, the convergence of *A* to some finite value is evident. Though the number of languages and texts analysed is not sufficient (and will never be sufficient) to yield a strong corroboration to this statement, we can conjecture that A may be a characteristic of the language and within language that of the style or of the genre. Comparing the intervals in which the *A* values lie:

English:	0.8833 – 0.9315
Hungarian:	0.8214 – 0.9407
Italian:	0.7102 – 0.8972
Romanian:	0.6688 – 0.8497
Latin:	0.6568 – 0.9028
German:	0.5094 – 0.8451
Indonesian:	0.4961 – 0.6442

we see that the differences are considerable. Even the work of one writer displays great differences (c.f. Eminescu in Romanian and Goethes Elegies in German). The great difference between Hungarian and Indonesian newspaper texts is rather a language, not a style problem. This field is open to further investigation.

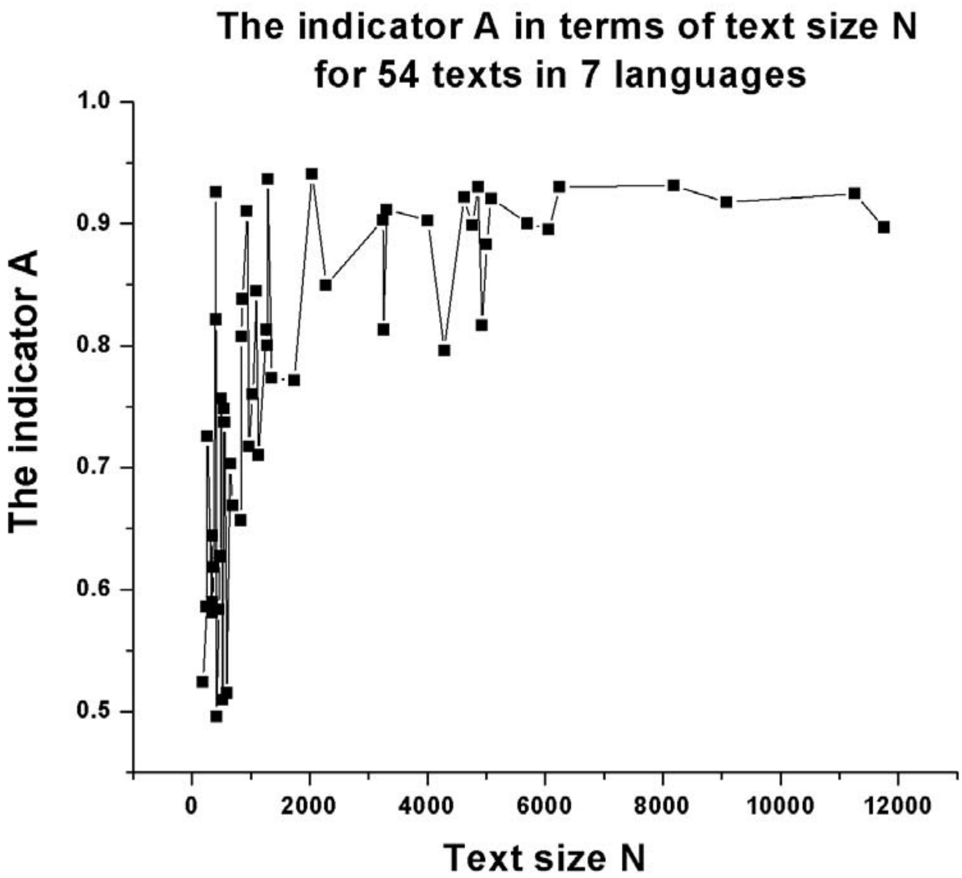

Figure 3. The relation between A and N

Further investigations should be concentrated (a) on one language, (b) within that language on different text length *N*, (c) different text sorts, (d) different historical epochs and (e) complete works of an author. In the course of the analysis possibly further factors will appear whose effect could be taken into account. As can easily be seen, this is rather a problem for a research project than for an isolated article.

3. The frequency spectrum

Consider now the analogous characteristic *B*, defined in formulas (4) to (6) based on the frequency spectrum. Here the interpretation is different. The smaller B, the more words are in the text which occur once, twice,…, i.e. the text is richer. In rich texts, the curve (the frequency spectrum) would lie near the upper straight line joining Q_1 and Q_2 and the triangle would be small compared with the triangle B_{max}. In Tables 2a to 2g one can find the relevant numbers concerning the frequency spectrum.

Table 2a
English texts (Nobel lectures)

Author	Text	*N*	*W*	*g*(1)	*k*	*B*
Sinclair Lewis	Lit 1930	5004	45	1076	7	0,8581
Bertrand Russell	Lit 1950	5701	52	1005	8	0,8558
Saul Bellow	Lit 1976	4760	43	972	7	0,8510
Pearl Buck	Lit 1938	9088	64	1071	10	0,8487
J.M. Buchanan Jr.	Econ 1986	4622	42	694	7	0,8450
Richard P. Feynman	Phys 1965	11265	70	737	11	0,8415
Frederick G. Banting	Med 1925	8193	57	881	11	0,8101
Linus Pauling	Peace 1963	6246	50	694	10	0,8033
John Macleod	Med 1925	4862	42	641	9	0,7924
George C. Marshall	Peace 1953	3247	33	621	8	0,7700
Ernest Rutherford	Chem 1908	5083	48	474	12	0,7427

Table 2b
German texts

Author	Text	N	W	g(1)	k	B
Anonym	Fabel - Mäuschen	545	15	186	4	0,7699
Droste-Hülshof, A.	Der Geierpfiff	965	18	380	5	0,7542
Fontane, Th.	Gorm Grymme	460	13	177	4	0,7330
Moericke, E.	Peregrina	593	12	305	5	0,7177
Goethe, J.W.v	Elegie 13	480	12	238	4	0,7147
Goethe, J.W.v	Elegie 5	184	8	108	3	0,6960
Goethe, J.W.v	Elegie 19	653	14	303	5	0,6791
Goethe, J.W.v.	Der Gott und die Bajadere	559	13	261	5	0,6513
Goethe, J.W.v	Elegie 2	251	10	142	4	0,6457
Anonym	Fabel - Zaunbär	845	16	223	6	0,6444
Krummacher, F.A.	Das Krokodil	500	12	221	5	0,6182
Lichtwer, M.G.	Die Rehe	518	12	216	5	0,6179
Heine, H.	Belsazar	263	9	133	4	0,6023
Schiller, F.v.	Der Taucher	1095	19	396	8	0,5935
Sachs, H.	Das Kamel	545	12	249	6	0,5257
Goethe, J.W.v	Elegie 15	468	10	233	7	0,3075

Table 2c
Romanian texts

Author	Text	N	W	g(1)	k	B
Eminescu, M.	Scrisoarea V - Satire V	1032	19	425	5	0,7683
Eminescu, M.	Scrisoarea II - Satire II	695	14	354	4	0,7608
Eminescu, M.	Luceafarul - Lucifer	1738	26	607	7	0,7501
Eminescu, M.	Scrisoarea III – Satire III	2279	27	909	8	0,7231
Eminescu, M.	Scrisoarea IV - Satire IV	1264	18	568	6	0,6971
Eminescu, M.	Scrisoarea I - Satire I	1284	18	574	7	0,6366

Table 2d
Indonesian (online) newspaper texts

Text	*N*	*W*	*g*(1)	*k*	*B*
Assagaf-Ali Baba	346	11	167	3	0,7880
BRI Siap Cetak	373	11	148	4	0,6803
Pengurus	347	11	131	5	0,5692
Pelni Jamin Tiket Tidak Habis	414	11	122	5	0,5669
Pemerintah	343	8	146	4	0,5512

Table 2e
Hungarian (online) newspaper texts

Text	*N*	*W*	*g*(1)	*k*	*B*
Kunczekolbász	413	9	251	3	0,7420
Orbán Viktor beszéde	2044	19	845	7	0,6596
Egyre több	936	13	510	5	0,6588
A nominalizmus forradalma	1288	14	639	6	0,6077
Népszavazás	403	7	260	4	0,4891

Table 2f
Italian texts

Author	Text	*N*	*W*	*g*(1)	*k*	*B*
Grazia Deledda	Canne al vento	3258	36	849	6	0,8513
Edmondo de Amicis	from: Il cuore	1129	23	356	5	0,8069
Alessandro Manzoni	I promessi sposi	6064	46	1605	10	0,7944
Silvio Pellico	Le mie prigioni	11760	65	2515	14	0,7917
Giacomo Leopardi	Canti	854	17	383	5	0,7395

Table 2g
Latin texts

Author	Text	*N*	*W*	*g(1)*	*k*	*B*
Apuleius	Fables, Book 1	4010	32	1879	7	0,8033
Cicero	Post reditum in senatu oratio	4285	36	1360	9	0,7655
Ovidius	Ars amatoria, liber primus	4931	33	2050	9	0,7461
Vergil	Georgicon liber primus	3311	21	1793	7	0,6967

Martialis	Epigrammata	1354	15	738	6	0,6362
Horatius	Sermones.Liber 1, Sermo 1	829	9	522	4	0,6195

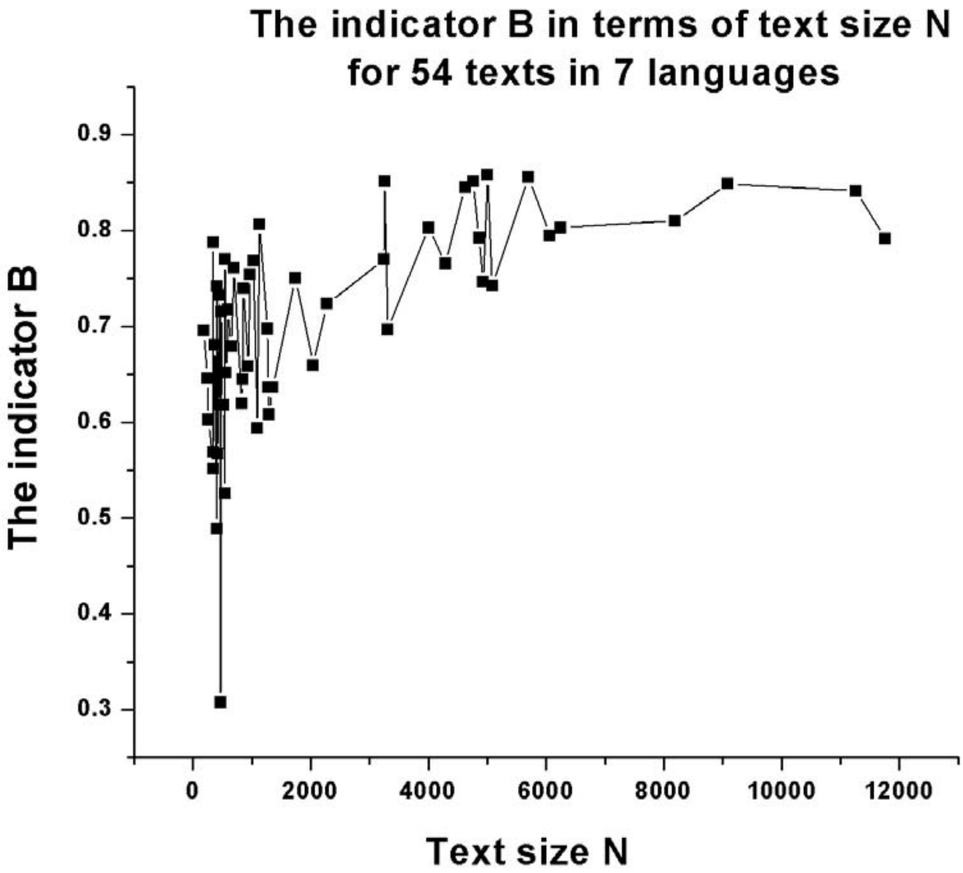

Figure 4. The relation between B and N

The relation of *B* to *N* can be seen in Figure 4. Again, the variability with small *N* is as expected, with greater *N* the coefficient is stable. Even if one would set up a hypothesis of dependence of B on N, it could be corroborated only for individual cases (like languages, genres etc.), i.e. eliminating all other factors except *N*. The languages behave differently:

English: 0.7427 – 0.8581
Italian: 0.7395 – 0.8513
Romanian: 0.6366 – 0.7683
Latin: 0.6195 – 0.8033
Indonesian: 0.5512 – 0.7880
Hungarian: 0.4891 – 0.7420
German: 0.3075 – 0.7699

but one cannot make statements about authors, only about texts. Even a unique genre has a great interval of values. Here, the newspaper texts of Hungarian and Indonesian are nearer to one another. Again, only further research can shed light upon the boundary conditions which led to the given indicators.

The coefficient B is also a proportion which can be processed statistically in the above mentioned way.

4. Another relationship

Since the study of relationships *A* vs. *N* and *B* vs. *N* must be postponed until many longer texts will be processed, we can look at the relationship between *A/B* and *N*. Let us call *A/B* the *wording indicator* of a text giving a complex picture of the play with words or rather word forms, their repetition and variation. As can be seen in Figure 5 we get a result as expected. This figure looks like a cross-section of an asymmetrical funnel with an almost "horizontal" axis centred at about A/B between 1.11 and 1.14. We suppose that, by increasing indefinitely the number of points in this graph (that is the wording number A/B of texts of various size N), we will rather more compactly fill up this funnel and its profile will get much clearer than only 54 points presently may suggest. In other words, while the funnel axis remain fixed at about, say, A/B = 1.125, the A/B scaling min/max limits get preliminarily closer as the text size N increases. Thus, as Fig. 5 shows, this min/max range decreases

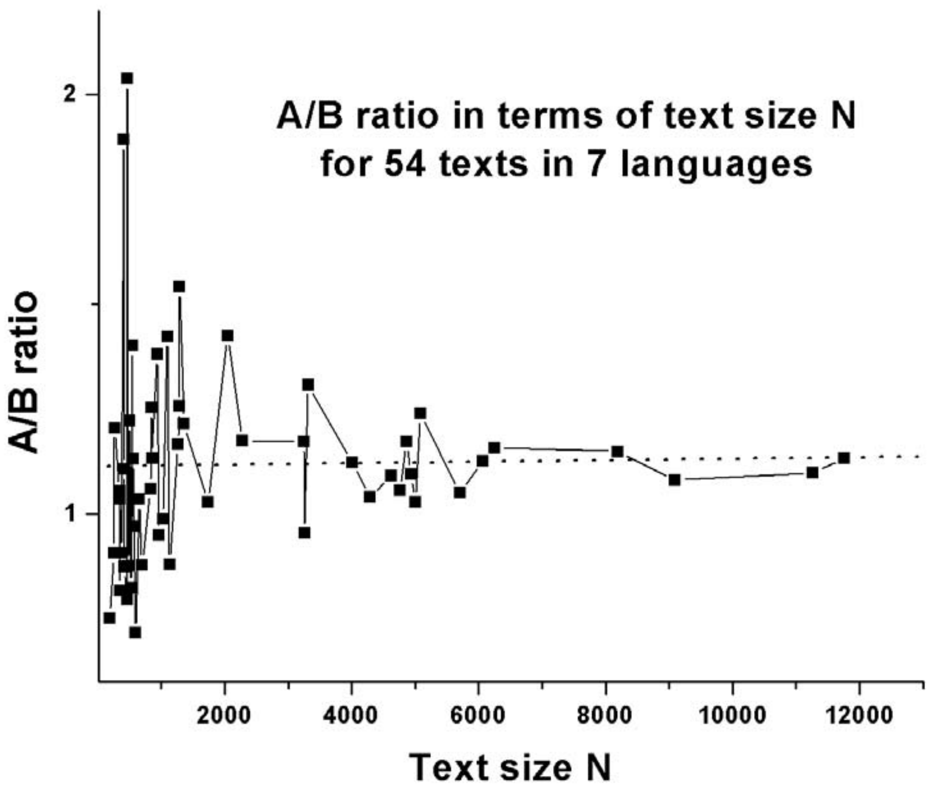

Figure 5. The relationship between the wording indicator *A/B* and text size *N*

from about	[0.7 - 2.05]	at N = 500
to about	[0.9 - 1.55]	at N = 1000
to about	[1 - 1.4]	at N = 2000
to about	[1 - 1.3]	at N = 3000

a.s.o.

Obviously, the easiest relative comparison/ranking one can make for texts of about the same size, while for texts of very different sizes one should take into consideration "the funnel neck narrowing". But only a future research can show whether the indicator stabilizes with increasing *N*.

Since both A and B are proportions, A/B is a ratio of two independent proportions. The variance of A/B can easily be derived as

(8) $$Var(A/B) = \frac{1}{B^2}Var(A) + \frac{A^2}{B^4}Var(B),$$

hence an asymptotic test criterion for the comparison of two texts or for the deviation from the expected value can easily be set up.

Conclusions

The present study is an exercise in methods, not in literary criticism. It shows that texts behave differently on account of different boundary conditions which can be due to language, style, genre, epoch, author´s age, author´s aim etc. but it cannot show which condition is present and to what degree in the given text. A thorough literary analysis would be necessary to deeper penetrate in this domain where we have to do with the cultural and psychological embedding of the author. Perhaps studies of this kind would enable us to make even a step in the quantification of cultural features. It cannot be excluded that even test persons reading the texts must be included in this research.

The indicators A, B and A/B seem to have a great dispersion whose decomposition could help us isolating the boundary conditions, the greatest enigma of any text research. Especially texts of greater length – yet not too great – should be analysed. We recommend not to surpass N = 10000 because longer text are not homogeneous and any indicators are some distortions resulting from mixing texts. Only shorter texts can be kept homogeneous.

All the above mentioned indicators are characteristics of frequency structuring of texts. As a matter of fact, we would obtain other results if we counted lemmas or morphemes. In principle, no way of counting is "more correct" but perhaps one of them would inspire us more than the other ones to establish some laws of frequency structuring. This is the only criterion of "better" or "worse" of any scientific method.

References

Orlov, J.K., Boroda, M.G., Nadarejšvili, I.Š. (1982). *Sprache, Text, Kunst. Quantitative Analysen.* Bochum: Brockmeyer.

Popescu, I.-I., Altmann, G. (2006). Some aspects of word frequencies. *Glottometrics 13, 23-46.*