Glottometrics 26, 2013, 50-95

Some statistics for sequential text properties

Ioan-Iovitz Popescu, Bucharest
Peter Zörnig, Brasilia
Peter Grzybek, Graz
Sven Naumann, Trier
Gabriel Altmann, Lüdenscheid

Abstract. The aim oft he article is the measurement and the modelling of some sequential properties of word length, sentence length and word frequency by means of arc length, Hurst exponent and the distances between equal entities. Some of them were computed for various languages and their preliminary form has been shown.

Keywords: Arc length, Hurst exponent, distance, word length, sentence length, word frequency

1. Introduction

Written language moves in a one-dimensional space. The segmental entities make up a simple straight line. But if we begin to measure the properties of some entities quantitatively and replace them by the measured values, the straight line of written symbols changes and obtains a more or less fractal form. The given property need not be constant, it can begin to oscillate irregularly. In quantified form the text can be considered a time series whose properties can be scrutinized. Languages, individual texts, text sorts and properties may display differences with regard to their time series behaviour. The only condition is that the given property is variable (not constant like e.g. in monosyllabisms).

Time series have a number of properties which can be studied by a number of methods making up a whole discipline, therefore, we restrict ourselves here to the study of the smoothness/roughness of such series, i.e. to the variation of the values comparing subsequent neighbours.

There are a number of indicators measuring the smoothness of time series (cf. any text book on time series, e.g. Pandit, Wu 1983; Hamilton 1994; Brockwell, Davis 2010; Percival, Walden 2010; Kitagawa, Gersch 1996). We shall restrict ourselves to some of them used already previously in linguistics.

If the neighbouring values of some property in the time series differ strongly, the oscillation curve begins to cover more of the two-dimensional space than a simple straight line. Regular oscillation can easily be captured by Fourier analysis (or other methods) in such a way that the parameters are linguistically interpretable. But the more irregular the oscillation becomes, the more components must be added to the Fourier polynomial, and in that case the linguistic interpretation could become fuzzy.

Since in linguistic sequences there is no regular oscillation (except for those constructed artificially, e.g. rhythm in poetry), we can use the concept of smooth-

ness/roughness for our purposes and characterize texts, units, properties and in some cases also languages. Various properties can be found in Köhler, Altmann (2008: Chapter 9). In the present article we restrict ourselves to some elementary entities. As is known, the number of linguistic properties is infinite (cf. Altmann 2006) and linguistic entities are conceptual creations arising on the basis of the state of the art in linguistics.

Volatility and persistence of time series are most often measured using Hurst's exponent, Minkowski-sausage, Lyapunov coefficient and other indicators known from time-series research. Needless to say, all this can be used also in text analysis as initiated by L. Hřebíček (2000). The computation of the variance shows the variation of the values in the whole series, but does not yield an image of the neighbouring steps. Hence we begin with a slightly different approach considering the arc length between subsequent values of a property measured on respective units. This approach is naturally subdivided in several steps: first we choose a delimitable unit, e.g. syllable, morpheme, word, clause, sentence, verse, strophe, etc.; then we consider one of the enormous number of properties of these units. In the next step we analyse a text, replace the entities by the measured values and obtain a numerical sequence. Before we perform the two steps and evaluate the sequence, we state a linguistically substantiated hypothesis, define an indicator expressing the overall behaviour of the text, and propose a test for comparison of levels, units, properties, etc. Since our data/texts are usually very long, the tests can be performed asymptotically.

Our procedures are, nevertheless, merely experimental and inductive. We choose a property and scrutinize its sequential behaviour in order to obtain a first image. A theoretical substantiation can be added only after many languages and many texts will be analyzed.

2. Arc length

Here we shall define the extent of oscillation of values using the simple arc length between neighbouring entities defined as

$$L = \sum_{i=1}^{n-1} [(x_i - x_{i+1})^2 + 1]^{1/2} \qquad (2.1)$$

which is the usual sum of Euclidean distances between the subsequent values. It has been used extensively in text analyses (cf. Popescu, Mačutek, Altmann 2009). It should not be used in cases where the x_i values vary in the interval [0;1] because the step itself would make up a much greater part of the distance than the values themselves. In those cases one can use e.g. the Manhattan metrics or some other variant of the Minkowski distance. In the next chapters we present some other methods.

Since this indicator depends on the text length, it has been relativized in different ways for different purposes, mostly in connection with word frequency data.

For the sake of easy manipulability one can consider the *mean arc length* defined as

$$\bar{L} = \frac{L}{n-1} \qquad (2.2)$$

where n is the number of units in the sequence (n-1 is the number of individual arcs). Since the variance can be computed empirically, there is no problem with text comparisons, confidence intervals, etc. In order to illustrate the problem we consider the sequence of syllabic lengths of the verses in the poem *Der Erlkönig* (E) by Goethe and obtain:

$$E_1 = (8,7,8,8,9,6,6,6,7,7,7,6,8,5,6,6,7,6,6,8,9,5,8,7,8,9,9,6,6,7,7,7). \quad (2.3)$$

The elements represent the numbers of syllables in the individual verses, n = 32. The values do not vary strongly, because in the poem verse length is rather stereotype, predetermined, nevertheless, there is some oscillation within the interval [5,9]. Using the definition (2.1), we obtain the arc length

$$L = [(8-7)^2 + 1]^{1/2} + [(7-8)^2 + 1]^{1/2} + [(8-8)^2 + 1]^{1/2} + \ldots + [(7-7)^2 + 1]^{1/2} + + [(7-7)^2 + 1]^{1/2} = 50.6291$$

hence the mean (according to 2.2) is

$$\bar{L} = 50.6291/31 = 1.6332.$$

The variance of L_i from their mean is $\frac{1}{31}\sum_{i=1}^{31}(L_i - \bar{L})^2 = 21.3126$. According to the Central Limit Theorem we now obtain the variance of $\bar{L}$ as Var($\bar{L}$) = $Var(L)/n$ = 21.3126/31 = 0.6875, i.e. the variance of the mean is equal to the variance of individual arc lengths divided by n..

If all elements of the sequence are equal, we obtain $\bar{L}$ = 1. Here we see that only a small part of the sequential differences add something to the mean arc. Hence, the sequence is rather smooth.

Consider now the indicator

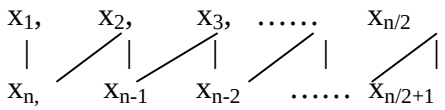
$$P = \frac{L - L_{min}}{1 + L_{mx} - L_{min}} \quad (2.4)$$

Where L_{max} is the maximal arc obtained under the given empirical conditions and L_{min} the minimal one. The minimal arc can be computed if we reorder the sequence in increasing (or decreasing) order. For our example it would be

$$E_2 = (5,5,6,6,6,6,6,6,6,6,6,6,7,7,7,7,7,7,7,7,7,8,8,8,8,8,8,8,8,9,9,9,9) \quad (2.5)$$

yielding $L_{min} = 27(1) + 4\sqrt{2} = 32.6569$.

In order to compute L_{max} one can proceed as follows: Order the numbers in non-decreasing order; if there are n elements (n being even), place the elements in the first row from x_1 to $x_{n/2}$, and the elements from the n^{th} to $n/2+1^{st}$, i.e. in reversed order in the second row. One obtains



(2.6)

Compute the arcs between the connected elements and add them to obtain a rough estimation of L_{max}. If the number of elements is odd, let the first row contain one element more, i.e. the last element in the first row is $x_{(n-1)/2+1}$ and the first element in the second row is again x_n. One computes also the distance between $x_{(n-1)/2+1}$ (which is the last element in the first row) and $x_{(n-1)/2}$ which is the last element in the second row.

This computation can easily be programmed and performed among others with Excel, for any text length. It is not necessary to check the individual permutations.

In our above example we obtain

5,5,6,6,6,6,6,6,6,6,6,6,7,7,7,7,
9,9,9,9,8,8,8,8,8,8,8,7,7,7,7,7,

Thus the maximum arc will be

$[(5-9)^2 + 1]^{1/2} + [(9-5)^2 + 1]^{1/2} + [(5-9)^2 + 1]^{1/2} + [(9-6)^2 + 1]^{1/2} + [(6-9)^2 + 1]^{1/2} + [(9-6)^2 + 1]^{1/2} + [(6-9)^2 + 1]^{1/2} + [(9-6)^2 + 1]^{1/2} + [(6-8)^2 + 1]^{1/2} + [(8-6)^2 + 1]^{1/2} + [(6-8)^2 + 1]^{1/2} + [(8-6)^2 + 1]^{1/2} + [(6-8)^2 + 1]^{1/2} + [(8-6)^2 + 1]^{1/2} + [(6-8)^2 + 1]^{1/2} + [(8-6)^2 + 1]^{1/2} + [(6-8)^2 + 1]^{1/2} + [(8-6)^2 + 1]^{1/2} + [(6-8)^2 + 1]^{1/2} + [(8-6)^2 + 1]^{1/2} + [(6-7)^2 + 1]^{1/2} + [(7-7)^2 + 1]^{1/2} + [(7-7)^2 + 1]^{1/2} + [(7-7)^2 + 1]^{1/2} + [(7-7)^2 + 1]^{1/2} + [(7-7)^2 + 1]^{1/2} + [(7-7)^2 + 1]^{1/2} + [(7-7)^2 + 1]^{1/2} + [(7-7)^2 + 1]^{1/2} = 3(4.1231) + 5(3.1623) + 14(2.2361) + 1.4142 + 8 = 68.8999.$

The above calculations can be summarized by the following formula:

$$L_{\max} = \begin{cases} \sum_{i=1}^{n/2} \sqrt{(x_i - x_{n+1-i})^2 + 1} + \sum_{i=1}^{n/2-1} \sqrt{(x_{i+1} - x_{n+1-i})^2 + 1} & \textit{for even } n \\ \sum_{i=1}^{(n-1)/2} \sqrt{(x_i - x_{n+1-i})^2 + 1} + \sum_{i=1}^{(n-1)/2} \sqrt{(x_{i+1} - x_{n+1-i})^2 + 1} & \textit{for odd } n \end{cases} \quad (2.7)$$

If e.g. n is even, the elements in the first and second sum correspond to the vertical and diagonal line segments in diagram (2.6), respectively.

The indicator P in (1.3) becomes P = (50.6291 – 32.6569)/(1 + 68.8999 – 32.6569) = 0.4826, hence we would consider the sequence as rather smooth.

Consider the behaviour of the indicator P. If all values of the sequence are equal, e.g. 2,2,2,2,2,…, then all values (L, L_{max} and L_{min}) are equal and we obtain 0/1 = 0. This sequence is extremely smooth. Here one sees why 1 has been inserted in the denominator: without 1 we would obtain 0/0 which is no definite value.

Now, take a maximally rough sequence in which there are only minimal and maximal values in regular succession, e.g. 1,10,1,10,1,10,…. In that case $L = L_{max}$ and $L_{max} - L_{min}$ is some increasing function of n, say = $k(n)$. Hence $P = k/(1+k)$. Taking the limit for $n \rightarrow \infty$ we obtain P = 1. Hence P is always between 0 and 1.

In the present article we shall consider sequences of word length, sentence length, and frequency, in different extent. The words will be replaced by their topical property and the smoothness/roughness of the sequence will be computed. Each computation will be accompanied by a hypothesis. Below, we add some further methods capturing the behaviour of the sequence.

2.1. Word length

Word length is the most frequent object of quantitative investigations because usually the data are readily available and one does not need determine the boundaries of syllables whose number represents the word length. Nevertheless, even here problems may arise e.g. with diphthongs, triphthongs. In some languages one counts also non-syllabic words or one considers them as clitics of the preceding or next word (cf. in Slavic languages the non-syllabic prepositions *s*, *z*, *v*, in Hungarian the conjunction *s* being the elliptic form of *és*, etc.). In strongly analytic languages the oscillation may be relatively small because words are not prolonged by affixation; as compared to this, in strongly synthetic languages, whether inflectional or agglutinative, words can attain a greater length and the irregular oscillation may be stronger. Hence we can preliminarily conjecture that the greater the word length roughness in text, the stronger is the synthetism of language.

Of course, the conjecture can be formulated in reverse order because it is not roughness which is the cause of synthetism but just conversely. We do not speak about causes in language but rather about links between properties, as is usual in dynamic systems. If we would consider synthetism as the independent variable, we would be forced to measure it in some way. Hence the above statement is merely a conjecture that can be tested.

In order to normalize *P*, and at the same time to perform the test for the significance of the smootheness/rougness of the data, we need its expectation and variance. We assume preliminarily that the expectation is 0.5, the mid of the range of *P*, because we have to do with very different entities. The variance is given as follows: for the given text L_{max} and L_{min} are some fixed values, hence

$$Var(P) = \frac{Var(L)}{(1 + L_{max} - L_{min})^2} \tag{2.8}$$

and the normalization yields

$$u = \frac{P - 0.5}{\sqrt{Var(P)}} \tag{2.9}$$

For illustration we compute (2.8) for the above mentioned verse length data and obtain

$P = 0.4826$,
$Var(L) = 21.3125$,
$Var(P) = 21.3125/(1 + 68.8999 - 32.6569)^2 = 0.01537$,
hence

$u = (0.4826 - 0.5)/\sqrt{0.01537} = -0.1404.$

We state that the text is smoother than expected ($u < 0$) but not significantly.

The respective values for 60 texts in 28 languages are presented in Table 2.2.

As can be seen, all word length sequences have a rather great roughness (viz. significantly positive u). One could order the texts according to u but there are great differences even within individual languages, referring rather to text size, text sort or stylistic differences, not to language type. Hence taking averages of P seems to be more adequate. However, at present, we cannot take averages in all languages because the number of texts in some of them is merely 1. Nevertheless, a preliminary order can be stated. Using Table 2.2 and taking the averages we obtain (in parenthesis is the number of analysed texts) the order presented in Table 2.1.

As can be seen, the conjecture concerning word length smoothness and synthetism can preliminarily be accepted: the smaller P, the stronger is the synthetism. However, further collecting of data is necessary. The result is very preliminary and shows merely the method. Nevertheless, one question remains open: why in some languages the regularity is greater than in other ones, i.e. what is the background mechanism cotrolling the *sequences* of word length?

Evidently, the languages are only partially ordered according to their degree of analytism/synthetism.

But even if the P-values seem to be quite similar, there may be significant differences between texts or languages. In order to state them, we simply perform the t-test for the difference of two means omitting languages in which only one text has been processed. The t-test with $n_1 + n_2 - 2$ degrees of freedom can be computed according to the formula

$$t = \frac{\bar{P}_1 - \bar{P}_2}{\hat{\sigma}_{\bar{P}_1 - \bar{P}_2}} \tag{2.10}$$

where

$$\hat{\sigma}_{\bar{P}_1 - \bar{P}_2} = \sqrt{\frac{\sum_{i=1}^{n_1}(P_{i1} - \bar{P}_1)^2 + \sum_{i=1}^{n_2}(P_{i2} - \bar{P}_2)^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{2.11}$$

and n_1, n_2 are the numbers of cases in the given language. The probability that t is greater or equal to the observed value can be found in the respective tables.

For the sake of illustration we perform the test for Slovak with $\bar{P}_{Sl}$ = 0.6135, sum of squared deviations = 0.000316, n = 2 and Sundanese with $\bar{P}_{Su}$ = 0.7097, sum of squared deviation = 0.000098, n = 2, yielding

$$t = \frac{|0.6135 - 0.7097|}{\sqrt{\frac{0.000316 + 0.000098}{2+2-2}\left(\frac{1}{2}+\frac{1}{2}\right)}} = 6.69$$

which is significant with 2 degrees of freedom (Sundanese has greater roughness than Slovak, the critical value is $t_{0.05,2} = 4.30$). The results will become more clear when the number of texts increases; in this form it is not quite true that Sundanese is more synthetic than Slovak.

Table 2.1
Oscillation of the word length arc in languages
(ordered by ascending mean of P)

Language	**mean P ascending**
Russian(1)	0,5907
Tamil(1)	0,5997
Slovenian(1)	0,6111
Slovak(2)	0,6135
Czech(5)	0,6137
Telugu(2)	0,6193
Malayalam(2)	0,6201
Latin(2)	0,6236
Indonesian(2)	0,6305
Serbian(2)	0,6444
Odia(2)	0,6475
Maninka(3)	0,6534
Hindi(2)	0,6582
German(5)	0,6675
Bulgarian(1)	0,6720
Hungarian(2)	0,6768
Welsh(2)	0,6777
Vai(3)	0,6799
Romanian(3)	0,6819
French(1)	0,6823
Macedonian(1)	0,7062
Italian(1)	0,7083
Sundanese(2)	0,7097
Bamana(4)	0,7247
Japanese(1)	0,7384
Kikongo(3)	0,7522
Akan(2)	0,7556
Tagalog(2)	0,7651

Table 2.2
Roughness in 60 texts from 28 languages

Language: Text	n	L	L_{min}	L_{max}	P	Var(L)	Var(P)	u
Akan: Agya Yaw Ne Akutu Kwaa	201	290,47	201,24	323,55	0,7236	0,2957	1,94E-05	50,71
Akan: Mma Nnsua Ade B□ne	143	218,62	143,66	237,84	0,7876	0,4536	5,01E-05	40,64
Bamana: Bamak□ sigicoya	1138	1739,97	1139,07	1966,32	0,7255	0,4910	7,16E-07	266,55
Bamana: Masadennin	2616	4057,15	2617,9	4559,65	0,7408	0,7347	1,95E-07	545,85
Bamana: Namak□r□ba halakilen	1406	1890,23	1406,66	2118,59	0,6783	0,2662	5,24E-07	246,35
Bamana: Sonsannin ani	2392	3615,73	2393,49	4013,4	0,7540	0,6195	2,36E-07	523,18
Bulgarian: Ostrovskij, Kak se kaljavaše stomanata (Chap. 1)	926	1644,85	927,07	1994,13	0,6720	0,5839	5,12E-07	240,47
Czech: Čulík, O čem jsou dnešní Spojené státy?	2003	3633,34	2006,14	4597,83	0,6276	0,8257	1,23E-07	364,10
Czech: Hvížďala, O předem zpackané prezidentské volbě	929	1615,07	930,49	2058,75	0,6062	0,6209	4,87E-07	252,23
Czech: Macháček, Slovenský dobrý příklad	340	599,02	341,07	751,56	0,6269	0,6063	3,58E-06	67,05
Czech: Spurný, Prekvapení v justici	288	499,95	289,07	632,99	0,6114	0,5822	4,89E-06	50,35
Czech: Švehla, Editorial, Voličův kalkul	288	482,14	291,23	610,38	0,5963	0,6727	6,56E-06	37,60
French: Dunkerque – La route des dunes (press)	1532	2558,39	1533,9	3034,36	0,6823	0,702	3,11E-07	326,74
German: Assads Familiendiktatur	1415	2587,27	1417,73	3228,75	0,6454	1,1336	3,45E-07	247,51
German: ATT0012 (press)	1148	2157,49	1150,31	2660,01	0,6667	1,2753	5,59E-07	223,00
German: Die Stadt des Schweigens	1567	2871,06	1569,73	3502,33	0,6730	1,1681	3,12E-07	309,52
German: Terror in Ost Timor	1398	2475,4	1400,31	2972,87	0,6832	0,9547	3,86E-07	295,07
German: Unter Hackern und Nobelpreisen	1363	2558,37	1365,31	3147,71	0,6690	1,2029	3,78E-07	274,77
Hindi: After the sanctionto love marriage (press)	1103	1648,69	1103,66	1895,7	0,6873	0,3134	4,98E-07	265,29
Hindi: The Anna Team on a cross-road (press)	860	1212,27	860,66	1418,53	0,6291	0,2214	7,09E-07	153,39
Hungarian: A nominalizmus Forradalma (press)	1314	2841,26	1316,31	3679,39	0,6451	1,3821	2,47E-07	291,68
Hungarian: Kunczekolbász (press)	458	1016,71	460,31	1244,57	0,7086	1,5538	2,52E-06	131,38
Indonesian: Pengurus PSM terbelah (press)	345	537,8	346,07	656,5	0,6156	0,3785	3,90E-06	58,54
Indonesian: Sekolah ditutup (press)	280	456,16	281,07	551,39	0,6453	0,463	6,29E-06	57,95

Italian: Il bosone di Higgs scoperto dal Cern (Internet)	2516	4974,2	2518,31	5984,39	0,7083	1,0795	8,98E-08	695,24
Japanese: Miki, Jinseiron Note	1805	3483,94	1809,06	4076,37	0,7384	1,1911	2,31E-07	495,45
Kikongo: Bimpa: Ma Ngo ya Ma Nsiese	823	1397,12	829,15	1621,03	0,7163	0,6927	1,10E-06	206,10
Kikongo: Lumumba speech	956	1557,76	957,07	1759,73	0,7474	0,4283	6,63E-07	303,88
Kikongo: Nkongo ye Kisi Kongo	768	1139,63	769,07	1235,41	0,7929	0,3269	1,50E-06	239,42
Latin: Cicero, In Catilinam I	1116	1948,25	1117,49	2470,53	0,6135	0,5908	3,22E-07	200,02
Latin: Cicero, In Catilinam II	3095	5632,4	3096,9	7097,23	0,6337	0,6637	4,15E-08	656,50
Macedonian: Ostrovskij, Kako se kaleše čelkiot (Chap. 1)	1123	2251,33	1124,07	2719,36	0,7062	0,8963	3,52E-07	347,63
Malayalam: Moralistic hooligans (press)	282	594,31	284,31	790,19	0,6116	1,3026	5,07E-06	49,56
Malayalam: No one should die (press)	288	668,42	290,31	890,86	0,6286	1,8581	5,13E-06	56,73
Maninka: Nko Doumbu Kende no.2 (press)	2076	3132,96	2077,07	3877,8	0,5860	0,4379	1,35E-07	234,27
Maninka: Nko Doumbu Kende no.7 (press)	1535	2394,57	1536,49	2814,04	0,6711	0,4746	2,90E-07	317,61
Maninka: Sìikán` (Constitution of Guinea, an excerpt)	1662	2950,33	1663,07	3492,93	0,7031	0,7119	2,12E-07	440,69
Odia: The Samaj, Bhuba-neshwar (28 June 2012), p. 4	348	549,27	349,49	662,51	0,6362	0,5059	5,13E-06	60,13
Odia: The Dharitri, Balasore (12th Feb, 2012), p. 10	630	1084,28	631,49	1317,72	0,6589	0,5642	1,19E-06	145,35
Romanian: Paler, excerpt from Aventuri solitare	891	1681,02	892,49	2008,51	0,7059	0,7763	6,22E-07	261,07
Romanian: Steinhardt, Jurnalul fericirii, Trei soluții	1511	2718,28	1512,49	3361,98	0,6516	0,8189	2,39E-07	310,02
Romanian: Popescu D.R., Vânătoarea regală	1006	1664,39	1007,07	1961,26	0,6882	0,5061	5,55E-07	252,63
Russian: Ostrovskij, Kak zakaljalas stal' (Chap. 1)	792	1319,67	793,49	1683,19	0,5907	0,5251	6,62E-07	111,55
Serbian: Ostrovskij, Kako se kalio čelik (Chap. 1,	994	1675,53	994,66	2050,33	0,6444	0,5376	4,81E-07	208,04
Slovak: Bachletová, Moja Dolná zem	872	1435,26	873,07	1770,07	0,6260	0,4523	5,61E-07	168,30
Slovak: Bachletová, Riadok v tlačive	924	1655,59	925,49	2139,51	0,6009	0,696	4,71E-07	146,94
Slovenian: Ostrovskij, Kako se je kalilo jeklo (Chap. 1)	977	1556,7	978,07	1923,95	0,6111	0,4304	4,80E-07	160,34
Sundanese: Agustusan (Online)	416	664,27	417,07	761,00	0,7167	0,438	3,68E-06	112,92
Sundanese: Aki Satimi (Online)	1283	2011,23	1283,66	2318,1	0,7027	0,3508	3,27E-07	354,31
Tagalog: Rosales, Kristal Na Tubig	1958	3794,27	1959,9	4309,66	0,7803	0,8705	1,58E-07	706,31
Tagalog: Hernandez, Limang Alas: Tatlong Santo	1738	3238,09	1739,9	3740,27	0,7486	0,7971	1,99E-07	557,24
Tagalog: Hernandez, Magpisan	1466	2838,63	1467,9	3255,29	0,7665	0,8667	2,71E-07	511,87
Tamil: Emu Bird Trading (press)	384	771,84	386,31	1028,18	0,5997	1,1346	2,75E-06	60,17

Telugu: Trailangaswamy (press)	295	616,92	297,31	810,27	0,6219	1,2091	4,58E-06	56,96
Telugu: Train Journey (press)	666	1299,12	668,73	1689,71	0,6168	1,1084	1,06E-06	113,41
Vai: Sa'bu Mu'a'	495	631,4	495,24	716,62	0,6123	0,1275	2,58E-06	69,93
Vai: Sherman, Mu ja vaa	3140	4079,9	3140,66	4579,51	0,6523	0,1571	7,58E-08	553,32
Vai: Vande	426	571,29	426,24	612,39	0,7750	0,1648	4,71E-06	126,80
Welsh: text 1 (gaenv)	985	1750,5	986,49	2094,86	0,6887	0,5934	4,82E-07	271,74
Welsh: text 2 (gasodl)	1002	1441,3	1002,66	1659,57	0,6667	0,2445	5,65E-07	221,82

The basic data were taken from Popescu et al. (2013) with the kind permission of E. Kelih, A. Rovenchak, A. Overbeck, H. Sanada, R. Smith, R. Čech, P. Mohanty and A. Wilson.

The results of pair-wise testing are presented in Table 2.3. As can be seen, there is a number of significant differences (grey). Though the number of used texts is very small, we can conjecture that not only texts may differ but also languages. A thorough investigation should be performed in such a way that first several texts of the same sort in one language must be analyzed, then the mean *P* could be used for comparison with another text sort in the same language, etc. This is a practically endless problem.

In our data (cf. Table 2.3), Tagalog seems to be an outlier among the other Indonesian languages (Malay, Sundanese), the group of Indo-European languages disintegrates, etc. But all these statements are merely conjectures that must be thoroughly tested. A classification of the results in Table 2.3 could be performed with the aid of many clustering methods but it is not our aim here.

Table 2.3
Testing the difference of mean $P(u)$ using the t-test

Language	Slo	Cze	Tel	Mal	Lat	Ind	Odi	Man	Hin	Ger	Hun	Wel	Vai	Rom	Sun	Bam	Kik	Aka
Czech 5	0,02																	
Telugu 2	0,46	0,55																
Malayalam 2	0,44	0,58	0,08															
Latin 2	0,63	0,87	0,41	0,27														
Indonesian 2	0,87	1,31	0,74	0,61	0,38													
Odia 2	2,01	2,88	2,43	1,94	1,58	0,91												
Maninka 3	0,87	1,49	0,76	0,73	0,65	0,49	0,13											
Hindi 2	1,41	2,42	1,33	1,26	1,13	0,85	0,34	0,10										
German 5	4,38	6,21	4,60	4,19	3,76	2,84	1,66	0,52	0,50									
Hungarian 2	1,85	3,22	1,80	1,73	1,60	1,32	0,87	0,46	0,43	0,47								
Welsh 2	3,85	5,49	5,18	4,15	3,64	2,56	1,91	0,53	0,63	0,86	0,03							
Vai 3	1,04	1,81	0,96	0,94	0,88	0,77	0,51	0,44	0,32	0,34	0,05	0,03						
Romanian 3	3,02	4,81	3,02	2,86	2,65	2,19	1,54	0,74	0,79	1,01	0,16	0,19	0,04					
Sundanese 2	6,69	8,90	12,12	8,15	7,03	4,82	4,66	1,24	1,72	3,83	1,01	2,46	0,47	1,30				
Bamana 4	4,28	6,91	4,24	4,13	3,96	3,57	3,00	2,03	2,18	3,55	1,52	1,83	0,99	1,81	0,60			
Kikongo 3	4,60	7,64	4,62	4,50	4,34	3,96	3,50	2,39	2,61	4,65	2,03	2,50	1,35	2,57	1,46	1,02		
Akan 2	4,14	7,20	4,25	4,10	3,94	3,55	3,19	2,00	2,25	4,44	1,75	2,30	1,12	2,34	1,40	0,98	0,09	
Tagalog 2	8,99	12,90	12,64	10,30	9,38	7,23	7,36	2,44	3,43	8,17	2,62	5,56	1,34	3,74	4,19	1,57	0,43	0,28

2.2. Sentence length (style)

Sentence length measured in terms of word numbers does not depend on language but on the communicative aim, on the style of the author, on text sort, on the age of the author, on the spontaneity of writing, etc. One can suppose that in text books for children the sentences are rather short; the same holds for poetry but not for prose where a sentence can consist of several hundreds of words. If no form restricts the writer - e.g. as in poetry, stage play, text-book, science, law, etc. - and (s)he writes spontaneously, the sentences can get longer. Punctuation did not exist from the beginning of writing, since it was introduced at a later time. Speech does not contain punctuation, sometimes one speaks a long time without any pause. This does not hold for stage plays where even monologues contain punctuation.

Thus we can set up several hypotheses concerning the roughness of sentence length and test them.

(1) The greater is the sentence length roughness, the more spontaneously the text has been written.

(2) A preliminary hypothesis concerning individual text sorts may be set up as follows: *text books for children* < *journalistic texts* < *poetry* < *stage play* < *law texts* < *scientific texts* < *prose.*

In the above ordering of text sorts, some of them are rather persistent than volatile, but the turning point is not yet known. Its finding is a matter of extensive testing.

Table 2.4 contains some computation of the indicator *P* in 15 texts of 9 languages. The order of languages does not correspond to the degree of their synthetism, hence *P* is a property of the given text. Texts in which *P* is maller than 0.5 have a certain „sentence rhythm".

A slightly more complex computation of sentence lengths may be performed in terms of clause numbers. However, the stating of the number of clauses cannot be made based on a general rule – which does not exist -, it is a problem of definition which may differ from language to language (and from linguist to linguist). Since we analyze only German data, we define clause as a construction containing a finite form of a verb, also auxiliary and modal. Hence the clause length of a sentence is simply the number of finite verbs in it. We used 20 German newspaper articles published in January 1999 in *Tageszeitung*, namely:

T1 Taz 22.01.1999: Kulturstadt Weimar: Winter mit fröhlicher Sonne.
T2 Taz 21.01.1999: Wer macht das Spiel?
T3 Taz 20.01.1999: Die Nerven behalten
T4 Taz 16.01.1999: Auf Elefanten Richtung Rhino!
T5 Taz 16.01.1999: Der Gefangene von Gaghan
T6 Taz 16.01.1999: Zeitschriften sind Originale
T7 Taz 09.01.1999: Die Friedhöfe an der Drina
T8 Taz 15.01.1999: Zwischen Finanzkrise und Handelskrieg
T9 Taz 15.01.1999: Das belgische Modell einer präventiven Ausländerfeindlichkeit
T10 Taz 15.01.1999: Zweierlei Recht im chilenischen Rechtsstaat
T11 Taz 15.01.1999: Frankreichs Front National entdeckt Osteuropa
T12 Taz 15.01.1999: Für einen neuen Stabilitätspakt
T13 Taz 15.01.1999: Kontrolle der Kapitalströme emerging markets - ein Gebot der Demokratie

Table 2.4
The arc of sentence length in 15 texts (in terms of word numbers)

Language	**Text**	**n**	**L**	**L_{max}**	**L_{min}**	**P**	**Var(L)**	**Var(P)**	**u**
Hungarian	A nominalizmus forradalma, press	63	703,42	982,87	96,73	0,6839	68,7281	8,73E-05	19.68
Indonesian	Pengurus, press	28	126,27	235,78	41,5	0,4341	13,481	3,54E-04	-3.51
Latin	Cicero, In Catilinam I	80	720,22	1225,11	101,36	0,5502	52,5753	4,16E-05	7.79
Romanian	Octavian Paler, Aventuri solitare (excerpt)	17	185,32	397,02	46,94	0,3942	89,365	7,25E-04	-3.93
Romanian	D.R. Popescu, Vânătoarea regală, Chapter 2	61	973,26	1367,43	157,99	0,6735	366,0799	2,50E-04	10.98
Romanian	N. Steinhardt, Jurnalul fericirii, Trei soluții	85	1260,5	1810,47	189,04	0,6604	446,615	1,70E-04	12.31
Russian	N. Ostrovskij, How the steel was tempered	76	605,21	840,72	108,33	0,6775	70,2603	1,31E-04	15.53
Slovak	Bachletová, Moja Dolna zem	92	617,45	892,5	118,38	0,6439	47,0381	7,83E-05	16.26
Slovak	Bachletová, Riadok v tlačive	78	641,52	836,31	92,12	0,7373	41,5035	7,47E-05	27.44
Slovenian	N. Ostrovskij, How the steel was tempered	84	754,91	1047,81	117,84	0,6843	79,5594	9,18E-05	19.24
Sundanese	Aki Satimi, press	147	673,99	1014,63	157,27	0,6020	10,5146	1,43E-05	27.00
Sundanese	Agustusan, press	53	209,97	280,28	57,38	0,6815	5,749	1,15E-04	16.95
Tagalog	Hernandez, Limang Alas, Tatlong Santo	104	894,34	1411,56	124,39	0,5977	44,6326	2,69E-05	18.84
Tagalog	Hernandez, Magpisan	111	965,66	1397,85	132,3	0,6580	48,1819	3,00E-05	28.83
Tagalog	Rosales, Kristal Na Tubig	139	1042,84	1717,3	171,91	0,5632	42,8283	1,79E-05	14.93

T14 Taz 13.01.1999: Für uns Serben wird hier kein Platz sein
T15 Taz 11.01.1999: Von Frust und Lust im samtenen Sweat-shop
T16 Taz 15.01.1999: Ethnische Definitionen als Machtpolitik
T17 Taz 15.01.1999: Bündnisse und Rivalitäten im Mittleren Afrika
T18 Taz 14.01.1999: Klick, klick, klick.
T19 Taz 11.01.1999: Wo Es war, soll Wir werden
T20 Taz 23.01.1999: Die Schlagerfamilie Ost.

For the German clause-variant (Table 2.5) we obtain a mean $P = 0.6086$. For the word-variant (Table 2.4) in a mixture of several languages it holds $P = 0.6161$. If these results turn out to be relatively constant, then we have found a very peculiar stability which may hold in the entire hierarchy (sentence – clause – word – syllable – sound duration). In order to state it, Sisyphean work is necessary. For the time being, we can state that this value tends to the value of the golden section minus 1 (0.6180)

Table 2.5
Arc of sentence lengths (in terms of clauses) in 20 German texts

Text	**n**	**L**	**L_{min}**	**L_{max}**	**P**	**Var(L)**	**Var(P)**	**u**
T1	148	226,9328	149,8929	294,5745	0,5288	0,5525	2,60E-05	5,65
T2	80	129,4372	81,4853	156,7301	0,6289	0.6184	1,06E-04	12,54
T3	112	188,1304	113,0711	220,2935	0,6936	0,6104	5,21E-05	26,81
T4	208	332,1408	209,4853	397,1550	0,6501	0,6508	1,83E-05	35,11
T5	246	361,5297	247,4853	437,9837	0,5955	0,4283	1,17E-05	27,95
T6	109	191,3951	110,8995	223,4260	0,7090	0,9889	7,67E-05	23,86
T7	107	176,4568	108,4853	223,7354	0,5847	0,7643	5,66E-05	11,26
T8	85	150,2687	86,0711	180,0518	0,6759	0,7251	8,04E-05	19,62
T9	97	148,8435	98,0711	188,5416	0,5551	0,4445	5,31E-05	7,56
T10	112	181,3166	113,0711	219,1969	0,6371	0,5075	4,42E-05	20,61
T11	95	137,2944	95,6569	176,1045	0,5112	0,2739	4,13E-05	1,75
T12	74	106,0632	74,6569	125,0160	0,6115	0,3458	1,31E-04	9,74
T13	120	170,7611	120,6569	213,9045	0,5316	0,2878	3,24E-05	5,56
T4	139	191,7779	140,0711	216,7953	0,6653	0,2736	4,53E-05	24,56
T15	105	165,7787	106,0711	201,4551	0,6195	0,4344	4,68E-05	17,47
T16	119	186,0022	120,8929	243,8176	0,5254	0,4684	3,05E-05	4,60
T17	110	163,2036	111,0711	204,6998	0,5509	0,4133	4,62E-05	7,49
T18	197	274,6528	197,6569	322,4205	0,6122	0,3032	1,92E-05	25,63
T19	79	131,3901	80,0711	163,9010	0,6050	0,5153	7,16E-05	12,40
T20	151	211,0678	152,0711	237,7794	0,6804	0,3759	5,00E-05	25,52

2.3. Frequency

The computation of frequency is simple, there is software available to perform this procedure. Having the frequencies of individual units (word forms or lemmas), we replace the respective words in the text by their frequency and compute the properties of the constructed sequence. We may start from two considerations: If there are many forms in language, then the number of seldom word-forms will be greater, many of them are placed in immediate neighbourhood, hence the arc will be smaller. This will be the case also with short texts. At the same time there are some few words (usually synsemantics) which occur quite frequently, hence the maximum arc will be longer and the minimum arc shorter. Hence P will be smaller.

Consider under these presuppositions 20 Slovak texts. Here we obtain the results as given in Table 2.6. The texts have different sizes which do not influence the value of P. The shortest text (n = 93) has P = 0.9319 while the longest text (n = 3704) has P = 0.8749. The texts were taken from http://quanta-textdata.uni-graz.at in 5 different text sorts as given in the following list.

Author	**Text sort**	**Text**	
1.anonymous	Agency (TASR)	Banskobystrický samosprávny kraj sa vzdáva zdravotníckych zariadení	03.11.2004
2.anonymous	Agency (TASR)	Bratislavskí taxikári zvyšujú poplatok za nástupenie o 100 %	09.11.2004
3.D. Dušek	Short story	1. máj 1977. In: *Kufor na sny*	1993
4.D. Dušek	Short story	Alfabet D.D.: *Kufor na sny*	1993
5.D. Hevier	Fairy tale	Kotrmelec a kotrmelec. In: *Futbal s papučou*	1989
6.D. Hevier	Fairy tale	Lyžicová Naháňačka. In: *Futbal s papučou*	1989
7.anonymous	Agency (TASR)	Čoskoro rozhodnú o investorovi pre SND	03.11.2004
8.Eugen Č.	Agency (TASR)	Ďakujem všetkým, ktorí nám dôverovali, verím v spravodlivosť	03.11.2004
9.B. Hochel	Short story	Kúpalisko	1997
10.B. Hochel	Short story	Muž, ktorému veľmi ukrivdili	1997
11.D. Hevier	Fairy tale	Krajina agord, Kap. 1, A	2001
12.D. Hevier	Fairy tale	Krajina agord, Kap. 1, B	2001
13.P. Holka	Short story	Kap. 1. *Normálny cvok*	1993
14.P. Holka	Short story	Kap. 2	1993
15.V. Šikula	Novel	Kap. 1. *Veterna ružica*	1995
16.V. Šikula	Novel	Kap. 2. *Veterna ružica*	1995
17.R. Sloboda	Novel	Kap. 1. *Pamäti*	1996
18.R. Sloboda	Novel	Kap. 2. *Pamäti*	1996
19.P. Vilikovský	Short story	Kap. 1. *Peší príbeh*	1992
20.P. Vilikovský	Short story	Kap. 2. *Peší príbeh*	1992

For all texts the value P deviates significantly from 0.5, and the value of u increases with the text size. The first three shortest texts are exceptions.

The mean values of P for the same text sorts are:

News = 0.7272 < *Fairy tale* = 0.7704 < *Short story* = 0.7939 < *Novel* = 0.82

As can be seen there is a clear hierarchy which can be left to literary scientists for interpretation. One could perform tests for the difference between text sorts but we are preliminarily content with this hierarchy.

The mean of all (n = 20) Slovak values is $\overline{P}_{Slk}$ = 0.7811, and the variance of the mean is $Var(\overline{P}_{Slk})$ = 0.0007125. The mean P is an indicator of written Slovak and since its empirical variance is known, it can be used for inter-language comparisons (s. below). The value of u increases with increasing n but there are some outliers which must be studied separately.

Table 2.6
The *P*-indicator in Slovak frequency data
(ordered according to increasing n)
(FT = fairy tale, NW = news, SS = short story, NO = novel)

Text sort	**n**	**L**	**L_{max}**	**L_{min}**	**P**	**Var(L)**	**Var(P)**	**u**
6. FT	95	145,74	160,28	95,24	0,7647	0,6134	1,41E-04	22,32
5. FT	123	288,48	386,64	125,82	0,6213	2,1312	3,11E-05	21,75
1. NW	229	456,51	670,11	230,49	0,5130	2,1451	1,10E-05	3,90
12. FT	267	983,65	1062,06	277,54	0,8989	21,7266	3,52E-05	67,23
11. FT	283	853,57	1033,72	286,14	0,7580	6,9087	1,23E-05	73,48
2. NW	293	686,41	812,69	294,9	0,7547	3,2351	1,20E-05	73,45
8. NW	351	991,92	1134,39	355,37	0,8161	6,8133	1,12E-05	94,45
19. SS	409	1443,77	1742,34	415,67	0,7744	11,2360	6,37E-06	108,67
20. SS	428	1731,6	2077,72	436,22	0,7887	15,8988	5,89E-06	118,91
7. NW	437	1329,85	1523,52	440,55	0,8204	8,4340	7,18E-06	119,59
3. SS	793	6877,33	7967,42	823,9	0,8473	142,7026	2,80E-06	207,70
13. SS	821	6405,93	8461,62	834,34	0,7304	60,5033	1,04E-06	225,94
14. SS	980	12081,5	14474,67	1014,47	0,8221	215,4821	1,19E-06	295,41
4. SS	1044	7281,84	8532,45	1061,9	0,8325	66,9222	1,20E-06	303,67
9. SS	1132	11036,96	13043,63	1165,91	0,8310	162,2079	1,15E-06	308,70
10. SS	1143	12302,14	16553,95	1183,01	0,7233	196,3127	8,31E-07	245,03
17. NO	1486	21620,58	24365,36	1541,63	0,8797	417,6505	8,02E-07	424,07
15. NO	1605	25912,97	33647,88	1649,83	0,7582	345,4501	3,37E-07	444,61
18. NO	3666	117620,74	134684,07	3795,34	0,8696	1902,8985	1,11E-07	1109,08
16. NO	5364	250592,66	322784,07	5529,3	0,7724	2830,8441	2,81E-08	1624,55

For Russian we used 20 texts as given below (taken from http://quanta-textdata.uni-graz.at) (here SP = Stage play, SS = Short story, NW = News, PL = Private letter, NO = Novel)

For these texts we obtain the results presented in Table 2.7. As can be seen, the value of u increases with increasing n. This is simply due to the fact that normality in

language is rather an exception. Linguistic data do not want to behave "normally", in the statistical sense of the word. Nevertheless, we use the tests as a first information.

	Author	**Text sort**	**Text**	**Year**
1	A.P. Čechov	SP	Svadba	1890
2	A.P. Čechov	SP	Čajka, 1. Act	1896
3	A.P. Čechov	SP	Čajka, 2. Act	1896
4	A.P. Čechov	SP	O vrede tabaka	1902
5	A.P. Čechov	SS	Chameleon	1884
6	A.P. Čechov	SS	Chirurgija	1884
7	L.N. Tolstoj	SS	Chozjain i rabotnik, Ch. 1	1895
8	L.N. Tolstoj	SS	Chozjain i rabotnik, Ch. 2	1895
9	E. Sal'nikova	NW	Lider ottepeli - žertva zastoja	Nezavisimaja gazeta, 01.11.2001
10	E. Sal'nikova	NW	Pereput'e zamyslov	Nezavisimaja gazeta, 12.05.2001
11	Olga Tropkina	NW	Podberezkin stal vtorym kandidatom	Nezavisimaja gazeta, 03.06.2000
12	Olga Tropkina	NW	Gotovjatsja novye zakony o vyborach	Nezavisimaja gazeta, 10.02.2001
13	L.N. Tolstoj	PL	to: A.A.Tolstaja	vom 19.9.1872
14	L.N. Tolstoj	PL	to: S.A.Tolstaja	vom 8.12.1884
15	L.N. Tolstoj	PL	to: S.A.Tolstaja	vom 28.02.1892
16	L.N. Tolstoj	PL	to: Nikolaj II	vom 16.01.1902
17	F.M. Dostoevskij	NO	Prestuplenie i nakazanie, Part I, Ch. 1	1886
18	F.M. Dostoevskij	NO	Prestuplenie i nakazanie, Part I, Ch. 2	1886
19	L.N. Tolstoj	NO	Anna Karenina, Book I, Ch. 1	1877
20	I.A. Gončarov	NO	Oblomov, Part I, Ch. 1	1858

It must be remarked that the newspaper texts were written in the present millennium. They are very short and represent a different text sort. In order to get a more thorough insight probably hundreds of texts must be analyzed and their size and origin must be taken into account. Our results show merely the method.

Considering again the mean P in individual text sorts:

Stage play (1,2,3,4) = 0.8176 < *Short story* (5,6,7,8) = 0.8234 < *Private letter* (13,14,15,16) = 0.8506 < *Novel* (17,18,19,20) = 0.8549 < *Comment/News* (9,10,11,12) = 0.8662

we obtain a first ordering of smoothness of frequencies in text sorts in Russian. The mean of all values is $\overline{P}_{Rus}$ = 0.8425 and the variance of P is $Var(P_{Rus})$ = 0.002224.

Table 2.7
The *P*-indicator in 20 Russian frequency data
(ordered according to ascending n)

Text	n	L	L_{max}	L_{min}	P	Var(L)	Var(P)	u
11	220	449,28	499,47	222,82	0,8156	3,2628	4,23E-05	48,52
09	251	694,16	718,85	255,0552	0,9447	7,681	3,56E-05	74,58
12	377	1295,03	1569,05	384,64	0,7680	13,1244	9,34E-06	87,69
19	702	6767,05	8039,61	726,75	0,8259	126,6025	2,37E-06	211,82
15	778	8590,44	10157,13	808,47	0,8323	164,2835	1,88E-06	242,42
10	785	5296,11	5629,94	807,2161	0,9306	99,4081	4,27E-06	208,32
13	879	12379,97	15273,9	912,3	0,7984	233,7948	1,13E-06	280,33
05	908	7326,17	8718,24	934,5	0,8211	104,086	1,72E-06	244,98
06	949	8623,27	10038,16	980,7	0,8437	155,2042	1,89E-06	249,90
14	1097	17306,11	18895,95	1158,63	0,9103	542,5134	1,72E-06	312,48
08	1453	21457,02	27407,91	1505,3372	0,7702	341,216	5,09E-07	378,95
07	1951	40905,22	47325,23	2040,91	0,8582	956,6879	4,67E-07	524,46
16	2175	47801,08	55137,98	2280,12	0,8612	1271,3319	4,55E-07	535,44
17	2595	76014,32	87041,45	2711,7	0,8692	1654,1425	2,33E-07	765,59
04	2656	54562,01	65429,5	2739,56	0,8266	813,3295	2,07E-07	718,01
03	2657	77408,03	92082,75	2743,56	0,8357	1167,4655	1,46E-07	877,84
01	2905	67799,73	85133,84	2979,7	0,7890	799,6347	1,18E-07	839,61
02	3389	101668,83	123370,02	3491,7	0,8190	1461,5572	1,02E-07	1000,19
20	3576	105276,39	125228,55	3680,21	0,8358	1262,087	8,54E-08	1149,06
18	5604	427782,29	480689,34	5946,08	0,8886	13758,4071	6,10E-08	1572,63

As can be seen, the mean *P* of Russian is slightly greater than that of Slovak. Using the given numbers we can test the difference by (2.2). First we multiply each *Var*(*P*) by *n* = 20 in order to obtain the sum of squares of deviations. For Slovak it is 0.01425 and for Russian it is 0.04448. Hence we obtain

$$t = \frac{|0.8425 - 0.7811|}{\sqrt{\frac{0.04448 + 0.01425}{20 + 20 - 2}\left(\frac{1}{20} + \frac{1}{20}\right)}} = 1.56$$

which is not significant, i.e. these two languages belong to the same frequency-roughness class.

For Slovenian, 20 texts were taken from http://quanta-textdata.uni-graz.at, too. They are as follows (*SN* – Short novel, *LN* – Letter novel, *OL* – Open letter, *Sc* = Science, *Hi* = History, *NS* = News).

	Autor	Text sort	Text	
1	Peter Kolšek	LN	Brina Švigelj - Visok hrib pravzaprav gora pri	1998
2	Peter Kolšek	LN	Brina Švigelj - Da od Zjutraj ko pri	1998
3	Matija Kočevar	SS	Izgubljene stvari	2001
4	Matija Kočevar	SS	Ko je vsega konec	2001
5	Milan Hladnik	Hi	Slovenska kmecka povest: In kakšen je bil konec	1990
6	Milan Hladnik	Hi	Slovenska kmecka povest: Iz česa vse je kmečka povest	1990
7	Boris Jež	NS	Kam z Jelinčičem!?	2002
8	Ivan Praprotnik	NS	Odhod z odra	2002
9	Grega Repovž	NS	Tako razlièna	2002
10	Ivan Cankar	SN	Hlapec jernej in njegova pravica, Ch. 1	1907
11	Ivan Cankar	SN	Zzodba o dveh mladih ljudeh, Ch. 1	1911
12	Fran Finžgar	SN	Strici, Ch. 1	1927
13	Fran Finžgar	SN	Strici, Ch. 2	1927
14	Nina Mazi	Sc	Tradicionalna medicina in WHO	1997
15	Dimitrij Zimsek	Sc	Klinična informatika	1997
16	Tone Škerlj	OL	Slovenski škofovski konferenci	13.03.2000
17	Tone Škerlj	OL	Mestni prostor in mestna uprava	27.01.1999
18	Tatjana Greif	OL	Odprto pismo Teološki fakulteti	2003
19	Marijan Pušavec	NO	Zbiralec nasmehov: Kral in angel	1991
20	Marijan Pušavec	NO	Zbiralec nasmehov: Zadnja škusnjava	1991

The results of computation are presented in Table 2.8. In Slovenian, the order of test sorts is

News = 0.7571 < *Open letter* = 0.7905 < *Short story* = 0.8077 < *Letter novel* = 0.8200 < *Novel* = 0.8358 < *Short novel* = 0.8540 < *Science* = 0.8629 < *History* = 0.8719.

Table 2.8
Computation of *P* in 20 Slovenian texts (ordered according to increasing *n*)

Text	**n**	**L**	**L_{max}**	**L_{min}**	**P**	**Var(L)**	**Var(P)**	**u**
T 2	225	504,82	584,06	228,37	0,7750	3,4991	2,75E-05	52,45
T 1	236	648,38	713,10	241,53	0,8609	7,6585	3,43E-05	61,63
T 16	296	1152,58	1386,18	306,13	0,7830	21,5403	1,84E-05	65,92
T 17	556	3878,35	4627,94	581,99	0,8145	103,1137	6,30E-06	125,35
T 18	563	2328,05	2843,10	574,37	0,7726	24,7419	4,80E-06	124,41
T 10	602	6980,76	8486,38	638,77	0,8080	225,9898	3,67E-06	160,83
T 9	625	4098,97	5261,43	647,16	0,7479	66,1066	3,10E-06	140,72
T 7	691	3798,75	4969,48	706,06	0,7252	33,6839	1,85E-06	165,49
T 8	775	8011,35	9835,30	810,81	0,7978	174,1371	2,14E-06	203,68

T 13	1023	26161,24	29862,88	1107,50	0,8712	1164,5566	1,41E-06	312,83
T 15	1199	9149,26	10763,63	1220,17	0,8308	81,3785	8,93E-07	349,95
T 14	1357	21099,12	23408,45	1418,94	0,8949	543,5198	1,12E-06	372,53
T 4	1578	31018,06	37409,64	1645,46	0,8213	599,4573	4,69E-07	469,29
T 20	1662	28755,15	34346,91	1702,42	0,8287	340,9218	3,20E-07	581,13
T 19	1679	26899,94	31593,85	1718,63	0,8429	361,7491	4,05E-07	538,56
T 12	1829	64957,88	73851,20	1960,11	0,8763	2489,1533	4,82E-07	542,21
T 3	2360	58583,59	73139,13	2439,42	0,7941	908,2377	1,82E-07	689,98
T 5	3304	103818,03	119803,24	3418,23	0,8626	1949,3819	1,44E-07	955,95
T 6	3554	115084,21	130131,18	3656,80	0,8810	1874,7870	1,17E-07	1112,96
T 11	4028	307502,60	356684,69	4325,71	0,8604	11578,7559	9,33E-08	1180,22

The mean P of all Slovenian texts is $\overline{P}$ = 0.8225, $Var(P)$ = 0.0021. Testing the difference according to (2.10) we obtain

t(Slovak, Russian) = 1.56
t(Slovak, Slovenian) = 3.37
t(Russian, Slovenian) = 0.003

hence only the difference between Slovak and Slovenian is significant.

3. Hurst exponent

Here we introduce another way of computing a characteristic of time series, namely the *Hurst exponent*. This measure of long time memory of time series was originally developed in hydrology and later applied to series in several other fields. It can also be applied in quantitative linguistics as shown by L. Hřebíček (2000). The Hurst exponent H takes usually values between 0 and 1, where H < 0.5 and H > 0.5 indicate volatility or the presence of a tendency, respectively. The most common method to determine H is by means of the rescaled range-statistic (*R/S*-statistic) which will be illustrated as follows.

Let us consider a time series $(x_1,x_2,\ldots,x_n)$. The *R/S* statistic consists of a sequence of values *RS(T)* which must be computed for T = 2, 3,...,n. For a given number T we define *RS(T)* as follows:

Calculate mean and standard deviation of the first T elements of the sequence by

$$\overline{x}_T = \sum_{i=1}^{T} x_i \qquad (3.1)$$

and

$$S_T = \sqrt{\frac{1}{T}\sum_{i=1}^{T}(x_i - \overline{x}_T)^2} \quad . \qquad (3.2)$$

Calculate the sums of centralized values

$$Y_t = \sum_{i=1}^{t} (x_i - \bar{x}_T) \tag{3.3}$$

for $t = 1,...,T$ and compute

$$RS(T) = \frac{1}{S_T} \left[\max_{1 \le t \le T} Y_t - \min_{1 \le t \le T} Y_t \right] . \tag{3.4}$$

We consider again the sequence (1.2) concerning verse length in *Der Erlkönig:* (8,7,8,8,9,6,6,6,7,7,7,6,8,5,6,6,7,6,6,8,9,5,8,7,8,9,9,6,6,7,7,7) which has the length n = 32. Setting for example T = 10 we obtain from (3.1) and (3.2):

$$\bar{x}_{10} := \frac{1}{10}(8+7+8+8+9+6+6+6+7+7) = 7.2$$

and

$$S_{10} = \sqrt{\frac{1}{10}\sum_{i=1}^{10}(x_i - 7.2)^2} = \sqrt{0.96} .$$

The centralized values are

$(x_1 - 7.2, x_2 - 7.2, \ldots, x_{10} - 7.2) =$

$= (0.8, -0.2, 0.8, 0.8, 1.8, -1.2, -1.2, -1.2, -0.2, -0.2)$

and the corresponding partial sums of (3.3) are

$(Y_1, Y_2,...,Y_{10}) = (0.8, 0.6, 1.4, 2.2, 4.0, 2.8, 1.6, 0.4, 0.2, 0.0).$

Here we obtain $Y_{\max} = 4$ and $Y_{min} = 0$, hence

$$RS(10) = \frac{Y_{\max} - Y_{\min}}{S_{10}} = \frac{4-0}{\sqrt{0.96}} = 4.0825 .$$

Performing these computations for **a**ll values T = 2, 3,...,n = 32 (sequence length) yields the observed values *RS(T)* in the second column of Table 3.1. In applied sciences, one usually fits the function $(T/2)^H$ to such data, where H is the Hurst exponent. However, other possibilities have been tested, too.

The calculated values of the fitted power function are shown in the third column of Table 3.1. We obtain $RS(T) = (T/2)^{0.7981}$ with $R^2 = 0.96$, i.e. $H = 0.7981$. Since H is considerably larger than 0.5, the linguistic process can be considered as persistent. The results from Table 3.1 are plotted in Figure 3.1.

Table 3.1
Computation of *RS*(*T*) for verse length in *Der Erlkönig*

T	RS(T)	Computed		T	RS(T)	Computed
1				17	5.5817	5.518059
2	1.0000	1.000000		18	5.7540	5.775619
3	1.4142	1.382107		19	6.0276	6.030304
4	1.7321	1.738832		20	6.8070	6.282296
5	1.5811	2.077798		21	7.6449	6.531756
6	2.1213	2.403252		22	6.2201	6.778829
7	2.7217	2.717883		23	6.8343	7.023644
8	3.4412	3.023536		24	7.0259	7.266318
9	3.7741	3.321551		25	7.5842	7.506959
10	4.0825	3.612941		26	8.2256	7.745665
11	4.3708	3.898496		27	8.8205	7.982523
12	4.8054	4.178850		28	8.3816	8.217617
13	5.3533	4.454522		29	7.9786	8.451021
14	4.6771	4.725941		30	8.1011	8.682806
15	5.0196	4.993471		31	8.2218	8.913035
16	5.3405	5.257421		32	8.3408	9.141771

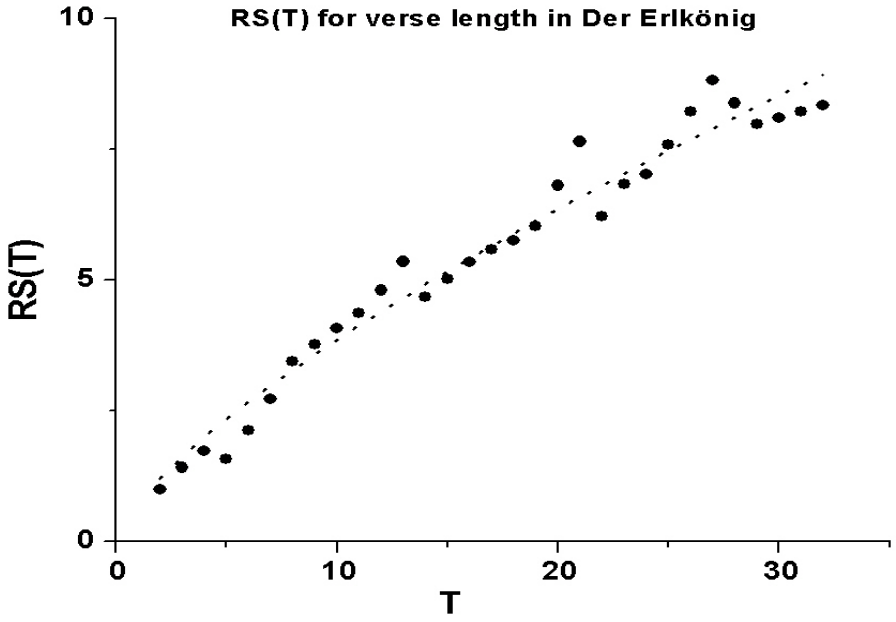


Figure 3.1. Plot of the *RS*(*T*) for verse length in *Der Erlkönig*

3.1. Word length

The computations of *H* for word length data are presented in Table 3.2.

Table 3.2
Computing the Hurst exponent for word length data

Language/Text	**n**	**H**	**R^2**	**P**
Akan: Agya Yaw Ne Akutu Kwaa	201	0,6776	0,9326	0,7296
Akan: Mma Nnsua Ade	143	0,6454	0,8296	0,7960
Bamana: Bamakɔ sigicogoya	1138	0,6439	0,9214	0,7264
Bamana: Masadennin	2616	0,6642	0,9087	0,7412
Bulgarian: Ostrovskij, Kak se kaljavaše …	926	0,7401	0,9598	0,6727
Czech: Čulík, O čem jsou dnešní Spojené státy?	2003	0,6276	0,8682	0,6279
Czech: Hvížďala, O předem zpackané prezidentské volbě	929	0,6878	0,8915	0,6068
Czech: Macháček, Slovenský dobrý příklad	340	0,7399	0,9201	0,6284
Czech: Spurný, Prekvapení v justici	288	0,5814	0,4141	0,6132
Czech: Švehla, Editorial, Voličův kalkul	288	0,7312	0,9323	0,5982
German: Assads Familiendiktatur	1415	0,6502	0,8449	0,6458
German: ATT0012 (Press)	1148	0,6350	0,9125	0,6671
German: Die Stadt des Schweigens	1567	0,6631	0,8192	0,6734
German: Terror in Ost Timor	1398	0,6719	0,8704	0,6837
German: Unter Hackern und Nobelpreisen	1363	0,6631	0,9108	0,6694
Hungarian: A Nominalizmus forradalma	1314	0,6545	0,7750	0,7095
Russian: Ostrovskij, Kak zakaljalas stal’	792	0,7233	0,9650	0,5914
Slovak: Bachletová, Moja dolná zem	872	0,6586	0,9444	0,6267
Slovak: Bachletová, Riadok v tlačive	924	0,6671	0,3282	0,6014

Evidently the Hurst exponent and the *P* indicator are not associated as shown in Figure 3.2. They represent different kinds of measurement.

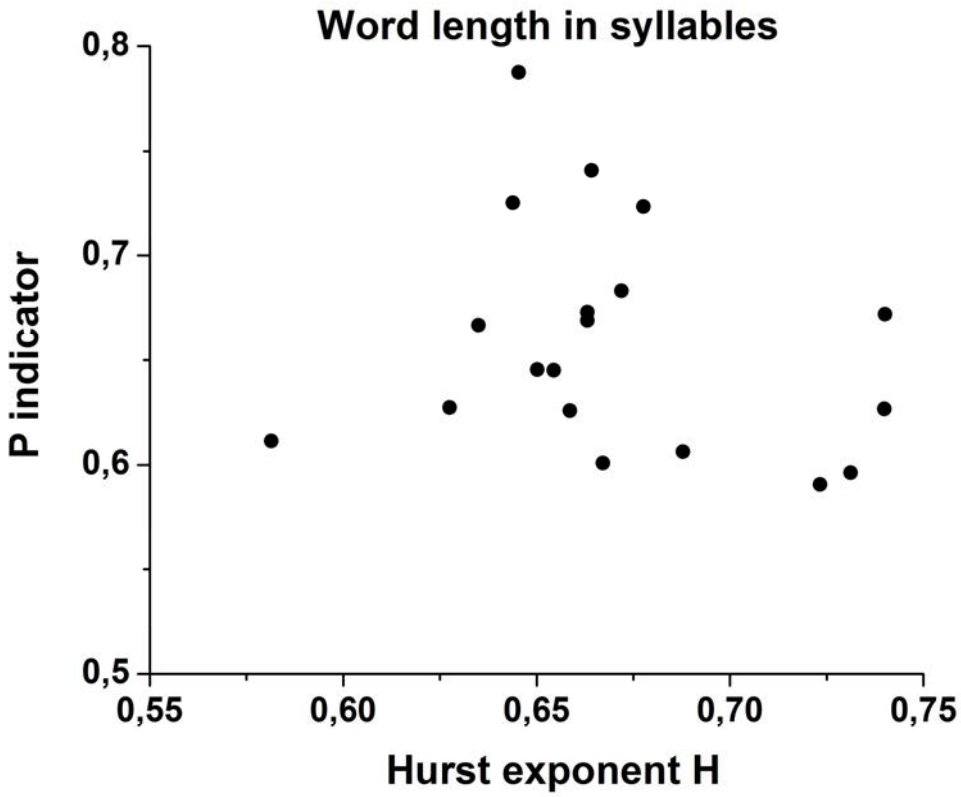


Figure 3.2. The independence of the indicators *H* and *P*

3.2. Frequencies

Let us now consider the form of the Hurst exponent in the frequency data. Besides "regular" results we obtain evident breaks in the text which can be interpreted as a loss of the internal rhythm, change of theme, a place where many corrections have been performed etc. Thus the breaks in the *RS*(*T*) curve can be interpreted textologically. For the texts in Slovak, Russian and Slovenian we obtain the results presented in Tables 3.3, 3.4 and 3.5.

Table 3.3
Hurst exponent in 20 Slovak texts

Text No	Sort	n	H	R^2
1	NW	229	fitting failed	
2	NW	293	0,6760	0,7625
3	SS	793	0,5426	0,1514
4	SS	1044	fitting failed	
5	FT	123	fitting failed	

6	FT	95	0,5597	0,6520
7	NW	437	0,5520	0,6239
8	NW	351	0,6822	0,9332
9	SS	1132	0,7171	0,9521
10	SS	1143	0,6223	0,9291
11	FT	283	0,7203	0,9353
12	FT	267	0,7079	0,9014
13	SS	821	0,6401	0,7272
14	SS	980	0,6320	0,8827
15	NO	1605	0,6409	0,8028
16	NO	5364	0,6428	0,9049
17	NO	1486	0,6101	0,9405
18	NO	3666	0,6473	0,8642
19	SS	409	0,6063	0,8371
20	SS	428	0,6429	0,6978

The mean is $\bar{H}$ = 0.6378 and the variance is *Var*(*H*) = 0.002715 (not taking into account the failed fittings, that is *n* = 17).

The ordering of text sorts is (according to their mean respective *H*) yields

Short story = 0.6262 < *Novel* = 0.6352 < *News* = 0.6367 < *Fairy tale* = 0.6626

which is, so to say, almost the reverse order as compared with *P* where we had *NS* < *FT* < *SS* < *NO*. More texts would surely lead to a better crystallisation of text sort.

As can be seen, one obtains different sequences. The first sort is monotonously increasing and can be well captured by the power function, for example text No 9 presented graphically in Figure 3.3.

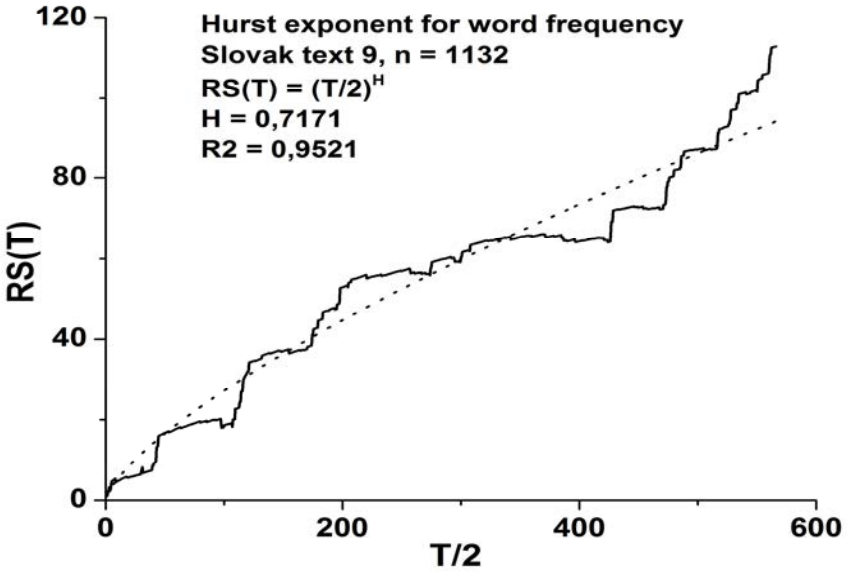


Figure 3.3. The RS(T) function for the Slovak text No. 9

A second alternative is represented by texts in which there is a turning point. Beginning from this point the *RS*(*T*) curve decreases (regularly or irregularly). This is the case e.g. in the Slovak text No 1 presented in Figure 3.4.

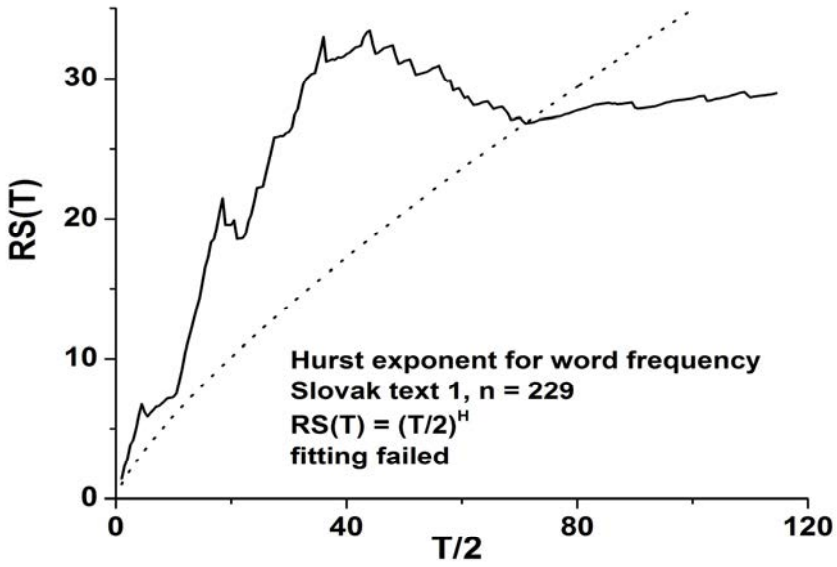


Figure 3.4. The RS(T) function in the Slovak text No 1

The third alternative represents a very irregular course of the curve testifying to the most probable pauses in writing or changes of the text. But one will surely find still other literary or textological "causes". An example for such a behaviour is the Slovak text no. 19, presented graphically in Figure 6.3. The general trend is conserved but at some places in the text there are positions which seem to represent a new begin of writing. Thus the Hurst exponent seems to be an interesting indicator of the dynamics of text generation.

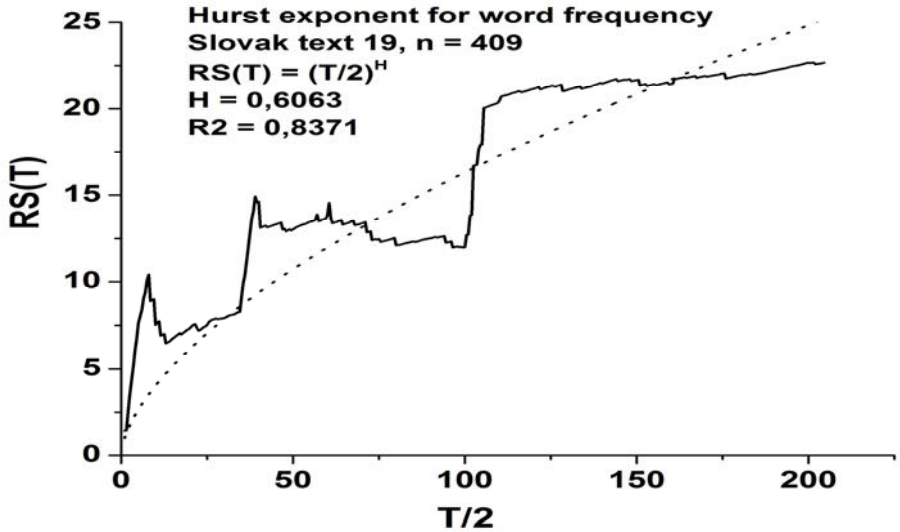


Figure 3.5. The RS(T) function in the Slovak text No 19

For the Russian texts we obtain the results presented in Table 3.4

Table 3.4
Hurst exponent in 20 Russian texts

Text	**Text sort**	**n**	**H**	**R^2**
1	SP	2905	0,5856	0,8990
2	SP	3389	0,6366	0,9152
3	SP	2657	0,6331	0,8391
4	SP	2656	0,8347	0,7816
5	SS	908	0,5254	0,5852
6	SS	949	0,6405	0,9342
7	SS	1951	0,5487	0,8764
8	SS	1453	0,7274	0,9539
9	NW	251	0,6811	0,9188
10	NW	785	fitting failed	
11	NW	220	0,6227	0,9315
12	NW	377	0,7099	0,9502
13	PL	879	0,6290	0,1603
14	PL	1097	0,5519	0,7721
15	PL	778	0,6519	0,9351
16	PL	2175	0,6853	0,8925
17	NO	2595	0,5820	0,8365
18	NO	5604	0,6157	0,9304
19	NO	702	0,6283	0,8378
20	NO	3576	0,6374	0,8842

In Russian we find another type of sequence which may, perhaps, display some partitions of the novel (cf. Figure 3.6).

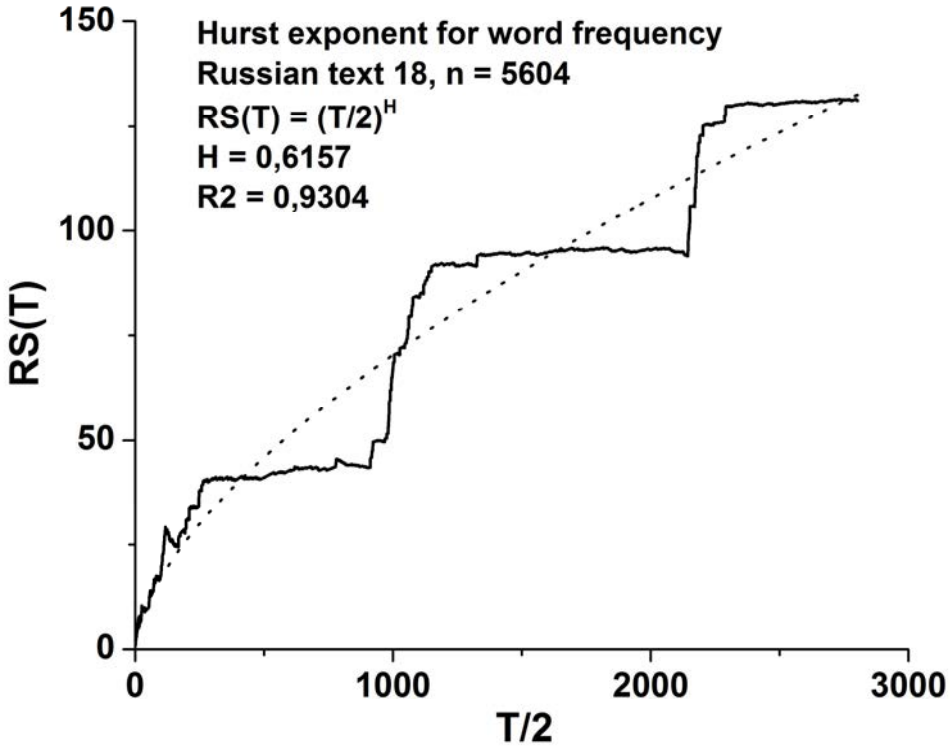


Figure 3.6. The RS(T) function in the Russian text No 18

The mean of *H* in Russian texts is $\bar{H}$ = 0.6383 and the variance is *Var*(*H*) = 0.00479. Here *n* = 19.

The order of text sorts according to their respective mean *H* is

Short story (0.6105) < *Novel* (0.6159) < *Private letter* (0.6295) < *News* (0.6712 < < *Stage play* (0.6725).

For the 20 Slovenian texts we obtain the results presented in Table 3.5. As can be seen, here the order of text sorts seems to be quite different to that based on *P*. Taking again the means of *H* from Table 3.5 we obtain

History (0.5551) < *Science* (0.5444) < *Letter novel* (0.6105) < *Open letter* (0.6120) < *Short novel* (0.6221) < *Short story* (0.6584) < *News* (0.6855) < *Novel* (0.6972).

The stability of this order must be tested using a number of other texts. A textological interpretation will be possible - perhaps - after analysing at least ten texts of each sort.

Table 3.5
Hurst exponent in 20 Slovenian texts

Text	**Text sort**	**n**	**H**	**R^2**
1	LN	236	0,6413	0,9431
2	LN	225	0,5797	0,9355
3	SS	2360	0,6584	0,9509
4	SS	1578	fitting failed	
5	Hi	3304	fitting failed	
6	Hi	3554	0,5551	0,5395
7	NS	691	fitting failed	
8	NS	775	0,6539	0,5147
9	NS	625	0,7171	0,9603
10	SN	602	0,6483	0,9503
11	SN	4028	0,6619	0,9005
12	SN	1829	0,6446	0,8254
13	SN	1023	0,5335	0,8379
14	Sc	1357	0,5335	0,8748
15	Sc	1199	0,5552	0,9134
16	OL	296	0,6235	0,8979
17	OL	556	0,5823	0,7711
18	OL	563	0,6301	0,8774
19	NO	1679	0,6491	0,9369
20	NO	1662	0,7452	0,9747

What is the relationship between P computed from the arc and *H* computed from normalized ranges? Are they correlated? As can easily be stated, there is no link between the two indicators. They characterize quite different properties of the text, though both capture the oscillation of the values. Even if we put together texts of the same sort in the three Slavic languages (there are only three such class up to now: SS, Nw, NO) we do not obtain any correlation. Hence the arc and the Hurst exponent are (preliminarily) independent of one another.

Computing the Hurst exponent for sentence lengths measured in terms of clause numbers we obtain for 20 German texts the results presented in Table 3.6.

Table 3.6
Hurst exponent in 20 German texts: sentence length in clauses

Text	**n**	**H**	**R^2**	**P**
T1	148	0,7244	0,7910	0,5288
T2	80	0,6343	0,7505	0,6289
T3	112	0,7296	0,7393	0,6936
T4	208	0,6296	0,8843	0,6501

T5	246	0,6195	0,3966	0,5955
T6	109	0,7178	0,2904	0,7090
T7	107	0,8157	0,8780	0,5847
T8	85	0,6428	0,5725	0,6759
T9	97	0,8059	0,8808	0,5551
T10	112	0,7220	0,9003	0,6371
T11	95	0,8444	0,9378	0,5112
T12	74	0,6700	0,8638	0,6115
T13	120	0,7778	0,5956	0,5316
T14	139	0,8545	0,9014	0,6653
T15	105	0,8703	0,7015	0,6195
T16	119	0,8766	0,9563	0,5254
T17	110	0,8377	0,9804	0,5509
T18	197	0,7487	0,9056	0,6122
T19	79	fitting failed		0,6050
T20	151	0,6929	0,5714	0,6804

The values of the determination coefficient are not always satisfactory but this is most probably a consequence of text corrections, our way of measurement or style. S can be shown in Figure 3.7., there is no relation between the Hurst exponent and the indicator P.

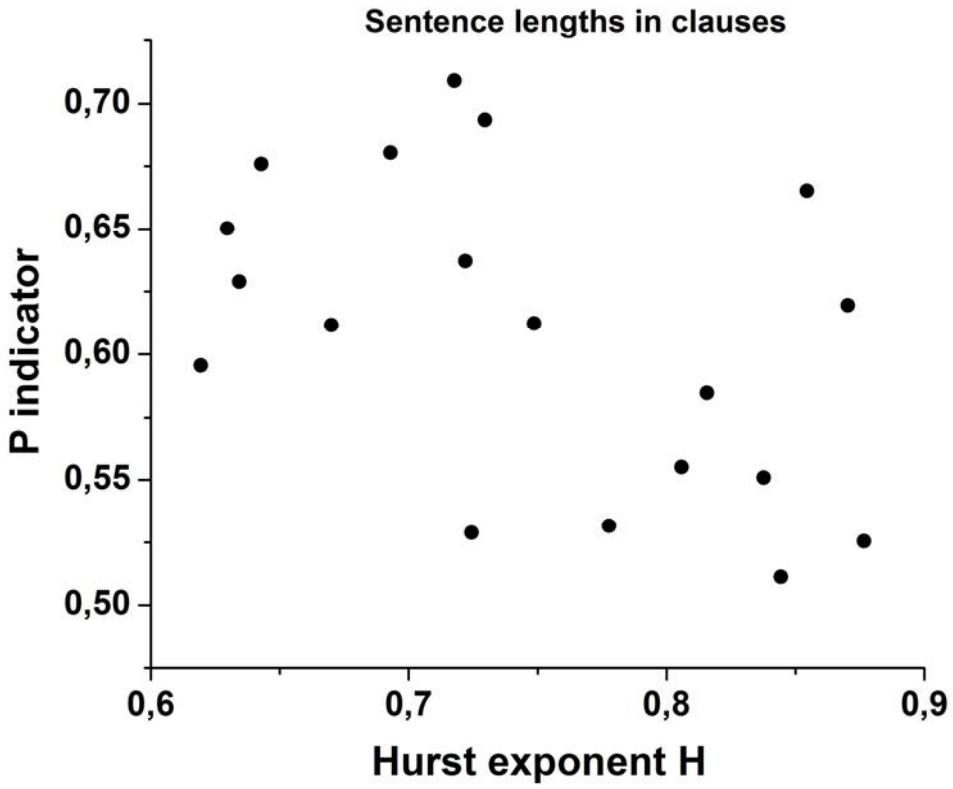


Figure 3.7. The independence of *H* and *P* indicators

4. Distances

In the analysis of a text we consider a sequence of numbers representing e.g. word length, sentence length or word frequency, we can measure the distance between equal numbers using different indicators. The simplest way is to consider the number of steps necessary to arrive at the next identical number. For example in (2.3) where we considered E_1 = (8,7,8,8,9,6,6,6,7,7,7,6,8,5,6,6,7,6,6,8,9,5,8,7,8,9,9,6,6,7,7,7) we begin with the first element, 8, and need 2 steps to meet the same number 8. The next 8 follows immediately after the sendond one, hence the distance is here 1.

If we count all distances and omit the distance of the last given element to the "next" (which would be infinity) we obtain the distribution of distances measured in this way. It is to be noted that this kind of measurement can be performed also with qualitative variables. In order to capture the course of f_x representing the number of distances x we may model it using the discrete or the continuous (c.f. Zörnig 2013) Zipf-Alekseev distribution. However, frequently the large distances have frequency 0 and in order to fit their distribution sequence one must pool many classes. Thus one uses mostly the continuous function for which the zero classes can be omitted. This is a "legal" procedure because the distance between identical elements can be measured in various (continuous) ways. Setting up a discrete or continuous model does not change the principal results because these concepts are chosen by the model builder, not by the reality. Thus we lean against the unified theory and conjecture that the relative rate of change of frequencies is proportional to the relative rate of change of distances, and the proportionality function is given as $g(x) = a + b \ln x$. We obtain the differential equation

$$\frac{df_x}{f_x} = \frac{a + b \ln x}{x} dx \tag{4.1}$$

The solution is

$$f_x = \frac{C}{x^{a + b \ln x}} \tag{4.2}$$

Which is an extended Zipf function. Would we consider (4.2) a discrete distribution, then C would be the normalizing constant. For our data concerning verse lengths we obtain the results presented in Table 4.1

Table 4.1
Distances between equally long verses in Erlkönig

x	1	2	3	4	5	6	7	8	9	16
$\mathbf{f_x}$	11	3	2	1	1	2	3	1	2	1
$\hat{f}_x$	10.92	3.37	2.11	1.66	1.47	1.37	1.33	1.32	1.33	1.68
a = 2.0384, b = 0.4924, C = 10.9218 , R^2 = 0.94										

As can be seen, the missing distances 10-15 are simply omitted. The formula expresses adequately the state of the affairs.

In the sequel we simply test the formula in order to obtain a background for such a procedure.

4.1. Word length

In order to test the adequateness of the above model, we fit it to the observed distances between words of identical length in texts from 28 languages. The parameter *C* is irrelevant because it merely expresses the frequency of the distance 1. The two other parameters could be used both for typological, text-sort analytic, stylistic and developmental analyses. Unfortunately the number of texts is too small in individual languages hence further research is necessary. But even at this preliminary stage one can observe e.g. differences between some Slavic languages for which the same text has been used. The interpretation must be postponed.

The fitting of the above model (4.2) to word length distances is presented in Table 4.2.

Table 4.2
Fitting the Zipf-Alekseev model to word length distances

Language/ Text	**Parameters**			
	a	**b**	**c**	**R^2**
Akan: Agya Yaw Ne Akutu Kwaa	-0,3018	1,6105	88,1164	0,99
Akan: Mma Nnsua Ade Bɔne	-0,4506	1,2125	50,2644	0,99
Bamana: Bamakɔ sigicoya	0,5827	0,5406	462,4441	0,997
Bamana: Masadennin	1,1021	0,3529	1229,1857	0,998
Bamana: Namakɔrɔba halakilen	1,2538	0,2606	706,4107	0,9995
Bamana: Sonsannin ani	0,7449	0,565	1087,639	0,999
Bulgarian: Ostrovskij, Kak se kaljavaše stomanata (Chap. 1)	-0,6538	0,8604	204,1388	0,996
Czech: Čulík, O čem jsou dnešní Spojené státy?	0,1993	0,39	505,9607	0,998
Czech: Hvížďala, O předem zpackané prezidentské volbě	0,1272	0,4291	234,5507	0,99
Czech: Macháček, Slovenský dobrý příklad	0,0655	0,4914	85,9735	0,99
Czech: Spurný, Prekvapení v justici	0,2114	0,3764	72,787	0,98
Czech: Švehla, Editorial, Voličův kalkul	-0,0215	0,5184	72,0338	0,98
French: Dunkerque – La route des dunes (press)	-0,0558	0,767	493,4571	0,996
German: Assads Familiendiktatur	0,1938	0,4734	398,6965	0,998
German: ATT0012 (press)	0,1766	0,4816	320,2555	0,998
German: Die Stadt des Schweigens	0,1145	0,5835	466,9158	0,996
German: Terror in Ost Timor	-0,0027	0,6448	4.102.169	0,998
German: Unter Hackern und Nobelpreisen	0,248	0,4769	399,852	0,998

Hindi: After the sanctionto love marriage (press)	-0,2918	0,8888	346,947	0,999
Hindi: The Anna Team on a cross-road (press)	-0,3026	1,0229	305,7838	0,998
Hungarian: A nominalizmus Forradalma (press)	-0,3578	0,5469	232,5982	0,998
Hungarian: Kunczekolbász (press)	-0,4872	0,5948	77,0436	0,98
Indonesian: Pengurus PSM terbelah (press)	0,5951	0,279	109,2435	0,99
Indonesian: Sekolah ditutup (press)	-0,3865	0,8586	74,5227	0,97
Italian: Il bosone di Higgs scoperto dal Cern (Internet)	-0,4921	0,7455	544,6497	0,998
Japanese: Miki, Jinseiron Note	-1,1231	1,2207	381,4454	0,99
Kikongo: Bimpa: Ma Ngo ya Ma Nsiese	-0,1335	0,6449	219,7708	0,99
Kikongo: Lumumba speech	-0,9993	1,0809	222,1013	0,997
Kikongo: Nkongo ye Kisi Kongo	-1,2297	1,6979	233,6042	0,99
Latin: Cicero, In Catilinam I	0,1416	0,4203	283,3235	0,99
Latin: Cicero, In Catilinam II	-0,3045	0,5902	654,3742	0,999
Macedonian: Ostrovskij, Kako se kaleše čelkiot (Chap. 1)	-0,9083	0,9126	204,2451	0,996
Malayalam: Moralistic hooligans (press)	0,0922	0,3201	51,777	0,96
Malayalam: No one should die (press)	-0,0804	0,3471	43,3842	0,95
Maninka: Nko Doumbu Kende no.2 (press)	0,7791	0,3617	833,7428	0,999
Maninka: Nko Doumbu Kende no.7 (press)	0,4372	0,4898	550,7852	0,997
Maninka: Sìikán` (Constitution of Guinea, an excerpt)	-0,0746	0,6721	481,1838	0,996
Odia: The Samaj, Bhuba-neshwar (28 June 2012), p. 4	0,2968	0,4666	106,3381	0,98
Odia: The Dharitri, Balasore (12th Feb, 2012), p. 10	-0,1116	0,5341	146,6132	0,98
Romanian: Paler, excerpt from Aventuri solitare	-0,9371	1,1645	204,4162	0,98
Romanian: Steinhardt, Jurnalul fericirii, Trei soluții	-0,3386	0,8373	412,3248	0,99
Romanian: Popescu D.R., Vânătoarea regală	-0,3802	0,904	293,4566	0,995
Russian: Ostrovskij, Kak zakaljalas stal' (Chap. 1)	0,0277	0,5064	206,283	0,996
Serbian: Ostrovskij, Kako se kalio čelik (Chap. 1)	-0,0527	0,5345	255,1807	0,99
Slovak: Bachletová, Moja Dolná zem	-0,0647	0,5329	216,1887	0,998
Slovak: Bachletová, Riadok v tlačive	0,0388	0,4578	205,8707	0,99
Slovenian, Kak zakaljalas	0,1864	0,5209	299,9342	0,996
Sundanese: Agustusan (Online)	0,0622	0,5837	122,0413	0,99

Sundanese: Aki Satimi (Online)	-0,2768	0,7341	344,2042	0,998
Tagalog: Rosales, Kristal Na Tubig	-2,078	1,7239	303,0685	0,98
Tagalog: Hernandez, Limang Alas: Tatlong Santo	-1,0358	1,0807	349,0266	0,996
Tagalog: Hernandez, Magpisan	-1,4802	1,2429	236,0223	0,99
Tamil: Emu Bird Trading (press)	0,4878	0,2059	88,9384	0,97
Telugu: Trailangaswamy (press)	0,0935	0,3672	59,7176	0,98
Telugu: Train Journey (press)	0,1276	0,3614	141,1524	0,99
Vai: Sa'bu Mu'a'	0,7317	0,5156	232,4315	0,996
Vai: Sherman, Mu ja vaa	0,7717	0,5211	1494,9202	0,999
Vai: Vande	-0,515	1,5147	176,3501	0,99
Welsh: text 1 (gaenv)	-0,786	0,9916	226,8039	0,99
Welsh: text 2 (gasodl)	0,0522	0,7892	469,4913	0,996

4.2. Sentence length

We shall omit sentence length measured in terms of word numbers and concentrate on those measured in clause numbers. We have merely 20 German newspaper data. In all of them, the Zipf-Alekseev model can be fitted with very good results. The data are presented in Table 4.3. Again, the parameter *C* depends merely on the first value (or rather text size) and is not relevant.

Table 4.3
Fitting the Zipf-Alekseev function to distances between equal sentence lengths (in clauses) in 20 German newspaper texts.

T1			T2			T3		
x	f_x	$\hat{f}_x$	x	f_x	$\hat{f}_x$	x	f_x	$\hat{f}_x$
1	48	48.11	1	24	23.94	1	36	36.08
2	27	25.94	2	13	13.59	2	21	21.13
3	14	16.28	3	9	8.98	3	16	13.00
4	11	11.17	4	9	6.44	4	3	8.53
5	11	8.12	5	5	4.87	5	7	5.88
6	7	6.15	6	2	3.83	6	7	4.22
7	4	4.81	7	2	3.09	7	1	3.12
8	3	3.85	8	2	2.55	8	2	2.37
9	2	3.14	9	2	2.14	9	3	1.84
11	1	2.19	19	2	0.62	12	1	0.94
12	2	1.86	25	1	0.37	15	1	0.54
13	1	1.59	35	1	0.19	19	1	0.28
14	1	1.38	56	1	0.07	20	1	0.25

17	2	0.93				27	2	0.10
19	2	0.74				37	1	0.04
20	1	0.67				41	1	0.03
21	1	0.60				57	1	0.01
24	1	0.45				59	1	0.01
30	1	0.27						
31	1	0.25						
a = 0.7296, b = 0.2334 c = 48.1140, R^2 = 0.99			a = 0.6857, b = 0.1885 c = 23.9384, R^2 = 0.97			a = 0.5030, b = 0.3877 c = 36.0815, R^2 = 0.96		

T4			T5			T6		
x	f_x	$\hat{f}_x$	x	f_x	$\hat{f}_x$	x	f_x	$\hat{f}_x$
1	74	74.30	1	84	82.66	1	25	24.59
2	41	38.53	2	36	44.86	2	18	20.03
3	19	23.44	3	36	28.55	3	17	14.48
4	17	15.66	4	29	19.87	4	10	10.49
5	11	11.13	5	13	14.64	5	9	7.76
6	9	8.26	6	10	11.23	6	4	5.86
7	7	6.34	7	5	8.88	7	7	4.51
9	5	4.01	8	6	7.19	8	2	3.53
10	3	3.28	9	2	5.92	9	1	2.81
11	2	2.72	10	6	4.96	11	1	1.85
12	2	2.28	12	1	3.61	12	2	1.52
13	2	1.94	13	4	3.12	14	1	1.06
14	1	1.66	15	2	2.40	16	1	0.77
17	1	1.09	17	1	1.89	22	1	0.33
24	3	0.49	23	1	1.03	25	1	0.23
32	1	0.24	29	1	0.63	45	1	0.03
59	1	0.05	31	1	0.55			
67	1	0.03						
88	1	0.01						
a = 0.7717, b = 0.2534 C = 74.2955, R^2 = 0.99			a = 0.7354, b =0.2112 C = 82.6562, R^2 = 0.96			a = 0.02225, b = 0.4593 C = 24.5914, R^2 = 0.96		

T7			T8			T9		
x	f_x	$\hat{f}_x$	x	f_x	$\hat{f}_x$	x	f_x	$\hat{f}_x$
1	32	32.30	1	18	17.82	1	36	35.60
2	20	18.83	2	15	15.62	2	14	16.70
3	13	11.96	3	10	11.34	3	11	10.47
4	5	8.15	4	14	8.10	4	13	7.43
5	5	5.84	5	3	5.87	6	4	4.51
6	3	4.35	6	2	4.33	7	2	3.72
7	6	3.33	7	3	3.25	9	1	2.69
8	2	2.62	8	3	2.49	10	2	2.35
9	2	2.09	9	2	1.93	11	1	2.07

10	3	1.70	10	1	1.52	12	2	1.84
12	2	1.17	12	1	0.98	14	1	1.50
13	1	0.99	16	1	0.45	22	1	0.81
15	1	0.72	17	1	0.38	23	1	0.76
18	2	0.47	20	2	0.23	34	1	0.44
19	1	0.42	36	1	0.03	38	1	0.37
35	1	0.09	40	1	0.02			
36	1	0.08						
a = 0.5634, b = 0.3104 C = 32.3002, R^2 = 0.97			a = -0.1891, b = 0.5462 C = 17.8153, R^2 = 0.89			a = 1.0535, b = 0.0554 C = 35.6046, R^2 = 0.96		

T10			T11			T12		
x	f_x	$\hat{f}_x$	x	f_x	$\hat{f}_x$	x	f_x	$\hat{f}_x$
1	24	24.69	1	30	30.10	1	28	27.68
2	32	29.48	2	21	20.52	2	10	12.98
3	13	17.40	3	12	12.33	3	12	8.04
4	10	9.02	4	7	7.56	4	8	5.63
5	4	4.61	5	5	4.81	5	3	4.23
6	6	2.40	6	3	3.17	6	1	3.32
7	4	1.28	7	2	2.16	7	2	2.70
8	1	0.71	8	2	1.51	10	1	1.65
10	3	0.23	9	1	1.08	14	1	1.01
11	1	0.14	10	1	0.79	16	1	0.83
13	2	0.05	14	2	0.26	25	1	0.42
16	3	0.01	15	1	0.20	41	1	0.19
22	1	0.001	18	1	0.10			
27	1	0.0003	32	1	0.01			
32	1	0.00007	36	1	0.01			
a = -1.2381, b = 1.4171 C = 24.689, R^2 = 0.94			a = 0.1093, b = 0.6400 C = 30.1031, R^2 = 0.99			a = 1.0348, b = 0.0826 C = 27.6784, R^2 = 0.95		

T13			T14			T15		
x	f_x	$\hat{f}_x$	x	f_x	$\hat{f}_x$	x	f_x	$\hat{f}_x$
1	43	43.22	1	63	63.10	1	25	24.02
2	32	30.59	2	30	29.06	2	19	22.61
3	13	16.00	3	12	14.09	3	18	16.58
4	9	8.28	4	8	7.47	4	18	11.76
5	6	4.43	5	5	4.26	5	8	8.40
8	1	0.86	6	3	2.58	6	2	6.10
9	2	0.53	7	1	1.63	7	3	4.51
11	2	0.22	8	1	1.07	9	1	2.58
12	1	0.15	9	1	0.73	13	1	0.99
15	1	0.05	10	2	0.51	19	1	0.31
25	1	0.003	12	1	0.26	23	1	0.16
28	1	0.001	15	1	0.11	26	1	0.10

32	1	0.0005	20	1	0.03	77	1	0.0009
34	1	0.0003	22	1	0.02			
36	1	0.0002	29	1	0.006			
			37	1	0.002			
			39	1	0.001			
a = -0.1951, b = 1.0009 C = 43.2204, R^2 = 0.99			a = 0.6987, b = 0.6060 C = 63.0995, R^2 = 0.996			a = -0.3396, b = 0.6165 C = 24.0244, R^2 = 0.92		

T16			T17			T18		
x	f_x	$\hat{f}_x$	x	f_x	$\hat{f}_x$	x	f_x	$\hat{f}_x$
1	36	36.53	1	42	41.81	1	88	87.44
2	31	27.82	2	17	18.72	2	26	31.18
3	9	15.60	3	14	11.29	3	23	17.67
4	12	8.57	4	6	7.77	4	14	11.99
5	4	4.84	5	9	5.76	5	10	8.96
6	3	2.83	6	4	4.48	6	7	7.10
7	5	1.71	8	1	2.99	7	5	5.86
8	2	1.07	10	1	2.16	8	5	4.97
9	2	0.69	11	1	1.88	9	5	4.32
11	2	0.30	13	2	1.46	13	2	2.81
13	1	0.15	14	1	1.31	14	1	2.58
18	1	0.03	15	1	1.18	18	1	1.96
20	1	0.02	19	1	0.81	19	1	1.84
25	2	0.005	23	1	0.60	20	1	1.75
35	1	0.0006	26	1	0.49	21	1	1.66
43	1	0.0002	41	1	0.23	25	1	1.38
			44	1	0.21	34	1	1.01
a = -0.2598, b = 0.9417 C = 36.5342, R^2 = 0.95			a = 1.1044, b = 0.07922 C = 41.8129, R^2 = 0.98			a = 1.5421, b = -0.0786 C = 87.4357, R^2 = 0.99		

T19			T20		
x	f_x	$\hat{f}_x$	x	f_x	$\hat{f}_x$
1	19	19.15	1	69	68.82
2	16	15.84	2	21	22.43
3	12	10.95	3	11	12.37
4	8	7.51	4	13	8.34
5	1	5.25	5	9	6.24
6	2	3.75	6	6	4.97
7	7	2.74	7	1	4.13
9	3	1.55	8	2	3.54
10	1	1.19	9	2	3.10
13	1	0.59	10	2	2.76
28	1	0.05	11	2	2.49
29	1	0.04	15	2	1.82
36	1	0.02	20	1	1.39

			33 43 44 54	1 1 1 1	0.92 0.76 0.75 0.66
a = -0.1265, b = 0.5785 C = 19.1548, R^2 = 0.91			a = 1.7125, b = -0.1371 C = 68.8191, R^2 = 0.99		

Here it is easier to study the relationship between the parameters *a* and *b* because we have 20 data sets taken from the same text sort (newspapers). If we order the two parameters according to increasing values of *a*, we obtain the results presented in Table 4.4, showing that $b = f(a)$ where *b* is a linear function of *a*. That means, the distances between identical sentence lengths are controlled by two mechanisms. The first controls the distances, the second guarantees an equilibrium between the parameters. The function expressed by the results in Table 4.4. is $b = 0.6448 - 0.5050a$ where $R^2 = 0.86$. Further data would yield still better results.

Table 4.4
The dependence of parameter *b* on parameter *a*
in German sentence length-distance relationship

a	**b**	**b = f(a)**
-1,2381	1,4171	1.2700
-0,3396	0,6165	0.8163
-0,2598	0,9417	0.7760
-0,1951	1,0009	0.7433
-0,1891	0,5462	0.7403
-0,1265	0,5785	0.7087
0,0223	0,4593	0.6335
0,1093	0,6400	0.5896
0,5030	0,3877	0.3908
0,5634	0,3104	0.3603
0,6857	0,1885	0.2985
0,6987	0,6060	0.2920
0,7296	0,2334	0.2764
0,7354	0,2112	0.2734
0,7717	0,2534	0.2551
1,0348	0,0826	0.1223
1,0535	0,0554	0.1128
1,1044	0,0792	0.0871
1,5421	-0,0786	-0.1339
1,7125	-0,1371	-0.2200

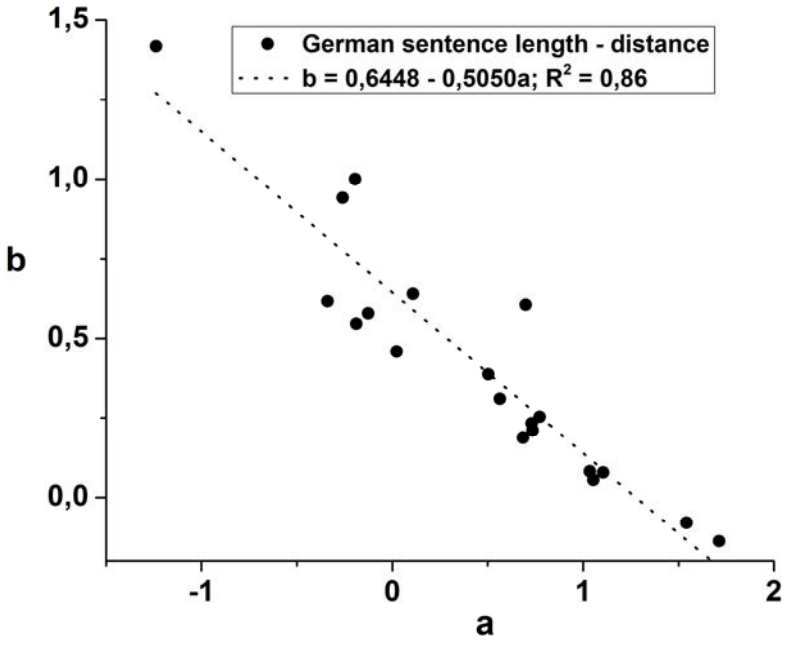


Figure 4.1. The dependence of parameter *b* on parameter *a* in Table 4.4

4.3. Frequencies

The distances between identical word frequencies present a rather different image. Frequency depends on thematic concentration, on vocabulary richness and other factors, and depends more on the given text than on frequencies which occur in corpuses. Hence one cannot expect a clearly expressed mutual dependence of parameters, quite the contrary. The frequencies may be very extreme, many of them are quite unique. Thus we expect a good fitting by the Zipf-Alekseev function but not a demonstrable relationship between the parameters.

The results of fitting (4.2) to the individual frequency vectors are presented in Table 4.5. All fits are very good, hence one can conjecture to have found the suitable function.

Table 4.5
Distances between equal frequencies fitted by the Zipf-Alekseev function (Slovenian)

Language/ Text	**Parameters**			
	a	**b**	**C**	**R^2**
T1	0,9781	0,0471	226,934	0,99
T2	0,1103	0,1871	377,1105	0,98

T3	0,8676	0,0972	182,6319	0,98
T4	0,8278	0,0761	139,5228	0,99
T5	1,0405	0,1604	208,0246	0,995
T6	0,7528	0,2848	149,2626	0,98
T7	0,8861	0,1138	230,508	0,99
T8	0,9338	0,2102	178,8466	0,99
T9	1,1223	0,0792	71,6603	0,99
T10	1,3713	0,0129	178,1989	0,99

Table 4.6
Ordered according to parameter *a*

a	**b**
0,1103	0,1871
0,7528	0,2848
0,8278	0,0761
0,8676	0,0972
0,8861	0,1138
0,9338	0,2102
0,9781	0,0471
1,0405	0,1604
1,1223	0,0792
1,3713	0,0129

Though it can be shown that parameter *b* decresases here with increasing *a*, the dependence is not linear but rather oscillating, hence no simple function can be used to captrure it. This need not hold for all languages, as can be shown below..

Table 4.7
Distances between equal frequencies fitted by the Zipf-Alekseev function (Slovak)

Text	**a**	**b**	**C**	**R^2**
T1	1,7279	-0,1064	99,9092	0,996
T2	1,1216	0,1879	112,8195	0,96
T3	0,8128	0,205	222,6503	0,99
T4	0,5194	0,2916	246,0265	0,99
T5	0,2425	0,253	24,1248	0,89
T6	2,1322	-0,2667	50,0155	0,998

T7	0,9093	0,2561	151,8341	0,99
T8	1,2987	0,0527	134,2203	0,995
T9	0,8963	0,2018	332,683	0,99
T10	0,8572	0,0655	234,9113	0,99
T11	1,4245	-0,006	108,6762	0,99
T12	1,3533	0,0232	103,4679	0,99
T13	0,3054	0,2708	144,8673	0,98
T14	0,7819	0,1503	234,3294	0,99
T15	0,3439	0,1664	204,4295	0,98
T16	0,0412	0,1821	368,2186	0,99
T17	0,4271	0,2894	312,9521	0,99
T18	0,3781	0,1754	489,6514	0,99
T19	0,4234	0,4662	121,6505	0,98
T20	0,211	0,5471	108,2932	0,96

If we consider the link between the parameters *a* and *b*, we may state that *b* can be expressed by a linear function of *a*, viz. $b = 0.3880 - 0.2686a$, $R^2 = 0.66$ The reordering according to increasing *a* is presented in Table 4.8. A graphical presentation can be found in Figure 4.2.

Table 4.8
The relationship between *a* and *b* in frequency distances in Slovak texts

a	**b**	**f(a)**
0.0412	0.1821	0,3769
0.2110	0.5471	0,3313
0.2425	0.2530	0,3229
0.3054	0.2708	0,3060
0.3439	0.1664	0,2956
0.3781	0.1754	0,2864
0.4234	0.4662	0,2743
0.4271	0.2894	0,2733
0.5194	0.2916	0,2485
0.7819	0.1503	0,1780
0.8128	0.205	0,1697
0.8572	0.0655	0,1578
0.8963	0.2018	0,1473
0.9093	0.2561	0,1438
1.1216	0.1879	0,0867
1.2987	0.0527	0,0392
1.3533	0.0232	0,0245
1.4245	-0.006	0,0054
1.7279	-0.1064	-0,0761
2.1322	-0.2667	-0,1847

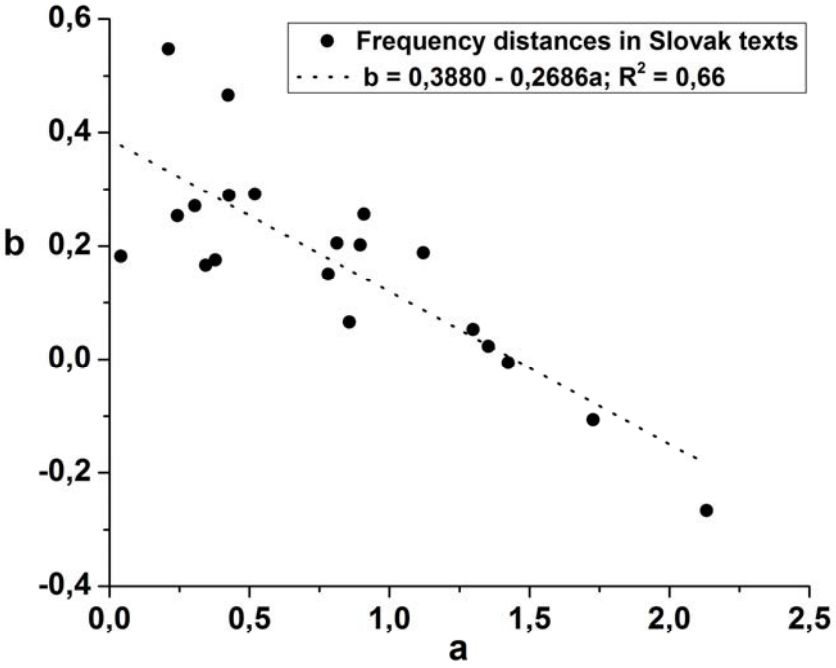


Figure 4.2. Dependence of *b* on *a* in frequency distances in Slovak texts

As could be expected, some of the texts display outliers whose character must be studied individually. It may be caused by the style of the author, by the theme, by the correction brought in after the text was ready, etc.

For the Russian data we obtain the results presented in Table 4.9.

Table 4.9
Distances between equal frequencies fitted by the Zipf-Alekseev function (Russian)

Text	**a**	**b**	**c**	**R^2**
T1	0,6464	0,0934	444,9898	0,99
T2	0,4194	0,1699	469,6606	0,99
T3	0,1792	0,2697	360,0277	0,98
T4	0,3580	0,2253	426,9983	0,99
T5	0,7201	0,1427	200,4294	0,97
T6	0,6948	0,2081	242,1476	0,99
T7	0,6315	0,1764	397,4242	0,99
T8	0,5626	0,1467	252,4396	0,99
T9	0,6437	0,9983	120,1178	0,99
T10	1,2262	0,1760	307,3182	0,99
T11	1,9545	-0,0997	112,4488	0,99
T12	1,5000	-0,0570	142.2132	0,99
T13	0,3863	0,2028	147,1140	0,98
T14	0,0583	0,4178	197,9598	0,97

T15	0,4529	0,2335	156,4488	0,97
T16	0,3969	0,1822	327,8049	0,99
T17	0,5885	0,1333	418.4794	0,99
T18	0,4989	0,1218	704,2576	0,99
T19	0,8346	0,1423	179,9826	0,99
T20	0,5189	0,1371	517,1111	0,99

Here merely one of the texts, namely T9 displays an outlier. Its cause could, perhaps, be found by analyzing the given text but this is not our aim. In any case, here *b* is linked with *a* in form $b = 0.2948 - 0.2043a$, $R^2 = 0.70$. The reordering of *a* and *b* values in presented in Table 4.10.

Table 4.10
Parameters *a* and *b* in Russian texts after reordering

Text	**a**	**b**	**b = f(a)**
T14	0,0583	0,4178	0,2829
T3	0,1792	0,2697	0,2582
T4	0,3580	0,2253	0,2217
T13	0,3863	0,2028	0,2159
T16	0,3969	0,1822	0,2137
T2	0,4194	0,1699	0,2091
T15	0,4529	0,2335	0,2023
T18	0,4989	0,1218	0,1929
T20	0,5189	0,1371	0,1888
T8	0,5626	0,1467	0,1799
T17	0,5885	0,1333	0,1746
T7	0,6315	0,1764	0,1658
T9	0,6437	0,9983	0,1633
T1	0,6464	0,0934	0,1627
T6	0,6948	0,2081	0,1529
T5	0,7201	0,1427	0,1477
T19	0,8346	0,1423	0,1243
T10	1,2262	0,1760	0,0443
T12	1,5000	-0,0570	-0,0117
T11	1,9545	-0,0997	-0,1045

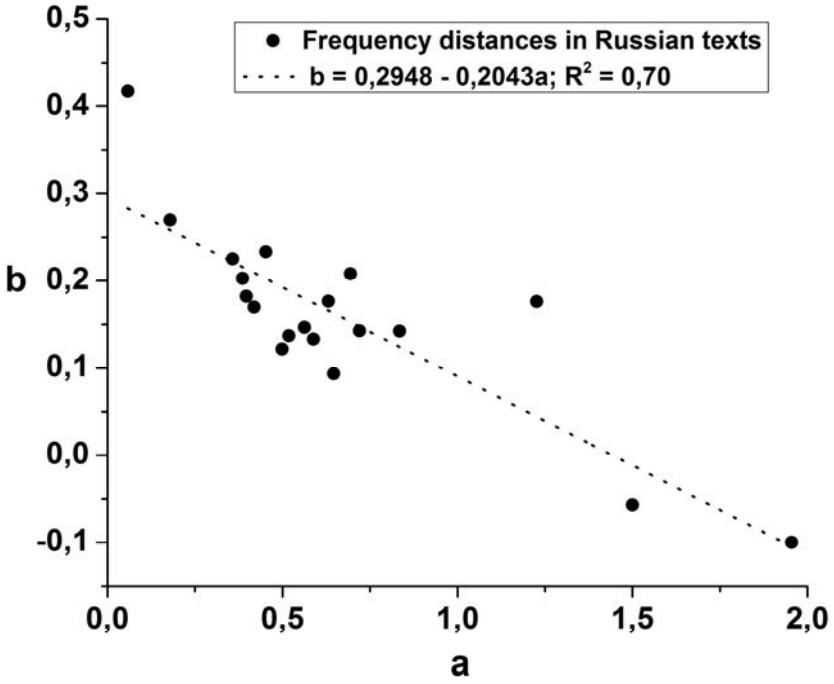


Figure 4.3. Relationship between a and b

Summary

Whatever property of text entites is scrutinized, i.e. measured and placed in the text replacing the original entity, the sequence itself may look chaotically forming rather a kind of fractal; nevertheless it conceals a number of stochastic regularities whose investigation is a task for the future. We merely selected some of them in order to show that they exist and abide by some regularities which after thorough testing may acquire the status of laws. Needless to say, a control cycle comprising all of them must be set up in order to see that they are linked with one another, and show the mathematical form of the link. Due to space restrictions, we merely showed the way directed to a deeper analysis. Arcs, Hurst exponents and distances have been analyzed in several languages for word length, sentence length and frequency representing only the surface of the text as a phenomenon. The next steps would be (1) the study of further languages in order to obtain a more extensive empirical basis, (2) the study of further properties taken from the infinite reservoir of language properties and (3) the linking of all these properties in control cycles in order to show that we have to do with a dynamic system. The study of these systems would be performed at a higher level using more complex mathematics.

References

Altmann, G. (2006). Fundamentals of quantitative linguistics. In: Genzor, J., Bucková, M. (eds.), *Favete linguis: 15-27*. Bratislava: Slovak Academic Press.

Brockwell, P.J., Davis, R.A. (2010). *Introduction to Time Series and Forecasting.* Berlin: Springer.

Çambel, A.B. (1993). *Applied chaos theory. A paradigm for complexity.* San Diego: Academic Press.

Hamilton, J.D. (1994). *Time series analysis.* Princeton University Press.

Hřebíček, L. (2000). *Variation in sequences*. Prague: Oriental Institute.

Kitagawa, G., Gersch W. (1996). Smoothness Priors Analysis of Time Series (Lecture Notes in Statistics 116), Springer, Berlin, Heidelberg, New York.

Köhler, R., Altmann, G. (2008, 2nd ed.). *Problems in Quantitative Linguistics, Vol. 1*. Lüdenscheid: RAM.

Mandelbrot, B.B. (1982). *The fractal geometry of nature.* New York: Freeman

Pandit, S.M., Wu, S.M. (1983). *Time series and system analysis with applications.* New York: Wiley.

Percival, D.B., Walden, A.T. (2010). *Wavelet Methods for Time Series Analysis.* Cambridge University Press.

Popescu, I.-I., Mačutek, J., Altmann, G. (2009). *Aspects of word frequencies*. Lüdenscheid: RAM.

Popescu, I.-I., Naumann, S., Kelih, E., Rovenchak, A., Overbeck, A., Sanada, H., Smith, R., Čech, R., Mohanty, P., Wilson, A., Altmann, G. (2013). Word length: aspects and languages. In: Köhler, R., Altmann, G. (eds.), *Issues in Quantitative Linguistics Vol. 3: 224-281*. Lüdenscheid: RAM-Verlag.

Zörnig, P. (2013). A continuous model for the distances between coextensive words in a text. *Glottometrics25, 54-68*.