Glottometrics 15, 2007, 71-81

Writer´s view of text generation

Ioan-Iovitz Popescu, Bucharest[1]
Gabriel Altmann, Lüdenscheid

Abstract: Generally, a "writer's view", defined by the angle between the ends of the word rank-frequency distribution, as seen from the *h*-point, should be limited in the interval [π/2, π]. However, as shown in the present paper with 176 texts from 20 languages, actually the lower limit appears to be the golden number φ = 1.618... , rather than π/2 = 1.57...

Keywords: h-point, rank-frequency distributions, golden section

The writer of a text abides by some mechanisms of writing. Writing and speaking is a human activity performed according to some rules, habits or laws. Some of the mechanisms are conscious, other ones are unconscious, the writers need not even be aware of their existence. What is conscious, is the given language, its grammar, the contents of the text, the approximate length of the text and, after some thinking, the order of events in which the narrative must be presented, etc. The writer can control consciously the sentence length, the choice of words and some other properties, but he cannot control consciously everything. Some of the unconscious processes may sometimes become conscious, especially when one speaks an artificially learned language. But even in a foreign language we must suppose the existence of unconscious control of different processes.

Here we shall examine only the process of increasing word frequencies in text and building the rank-frequency distribution. As the text increases in length, new words are added occurring at first just once (new types), or some words are repeated getting higher frequency of occurrence. Predictions about a new occurrence of a certain word can be made only probabilistically and are not very reliable. While both the words at lower ranks continuously increase their frequency and new words (*hapax legomena*) are added, the beginning of the frequency curve increases vertically and the tail of the curve increases horizontally. Besides, the words at not extreme ranks steadily change their (rank) positions, thus the frequency curve seems to be in a steady flux. Nevertheless, there are some more solid entities in rank-frequencies of words conserving their properties in spite of the steady motion accompanying the increase of text. We mention here the distribution itself – being of Zipfian type or its generalizations; Ord's criteria – being characteristic for languages; and the type-token curves, etc. (cf. Popescu et al. 2007). Other properties arise through self-organization, e.g. building of binary runs of F-segments in long sentences (cf. Uhlířová 2007).

Here we shall report a peculiar convergence property of the rank-frequency distribution, $f = f(r)$, by observing the development of the so-called *h*-point during the increase of text length. As is known (s. Popescu 2006; Popescu, Altmann 2006; Popescu et al. 2007), the *h*-point is defined as the point at which $r = f(r)$, i.e. where the rank is equal to frequency. As a rule, the actual rank-frequency distributions are discrete and this is why sometimes the *h*-point is situated between two ranks; but its exact value can easily be computed in different ways. In graphical terms, the fitted graph of $f(r)$ has the *h*-point in common with the line $y = x$. An analytical method consists in the best fitting with a Zipf distribution, $f(r) = c / r^a$ (with *a* and

[1] Address correspondence to: iovitzu@gmail.com

c constants), or with a Mandelbrot distribution $f(r) = c/(b+r)^a$ (with *a, b,* and *c* constants), from which we get $h_Z = c^{1/(1+a)}$ or, respectively the equation $h_M = c/(b+h_M)^a$. Obviously, the *h*-point of actual discrete distributions is closely related to the mathematical fixed point of continuous functions, and is defined by the same rule $r = f(r)$.

Writing the text, the *h*-point automatically increases with increasing *N*. However, it has been shown that the ratio $a = N/h^2$ is very stable for languages and depends on their degree of analyticity/syntheticity.

With the first written words, the *h*-point is 1 and it remains so until one of the written words is repeated (Fig. 1). This usually happens after two-three sentences. The first repeated word creates a triangle between the points $P_1(V, 1)$ – *V* being the vocabulary of the text – $P_2(1, f_1)$ – f_1 being the frequency of the most frequent word – and $P_3(h, h)$, which begins to increase (Fig. 2). The *h*-point increases with increasing *N* (Fig. 3 and 4) but we suppose that something in this growth remains constant and is subconsciously controlled by the writer. Imagine that the writer "sits" at the *h*-point and controls the consistent development of the upper part of the rank-frequency curve (between f_1 and *h*) and its lower part (between *h* and *V*). One can suppose that a "normal" text arises if the angle of the triangle does not surpass a minimum and a maximum value. Perhaps there is a general constant or different constants for individual languages and genres, but we shall study the problem generally.

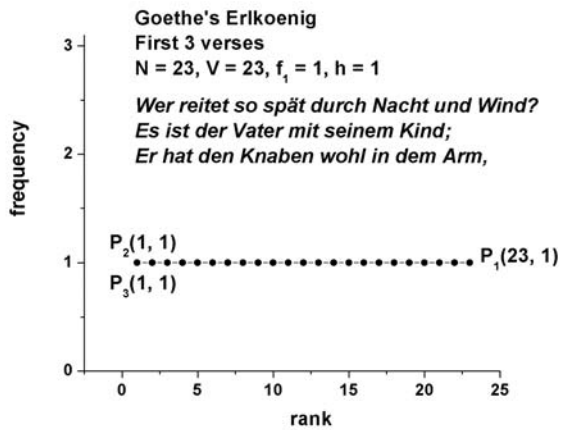

Figure 1

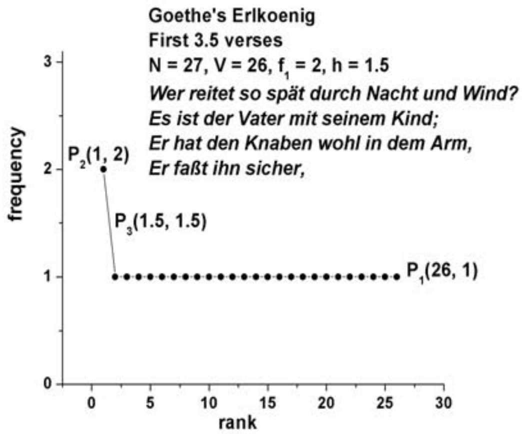

Figure 2

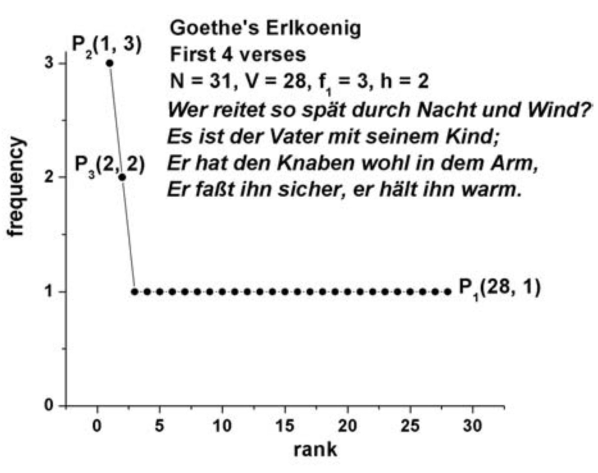

Figure 3

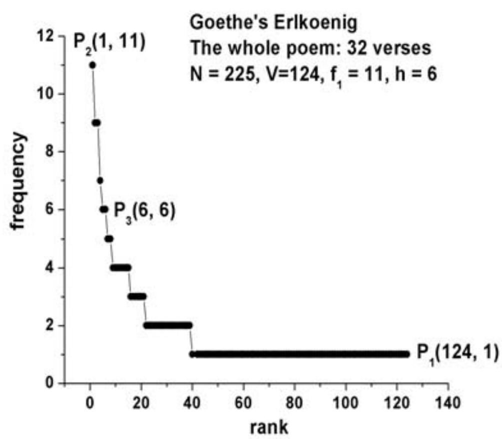

Figure 4

Figures 1 to 4 (Figures 2 to 4 show the developing h-point)

Let us consider a word rank-frequency distribution represented, as usually (see Figure 4), in a Cartesian *Oxy* system of coordinates, *Ox* for ranks (with the maximum value V at the point P_1) and *Oy* for frequencies (with the maximum value f_1 at the point P_2). Let us also have the *h*-point located at the point P_3. To summarize, we have the three characteristic points of the distribution given by their coordinates as follows:

$P_1\ (V, 1)$
$P_2\ (1, f_1)$
$P_3\ (h, h)$

We will consider the angle seen from the *h*-point P_3 and formed by the directions towards the distribution end P_1 and top P_2. In other words, we want an expression of the angle between the following two vectors:

(a) the vector **a** (a_x, a_y), directed from the *h*-point P_3 towards the distribution top P_2, having the components

$a_x = - h$
$a_y = f_1 - h$

and the absolute value (modulus)

(1) $$a = (a_x^2 + a_y^2)^{1/2} = (h^2 + (f_1 - h)^2)^{1/2}$$

(b) the vector **b** (b_x, b_y), directed from the *h*-point P_3 towards the distribution end P_1, having the components

$b_x = V - h$
$b_y = - h$

and the absolute value (modulus)

(2) $$b = (b_x^2 + b_y^2)^{1/2} = ((V - h)^2 + h^2)^{1/2}$$

In order to get the angle between the above vectors ***a*** and ***b*** we shall use the well known dot product formula of vector calculus

(3) $$\cos \alpha = \mathbf{a.b}/ab$$

where

(4) $$\mathbf{a.b} = a_x b_x + a_y b_y$$

that is

(5) $$\cos \alpha = \frac{a_x b_x + a_y b_y}{[(a_x^2 + a_y^2)^{1/2}][(b_x^2 + b_y^2)^{1/2}]}$$

which, in our particular case, becomes

(6) $$\cos \alpha = \frac{-[h(f_1 - h) + h(V - h)]}{[h^2 + (f_1 - h)^2]^{1/2}[h^2 + (V - h)^2]^{1/2}}$$

As expected, all numerical cosine values are negative because the corresponding angles are located in the second quadrant (from π/2 to π radians, respectively from 90 to 180 degrees). Let us illustrate the computation for a text having length N = 761, V = 400, h =10 and f_1 = 40. Inserting these numbers in (6) we obtain

$$\cos\alpha = \frac{-[10(40-10)+10(400-10)]}{[10^2+(40-10)^2]^{1/2}[10^2+(400-10)^2]^{1/2}} = -0.3404$$

hence **α** = 1.9182 radians.

In order to study the behavior of the angle **α** in radians, which we call "writer´s view", we used the data published in Popescu et al. (2007) containing 176 texts in 20 languages as presented in Table 1. Here B = Bulgarian, Cz = Czech, E = English, G = German, H = Hungarian, Hw = Hawaiian, In = Indonesian, I = Italian, Kn = Kannada, Lt = Latin, Lk = Lakota, M = Maori, Mq = Marquesan, Mr = Marathi, R = Romanian, Rt = Rarotongan, Ru = Russian, Sl = Slovenian, Sm = Samoan, T = Tagalog. Ordering the table according to increasing text length (N) we obtained a surprising result: The angle called "writer´s view" converges to the golden section known from several arts, different domains of aesthetics and sciences, namely to

$$\varphi = \frac{1+\sqrt{5}}{2} = 1.6180\ldots,$$

as can be seen in Table 1 and in Figure 5.

Table 1
The development of "writer's view" in increasing text size (176 texts from 20 languages)

Text ID	N	V	f(1)	h	cos α	α rad
G 14	184	129	10	5	-0.735	2.3965
Lk 04	219	116	18	6	-0.4953	2.0889
G 17	225	124	11	6	-0.7997	2.4977
G 12	251	169	14	6	-0.629	2.2511
G 07	263	169	17	5	-0.4126	1.9961
B 08	268	179	10	6	-0.8508	2.5883
Hw 01	282	104	19	7	-0.5647	2.1709
In 04	343	213	11	5	-0.6585	2.2896
Lk 01	345	174	20	8	-0.5941	2.207
In 03	347	194	14	6	-0.6252	2.2462
B 02	352	201	13	8	-0.8692	2.6244
In 02	373	209	18	7	-0.5658	2.1722
In 01	376	221	16	6	-0.5382	2.1391
H 03	403	291	48	4	-0.1044	1.6754
B 05	406	238	19	7	-0.5298	2.1292
H 05	413	290	32	6	-0.2454	1.8187
In 05	414	188	16	8	-0.7378	2.4006

Sm 05	447	124	39	11	-0.4541	2.0422
Mq 02	457	150	42	10	-0.3655	1.945
Cz 09	460	259	30	6	-0.2655	1.8395
G 13	460	253	19	8	-0.6143	2.2322
G 11	468	297	18	7	-0.5571	2.1617
G 10	480	301	18	7	-0.5568	2.1613
B 04	483	286	21	8	-0.5484	2.1512
G 03	500	281	33	8	-0.3325	1.9098
B 03	515	285	15	9	-0.8497	2.5862
G 16	518	292	16	8	-0.7267	2.3844
Cz 04	522	323	27	7	-0.3512	1.9296
G 04	545	269	32	8	-0.3451	1.9232
G 06	545	326	30	8	-0.3653	1.9447
B 09	550	313	20	9	-0.6559	2.2861
B 10	556	317	26	7	-0.3668	1.9464
B 07	557	324	19	8	-0.6085	2.2249
G 05	559	332	30	8	-0.3648	1.9443
G 15	593	378	16	8	-0.7222	2.3778
Sm 03	617	140	45	13	-0.4688	2.0587
Rt 04	625	181	49	11	-0.3395	1.9172
G 09	653	379	30	9	-0.4162	2.0000
Cz 08	677	389	31	8	-0.3483	1.9265
B 06	687	388	28	9	-0.4494	2.0369
R 06	695	432	30	10	-0.4683	2.0581
Sm 04	736	153	78	12	-0.2617	1.8356
Ru 01	753	422	31	8	-0.3467	1.9249
Sl 01	756	457	47	9	-0.2500	1.8234
B 01	761	400	40	10	-0.3404	1.9182
Lk 03	809	272	62	12	-0.2780	1.8525
Lt 06	829	609	19	7	-0.5139	2.1105
G 02	845	361	48	9	-0.2497	1.8232
Rt 02	845	214	69	13	-0.2885	1.8635
I 03	854	483	64	10	-0.2028	1.775
Rt 03	892	207	66	13	-0.3026	1.8782
H 04	936	609	76	7	-0.1125	1.6835
G 08	965	509	39	11	-0.3861	1.9672
Rt 01	968	223	111	14	-0.2087	1.781
Cz 02	984	543	56	11	-0.2575	1.8312
Cz 05	999	556	84	9	-0.1355	1.7067
R 05	1032	567	46	11	-0.3186	1.8951
Cz 01	1044	638	58	9	-0.1947	1.7668

Kn 004	1050	720	23	7	-0.4098	1.993
Rt 05	1059	197	74	15	-0.3252	1.902
G 01	1095	530	83	12	-0.1894	1.7614
I 05	1129	512	42	12	-0.3936	1.9753
Cz 10	1156	638	50	11	-0.2883	1.8632
Sm 02	1171	222	103	15	-0.2388	1.812
M 02	1175	277	127	15	-0.1892	1.7611
R 03	1264	719	65	12	-0.2373	1.8104
R 04	1284	729	49	10	-0.2618	1.8357
H 02	1288	789	130	8	-0.0757	1.6465
M 04	1289	326	137	15	-0.1697	1.7413
Kn 013	1302	807	35	10	-0.383	1.9638
Lt 05	1354	909	33	8	-0.3132	1.8894
Sl 02	1371	603	66	13	-0.2596	1.8334
M 03	1434	277	128	17	-0.2156	1.7881
Sm 01	1487	267	159	17	-0.186	1.7578
Mq 03	1509	301	218	14	-0.117	1.6881
T 01	1551	611	89	14	-0.2065	1.7788
Cz 06	1612	840	106	13	-0.154	1.7254
Lk 02	1633	479	124	17	-0.1931	1.7651
R 01	1738	843	62	14	-0.2962	1.8715
T 02	1827	720	107	15	-0.1819	1.7537
Hw 02	1829	257	121	21	-0.2914	1.8665
Mr 035	1862	1115	29	11	-0.5299	2.1293
Sl 03	1966	907	102	13	-0.1589	1.7304
Cz 07	2014	862	134	15	-0.1426	1.7139
H 01	2044	1079	225	12	-0.0675	1.6383
T 03	2054	645	128	19	-0.2015	1.7737
M 01	2062	398	152	18	-0.1799	1.7517
R 02	2279	1179	110	16	-0.1813	1.7531
E 01	2330	939	126	16	-0.1611	1.7326
Mq 01	2330	289	247	22	-0.1787	1.7505
Ru 02	2595	1240	138	16	-0.143	1.7143
Cz 03	2858	1274	182	19	-0.1308	1.702
Mr 002	2922	1186	73	18	-0.3256	1.9025
Mr 149	2946	1547	47	12	-0.3317	1.9089
E 02	2971	1017	168	22	-0.1708	1.7425
Mr 001	2998	1555	75	14	-0.2325	1.8055
Mr 007	3162	1262	80	16	-0.255	1.8286
Kn 003	3188	1833	74	13	-0.2154	1.7879
E 03	3247	1001	229	19	-0.1094	1.6804

I 04	3258	1237	118	21	-0.2284	1.8013
Lt 01	3311	2211	133	12	-0.1041	1.6751
Mr 293	3337	2006	41	13	-0.427	2.012
Mr 043	3356	1962	44	16	-0.5033	2.0982
Mr 150	3372	1523	64	16	-0.3263	1.9032
Mr 029	3424	1412	28	17	-0.8461	2.5795
Mr 034	3489	1865	40	17	-0.6018	2.2165
Sl 04	3491	1102	328	21	-0.0876	1.6585
Hw 03	3507	521	277	26	-0.1551	1.7265
Mr 052	3549	1628	89	17	-0.24	1.8132
Mr 154	3601	1719	68	17	-0.3257	1.9025
M 05	3620	514	234	26	-0.1767	1.7484
Mr 016	3642	1831	63	18	-0.3806	1.9612
Mr 006	3735	1503	120	19	-0.1974	1.7695
Mr 294	3825	1931	85	17	-0.2511	1.8247
Mr 296	3836	1970	92	18	-0.2453	1.8186
Mr 021	3846	1793	58	20	-0.4757	2.0666
Ru 03	3853	1792	144	21	-0.18	1.7518
Mr 020	3943	1825	62	19	-0.4138	1.9974
Mr 291	3954	1957	86	18	-0.2649	1.8389
Lt 02	4010	2334	190	18	-0.1118	1.6828
Mr 290	4025	2319	42	17	-0.5684	2.1754
Mr 288	4060	2079	84	17	-0.2539	1.8275
Mr 018	4062	1788	126	20	-0.1965	1.7686
Mr 038	4078	1607	66	20	-0.4103	1.9935
Mr 022	4099	1703	142	21	-0.1833	1.7551
Mr 027	4128	1400	92	21	-0.2982	1.8736
Mr 003	4140	1731	68	20	-0.3954	1.9773
Kn 012	4141	1842	58	19	-0.4473	2.0346
Mr 023	4142	1872	72	20	-0.369	1.9488
Mr 026	4146	2038	84	19	-0.2896	1.8646
Mr 017	4170	1853	67	19	-0.3777	1.9581
Mr 046	4186	1458	68	20	-0.3974	1.9795
Mr 036	4205	2070	96	19	-0.2486	1.822
Mr 024	4255	1731	80	20	-0.3273	1.9042
Lt 04	4285	1910	99	20	-0.2557	1.8293
Kn 017	4316	2122	122	18	-0.179	1.7507
Mr 033	4339	2217	71	19	-0.3513	1.9298
Kn 011	4541	2516	63	17	-0.353	1.9316
Mr 297	4605	2278	88	18	-0.2567	1.8305
E 04	4622	1232	366	23	-0.0859	1.6568

Mr 015	4693	1947	136	21	-0.1904	1.7623
Kn 016	4735	2356	93	18	-0.2409	1.814
E 05	4760	1495	297	26	-0.1131	1.6841
Mr 292	4765	2197	88	19	-0.2739	1.8482
Mr 289	4831	2312	112	19	-0.2083	1.7806
Mr 151	4843	1702	192	23	-0.1484	1.7198
E 06	4862	1176	460	24	-0.0757	1.6466
Kn 005	4869	2477	101	16	-0.1914	1.7634
Mr 295	4895	2322	97	20	-0.2598	1.8336
Lt 03	4931	2703	103	19	-0.2275	1.8003
Mr 005	4957	2029	172	19	-0.1326	1.7038
E 07	5004	1597	237	25	-0.1329	1.7041
E 08	5083	985	466	26	-0.086	1.6569
Mr 031	5105	2617	91	21	-0.2951	1.8703
Mr 028	5191	2386	86	23	-0.3521	1.9306
Mr 032	5195	2382	98	23	-0.3025	1.8781
Mr 040	5218	2877	81	21	-0.3373	1.9148
Kn 006	5231	2433	74	20	-0.3551	1.9338
Mr 010	5394	1650	217	27	-0.1571	1.7286
Mr 008	5477	1807	190	27	-0.1784	1.7501
Mr 030	5504	2911	86	20	-0.2966	1.8719
Sl 05	5588	2223	193	25	-0.1584	1.7299
E 09	5701	1574	342	29	-0.1109	1.682
Ru 04	6025	2536	228	25	-0.1321	1.7033
I 02	6064	2203	257	25	-0.1185	1.6896
Mr 009	6206	2387	93	26	-0.372	1.952
E 10	6246	1333	546	28	-0.0754	1.6463
Mr 004	6304	2451	314	24	-0.0923	1.6633
Hw 05	7620	680	416	38	-0.1586	1.7301
Hw 04	7892	744	535	38	-0.1297	1.7009
E 11	8193	1669	622	32	-0.0737	1.6445
E 12	9088	1825	617	39	-0.0891	1.66
E 13	11265	1659	780	41	-0.0807	1.6516
I 01	11760	3667	388	37	-0.115	1.686
Hw 06	12356	1039	901	44	-0.0953	1.6663
Ru 05	17205	6073	701	41	-0.0688	1.6396

In Figure 5 we see that some texts approximate this value even if they are short, but in any case, long texts tend to it. For instance, the latter case is best demonstrated by Goethe’s Faust as shown in the table below

Text	N	V	f(1)	h	cos α	α rad
Faust 1.	30625	6303	918	64	-0.0850	1.6559
Faust 1. and 2.	75050	13341	2089	90	-0.0518	1.6226

This is, perhaps, the way how to attain harmonic proportions in texts. It may be expected that further texts with $N \approx 2500$ had filled the gap in Figure 5 between the two branches of the points following most probably a power curve.

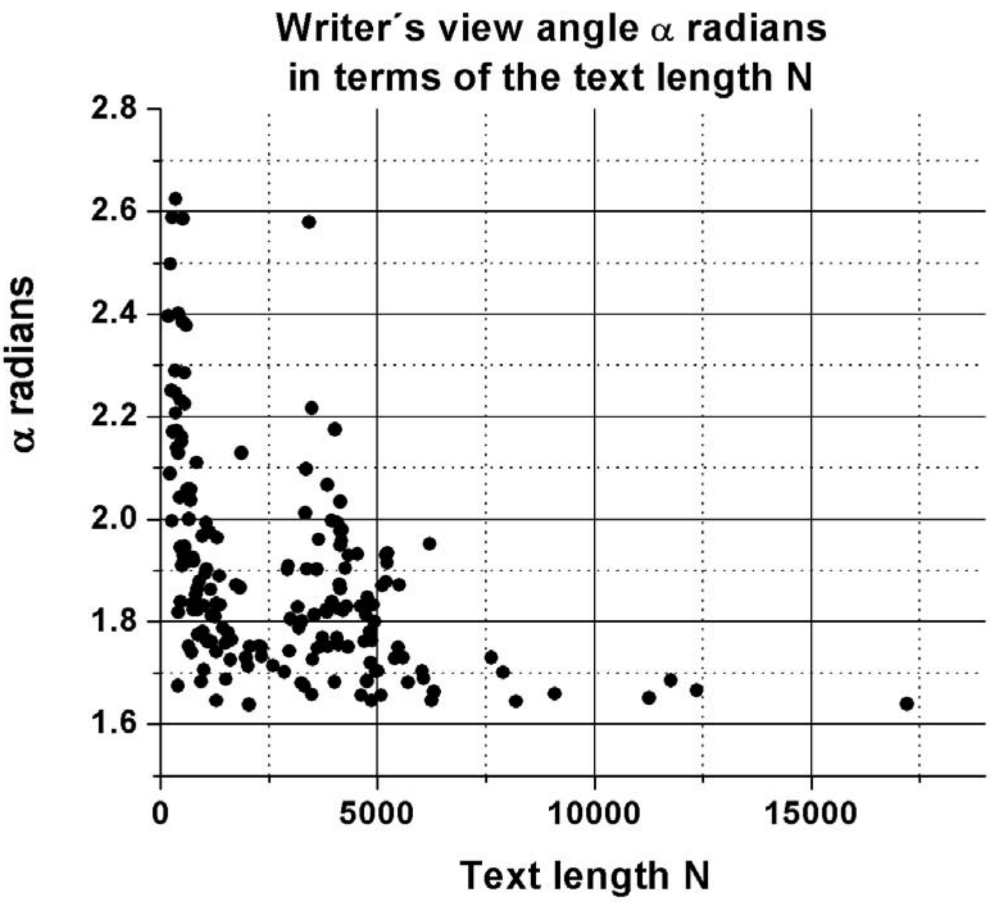

Figure 5. Illustrating the dependence of the "writer's view" *α* radians in terms of the text length *N* for 176 texts in 20 languages. All data lay within the interval φ = 1.618... to π = 3.14...

The theoretical upper angle is π = 3.14..., however, the actual boundary is not precisely known. Many shorter texts (e.g. poems) must be analysed in order to venture a well grounded statement. One example is given in continuation with cumulative sequences from Goethe's Erlkönig indicating for the corresponding upper boundary an average value of about 3 (Table 2).

Though we believe that the golden section is present in texts in some way and can be found by different methods, the fact that the rank-frequency distribution and its *h*-point can make it at least visible, is a good argument for further study of word frequencies.

Table 2
Development of α rad in Goethe´s Erlkönig

Goethe's Erlkönig	N	V	f(1)	h	cos α	α rad
first 10 words	10	10	1	1	-0.9939	3.0309
first 20 words	20	20	1	1	-0.9986	3.0890
first 30 words	30	27	3	2	-0.9272	2.7578
first 60 words	60	45	3	3	-0.9975	3.0703
first 100 words	100	69	4	4	-0.9981	3.0801
first 140 words	140	88	7	5	-0.9491	2.8213
first 220 words	220	122	11	6	-0.8003	2.4985

The texts in Table 1 are all of moderate length, allowing us to suppose that they were written "in one go". It is known that several writers were able to write a book in one day. Practically, longer texts are mixtures of texts even if they were written by the same author. The smooth process of writing is interrupted in many places and a new "vocabulary regime" can distort all ratios (e.g. TTR, the **α** angle, the thematic concentration etc.). Hence their investigation is not very prolific. Nevertheless, we ventured an experiment and took some very long texts, even translations (which are on many grounds forbidden in linguistics) and brought them to a common statement about the mean of the α radian. As individual texts they have no expressiveness – because they are mixed – but representing a statistical sample they can tell us something about the unweighted mean of the sample. In Table 3 we show some randomly chosen well known mixtures. Though with some texts α rad is below 1.60, their unweighted average is 1.61. Though we do not consider this result as corroborating our hypothesis, it is nevertheless enlightening. The texts presented in Table 1 and 3 are shown in Figure 6 where the long "texts" are better differentiated.

Further research can help us to decide whether the convergence is caused also by language or genre and help us to decipher the form of the convergence. It can help us also to diagnose text mixtures.

Table 3
Mixed texts

Author		Text	N	V	f(1)	h	cos α	α rad
Goethe	German	Faust 1.	30625	6303	918	64	-0.08496	1.6559
Goethe	German	Faust 1. and 2.	75050	13341	2089	90	-0.05176	1.6226
Milton	English	Paradise Lost	79879	10211	3330	98	-0.03999	1.6108
The Evangelists	English tr.	The Gospels	83932	3501	5669	112	-0.05316	1.6240
Conan Doyle	English	Sherlock Holmes	104230	8324	5601	112	-0.03403	1.6048
Homer	English tr.	The Odyssey	117386	6800	5875	137	-0.04442	1.6152
Homer	English tr.	The Iliad	152455	7776	9945	150	-0.03497	1.6058
Moses	English tr.	The Pentateuch	156872	4797	13667	150	-0.04335	1.6142
The Bible	English tr.	New Testament	180573	6005	10976	160	-0.04215	1.6130
Dickens	English	Great Expectations	185104	11376	8139	161	-0.03453	1.6053
Dostoevsky	English tr.	Crime and Punishment	202853	10728	7768	174	-0.03938	1.6102

Joyce	English	Ulysses	263324	29457	14905	169	-0.01724	1.5880
Dickens	English	David Copperfield	360779	17225	13918	210	-0.02766	1.5985
Tolstoy	English tr.	War and Peace	561723	20094	34391	255	-0.02032	1.5911
The Bible	English tr.	Old Testament	610051	10751	52934	270	-0.03088	1.6017
The Bible	English tr.	Old & New Testament	790624	12698	63910	294	-0.02832	1.5991

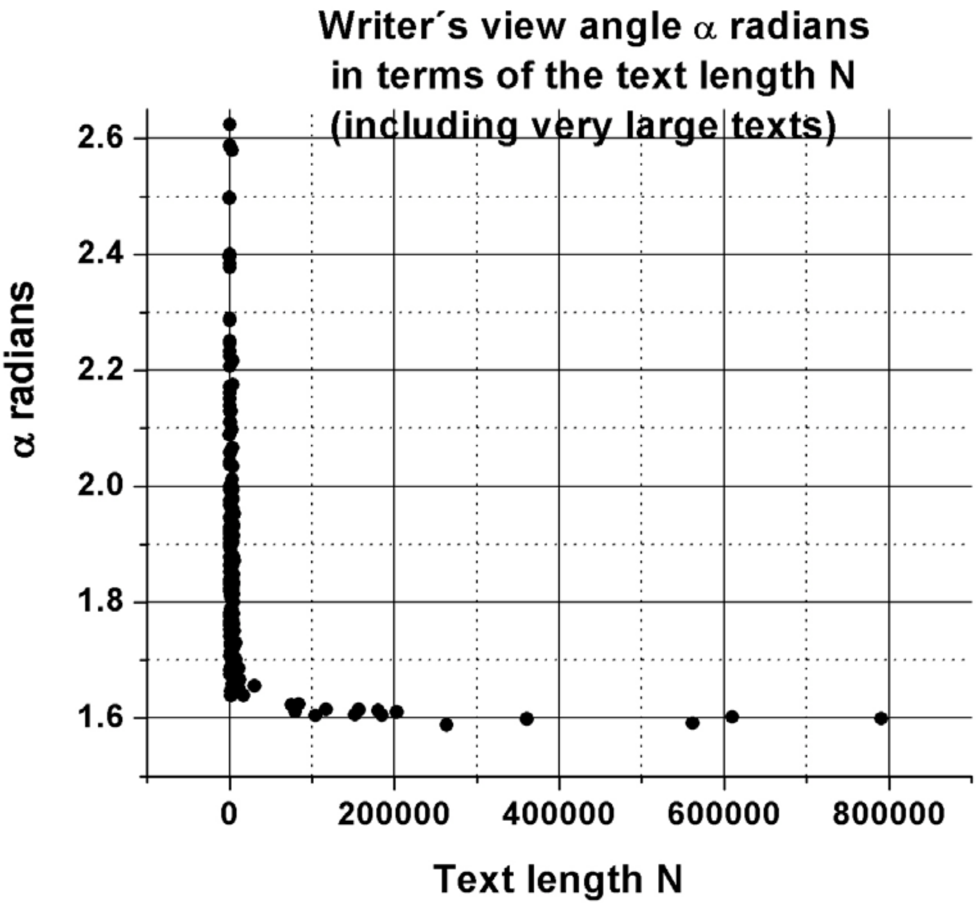

Figure 6. The course of α rad especially for very long texts

References

Goethe's *Faust*, German version from http://www.gutenberg.org/wiki/Main_Page

Livio, M. (2002). *The Golden Ratio: The Story of Phi, The World's Most Astonishing Number*. New York: Broadway Books.

Popescu, I.-I. (2006). Text ranking by the weight of highly frequent words. In: Grzybek, P., Köhler, R. (eds.), *Exact methods in the study of language and text: 553-562*. Berlin: de Gruyter.

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

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. (2007). *Word frequency studies.* (in press)

Uhlířová, L. (2007). Word frequency and position in sentence. *Glottometrics 14, 2007, 1-21.*