Unified modeling of length in language

Ioan-Iovitz Popescu
Karl-Heinz Best
Gabriel Altmann

2014

Studies in quantitative linguistics

Editors
Fengxiang Fan (fanfengxiang@yahoo.com)
Emmerich Kelih (emmerich.kelih@uni-graz.at)
Reinhard Köhler (koehler@uni-trier.de)
Ján Mačutek (jmacutek@yahoo.com)
Eric S. Wheeler (wheeler@ericwheeler.ca)

1. U. Strauss, F. Fan, G. Altmann, *Problems in quantitative linguistics 1*. 2008, VIII + 134 pp.
2. V. Altmann, G. Altmann, *Anleitung zu quantitativen Textanalysen. Methoden und Anwendungen.* 2008, IV+193 pp.
3. I.-I. Popescu, J. Mačutek, G. Altmann, *Aspects of word frequencies*. 2009, IV +198 pp.
4. R. Köhler, G. Altmann, *Problems in quantitative linguistics 2.* 2009, VII + 142 pp.
5. R. Köhler (ed.), *Issues in Quantitative Linguistics.* 2009, VI + 205 pp.
6. A. Tuzzi, I.-I. Popescu, G.Altmann, *Quantitative aspects of Italian texts*. 2010, IV+161 pp.
7. F. Fan, Y. Deng, *Quantitative linguistic computing with Perl.* 2010, VIII + 205 pp.
8. I.-I. Popescu et al., *Vectors and codes of text*. 2010, III + 162 pp.
9. F. Fan, *Data processing and management for quantitative linguistics with Foxpro***.** 2010, V + 233 pp.
10. I.-I. Popescu, R. Čech, G. Altmann, *The lambda-structure of texts*. 2011, II + 181 pp
11. E. Kelih et al. (eds.), *Issues in Quantitative Linguistics Vol. 2*. 2011, IV + 188 pp.
12. R. Čech, G. Altmann, *Problems in Quantitative linguistics 3*. 2011, VI + 168 pp.
13. R. Köhler, G. Altmann (eds.), *Issues in Quantitative Linguistics Vol 3.* 2013, IV + 403 pp.
14. R. Köhler, G. Altmann, *Problems in Quantitative Linguistics 4.* 2014, VI + 148 pp.
15. K.-H.Best, E. Kelih (eds.), *Entlehnungen und Fremdwörter: Quantitative Aspekte.* 2014, IV + 163 pp.
16. I.-I. Popescu, K.-H. Best, G. Altmann, *Unified modeling of length in language*. 2014, II + 124 pp.

ISBN: 978-3-942303-26-2

RAM-Verlag
Stüttinghauser Ringstr. 44
D-58515 Lüdenscheid
RAM-Verlag@t-online.de
http://ram-verlag.de

Preface

The aim of this volume is to find a unified model of length distribution of any unit in language. It is merely a trial leaning against published results. It is impossible to perform the analysis for "all possible" units because only some of them are known and used, the definition of new ones is the normal policy in science. And even if a new unit is defined, it takes a special place in the hierarchy of other units which must be defined, too. Hence, this enterprise is endless.

We used published data from ca 50 languages, their dialects and historical epochs. It was impossible to use everything that has ever been published, it is already a separate discipline within linguistics. The majority of data has been elaborated in several projects performed in Göttingen where also student took part in the work (cf. http://www.gwdg.de/~kbest/projekt.htm). Besides three volumes quoted in the references (cf. Best 1997, 2001; Grzybek 2006) there are a number of publications in different journals.

Our main aim was to find a law of length and its special forms at individual language levels, that is, to avoid the search for ever new models whose validity is always merely local: they hold for the given level in the given language and taken together form an enormous family of distributions. In the unified model there are merely differences in the parameters, and the parameters themselves are part of a dynamic system displaying self-regulation.

We hope that researchers will test it in further languages and on different linguistic levels.

Ioan-Iovitz Popescu
Karl-Heinz Best
Gabriel Altmann

Contents

1. Introduction

Linguistic units are abstractions depending on definitions and criteria. Even if one generally accepts the existence of some units, their definitions may differ with different researchers and the partitions of texts need not be unique. The criteria are not "given", they are our mental constructions. Nevertheless, one can always try to analyze texts or a dictionary using ad hoc or traditional or quite new units and search for some regularities. Not all entities in language have physical length, e.g. polysemy or polytexty are sets having a size (cardinal number), but material units can always be analyzed in terms of length. Here we shall use the existing literature and measure the length of syllables, morphs, words, compounds, rhythmic units, verses, sentences, and speech acts. One can add further units, e.g. vowels whose duration is a phonological feature, but measurement should be restricted to one speaker, otherwise idiolectal differences may distort the image. Similarly, sentence length can be measured only in written texts because in spoken texts there are no unequivocal markings of sentence length (there is no punctuation). Further, the measurement of length of a unit should always be performed in terms of the number of immediate constituents. If possible, measurement in time units should be omitted because it varies each time the same speaker says the same thing. For general purposes one can obtain the units and their components as follows:

Unit	**Immediate constituent**
Character	Component
Syllable	Phoneme, phone
Morph	Phoneme, phone
Word	Syllable, morph, mora
Compound	Stems
Rhythmic unit	Syllable
Verse	Syllable, word
Phrase	Word
Clause	Phrase
Sentence	Clause
Speech act chain	Speech act
Dialogue	Speech act chain

Units of whatever kind can be classified according to various criteria (e.g. words in nouns, verbs, adverbs …; phrases in noun and verb phrases, etc.) and the individual classes display their own properties and hierarchies whose ends depend only on the state of affairs in the science. New units can be considered any time, as it is usual in all sciences.

Length is, of course, linked with all other collateral properties of the given unit as has been sufficiently shown in language synergetics (cf. Köhler 1986,

2005). They will not interest us here. Our aim is to show that the background mechanism of all the above relations is the same. Deviations occur if there are some boundary conditions causing them, cf. e.g. Pustet, Altmann (2005) for Lakota. But they appear also as soon as one omits the level of immediate constituents and performs the measurement in terms of "lower" units, e.g. sentence length in terms of words whose problems have been shown several times by the research team in Graz (cf. Grzybek 1998, 1999, 2000, 2010, 2011; Grzybek, Kelih, Stadlober 2008; Grzybek, P., Schlatte, R. 2002; Grzybek, Stadlober, Kelih 2007; Kelih, Grzybek 2004, 2005; Kelih, Grzybek, Antić, Stadlober 2006).

Not all units have been studied with the same intensity. Most works concern word length; none is devoted to the study of clause length in terms of phrase numbers, or phrases in number of words, perhaps because the grammatical analysis strongly depends on the given "linguistic school" and there are quite different definitions.

In the present study we shall analyze individual stages of the above hierarchy and try to find a general model adequate for all of them.

When setting up a theory, one always strives for a unified background and representation. Sometimes, one is forced to admit different models and in those cases one strives for finding the forces or factors that cause the difference or for unifying the theoretical background mathematically. One admits boundary conditions which always cause different values of parameters but sometimes require even adding a new parameter or modifying some classes of phenomena.

Adequacy of data

If a model does not seem to be adequate, our first question must concern the adequacy of data. The measurements may be invalid or irrelevant, e.g. measuring sentence length in terms of phoneme numbers. We must base our investigation on the conjecture that in language everything abides by laws, just as in physics. The data are created (not collected) on the basis of the hypothesis derived from the background theory. The data must be relevant for its testing. In the inductive approach, one considers some "given" linguistic matter and tries to find a function expressing the state of the affairs. One usually chooses the "best" model depending on the available software. Later on, the result will be either subsumed under a theory or something else will be derived from the existing theory.

There are two kinds of data in linguistics:

Systemic ones taken from the dictionary or from a collection in which each entity occurs only once. But dictionaries exist only for some selected units, e.g. words, but not for those that exist only in texts. One may consider all units or perform random, systematic or authoritative sampling. The random sampling may, again, be performed in different ways: one creates or chooses random numbers and collects the entities occurring on the x-th place on the y-th page; or one takes randomly a word on each r-th page. For the systematic sampling one can take e.g. the last word on each page; for authoritative sampling one considers

only words based on the decisions or restrictions of the investigator, e.g. only nouns on the first fifty pages of the dictionary.

Pragmatic sampling is based on evaluation of texts. It is recommended to evaluate each text separately. Herewith one avoids the creation of mixed data which cannot be used for testing a hypothesis. The evaluation of a corpus as a whole is reasonable for finding grammatical rules which are important for learning a language but it cannot help in all cases of testing. In all of the above mentioned cases the size of the sample should be sufficient (= not too small) even if no statistician can answer the question "what is sufficiently large?"

Correctness of testing

The second question concerns the correctness of testing. Usually, for frequencies one applies the chi-square test for goodness-of-fit. Though it has been modified several times and also replaced by different coefficients which took into account also the sample size and the degrees of freedom, the problems of testing remain. Some of them are well known, e.g. (a) Chi-square increases with sample size; that means, large samples do not improve the testing; on the contrary, some determination coefficients (Cramér, Tschuproff, Pearson, etc.) decrease with increasing sample size and improve the testing. (b) It is not appropriate for small class frequencies. The classical $f_x > 5$ has been reduced to $f_x > 1$. Some classes must frequently be pooled especially with rank distributions. This can often cause a non-testability of the hypothesis because of the reduction of degrees of freedom to zero. (c) Frequently, the small classes (i.e. the unimportant ones) yield deviations that are too large and this leads to the rejection of the model. (d) The chi-square does not yield symmetric results, i.e. the same difference yields two different results according to whether the expected value is greater or smaller than the observed one. But even if all conditions are ideally fulfilled one should not believe that positive testing means finding the "truth". It simply means that our hypothesis "may be preliminarily accepted". We should not forget that our hypothesis, the model, the data, the test and the interpretation of the test are our mental constructs which allow us to get better orientation in the "constructed" reality. The reality is neither probabilistic nor deterministic, there are no discrete or continuous phenomena, all this is merely our endeavor to get a better orientation in the (practical or scientific) object of observation. Hence truth has nothing to do with the way of our model building. A phenomenon can be modeled discretely or continuously, applying a probability distribution or a simple function or sequence, a graph or other construct, and the model can be tested using an exact significance level or simply using a measure of deviation. The interpretation of the deviation is, again, nothing objective but depends on our decision. In statistics, one uses usually the $\alpha = 0.05$ level, in psychology one interprets the different values of the determination coefficient quite differently.

Even if we try to avoid text mixing, every text may contain elements which were inserted in the text after a pause in writing or *a posteriori* as a kind of cor-

rection, or even by editors who "repaired" the text. We do not know and in most cases we cannot ask the authors. But if we mix several texts of a corpus, we may be sure that there is some problem in our modeling. The corpus is no approximation to a population because there are no populations in language (Orlov, Boroda, Nadarejšvili 1982). The texts of a corpus are taken from different periods, from different authors, they have different lengths, different contents, belong to different text sorts, etc. There is no sense in measuring e.g. the mean size or weight of animals taking a mouse, a cat, a dog and a cow, or that of material objects wandering through the galaxy.

The hypothesis

The third question has a theoretical character and concerns the hypothesis itself. After one hundred years of intensive research in quantitative linguistics, one knows that on the inductive way one can always use software which yields several (good) solutions. Which of them is "correct" or "preliminarily acceptable"? Still worse is the situation in the discipline concerned with classification: which of the ca 500 methods is the "right" one? Our decisions should be done after positive answering of at least one of the following problems: (1) Was the answer from the software the only adequate one or were there several solutions for different samples? (2) Can the hypothesis be substantiated linguistically, i.e. can the process behind the formula be interpreted linguistically? (3) Does the hypothesis unify data from different languages or texts? (4) Can the hypothesis be derived from a more general theory, i.e. does it lean against previous knowledge? (5) Is the hypothesis in some connection with other properties of the given entity (here length) as it is done in Köhler's control cycle? That is, is it subsumed under a system of hypotheses or is it quite isolated? All positive answers lead to a stronger foundation for the hypothesis and render its validity more probable. Of course, this way has no end. And even if we attained a reliable and well founded result, a new paradigm, a revolution, may change everything, as happened many times even in physics. But all these circumstances cannot slow down the development of a discipline. Scientific paradigms come and go, they express a topical world view. The more abstract a discipline is, the more paradigms there are.

In the present book we shall try to model length distributions in many texts and languages using published data as far as they were available to us. As far as we could state, about 70 languages have been analyzed up to now from this point of view, the majority of studies concern word length. All researchers used on principle a discrete distribution and arrived at an extensive family consisting of simple, displaced, truncated, extended, modified and mixed distributions of the same family (binomial, geometric, Cohen-negative binomial, Cohen-Poisson, Consul-Jain-Poisson, Conway-Maxwell-Poisson, Dacey-Poisson, Fucks-Poisson, Fucks-Gačečiladze-Poisson, Hirata-Poisson, hyper-Pascal, hyper-Poisson, log-normal (being continuous!), Meyer-Thomas, negative binomial, Palm-Poisson,

Pandey-Poisson, Poisson, Poisson-uniform, Pólya, Singh-Poisson, Feller's generalization of Poisson, cf. Popescu et al. 2013). The differences may represent boundary conditions depending on the entity, language, text-sort, style, age of the writer, education, etc. but they were not named in the literature because their finding presupposed the investigation of all these factors - a task for which one needs teams of scientists in every language. A thorough survey of the history of word length research can be found in Grzybek (2006).

Here we shall try to go another way. We ignore the requirement of considering the data as given in discrete or continuous form – a requirement depending on the way of measurement and the presentation of the data, e.g. in natural, real numbers, intervals or percentages – and start from the general assumption that the relative rate of change of the dependent variable (here the frequency) is proportional to the rate of change of the independent variable (here the length), that is

$$\frac{dy}{y} = \frac{g(x)}{h(x)}dx\,. \tag{1}$$

Hypotheses of this kind are quite reasonable. Here $g(x)$ consists of the parameter representing the status quo in the given language and the force of the speaker/writer controlling the changes, i.e. we write $g(x) = A + B \ln x$. One could take another function of x but the logarithm seems to be an acceptable solution. The speaker/hearer perceives the length intuitively; it is usually not possible for him to tell the length of a long unit without explicit counting. Making changes, i.e. performing self-organization, must be done slowly, not in jumps, otherwise the transfer of information could be disturbed. The function $h(x)$ should contain a plain function of x and capture the equilibrating effort of the hearer/community. We set $h(x) = Cx$. We obtain

$$\frac{dy}{y} = \frac{A + B\ln x}{Cx}dx. \tag{2}$$

Solving by integration and reparametrizing the constants we obtain

$$y = cx^{a+b\ln x} \tag{3}$$

which represents the Zipf-Alekseev function. As a matter of fact, it is a simple modification of Zipf's law. Formula (2) is a special case of the unified theory (cf. Wimmer, Altmann 2005). The function can be changed in a discrete distribution if one sets a finite range and considers c the normalizing constant. Another method is shown in Mačutek, Altmann (2007). For our purposes we shall work with a continuous function. A slightly more complex formula with four parameters is shown in Pande, Dhami (2012).

If we succeed in applying the formula to any level of linguistic entities, we arrive at an enormous simplification. Though simplification is a conscious deviation from the complex reality, one may attain with it one of the aims of science: easier orientation in the reality. The question of truth does not play here any role; no model of any complexity can warrant the reaching of truth. In all sciences there is a hierarchy of dependencies and none of the levels is the highest or lowest one. There is always the possibility to make a step up or down in the infinite chain of explanations.

2.Syllable length

Ignoring time as a measurement unit, the only possibility to measure syllable length is in terms of phonemes or phones. Though even here there are sometimes problems with distinguishing between "diphthong or two phonemes", e.g. in Slovak or Italian, or triphthongs in Roumanian, or affricates in different languages, the researchers have fixed criteria which may be accepted. Thus we apply the given model to syllable length in several languages. We do not present the raw data. The interested reader can find it in the cited references.

Except for Maori, Marquesan and Roumanian (given in the Appendix) we used only published results in order to enable the reader to check each result. The tables contain the parameters of function (3) and the determination coefficient R^2. The origin and the names of texts are given in the cited references.

Table 2.1
Modeling syllable length in German press texts

Syllable length (data from Cassier 2001)				
Text	**a**	**b**	**c**	**R^2**
T 1	9.2606	-5.4356	1.5107	0.9994
T 2	16.8762	-9.2407	0.0712	0.9953
T 3	12.5753	-6.7749	0.3486	0.9975
T 4	10.5977	-5.9235	0.8024	0.9966
T 5	15.1357	-8.0457	0.1128	0.9830
T 6	13.7559	-7.5137	0.2727	0.9981
T 7	11.7848	-6.4453	0.6276	0.9955
T 8	13.4224	-7.4223	0.3064	0.9946
T 9	12.5148	-6.3462	0.1720	0.9847
T 10	16.3054	-8.4620	0.0439	0.9962
T 11	16.9044	-9.0909	0.0566	0.9966
T 12	16.4111	-9.1428	0.0659	0.9957
T 13	15.2507	-8.4026	0.1381	0.9951
T 14	17.6345	-9.3400	0.0258	0.9971
T 15	13.9206	-7.3693	0.1801	0.9954
T 16	14.4660	-7.5969	0.0817	0.9890
T 17	10.0960	-5.6897	1.1181	0.9870
T 18	15.5145	-8.2367	0.0862	0.9951
T 19	12.5508	-6.9233	0.4136	0.9970
T 20	9.5152	-5.5244	1.4214	0.9990
T 21	8.3862	-4.8160	2.5064	0.9999

The results are very satisfactory. The relation between the parameters *a* and *b* is visualized in Figure 2.1. The existence of this link is a sign of self-regulation.

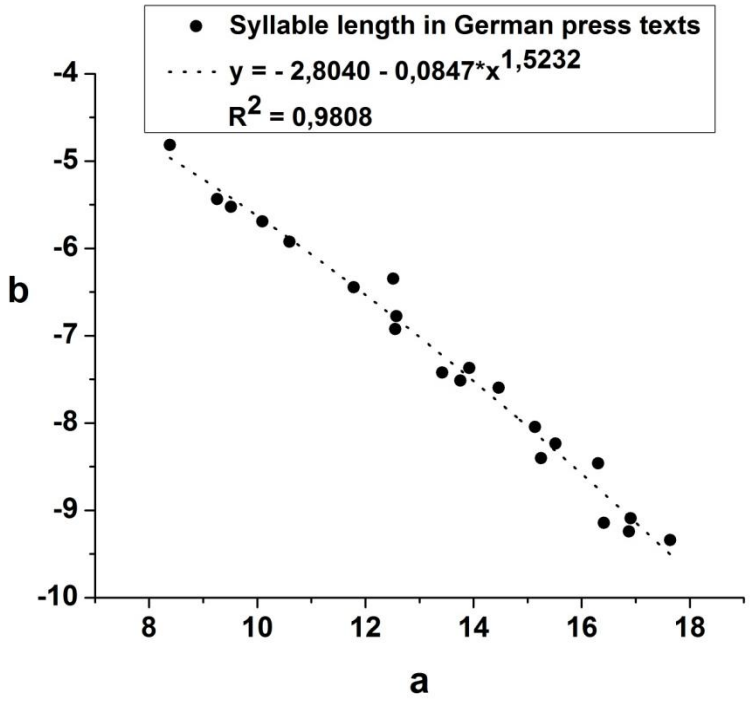


Figure 2.1. Syllable length in phonemes (German press texts)

German prose texts were taken by Best (2010) from *Sudelbücher* by G. Chr. Lichtenberg. The fitting is presented in Table 2.2. We took merely those texts that were analyzed by Best (2010).

Table 2.2
Modelling syllable length in German prose texts

Syllable length (data from Best 2010)				
Text	**a**	**b**	**c**	**R^2**
H 10	14.5982	-7.9794	0.1418	0.9994
H 13	10.5689	-5.8661	0.8214	0.9869
H 14	18.1982	-9.5640	0.0220	0.9998
H 15	11.9794	-6.7084	0.4607	0.9974
H 19	20.9595	-11.4814	0.0157	0.8623
H 52	19.6104	-10.4844	0.0140	0.9989
H 53	13.2422	-7.3827	0.3303	0.9980
H 66	14.0940	-7.7383	0.2571	0.9871
H 125	19.0812	-10.4190	0.0339	0.9927
H 134	14.7798	-7.9040	0.1333	0.9863
H 135	16.0346	-8.6397	0.0521	0.9929

H 138	15.1169	-7.9953	0.0800	0.9925
H 146	8.7619	-5.3999	3.6315	0.9971
H 147	19.6072	-10.9089	0.0185	0.9921
H 148	17.5328	-9.6176	0.0525	0.9873
H 150	16.8083	-8.9238	0.1453	0.9805
H 151	13.4051	-7.3874	0.5556	0.9958
H 155	8.0589	-4.9632	5.2676	0.9946
H 181	16.9831	-9.0555	0.0665	0.9873
H 191	14.5536	-7.9859	0.1411	0.9690

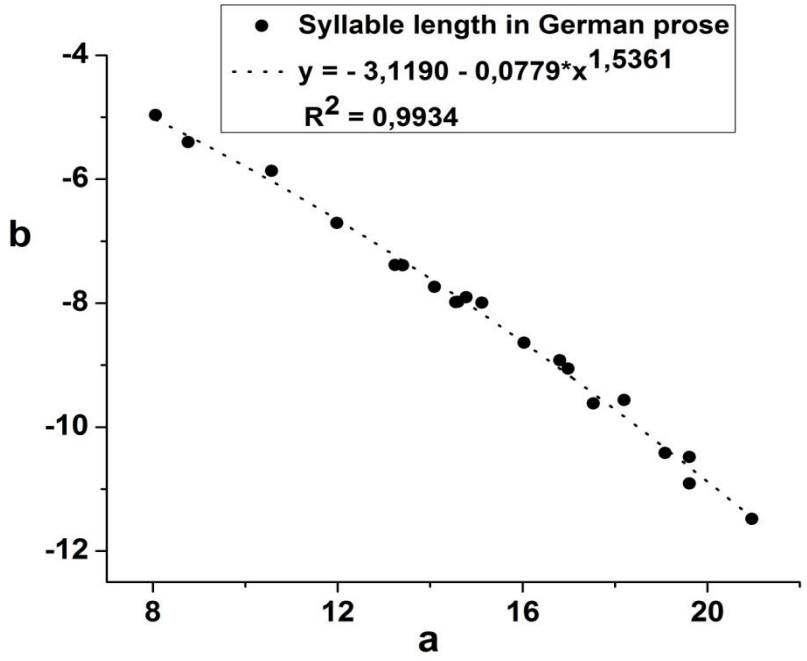


Figure 2.2. Syllable length in German (prose texts)

Taking the two tables together we obtain the result as presented in Figure 2.3. Evidently, increasing the sample would lead to a further smoothening of the results.

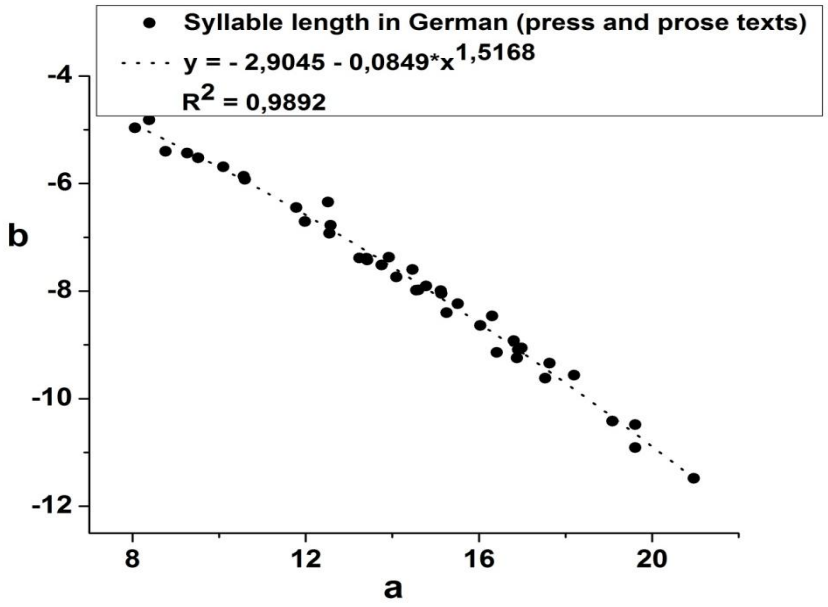


Figure 2.3. Syllable length in German (press and prose texts)

Data concerning syllable length can be obtained for each language mechanically. There are different ready-made programs but for our purposes the given corroboration of the supposed law is sufficient. As compared with other units (see below), the parameters *a* are very large and *b* very small. This is, perhaps, the first hint to the difference in self-regulation at the phonetic and grammatical levels.

3. Morph length

Measuring morph length, there is only one possibility: to count its phonemes or phones. Though morphemes themselves are semantic entities, their material appearance is given by morphs expressed by phonemes or phones. We do not consider the grammatical categories expressed by morphemes because these are purely semantic entities and may be represented also by a zero morpheme. Below, we use the data from German prose and press texts (Best 2000).

There is a problem of counting concerning introflection. In most cases it merely changes the morpheme to a special morph, e.g. German *sprech-en, ge-sproch-en* but in Semitic languages there are alternative possibilities: to take the root and the inflection as a whole, or to consider both separately. This must be decided depending on the aim of the analysis.

Table 3.1
Morph lengths in German prose

Morph length (data from Best 2000)				
Text	**a**	**b**	**c**	**R^2**
T 1	2.1396	-2.0086	60.6850	0.9913
T 2	1.4776	-1.6866	79.8292	0.9938
T 3	1.9233	-1.7456	130.4088	0.9789
T 4	1.7846	-1.7427	62.1041	0.9894
T 5	1.6673	-1.7614	134.9152	0.9966
T 6	1.2803	-1.2564	49.2696	0.9837
T 7	2.1746	-2.3411	75.6088	0.9872
T 8	2.3770	-1.9066	30.9205	0.9980
T 9	1.0418	-1.1949	54.4637	0.9811
T 10	1.7028	-1.6695	45.9020	0.9626
T 11	1.6676	-1.6470	70.6933	0.9964
T 12	1.2884	-1.4877	65.3397	0.9706
T 13	1.6903	-1.6643	48.5775	0.9919
T 14	1.5124	-1.6133	130.4373	0.9903
T 15	1.6379	-1.7030	68.9756	0.9827
T 16	2.1366	-1.8907	50.3956	0.9877
T 17	1.5246	-1.6042	79.6040	0.9873
T 18	1.1050	-1.3229	99.6505	0.9429

Table 3.2
Morph length in German press texts

Morph length (data from Best 2001)				
Text	**a**	**b**	**c**	**R^2**
T 1	1.1798	-1.2483	23.4041	0.9563
T 2	1.9094	-1.8762	28.6498	0.9470
T 3	2.5971	-2.1052	17.6476	0.9939
T 4	1.8879	-1.4662	13.5822	0.7591
T 5	2.8409	-2.2976	35.1557	0.9961
T 6	1.7849	-1.6520	34.9830	0.9402
T 7	1.9243	-1.7352	44.5549	0.9940
T 8	1.6434	-1.5859	30.0327	0.9922
T 9	2.0302	-1.7857	22.9729	0.9772
T 10	2.4320	-1.9111	22.1827	0.9854
T 11	1.7973	-1.6059	27.3052	0.9843
T 12	1.4277	-1.4714	68.2962	0.9727
T 13	2.2257	-2.1029	56.3698	0.9916
T 14	1.8091	-1.7109	49.0858	0.9974
T 15	2.1226	-1.8146	39.6425	0.9917
T 16	1.2502	-1.3608	59.4852	0.9671
T 17	1.4936	-1.4243	34.7536	0.9833
T 18	2.1423	-1.8671	22.3135	0.9910
T 19	1.9345	-1.7528	19.6459	0.9136
T 20	2.1888	-1.7247	28.5417	0.9188
T 21	1.8692	-1.6100	38.7842	0.9831

Putting the two tables together we obtain the relationship of the two parameters as presented in Figure 3.1.

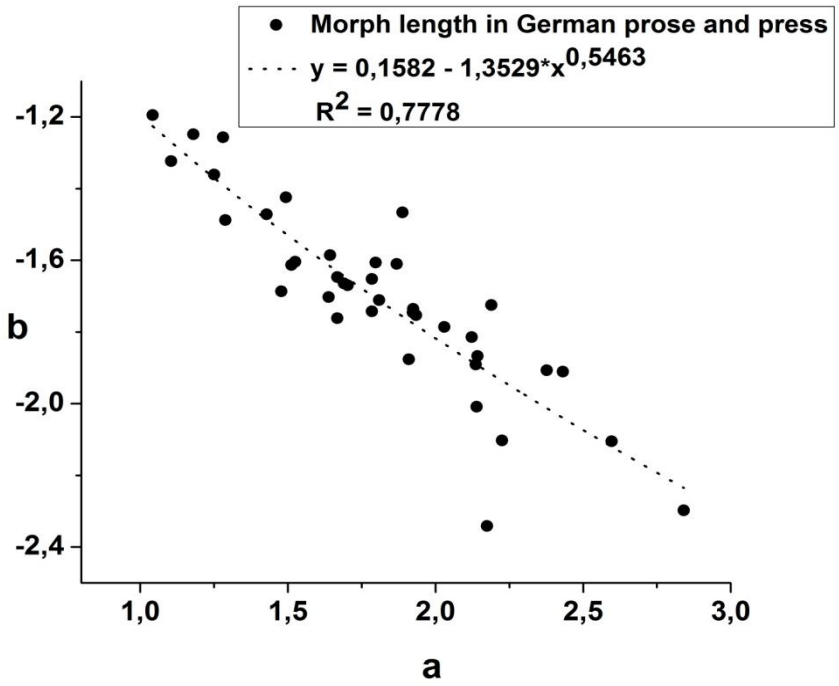


Figure 3.1. Morph length relations from Tables 3.1 and 3.2

The case of morph length in Lakota is quite complex. It has been shown (cf. Pustet, Altmann 2005) that the distribution is multimodal, a fact that can be substantiated by the grammar and captured by the Gegenbauer distribution. That means, there is a boundary condition which must be taken into account but in that case, we would obtain a second order differential equation. For our present purpose we omit this procedure. Here grammar/semantics and the pure material form do not display a simple coordination.

4. Word length

As shown above, word is a formal, grammatical and semantic entity, hence its length can be measured either in terms of syllable, mora, morpheme, morph or sememe numbers. The literature is extremely rich.

Here we shall check the validity of (3) using very extensive data taken from literature. For each of the data, we compute the parameters and the determination coefficient and study merely the form of the parameters and their development. The results are presented in Table 4.2 and following. The languages are not ordered, a kind of ordering is shown in Table 4.1. Unfortunately, the table shows merely those languages for which data could be found.

Below, we add the image of the dependence of the parameter *b* on parameter *a*. Here the self-regulation will be quite evident.

Table 4.1
A kind of ordering of the languages used

Language family	**Subfamily**	**Language**	**Language periods**	**Dialects**
Andean		Quechua		
Sino-Tibetan		Chinese		
Eskimo-Aleut		Inuktitut		
Finno-Ugric				
	Balto-Finnic			
		Estonian		
		Finnish		
	Sami			
	Ugric			
		Hungarian		
		Vogul/Mansi		
	Volgaic			
		Cheremis/Mari		
		Erzja-Mordvin		
Indo-European				
	Baltic			
		Latvian		
	Celtic			
		Gaelic		
		Welsh		
	Germanic			
		Dutch		

		English		
			Early Modern English	
			Modern English	
		Faeroese		
		German		
			Old High German	
			Middle High German	
			Early New High German	
			New High German	
				Palatine
		Gothic		
		Icelandic		
			Old Icelandic	
			Modern Icelandic	
		Low German		
		Swedish		
	Greek			
		Greek Koine		
	Latin			
	Romanic			
		French		
		Italian		
		Portuguese		
		Roumanian		
		Spanish		
	Slavic			
		Belorussian		
		Bulgarian		
		Czech		
		Low Sorbian		
		Old Church Slavonic		
		Polish		
		Russian		
		Slovak		
		Slovenian		
		Ukrainian		

Korean				
Japanese				
Polynesian				
		Maori		
		Marquesan		
Semitic				
		Arabic		
		Old Hebrew		
Turkic				
		Turkish		
		Uzbek		

As can be seen in this table, the research has not been performed systematically. Every researcher used the language for which he was specialized and had sufficient material in order to create data. There is much space for future research.

Visualizing the link between the parameters *a* and *b* we shall shift *a* in all cases into the domain of positive values. This technique does not change the link.

Andean: Quechua

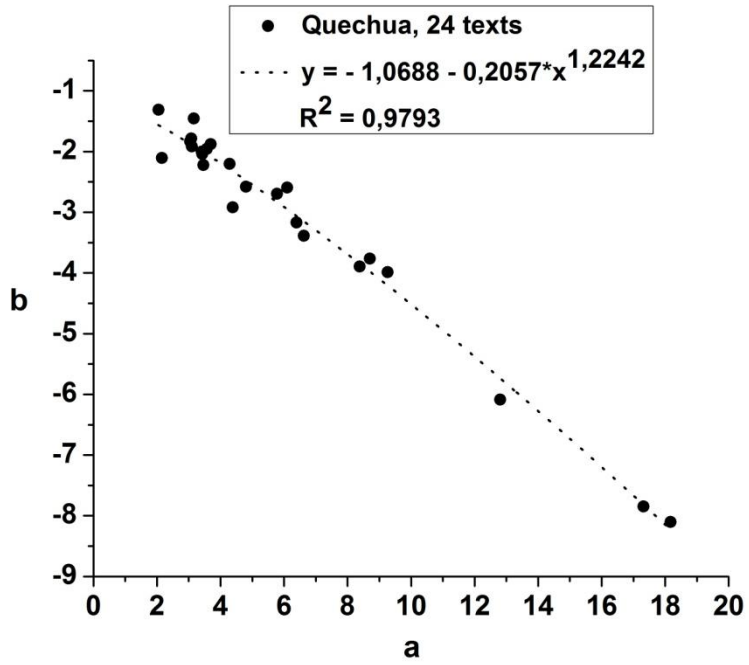
Table 4.2
Fitting the Zipf-Alekseev function to word length distribution in Quechua

Quechua (data from Best, Medrano 1997)					
Text	**a**	**b**	**c**	**k**	**R^2**
Text 1	3.6959	-1.8765	3.4293	3	0.9101
Text 2	8.7001	-3.7633	0.1033	3	0.9860
Text 3	9.2595	-3.9859	0.0754	-	0.8813
Text 4	17.3189	-7.8481	0.0027	3	0.9923
Text 5	12.8088	-6.0885	0.0468	3	0.9737
Text 6	18.1682	-8.1042	0.0017	3	0.8903
Text 7	6.6237	-3.3891	1.2564	3	0.9814
Text 8	4.3877	-2.9198	5.2377	-	0.8309
Text 9	3.0938	-1.9150	9.4058	5	0.8600
Text 10	3.1610	-1.4540	2.8596	5	0.9401
Text 11	2.0473	-1.3102	5.7185	-	0.8445
Text 12	3.4656	-2.2242	4.9201	4	0.8755
Text 13	2.1563	-2.1049	14.0110	2	0.9790
Text 14	6.3962	-3.1701	2,8264	-	0.9968
Text 15	3.4054	-2.0027	9.1172	-	0.8347
Text 16	6.1037	-2.5933	1.0887	-	0.9521

In some Quechua texts there is a local minimum in the mid of the distribution which cannot be captured by the above formula. The given class must be expressed by a special part of the formula, e.g.

$$(4) \qquad f(x) = \begin{cases} cx^{a+b\ln x}, & if \;\; x \neq k \\ \alpha, & if \;\; x = k \end{cases}$$

where *k* is a length class which may differ in individual texts. Since for the given class the difference between observed and computed (i.e. substituted α) value is zero, the determination coefficient does not change. For the individual texts, the class (*k*) is given in the table. In most cases *k* = 3. Because of this local minimum, Best and Medrano (1997) use a modified Hyperpoisson distribution. The given frequencies display a bell-shaped distribution with the given local minimum. It is possible that this phenomenon occurs only in short texts.

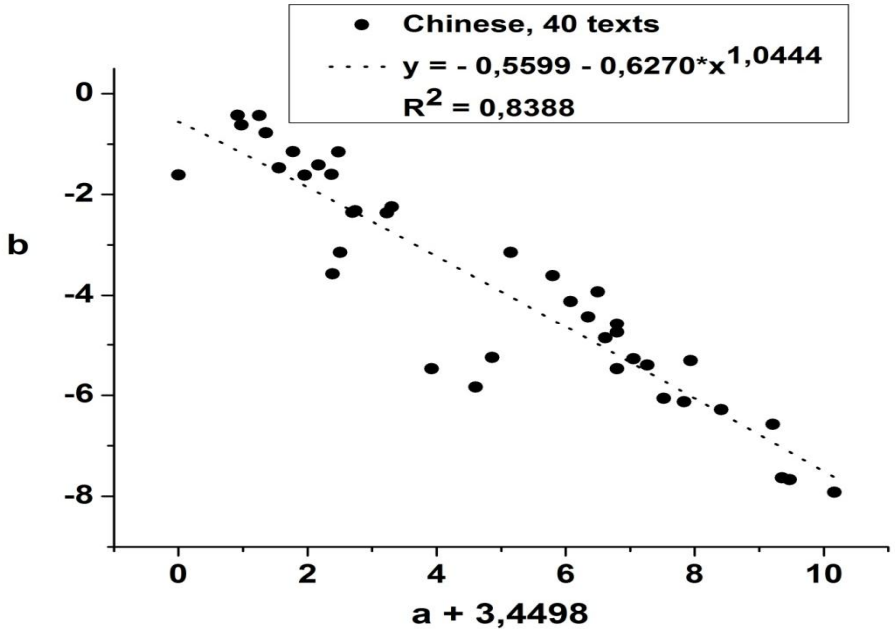


Sino-Tibetan: Chinese

Chinese (data from Zhu, Best 1997)				
Text	**a**	**b**	**c**	**R^2**
Text 2	3.3447	-5.4742	56.0000	1.0000
Text 3	6.7129	-7.9213	36.0006	0.9990
Text 4	4.0732	-6.0615	59.0031	0.9914
Text 5	2.8976	-4.4307	53.0012	0.9999
Text 6	3.3466	-4.5739	61.0125	0.9972

Text 7	4.4850	-5.3153	31.0000	1.0000
Text 8	6.0232	-7.6718	19.0031	0.9265
Text 9	3.3442	-4.7485	26.0050	0.9948
Text 10	5.7602	-6.5767	20.0000	1.0000
Text 11	3.0475	-3.9322	28.0000	1.0000
Text 12	5.9050	-7.6324	32.0008	0.9977
Text 13	4.9625	-6.2811	34.0001	0.9993
Text 14	2.3505	-3.6112	30.0016	0.9999
Text 15	2.6269	-4.1236	81.0004	0.9999
Text 16	3.1625	-4.8609	45.0013	0.9997
Text 17	3.8137	-5.3991	78.0073	0.9962
Text 18	4.3832	-6.1262	30.0025	0.9900
Text 19	3.6024	-5.2769	55.0102	0.9807
Text 20	1.6995	-3.1470	25.0123	0.9926
Data from Best, Zhu 2001				
Text 1	-2.0946	-0.7756	576.9977	1.0000
Text 2	-0.7103	-2.3246	674.9990	1.0000
Text 3	-1.8917	-1.4718	797.0025	1.0000
Text 4	1.1567	-5.8396	675.0000	1.0000
Text 5	-3.4498	-1.6098	900.0000	1,0000
Text 6	1.4093	-5.2482	239.0000	1.0000
Text 7	-1.0608	-3.5729	267.0001	1.0000
Text 8	-1.0608	-3.5729	267.0001	1.0000
Text 9	-2.4728	-0.6225	352.9962	1.0000
Text 10	-0.1449	-2.2433	178.9944	0.9998
Text 11	-0.9686	-1.1565	164.9802	0.9993
Text 12	-0.2185	-2.3633	181.0012	1.0000
Text 13	-1.0750	-1.6037	172.0087	0.9993
Text 14	-1.2807	-1.4134	177.9942	0.9997
Text 15	-0.7533	-2.3558	162.0003	1.0000
Text 16	-1.6726	-1.1490	875.0318	0.9996
Text 17	-2.1940	-0.4312	787.0385	0.9998
Text 18*	-2.5338	-0.4290	763.9903	1.0000
Text 19	-1.4944	-1.6156	832.0038	1.0000
Text 22	0.4724	-5.4740	880.0000	1.0000
Text 23	-0.9449	-3.1456	890.0001	1.0000

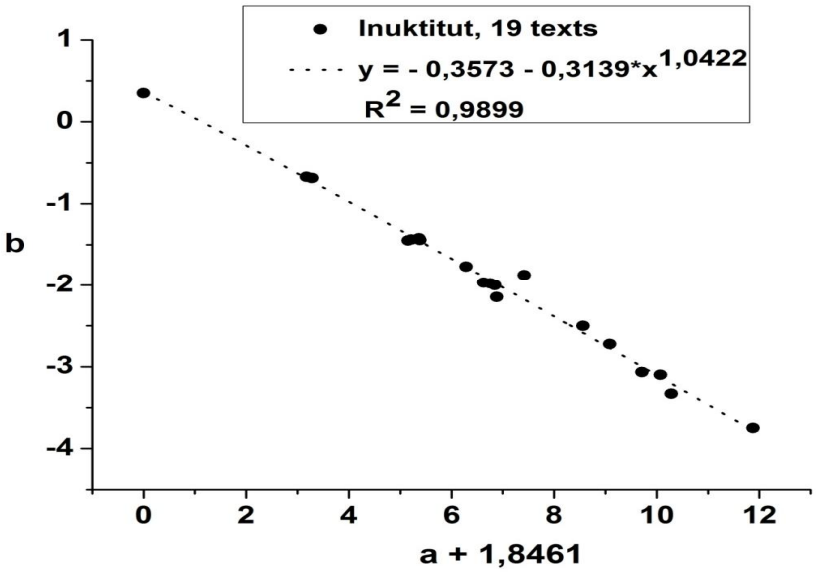
Text 18* is a mixture of two letters. Text 1 in Zhu, Best (1997) and Texts 20 and 21 in Best, Zhu (2001) contain merely 2 classes.



Eskimo: Aleut: Inuktitut

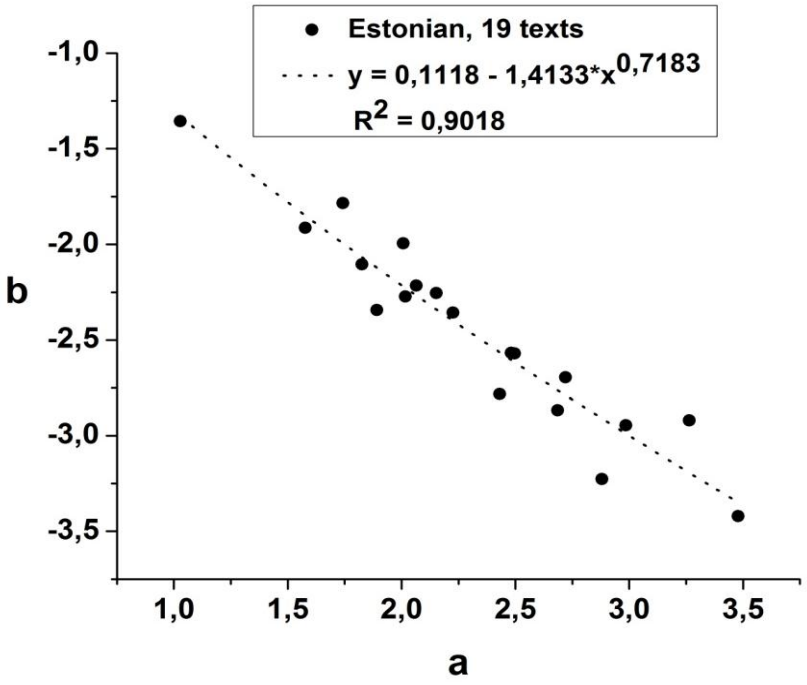
Inuktitut (data from Meyer 1997)				
Text	**a**	**b**	**c**	**R^2**
Text 1	7.2361	-2.7244	0.4083	0.9857
Text 2	5.5704	-1.8743	0.1562	0.5115
Text 3	4.9133	-1.9775	1.2330	0.8315
Text 5	7.8669	-3.0667	0.1549	0.9275
Text 6	8.2244	-3.1002	0.0857	0.9916
Text 7	6.7162	-2.5011	0.2177	0.9023
Text 8	10.0275	-3.7485	0.0411	0.9246
Text 9	8.4359	-3.3288	0.1254	0.9708
Text 10	1.3287	-0.6694	10.8123	0.8923
Text 11	3.5173	-1.4213	2.0869	0.7560
Text 12	1.4365	-0.6868	8.7682	0.6524
Text 13	3.3081	-1.4488	4.1634	0.7970
Text 14	3.5373	-1.4428	3.8216	0.8813
Text 15	4.7816	-1.9698	0.9793	0.8289
Text 16	4.9980	-1.9975	1.0144	0.9362
Text 17	-1.8461	0.3529	68.2935	0.5529
Text 18	4.4387	-1.7709	1.2246	0.8953
Text 19	3.3686	-1.4364	1.4357	0.2982
Text 21	5.0349	-2.1452	3.3814	0.8227

In Inuktitut, Text 2 and Text 20 in Meyer (1997) are missing. In three texts (2, 17, 19), the given model does not fit sufficiently. The cause must be sought by specialists.



Finno-Ugric: Balto-Finnic: Estonian

Estonian (data from Bartens, Best 1996) (L = Letters; P = Prose)				
Text	**a**	**b**	**c**	**R^2**
L 1	2.6854	-2.8676	242.1009	0.9999
L 2	2.4317	-2.7826	131.2663	0.9980
L 3	3.4779	-3.4200	62.3156	0.9954
L 4	2.8806	-3.2271	94.2335	0.9943
L 5	2.0171	-2.2731	78.0197	0.9996
L 6	2.4823	-2.5673	63.2901	0.9936
L 7	1.7429	-1.7846	51.1078	0.9985
L 8	1.8922	-2.3436	60.9204	0.9991
L 9	2.9854	-2.9452	33.1100	0.9984
L 10	2.0653	-2.2159	88.0509	0.9658
P 1	2.1522	-2.2546	258.5455	0.9823
P 2	1.5768	-1.9133	256.5598	0.9963
P 3	1.8263	-2.1045	320.9233	0.9994
P 4	2.4969	-2.5701	434.5408	0.9996
P 5	2.7200	-2.6946	248.8283	0.9997
P 6	2.0081	-1.9949	113.5174	0.9963
P 7	1.0283	-1.3556	203.6699	0.9820
P 8	3.2635	-2.9195	63.5321	0.9978
P 9	2.2257	-2.3573	68.0265	0.9598

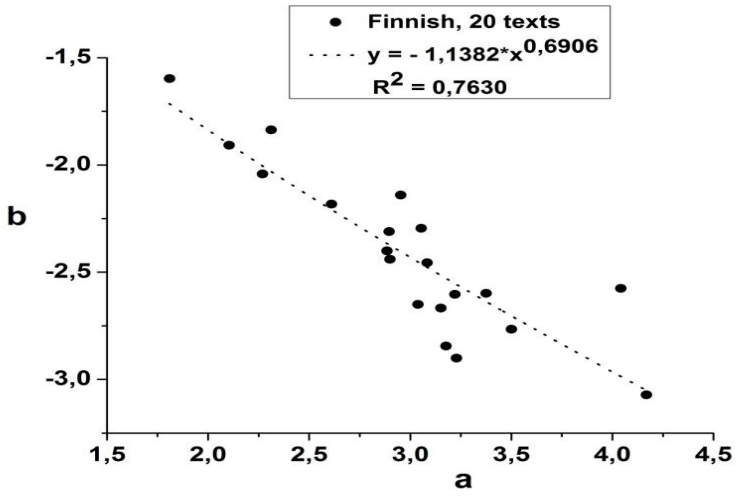


Finnish

Finnish (data from Müller 2003) [Letters, E-mails]				
Text	**a**	**b**	**c**	**R^2**
L 1	3.1525	-2.6673	139.1631	0.9955
L 2	3.3761	-2.5983	39.3097	0.9985
L 3	2.8851	-2.4010	33.9888	0.9859
L 4	1.8095	-1.5969	20.1138	0.9881
L 5	3.0549	-2.2956	68.3565	0.9938
L 6	2.3130	-1.8363	46.1428	0.9952
L 7	3.0393	-2.6501	161.5140	0.9979
L 8	2.8955	-2.3110	195.7951	0.9995
L 9	3.5011	-2.7663	81.6073	0.9936
L 10	2.6113	-2.1826	225.5978	0.9995
L 11*	3.2288	-2.9005	52.7311	0.9958
L12	4.1688	-3.0717	6.2425	0.9910
L 13	2.9007	-2.4402	22.6884	0.9843
L 14	2.9531	-2.1407	13.6464	0.9895
L 15	3.0839	-2.4560	101.2387	0.9924
L 16	2.2703	-2.0416	31.9975	0.9857
L 17	4.0419	-2.5760	2.7755	0.9926
L 18	2.1041	-1.9084	65.2719	0.9960
L 19	3.1769	-2.8443	27.1715	0.9958
L 20	3.2214	-2.6029	26.9988	0.9942

Data L 11 was identical with L 13 in the given article; we used a different Finnish set L 11* written in a letter by a native Finnish student. The numbers are as follows:

1	2	3	4	5	6	7
53	122	58	12	8	3	1

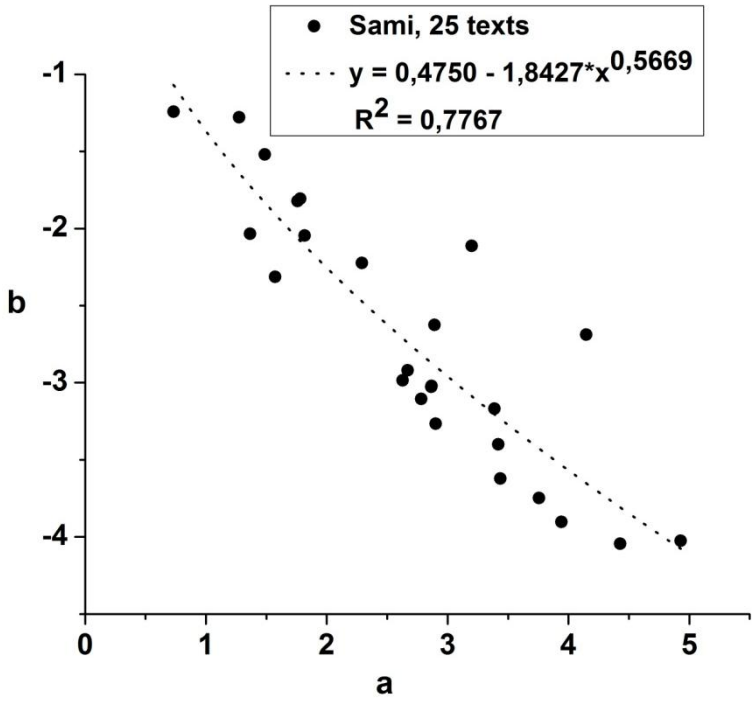


Sami

Sami (data from Bartens, Best 1997)				
Text	**a**	**b**	**c**	**R^2**
Text 1	1.8180	-2.0454	38.9030	0.9962
Text 2	1.7573	-1.8211	39.6999	0.9907
Text 3	1.5717	-2.3142	122.9529	0.9994
Text 4	2.6288	-2.9840	47.1369	0.9954
Text 5	1.3660	-2.0337	79.0492	0.9993
Text 6	4.1483	-2.6888	6.6867	0.9239
Text 7	4.4282	-4.0447	14.3782	0.9448
Text 8	4.9297	-4.0256	9.2392	0.9915
Text 9	0.7319	-1.2431	28.0597	0.9933
Text 10	2.9023	-3.2651	29.1685	0.9703
Text 11	3.9415	-3.9022	198.2361	0.9815
Text 12	3.7559	-3.7471	167.0858	0.9633
Text 13	3.4192	-3.3996	207.8648	0.9849
Text 14	2.8639	-3.0268	302.2703	0.9941
Text 15	3.4364	-3.6219	347.6177	0.9897
Text 16	1.7805	-1.8060	83.7897	0.9193
Text 17	1.2754	-1.2779	85.6934	0.8143
Text 18	1.4874	-1.5201	69.9306	0.9347
Text 19	2.2916	-2.2229	85.9449	0.9385
Text 20 (k=3)	3.1983	-2.1129	24.0411	0.9715

Text 21	2.6684	-2.9195	58.4737	0.9740
Text 22	2.8915	-2.6258	57.1572	0.9841
Text 23	2.8681	-3.0211	128.4124	0.9685
Text 24	3.3882	-3.1693	101.5748	0.9479
Text 25	2.7829	-3.1053	190.0998	0.9818

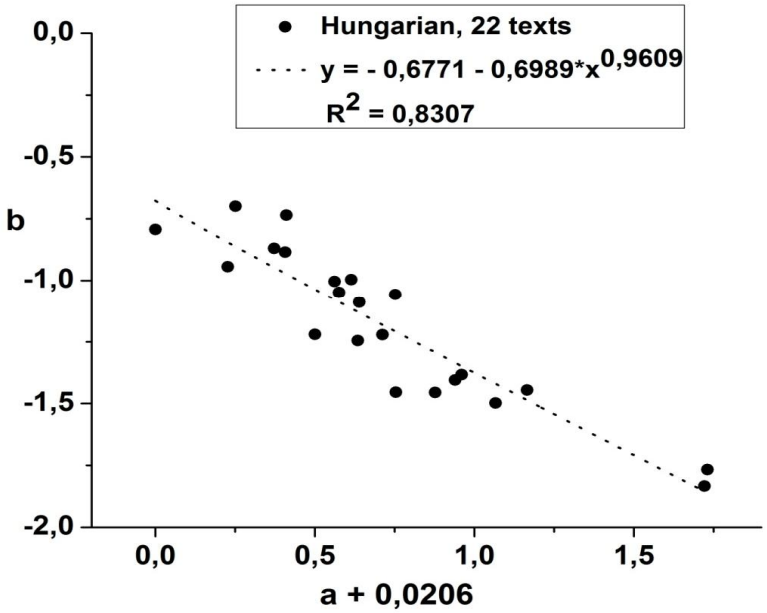
Some Sami texts display a local minimum in the middle of the distribution but it can be captured with the usual formula.



Ugric: Hungarian

Hungarian (data from Bartens, Zöbelin 1997)				
Text	**a**	**b**	**c**	**R^2**
Text 1	1.7099	-1.7674	80.8598	0.9787
Text 2	1.7005	-1.8332	67.6254	0.9854
Text 3	1.1443	-1.4456	27.9876	0.9602
Text 4	0.6912	-1.2212	79.7800	0.9352
Text 5	-0.0206	-0.7928	42.7749	0.9403
Text 6	0.8567	-1.4550	120.3630	0.9896
Text 7	0.9190	-1.4053	38.7302	0.9846
Text 8	0.5938	-0.9965	73.0715	0.9233
Text 9	0.4793	-1.2202	116.3399	0.9758
Text 10	0.5412	-1.0046	57.1171	0.9941
Text 11	0.3516	-0.8696	64.5796	0.8811

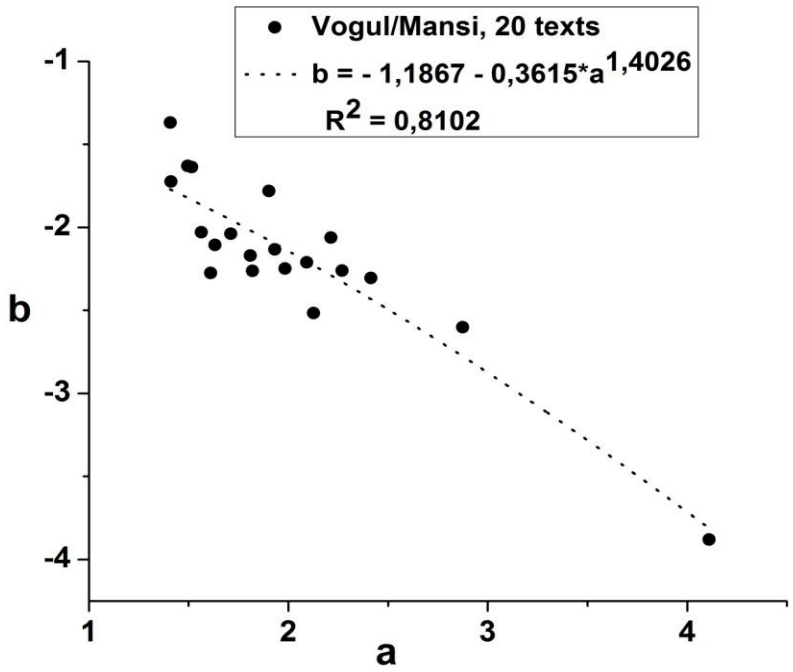
Text 12	1.0460	-1.4977	241.2676	0.9653
Text 13	0.7329	-1.4545	694.3299	0.9963
Text 14	0.6135	-1.2459	328.0364	0.9875
Text 15	0.9391	-1.3838	284.0703	0.9771
Text 16	0.6182	-1.0893	307.3109	0.9657
Text 17	0.7311	-1.0578	281.8609	0.9642
Text 18	0.2059	-0.9441	209.1957	0.9720
Text 19	0.3867	-0.8852	272.6841	0.9733
Text 20	0.5551	-1.0488	428.9556	0.9884
Text 21	0.3899	-0.7357	304.1261	0.9616
Text 22	0.2309	-0.6993	247.8644	0.9425



Vogul/Mansi

Vogul/Mansi (data from Kahl 2002)				
Text	**a**	**b**	**c**	**R^2**
T 1	1.7122	-2.0383	160.5271	0.9881
T 2	1.8208	-2.2623	110.1586	0.9967
T 3	1.6321	-2.1054	106.4005	0.9775
T 4	2.2695	-2.2611	77.1032	0.9860
T 5	2.1267	-2.5165	237.9392	0.9995
T 6	1.4953	-1.6295	111.1376	0.9893
T 7	1.6107	-2.2740	111.3664	0.9706
T 8	2.4136	-2.3044	33.1715	0.9871
T 9	1.4083	-1.3692	35.0093	0.7764

T 10	2.8736	-2.6017	32.3913	0.9909
T 11	1.9317	-2.1329	98.4976	0.9940
T 12	1.8095	-2.1691	90.4714	0.9699
T 13	2.0917	-2.2111	188.9252	0.9828
T 14	2.2143	-2.0614	154.3277	0.9782
T 15	1.5160	-1.6364	98.3659	0.9945
T 16	1.9035	-1.7815	35.5063	0.9813
T 17	1.4113	-1.7248	81.9140	0.9852
T 18	1.9836	-2.2473	91.3263	0.9899
T 19	4.1105	-3.8800	120.9921	0.9930
T 20	1.5639	-2.0288	210.4112	0.9967

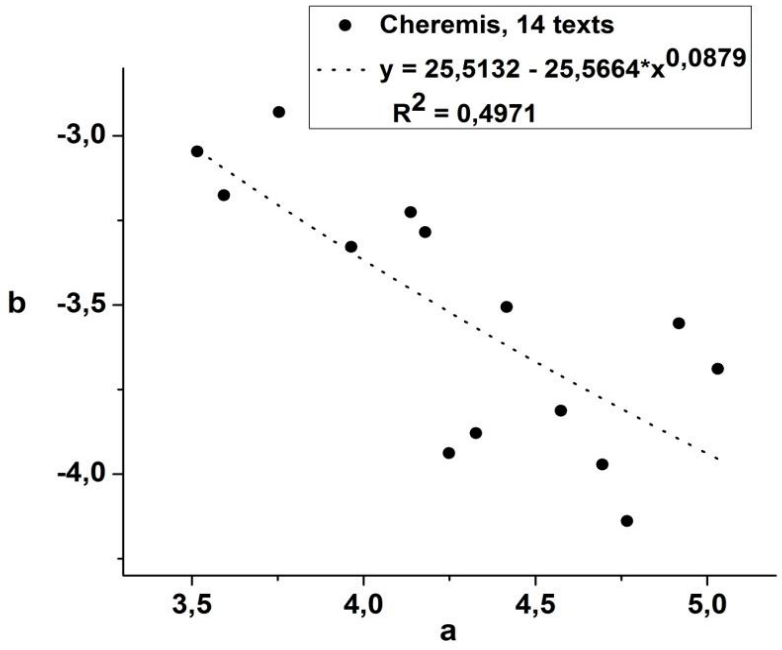


Volgaic: Cheremis/Mari

Cheremis (data from Bartens, Best 1997a)				
Text	**a**	**b**	**c**	**R^2**
T 1	3.7537	-2.9301	14.4815	0.9916
T 2	5.0308	-3.6894	25.3255	0.9930
T 3	4.9171	-3.5549	26.8582	0.9949
T 4	4.5739	-3.8133	105.6672	0.9980
T 5	4.4162	-3.5067	68.7478	0.9993
T 6	3.9639	-3.3288	103.9502	0.9983
T 7	3.5161	-3.0468	67.5901	0.9844

T 8	4.7665	-4.1390	26.8854	0.9992
T 9	3.5937	-3.1762	101.6176	0.9994
T 11	4.6951	-3.9719	70.3879	0.9993
T 12	4.2479	-3.9387	62.0944	0.9996
T 13	4.3268	-3.8793	57.3313	0.9982
T 14	4.1787	-3.2853	26.9320	0.9873
T 15	4.1365	-3.2260	86.9930	0.9977

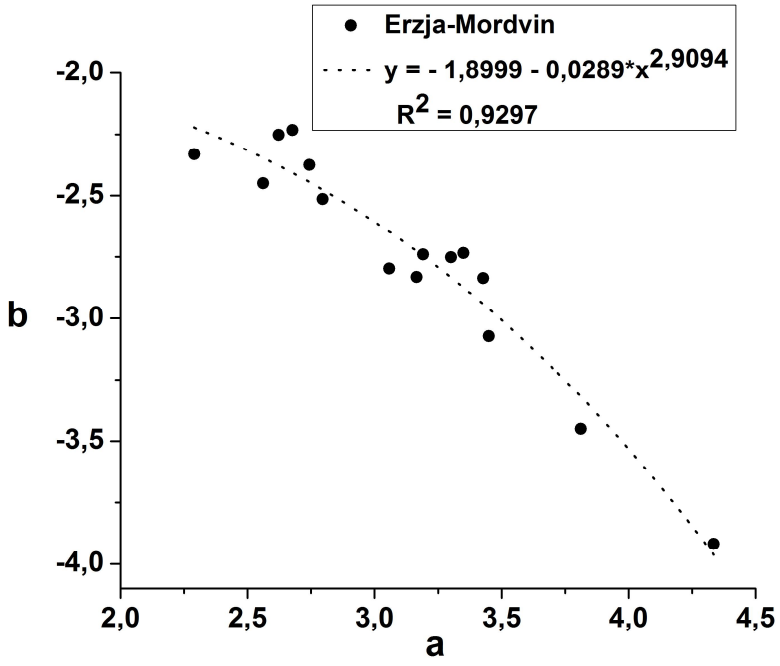
Text 10 and 11 are identical in the reference.



Erzja-Mordvin

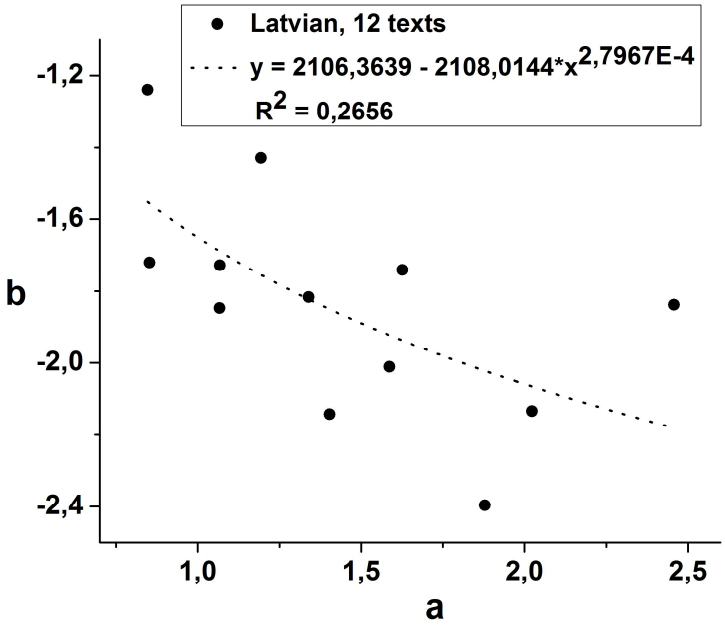
Erzja-Mordvin (data from Bartens, Best 1997)				
Text	**a**	**b**	**c**	**R^2**
T 1	2.7433	-2.3744	55.4445	0.9759
T 2	4.3353	-3.9181	38.0491	0.9999
T 3	3.0577	-2.7972	59.7199	0.9986
T 4	2.5619	-2.4497	53.4972	0.9947
T 5	3.4502	-3.0721	42.2294	0.9981
T 6	2.6770	-2.2324	37.3593	0.9581
T 7	2.7961	-2.5140	66.9645	0.9990
T 8	2.6222	-2.2521	72.8305	0.9846
T 9	2.2902	-2.3314	136.0134	0.9957
T 10	3.8119	-3.4501	51.3165	0.9964

T 11	3.3502	-2.7339	75.0036	0.9966
T 12	3.3008	-2.7512	48.8127	0.9930
T 13	3.4273	-2.8362	89.1187	0.9942
T 14	3.1912	-2.7401	98.0267	0.9979
T 15	3.1655	-2.8315	88.3660	0.9982



Indo-European: Baltic: Latvian

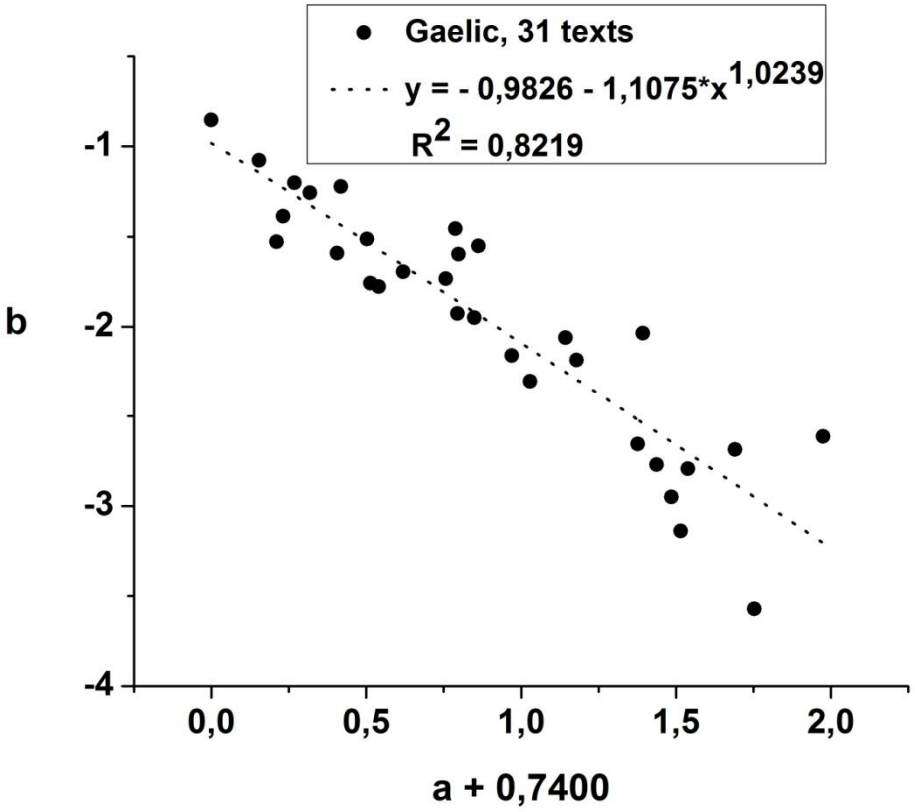
Latvian (data from Rottmann 2003)				
Text	**a**	**b**	**c**	**R^2**
T 1	2.0230	-2.1338	352.1521	0.9922
T 2	1.4037	-2.1429	555.0019	0.9957
T 3	1.0669	-1.8472	594.8423	0.9834
T 4	2.4573	-1.8384	183.1395	0.8442
T 5	1.1931	-1.4286	291.9258	0.9741
T 6	0.8470	-1.2392	316.7605	0.9785
T 7	1.3394	-1.8166	403.6860	0.9958
T 8	1.0681	-1.7279	398.6565	0.9946
T 9	1.8791	-2.3965	331.3196	0.9971
T 10	0.8527	-1.7206	448.2027	0.9954
T 11	1.6260	-1.7406	286.7612	0.9531
T 12	1.5871	-2.0113	301.7722	0.9822



Celtic: Gaelic

Gaelic (data from Drechsler 2001)				
Text	**a**	**b**	**c**	**R^2**
T 1	-0.2369	-1.5136	229.9107	0.9977
T 2	0.7449	-2.9488	295.0082	0.9998
T 3	1.0131	-3.5712	394.0006	1.0000
T 4	0.7752	-3.1390	132.0000	1.0000
T 5	-0.2259	-1.7575	207.9703	0.9992
T 6	0.6531	-2.0357	260.8593	0.9975
T 7	0.4386	-2.1844	162.0110	0.9998
T 8	-0.3217	-1.2219	70.8.3907	0.9944
T 9	-0.5292	-1.5285	240.9925	0.9999
T 10	0.7984	-2.7928	373.9990	1.0000
T 11	-0.1999	-1.7772	132.0046	0.9999
T 12	0.2893	-2.3023	212.9915	0.9999
T 13	-0.3345	-1.5914	344.9753	0.9998
T 14	0.6973	-2.7703	613.9965	1.0000
T 15	-0.7400	-0.8536	970.9146	0.9939
T 16	-0.5079	-1.3871	539.8688	0.9981
T 17	0.9501	-2.6847	144.9882	0.9996
T 18	0.4034	-2.0600	101.9877	0.9996
T 19	0.1228	-1.5516	1129.9242	0.9954
T 20	-0.5859	-1.0780	772.5528	0.9982

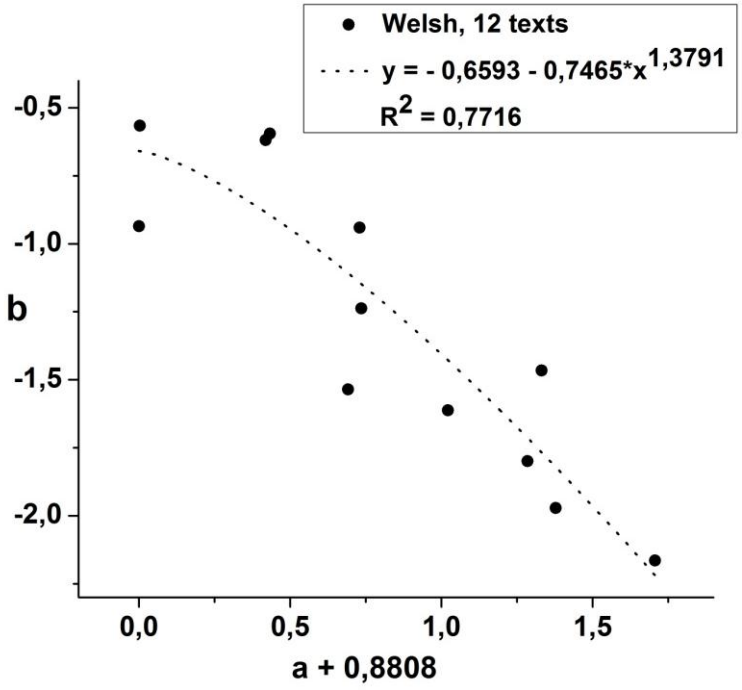
T 21	0.0169	-1.7331	677.8162	0.9986
T 22	-0.4216	-1.2576	651.0188	0.9996
T 23	-0.1213	-1.6951	955.9780	1.0000
T 24	0.0542	-1.9270	923.9161	0.9998
T 25	0.6368	-2.6550	559.9875	0.9999
T 26	-0.4722	-1.2015	820.4509	0.9961
T 27	0.2301	-2.1599	839.9968	1.0000
T 28	1.2351	-2.6133	384.9740	0.9999
T 29	0.0582	-1.5964	178.9501	0.9994
T 30	0.1090	-1.9492	576.9276	0.9996
T 31	0.0481	-1.4560	130.9739	0.9996



Welsh

Welsh (data from Wilson 2003)				
Text	**a**	**b**	**c**	**R^2**
T 1	-0.8808	-0.9352	201.9376	0.9982
T 2	-0.1894	-1.5355	200.9781	0.9997
T 3	0.8252	-2.1650	76.9773	0.9988
T 4	-0.8780	-0.5655	113.8342	0.9855

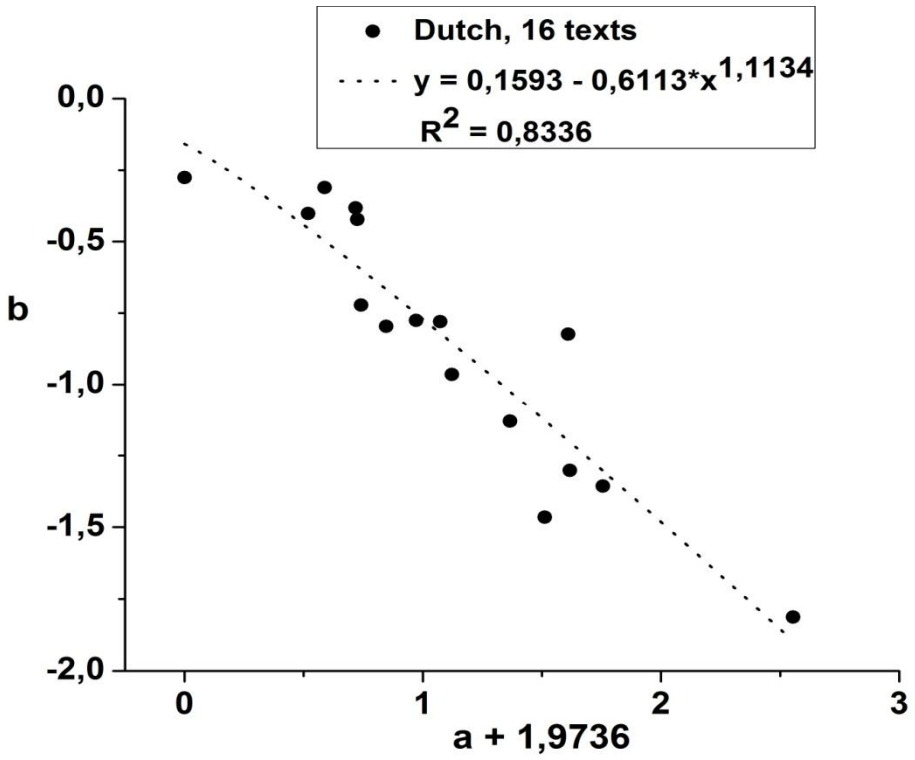
T 5	-0.1454	-1.2372	106.0110	0.9999
T 6	0.4503	-1.4661	123.6672	0.9864
T 7	0.4970	-1.9717	182.9854	0.9999
T 8	0.4035	-1.7998	178.9301	0.9981
T 9	0.1408	-1.6127	124.9544	0.9988
T 10	-0.4617	-0.6189	83.7618	0.9893
T 11	-0.1515	-0.9405	111.5513	0.9751
T 12	-0.4474	-0.5949	126.5200	0.9550



Gemanic: Dutch

Dutch (data from Rheinländer 2001)				
Text	**a**	**b**	**c**	**R^2**
T 1	-0.8505	-0.9643	138.9669	0.9994
T 2	-0.4611	-1.4650	69.9900	0.9993
T 3	-1.2320	-0.7225	79.9967	0.9999
T 4	-1.1269	-0.7963	96.9832	0.9994
T 5	0.5814	-1.8132	49.9683	0.9951
T 6	-0.3633	-0.8240	98.9262	0.9974
T 7	-0.2173	-1.3569	111.9842	0.9997
T 8	-1.2483	-0.4231	83.0304	0.9986
T 9	-1.0017	-0.7753	84.9840	0.9997
T 10	-1.2557	-0.3821	68.9190	0.9959

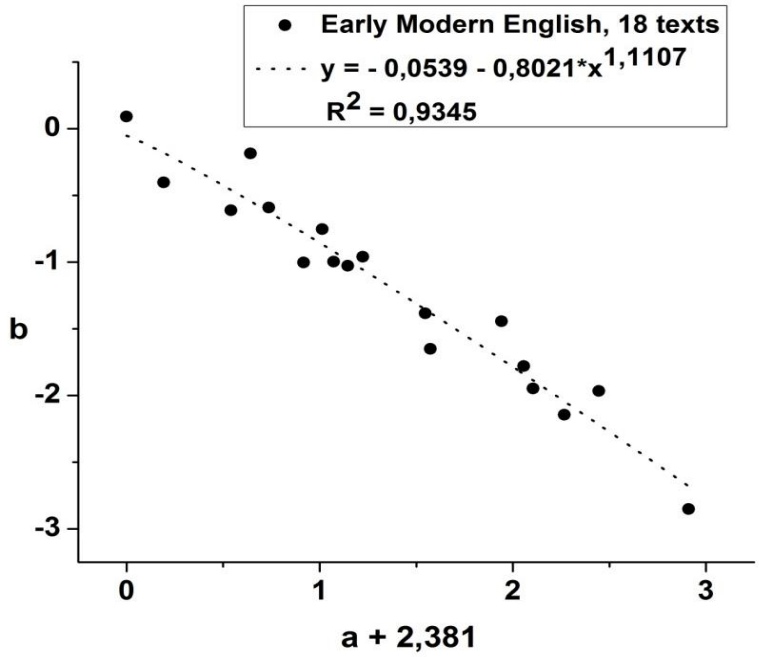
T 11	-0.3561	-1.3028	79.9675	0.9982
T 12	-0.9004	-0.7796	118.9409	0.9987
T 13	-1.9736	-0.2765	75.0114	0.9992
T 14	-1.4541	-0.4020	150.9139	0.9979
T 15	-0.6076	-1.1304	86.9873	0.9998
T 16	-1.3859	-0.3117	84.9404	0.9949



English: Early Modern English

Early Modern English (data from Zuse 1996)				
Text	**a**	**b**	**c**	**R^2**
L 1	-0.4396	-1.4448	197.0295	0.9993
L 2	-0.3238	-1.7796	208.0200	0.9993
L 3	-0.1150	-2.1444	157.9957	0.9999
L 4	-1.4636	-1.0032	210.9898	0.9998
L 5	-2.3810	0.0909	185.9522	0.9980
L 6	-1.7389	-0.1868	136.9739	0.9994
L 7	0.5286	-2.8527	161.0014	1.0000
L 8	-0.2768	-1.9492	166.9979	1.0000
L 9	-0.8082	-1.6514	301.9996	1.0000
L 10	0.0635	-1.9664	160.0000	1.0000
L 11	-1.3091	-0.9965	167.9998	1.0000

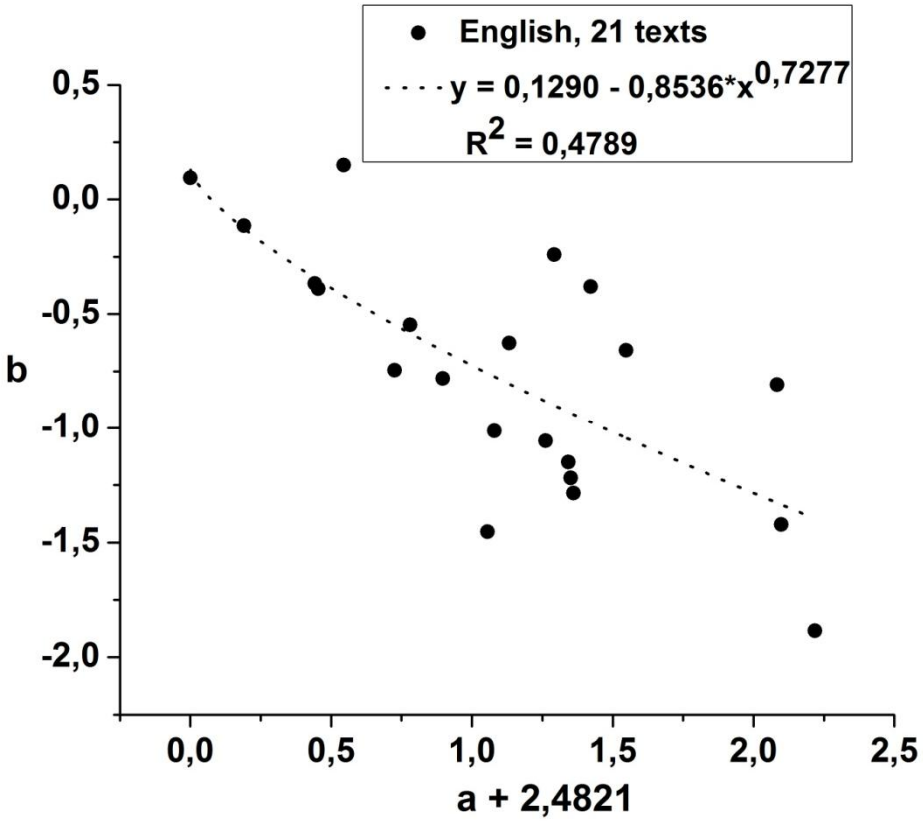
L 12	-1.2348	-1.0280	104.9893	0.9994
L 13	-1.3678	-0.7526	166.9944	0.9998
L 14	-1.1575	-0.9599	117.9813	0.9994
L 15	-0.8348	-1.3845	159.0087	0.9998
L 16	-2.1877	-0.4040	213.9763	0.9994
L 17	-1.8397	-0.6132	204.9740	0.9993
L 18	-1.6450	-0.5912	204.9852	0.9998



Modern English

English (data from Hasse, Weinbrenner 1997)				
Text	**a**	**b**	**c**	**R^2**
T 1	-1.4027	-1.0136	249.0056	1.0000
T 2	-1.4267	-1.4537	275.0042	0.9999
T 3	-1.3498	-0.6260	58.9910	0.9995
T 4	-2.4821	0.0941	345.9708	0.9999
T 5	-1.1211	-1,2844	290.9917	1.0000
T 6	-0.2644	-1.8849	133.9938	0.9998
T 7	-1.1401	-1.1492	316.9616	0.9992
T 8	-1.1314	-1.2182	251.9966	1.0000
T 9	-1.5853	-0.7822	292.9967	1.0000
T 10	-1.7018	-0.5472	645.9229	0.9997
T 11	-0.3838	-1.4214	286.0155	0.9995
T 12	-1.2205	-1.0563	619.9909	1.0000
T 13	-2.2920	-0.1139	172.9842	0.9996

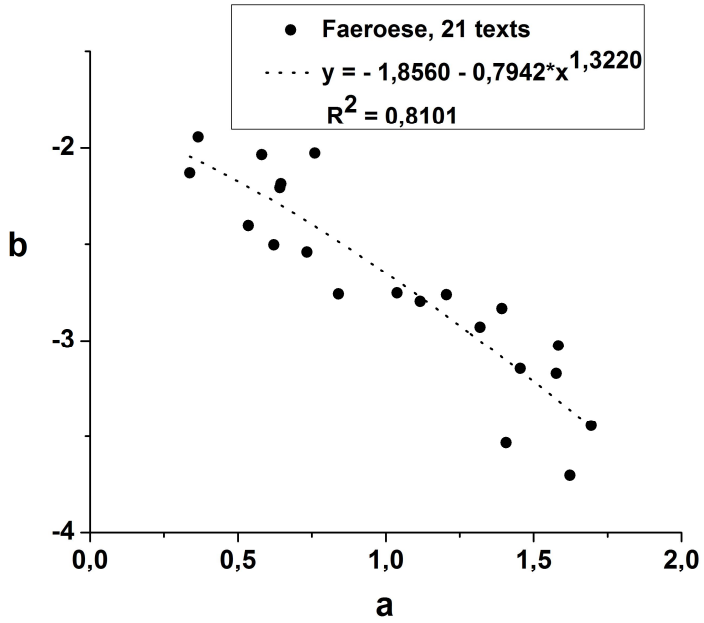
T 14	-1.7565	-0.7455	376.9996	1.0000
T 15	-2.0282	-0.3899	166.0096	0.9998
T 16	-2.0394	-0.3671	326.9725	0.9998
T 17	-0.9348	-0.6580	1038.2109	0.9986
T 18	-1.1897	-0.2412	324.0533	0.9904
T 19	-1.9373	0.1500	555.4634	0.9877
T 20	-1.0601	-0.3800	713.8416	0.9911
T 21	-0.3989	-0.8091	447.9985	0.9991



Faeroese

Faeroese (data from Best, Kaspar 1998 and 2001)				
Text	**a**	**b**	**c**	**R^2**
Letter 1	1.1161	-2.7969	254.0369	0.9987
Letter 2	0.6452	-2.1849	169.9936	1.0000
Letter 3	0.5350	-2.4060	151.0110	0.9997
Letter 4	1.6228	-3.6999	117.0083	0.9979
Letter 5	1.0385	-2.7534	175.0238	0.9988
Letter 6	0.3375	-2.1296	189.0124	0.9998
Letter 7	1.4070	-3.5336	140.0018	0.9999
Letter 8	0.3658	-1.9439	143.0017	1.0000

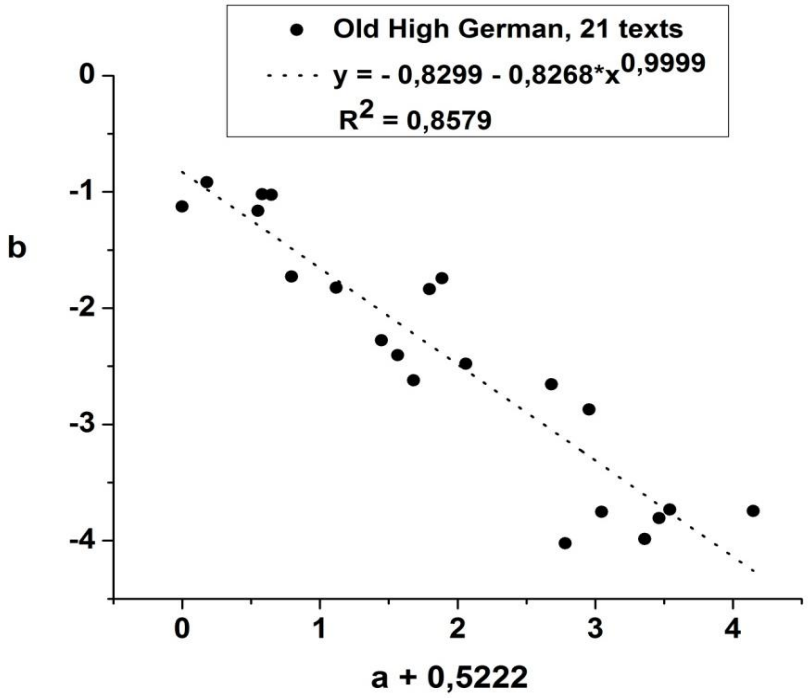
Letter 9	1.5765	-3.1718	140.0042	0.9998
Letter 10	1.2049	-2.7628	125.9979	1.0000
Letter 11	0.7337	-2.5410	144.9888	0.9996
Letter 12	1.4544	-3.1455	133.9982	1.0000
Letter 13	0.8400	-2.7593	159.9935	0.9998
Letter 14	0.6416	-2.2058	249.9854	0.9999
Letter 15	1.3192	-2.9301	153.0621	0.9934
Letter 16	0.7601	-2.0274	204.0590	0.9990
Letter 17	0.6216	-2.5040	281.0152	0.9998
Letter 18	1.3930	-2.8346	203.0497	0.9988
Letter 19	0.5810	-2.0347	252.0176	0.9999
Letter 20	1.5836	-3.0259	376.9957	1.0000
Letter 21	1.6946	-3.4451	168.0118	0.9995



German: Old High German

Old High German (data from Best 1996)				
Text	**a**	**b**	**c**	**R^2**
T1	1.5380	-2.4769	46.0518	0.9945
T2	2.2595	-4.0230	39.0068	0.9926
T3	-0.5222	-1.1268	34.0161	0.9966
T4	2.9422	-3.8056	21.0121	0.9963
T5	1.3655	-1.7437	8.9943	0.9996
T6	1.2728	-1.8363	15.0000	1.0000
T7	0.9256	-2.2761	22.0000	1.0000
T8	0.0274	-1.1631	36.9395	0.9943

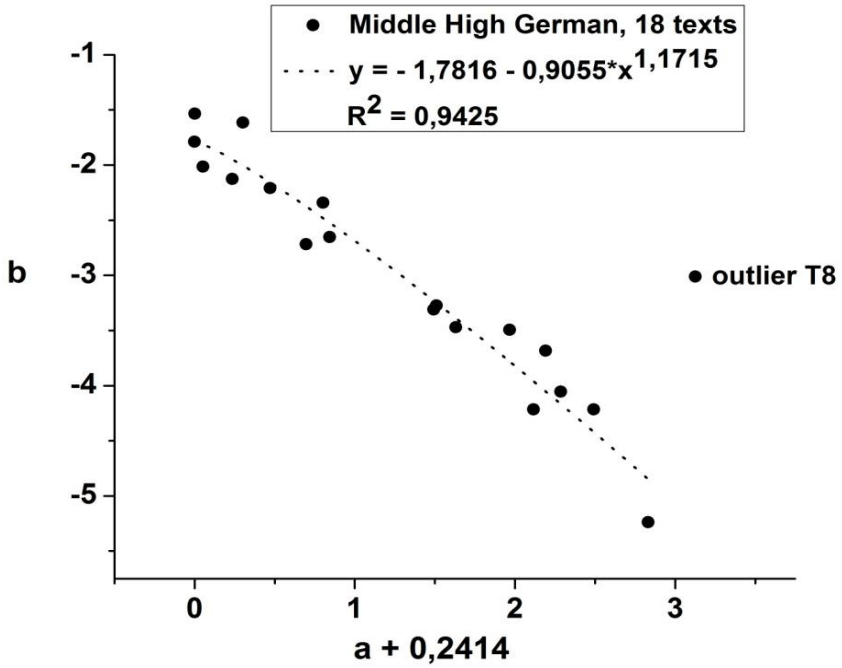
T9	3.0179	-3.7326	23.0000	1.0000
T10	2.8365	-3.9852	22.0000	1.0000
T11	-0.3435	-0.9172	11.9922	0.9976
T12	0.1245	-1.0236	21.0000	1.0000
T13	0.0582	-1.0199	31.8456	0.9323
T14	3.6252	-3.7447	35.1186	0.9477
T15	0.2720	-1.7283	19.0000	1.0000
T16	2.1587	-2.6561	16.9666	0.9937
T17	1.0424	-2.4031	135.9676	0.9993
T18	1.1589	-2.6200	182.9950	0.9999
T19	0.5965	-1.8250	220.8816	0.9987
T20	2.4320	-2.8714	65.1243	0.9931
T21	2.5247	-3.7508	22.0178	0.9829



Middle High German

Middle High German (data from Best 1996)				
Text	**a**	**b**	**c**	**R^2**
T1	-0.2405	-1.5337	177.9918	1.0000
T2	2.0440	-4.0555	131.0024	0.9998
T3	0.2307	-2.2091	115.9933	0.9997
T4	2.2507	-4.2163	137.0001	1.0000
T5	0.0602	-1.6135	83.0120	0.9996
T6	0.5606	-2.3411	119.0008	1.0000
T7	1.7252	-3.4935	118.0189	0.9963

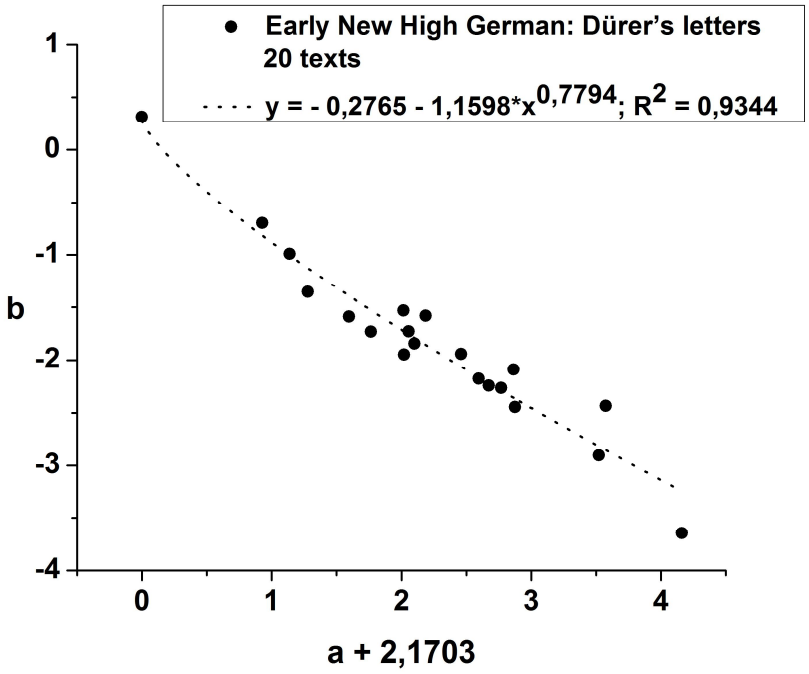
T8	3.1265	-3.0114	53.9974	1.0000
T9	-0.1879	-2.0155	42.0000	1.0000
T10	1.2684	-3.2725	116.0000	1.0000
T11	0.6022	-2.6534	165.0000	1.0000
T12	0.4540	-2.7181	97.0001	1.0000
T13	2.5896	-5.2384	108.0011	0.9989
T14	1.2509	-3.3092	135.9991	1.0000
T15	1.8749	-4.2158	62.0000	1.0000
T16	1.3890	-3.4709	87.0006	1.0000
T17	1.9503	-3.6845	44.0048	0.9982
T18	-0.2414	-1.7883	147.0218	0.9986
T19	-0.0055	-2.1249	282.9951	0.9999



Early New High German

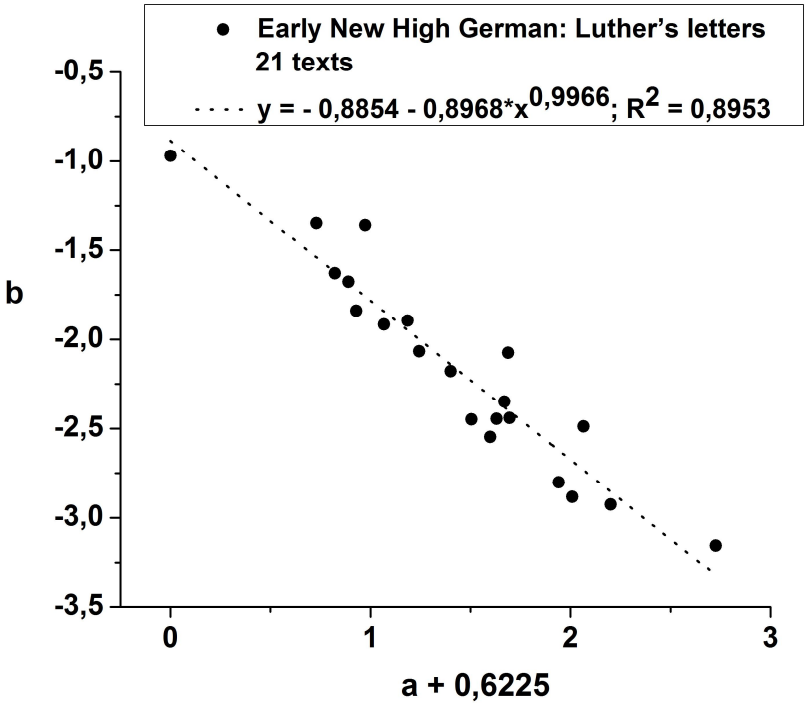
Early New High German: Dürer's letters (data from Ammermann 2001)(16th century				
Text	**a**	**b**	**c**	**R^2**
L 1	-0.0700	-1.8439	275.9709	0.9995
L 2	-0.4048	-1.7327	395.9943	1.0000
L 3	-1.2421	-0.6972	249.9669	0.9996
L 4	-1.0324	-0.9901	384.9466	0.9994
L 5	-0.5754	-1.5899	428.9925	1.0000

L 6	-0.8924	-1.3553	214.9797	0.9994
L 7	-0.1511	-1.9472	249.0011	1.0000
L 8	0.5986	-2.2641	139.0069	0.9999
L 9	1.9908	-3.6474	87.0066	0.9993
L 10	-0.1137	-1.7283	307.9929	0.9999
L 11	0.2885	-1.9422	383.9590	0.9998
L 12	0.4249	-2.1768	239.9813	0.9997
L 13	0.6933	-2.0892	140.9426	0.9974
L 14	0.0168	-1.5804	263.9121	0.9990
L 15	0.7039	-2.4452	443.9752	0.9999
L 16	0.5018	-2.2412	87.0034	0.9996
L 17	-2.1703	0.3132	128.8912	0.9922
L 18	1.3502	-2.9048	38.0005	1.0000
L 19	-0.1552	-1.5318	158.9694	0.9992
L 20	1.4033	-2.4336	55.9999	0.9994



Early New High German: Luther's letters (data from Kuhr, Müller 1997) (16th century)				
Text	**a**	**b**	**c**	**R^2**
L 1	2.1042	-3,1554	108.0056	0.9999
L 2	1.0468	-2.3494	271.9129	0.9992
L 3	1.3177	-2.8012	155.9753	0.9992
L 4	1.3864	-2.8826	249.0045	1.0000

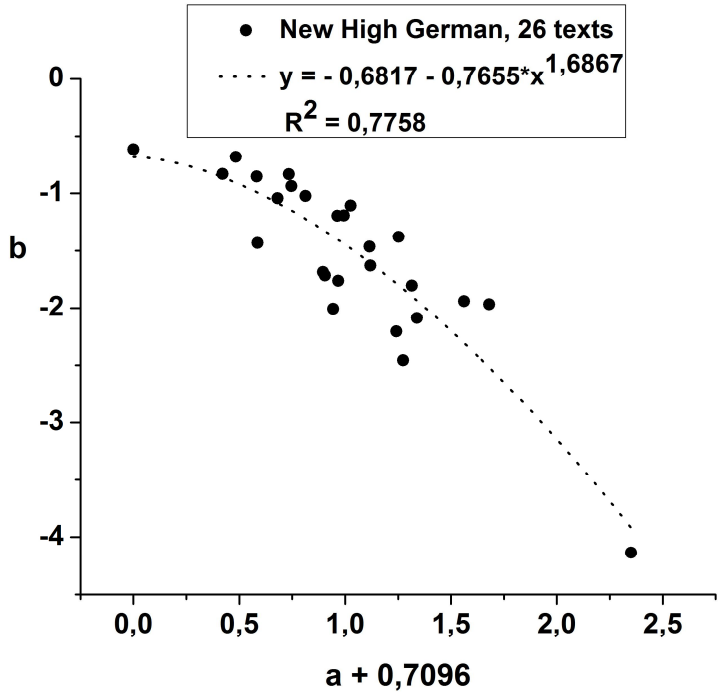
L 5	0.9757	-2.5472	321.9751	0.9998
L 6	1.0663	-2.0749	215.6184	0.9903
L 7	0.3511	-1.3577	112.7647	0.9832
L 8	1.0733	-2.4413	153.9699	0.9985
L 9	0.1075	-1.3448	425.4139	0.9950
L 10	0.2680	-1.6768	278.8924	0.9991
L 11	1.0077	-2.4455	790.9779	1.0000
L 12	0.8825	-2.4483	102.0003	0.9998
L 13	0.6213	-2.0675	183.9175	0.9980
L 14	0.7785	-2.1781	219.9460	0.9992
L 15	1.5776	-2.9264	107.9984	0.9999
L 16	1.4424	-2.4877	124.0020	1.0000
L 17	-0.6225	-0.9694	71.0052	0.9998
L 18	0.4457	-1.9152	221.9526	0.9996
L 19	0.5634	-1.8945	131.9471	0.9990
L 20	0.3063	-1.8389	178.0025	1.0000
L 21	0.2000	-1.6286	158.9523	0.9992



New High German

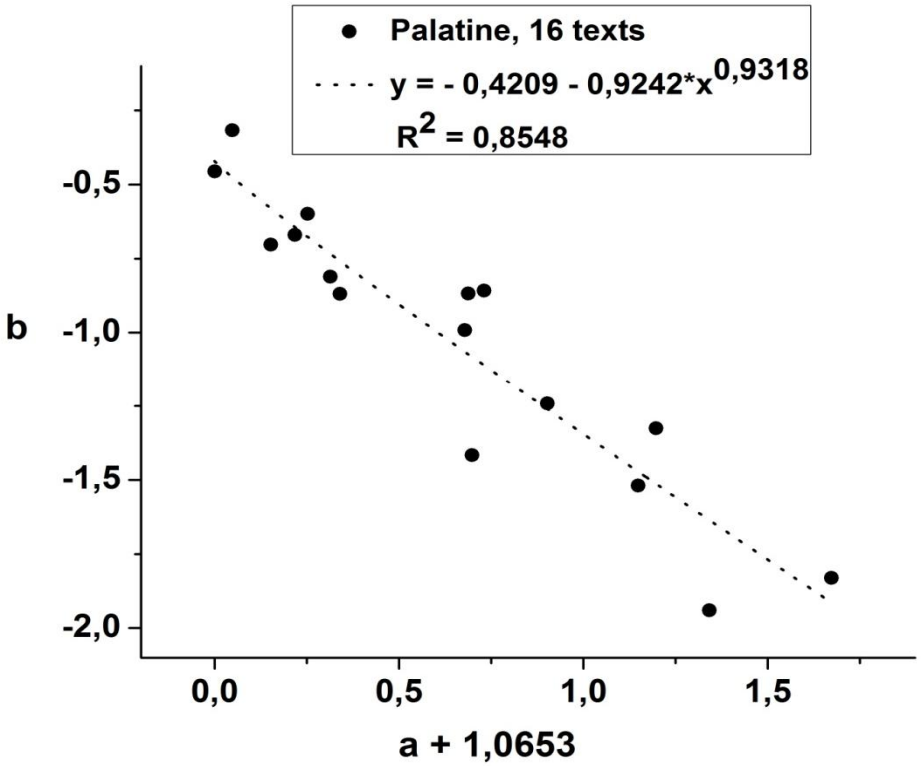
New High German (data from Altmann, Best 1996)				
Text	**a**	**b**	**c**	**R^2**
T1	0.0242	-0.8324	381.3486	0.9948
T2	0.3168	-1.1062	302.0792	0.9947
T3	-0.7096	-0.6136	459.1463	0.9985

T4	0.2840	-1.1908	535.5632	0.9981
T5	0.5420	-1.3778	439.9626	0.9964
T6	-0.2882	-0.8301	523.9097	0.9992
T7	-0.1280	-0.8521	455.7636	0.9954
T8	0.0363	-0.9364	351.2876	0.9982
T9	-0.0286	-1.0418	316.4470	0.9968
T10	-0.2263	-0.6807	307.7247	0.9940
T11	0.1033	-1.0228	586.9698	0.9967
T12	0.4052	-1.4653	485.8029	0.9991
T13	0.4089	-1.6300	739.3852	0.9988
T14	0.1950	-1.7147	265.9431	0.9996
T15	0.5316	-2.2042	428.0369	0.9997
T16	0.8523	-1.9401	338.2102	0.9987
T17	0.2338	-2.0050	151.0186	0.9995
T18	0.6055	-1.8051	164.0136	0.9989
T19	0.6297	-2.0864	219.0646	0.9988
T20	0.9700	-1.9672	133.9949	0.9996
T21	-0.1238	-1.4330	58.9837	0.9980
T22	0.2540	-1.1937	76.9084	0.9981
T23	0.5640	-2.4564	217.9988	1.0000
T24	0.2581	-1.7631	579.8708	0.9995
T25	1.6402	-4.1350	152.0013	0.9998
T26	0.1862	-1.6874	259.0910	0.9989



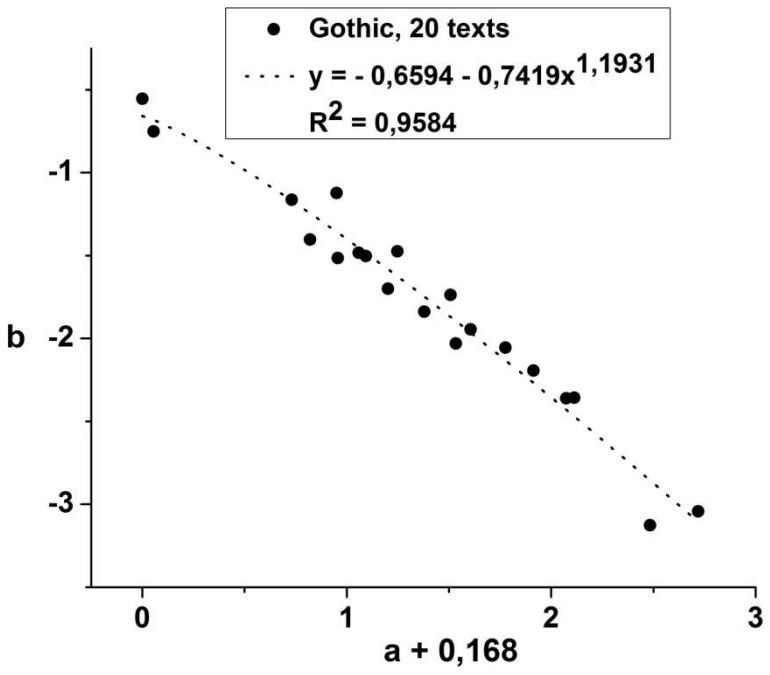
Palatine

Palatine (data from Kiefer 2001)				
Text	**a**	**b**	**c**	**R^2**
A	-0.3672	-1.4169	264.9990	0.9999
B	0.2762	-1.9399	251.1139	0.9970
C	-0.1629	-1.2417	208.0567	0.9992
D	-0.3350	-0.8580	221.5220	0.9994
E	-1.0653	-0.4557	204.4467	0.9891
F	-0.3778	-0.8683	228.6752	0.9975
G	-0.7519	-0.8112	241.8855	0.9995
H	0.6071	-1.8308	193.0109	0.9980
I	0.1315	-1.3257	185.1723	0.9931
J	-0.7255	-0.8691	274.7299	0.9957
K	-0.8132	-0.5988	249.5780	0.9953
L	-0.9126	-0.7030	235.8044	0.9978
M	0.0834	-1.5196	255.0817	0.9992
N	-0.8480	-0.6700	244.6309	0.9936
O	-0.3873	-0.9915	239.1919	0.9933
P	-1.0177	-0.3167	172.4559	0.9816



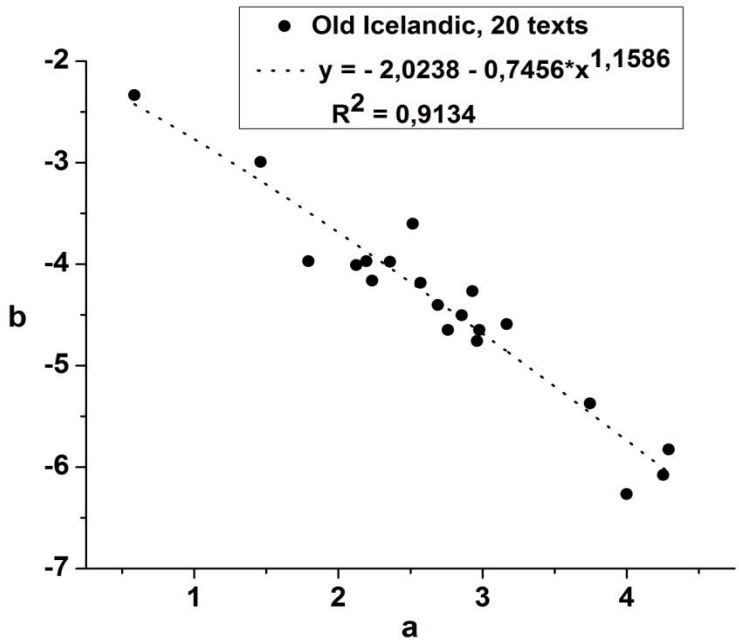
Gothic

Gothic (data from Kiyko 2007)				
Text	**a**	**b**	**c**	**R^2**
T 1	1.4385	-1.9457	87.8783	0.9976
T 2	0.8911	-1.4846	54.1208	0.9883
T 3	0.7888	-1.5155	223.5498	0.9973
T 4	1.6087	-2.0559	107.2415	0.9931
T 5	-0.1117	-0.7519	115.3964	0.9764
T 6	1.3396	-1.7373	87.8732	0.9983
T 7	0.6532	-1.4037	217.6526	0.9774
T 8	1.2108	-1.8391	103.0507	0.9921
T 9	1.3653	-2.0316	115.0522	0.9990
T 10	-0.1680	-0.5547	95.0080	0.9357
T 11	0.7824	-1.1242	74.7692	0.9304
T 12	1.0335	-1.7000	218.0678	0.9857
T 13	0.9251	-1.5037	106.0392	0.9998
T 14	1.9439	-2.3576	97.1276	0.9963
T 15	1.7458	-2.1935	118.7956	0.9982
T 16	0.5626	-1.1641	116.1245	0.9706
T 17	1.0805	-1.4750	78.8991	0.9749
T 18	1.9060	-2.3610	93.1573	0.9959
T 19	2.5508	-3.0438	82.1302	0.9971
T 20	2.3151	-3.1265	183.2093	0.9951



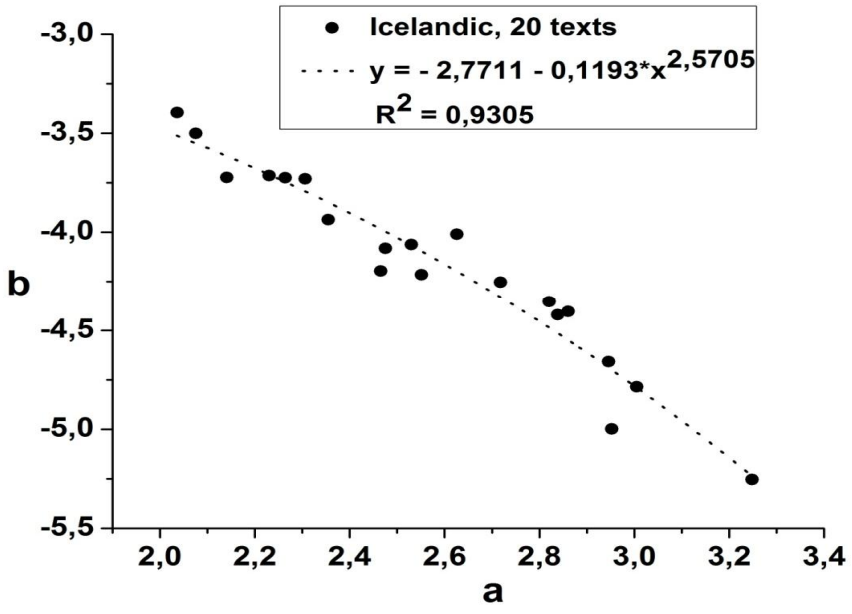
Icelandic: Old Icelandic

Old Icelandic (data from Best 1996)				
Text	**a**	**b**	**c**	**R^2**
Text 1	2.9785	-4.6503	385.9125	0.9997
Text 2	4.2521	-6.0776	360.0007	1.0000
Text 3	3.1668	-4.5910	283.0000	1.0000
Text 4	3.7448	-5.3743	144.0049	0.9995
Text 5	4.2911	-5.8279	425.0093	0.9998
Text 6	2.9612	-4.7580	125.0037	0.9996
Text 7	2.8551	-4.5061	376.9966	1.0000
Text 8	2.5162	-3.6035	495.9415	0.9997
Text 9	2.6895	-4.4032	180.0000	1.0000
Text 10	3.9984	-6.2675	549.0052	0.9992
Text 11	2.9306	-4,2660	229.0224	0.9993
Text 12	2.3574	-3.9779	58.0050	0.9991
Text 13	2.5697	-4.1864	428.0005	1.0000
Text 14	2.1250	-4.0093	110.0102	0.9978
Text 15	1.4599	-2.9926	151.9422	0.9969
Text 16	1.7927	-3.9712	191.0074	0.9968
Text 17	2.7608	-4.6497	230.0041	0.9998
Text 18	0.5859	-2.3366	364.9645	0.9996
Text 19	2.1936	-3.9730	143.0021	0.9999
Text 20	2.2344	-4.1620	474.0032	1.0000



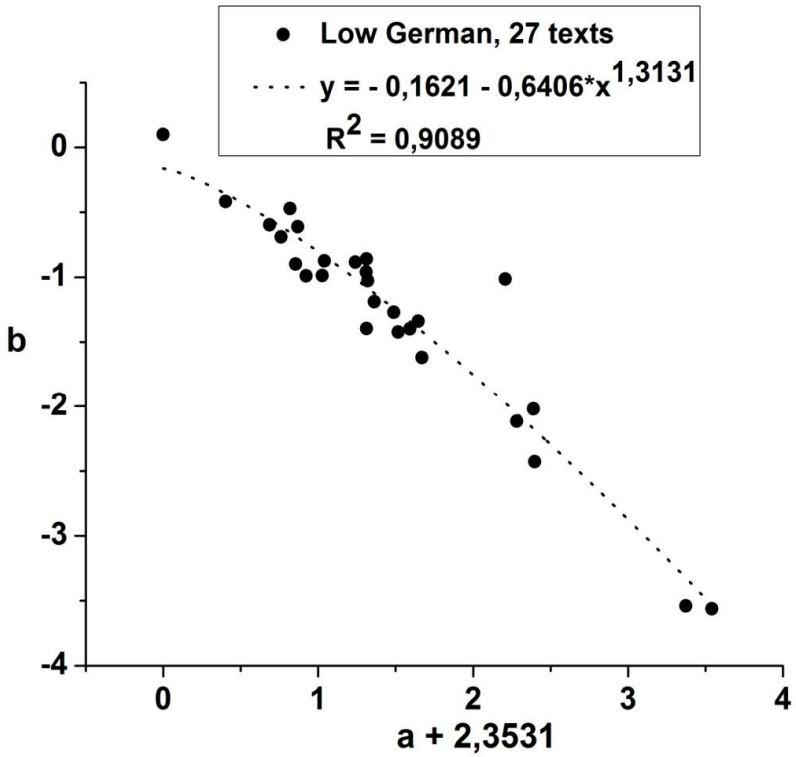
Modern Icelandic

Icelandic (data from Best, Brynjólfsson 1997)				
Text	**a**	**b**	**c**	**R^2**
Text 1	2.6263	-4.0100	140.0157	0.9991
Text 2	3.2480	-5.2532	151.0034	0.9994
Text 3	2.0361	-3.3957	189.0472	0.9986
Text 4	2.5301	-4.0623	185.0192	0.9985
Text 5	2.1413	-3.7237	207.0474	0.9957
Text 6	2.7182	-4.2533	253.0370	0.9977
Text 7	2.9456	-4.6585	186.0236	0.9961
Text 8	2.0754	-3.5006	252.0555	0.9984
Text 9	2.2303	-3.7143	171.0390	0.9971
Text 10	2.4653	-4.1963	265.0138	0.9995
Text 11	2.8603	-4.4042	193.0145	0.9990
Text 12	2.4754	-4.0821	212.0178	0.9992
Text 13	3.0051	-4.7849	300.0342	0.9950
Text 14	2.2643	-3.7245	236.0782	0.9950
Text 15	2.5514	-4.2154	239.0173	0.9991
Text 16	2.3547	-3.9363	198.0253	0.9987
Text 17	2.8203	-4.3538	269.0587	0.9955
Text 18	2.9527	-4.9980	295.0088	0.9990
Text 19	2.8388	-4.4209	195.0310	0.9964
Text 20	2.3063	-3.7299	337.0371	0.9996



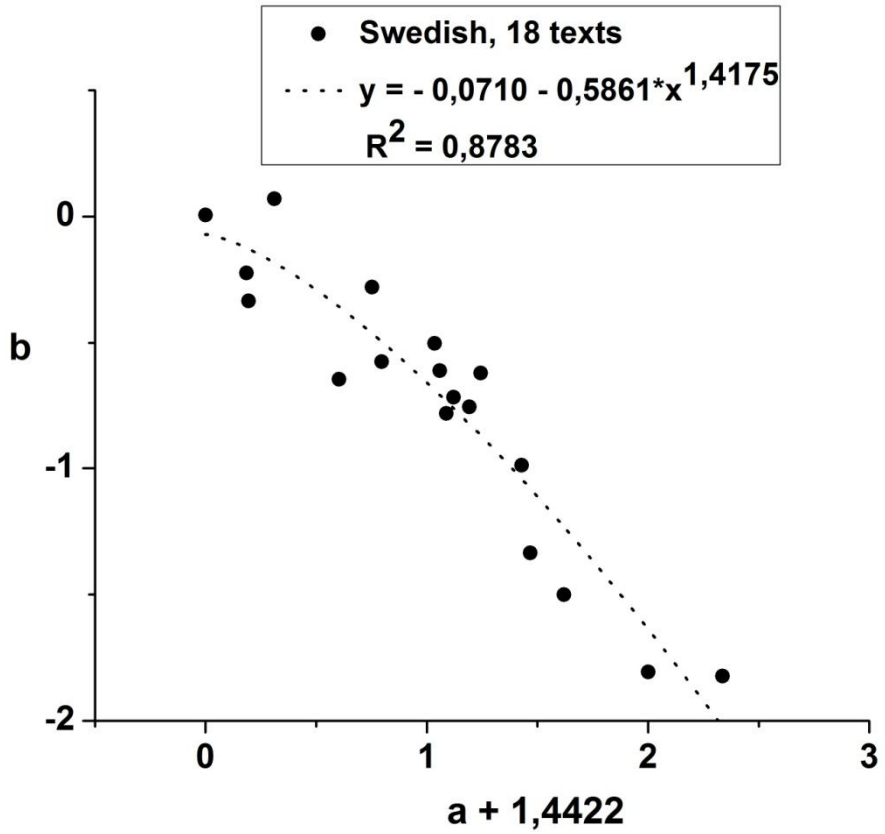
Low German

Low German (data from Ahlers 2001)				
Text	**a**	**b**	**c**	**R^2**
Letter 1	-1.4297	-0.9892	225.9929	1.0000
Letter 2	-0.6822	-1.6199	213.0025	1.0000
Letter 3	-1.0407	-1.3961	209.0025	1.0000
Letter 4	-0.7064	-1.3399	154.0255	0.9994
Letter 5	-1.1120	-0.8846	172.9819	0.9997
Letter 6	-0.0705	-2.1150	200.0036	1.0000
Letter 7	-1.9508	-0.4164	208.9899	0.9999
Letter 8	-0.8654	-1.2701	201.9871	0.9999
Letter 9	1.0206	-3.5393	170.0015	0.9999
Letter 10	-1.4836	-0.6117	211.9756	0.9996
Letter 11	0.0448	-2.4272	177.0066	0.9997
Letter 12	-0.7616	-1.3983	193.9752	0.9993
Letter 13	-1.5348	-0.4704	147.9595	0.9993
Letter 14	-0.8373	-1.4234	184.0014	1.0000
Letter 15	-0.1463	-1.0140	149.9981	0.9998
Letter 16	-1.4977	-0.8974	204.0026	1.0000
Letter 17	-2.3531	0.1001	182.9331	0.9985
Letter 18	-1.3120	-0.8732	177.0070	1.0000
Letter 19	-1.6659	-0.5961	158.9714	0.9993
Letter 20	-1.0420	-0.9595	171.9604	0.9992
Letter 21	-1.0329	-1.0254	189.9585	0.9990
Letter 22	0.0368	-2.0202	200.9918	0.9999
Letter 23	-1.3267	-0.9874	209.0068	0.9999
Letter 24	-0.9902	-1.1890	204.9850	0.9997
Letter 25	-1.5915	-0.6904	218.9921	0.9999
Letter 26	-1.0405	-0.8600	159.9734	0.9994
Letter 27	1.1869	-3.5624	180.0000	1.0000



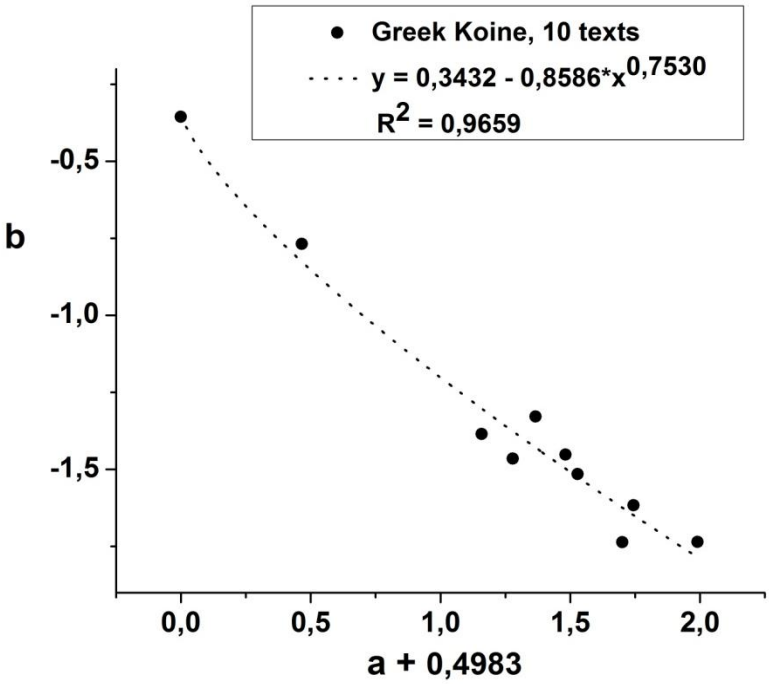
Swedish

Swedish (data from Best 1996)				
Text	**a**	**b**	**c**	**R^2**
T 1	-1.2475	-0.3342	93.8618	0.9949
T 2	-1.1318	0.0695	32.5086	0.8973
T 3	-0.3533	-0.7794	79.8596	0.9931
T 4	-0.2502	-0.7538	129.0820	0.9930
T 5	-1.2574	-0.2237	97.6190	0.9792
T 6	0.5583	-1.8072	303.0936	0.9995
T 7	-0.3207	-0.7150	175.3093	0.9903
T 8	-0.0131	-0.9885	105.7064	0.9933
T 9	0.1776	-1.5016	41.0529	0.9855
T 10	-0.4062	-0.5021	27.6853	0.9489
T 11	-0.6890	-0.2792	30.6603	0.9623
T 12	-0.1983	-0.6198	121.4856	0.9882
T 13	-0.8395	-0.6439	131.9046	0.9990
T 14	-1.4422	0.0052	94.0084	0.9992
T 15	0.8942	-1.8235	48.1236	0.9826
T 16	-0.3837	-0.6108	108.4687	0.9886
T 17	-0.6456	-0.5745	247.3422	0.9956
T 18	0.0251	-1.3355	227.9848	0.9998



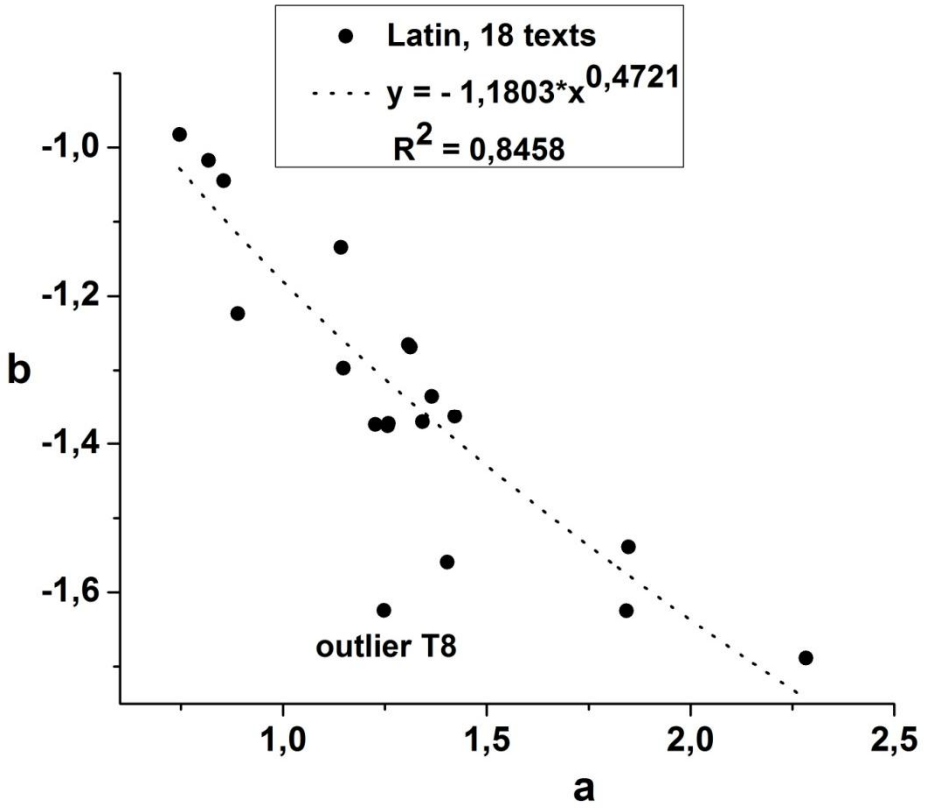
Greek: Greek Koine

Greek Koine (data from Egbers, Groen, Podehl, Rauhaus 1997)				
Text	**a**	**b**	**c**	**R^2**
T 1	1.4924	-1.7358	71.3596	0.9781
T 2	-0.4983	-0.3562	34.7531	0.9403
T 3	0.9835	-1.4524	55.6585	0.9865
T 4	0.7800	-1.4649	61.1570	0.9895
T 5	1.2462	-1.6167	52.5554	0.9619
T 6	-0.0328	-0.7680	65.6027	0.9869
T 7	0.8680	-1.3286	67.8896	0.9964
T 8	1.2026	-1.7360	67.9993	0.9991
T 9	0.6609	-1.3847	152.1919	0.9974
T 10	1.0297	-1.5156	287.4383	0.9966



Latin

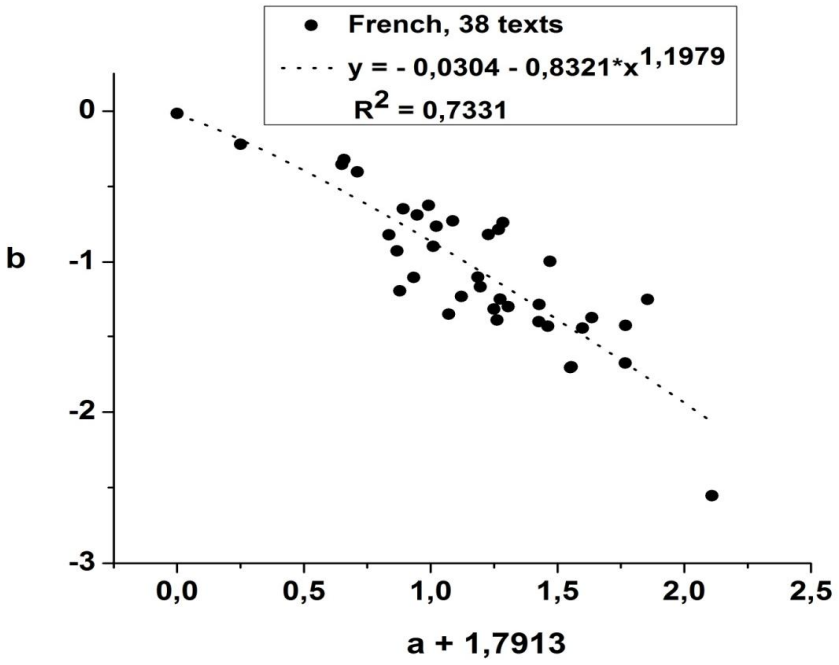
Latin (data from Röttger, Schweers 1997)				
Text	**a**	**b**	**c**	**R^2**
T 1	1.8476	-1.5391	213.3802	0.9150
T 2	1.3130	-1.2688	285.8083	0.9488
T 3	0.8549	-1.0449	312.4947	0.8965
T 4	1.1483	-1.2968	31.2407	0.9760
T 5	0.8890	-1.2234	37.9773	0.9991
T 6	1.2269	-1.3742	28.0150	0.9964
T 7	1.4033	-1.5594	43.9834	0.9994
T 8	1.2479	-1.6245	27.9053	0.9926
T 9	1.2585	-1.3726	38.4991	0.9763
T 10	0.7464	-0.9827	42.9538	0.8066
T 11	1.3428	-1.3702	49.8612	0.9370
T 12	1.3648	-1.3350	55.1596	0.8832
T 13	1.2573	-1.3762	34.8688	0.9222
T 14	1.3080	-1.2652	49.5994	0.8064
T 15	1.1426	-1.1343	41.5818	0.7281
T 16	1.8425	-1.6250	30.5502	0.9502
T 17	1.4219	-1.3631	33.3393	0.8975
T 18	0.8172	-1.0174	108.4602	0.9248
T 19	2.2833	-1.6888	114.4481	0.9061



Romanic: French

French (data from Dieckmann, Judt 1996)				
Text	**a**	**b**	**c**	**R^2**
T 1	-0.7685	-0.7636	425.5808	0.9962
T 2	-1.1332	-0.3220	487.4611	0.9755
T 3	-1.1409	-0.3540	433.3522	0.9931
T 4	-1.0800	-0.4024	371.5617	0.9932
T 5	-1.7913	-0.0165	438.8354	0.9803
T 6	-0.7036	-0.7273	372.3255	0.9935
T 7	-1.5404	-0.2207	523.9265	0.9908
T 8	-0.7992	-0.6254	676.4548	0.9908
T 9	0.0647	-1.2493	166.9611	0.9983
T 10	-0.5177	-1.2460	241.9342	0.9992
T 11	-0.2360	-1.7008	389.9813	0.9999
T 12	-0.5229	-0.7853	349.4426	0.9941
T 13	-0.8450	-0.6894	414.6611	0.9975
T 14	-0.5412	-1.3130	620.9238	0.9998
T 15	-0.7808	-0.8960	315.8860	0.9991
T 16	-0.0244	-1.6764	862.7403	0.9987

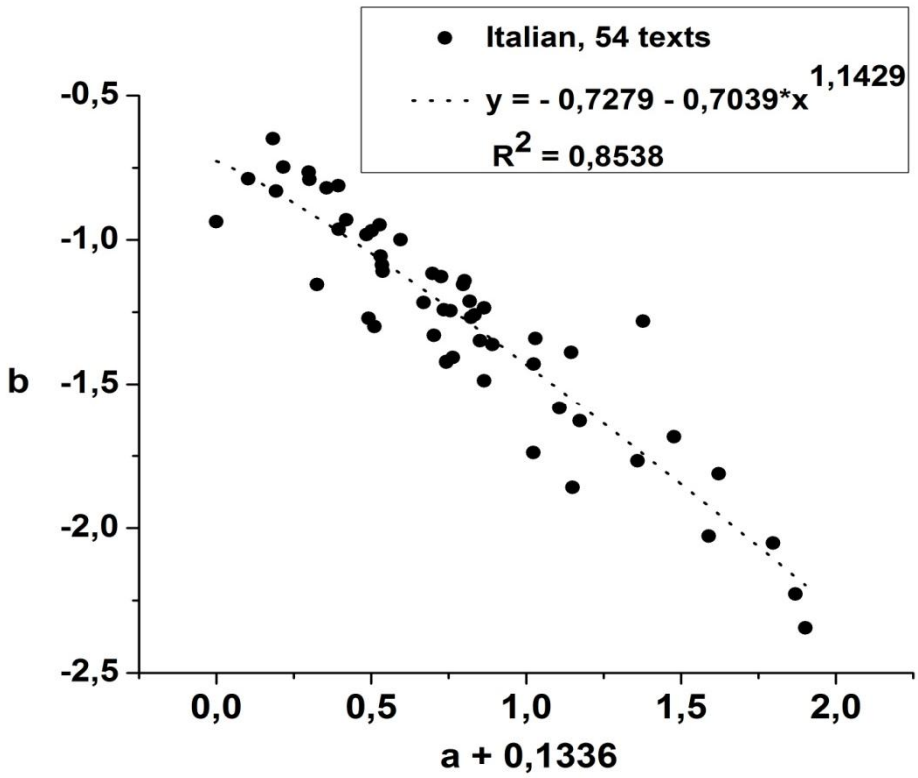
Data from Feldt, Janssen, Kuleisa 1997 (letters and journalism)				
T 1	-0.5289	-1.3853	689.8811	0.9994
T 2	-0.3628	-1.2827	417.9536	0.9997
T 3	-0.0230	-1.4216	616.7359	0.9992
T 4	-0.5634	-0.8181	324.5833	0.9970
T 5	-0.3212	-0.9955	272.8532	0.9988
T 6	-0.5061	-0.7387	209.7438	0.9970
T 7	-0.7198	-1.3467	619.9941	1.0000
T 8	-0.9129	-1.1915	266.9986	0.9999
T 9	0.3186	-2.5537	303.0136	0.9996
T 10	-0.9554	-0.8197	528.7345	0.9980
T 11	-0.1557	-1.3689	203.9228	0.9991
T 12	-0.3289	-1.4264	162.9702	0.9996
T 13	-0.8576	-1.1032	913.8195	0.9994
T 14	-0.6705	-1.2289	487.8746	0.9992
T 15	-0.9240	-0.9264	354.8768	0.9988
T 16	-0.3650	-1.3961	346.8704	0.9980
T 17	-0.4856	-1.2968	340.9524	0.9997
T 18	-0.5959	-1.1648	457.9805	0.9998
T 19	-0.2398	-1.7060	386.9717	0.9999
T 20	-0.1930	-1.4380	321.0170	0.9995
T 21	-0.8997	-0.6489	323.9192	0.9994
T 22	-0.6048	-1.1005	253.8944	0.9996



Italian

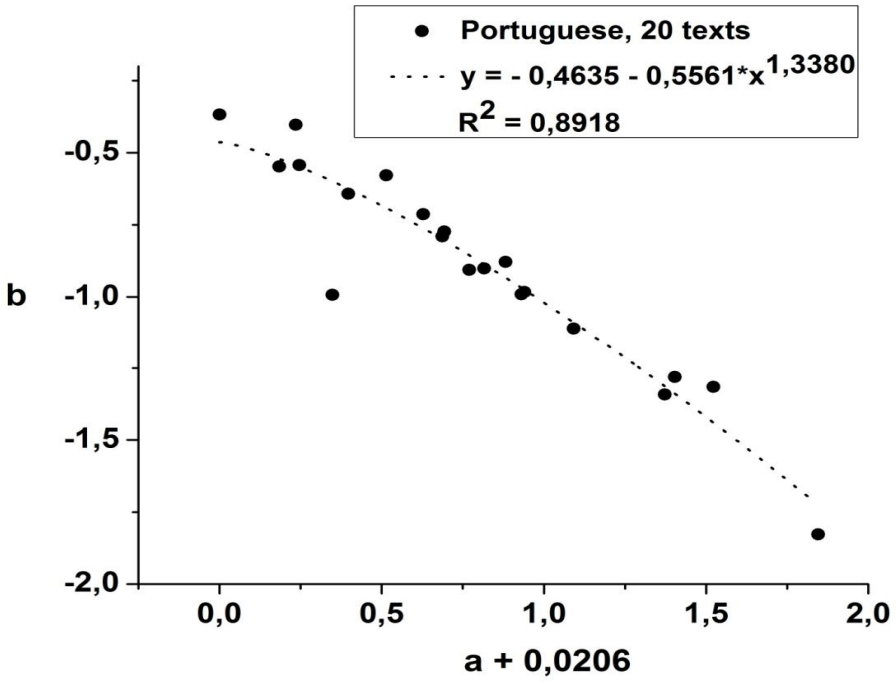
Italian (data from Hollberg 1997)				
Text	**a**	**b**	**c**	**R^2**
T 1	0.2852	-0.9299	251.8987	0.9835
T 2	0.2226	-0.8198	374.0524	0.9605
T 3	0.3507	-0.9813	328.1604	0.9778
T 4	0.0595	-0.8313	271.0504	0.9834
T 5	0.6681	-1.1409	433.6205	0.9694
T 6	1.2443	-1.2815	158.0940	0.9707
T 7	0.2598	-0.8128	219.4676	0.9249
T 8	0.1644	-0.7658	253.0041	0.9518
T 9	0.7003	-1.2589	285.4302	0.9892
T 10	0.8966	-1.3415	262.5641	0.9647
(data from Gaeta 1994)				
ALFI 1	0.6007	-1.2413	369.6589	0.9953
ALFI 2	0.4039	-1.1084	458.8565	0.9814
ALFI 3	0.3575	-1.2712	222,0206	0.9972
ALFI 4	-0.1336	-0.9370	219.4017	0.9944
ALFI 5	0.1914	-1.1540	120.8255	0.9980
CALVI 1	0.8911	-1.4293	333.0353	0.9943
CALVI 2	0.5639	-1.1158	250.9551	0.9772
CALVI 3	0.3929	-0.9478	338.2143	0.9613
CARD 1	0.5352	-1.2162	205.8416	0.9895
CARD 2	0.6086	-1.4227	160.2409	0.9462
DUIZ	0.3675	-0.9686	433.6232	0.9594
CLERI	0.5920	-1.1271	298.4210	0.9733
GAGLI	0.1666	-0.7907	235.9178	0.9756
PAST	0.6839	-1.2122	348.6791	0.9879
MORO 1	0.6892	-1.2676	151.3916	0.9876
MORO 2	0.7311	-1.2353	512.9692	0.9669
BUZZ 1	0.6226	-1.2452	421.1641	0.9847
BUZZ 2	0.4611	-0.9989	343.3427	0.9729
BUZZ 3	1.3438	-1.6842	360.7510	0.9950
PASO 1	0.9736	-1.5838	286.2089	0.9837
PASO 2	0.3969	-1.0563	125.5080	0.9191
BALE 1	0.6630	-1.1549	69.9513	0.9463
BALE 2	1.2263	-1.7678	79.3112	0.9916
BALE 3	0.7581	-1.3622	349.8682	0.9946
BOCCA	0.0829	-0.7479	333.2450	0.9673
MANZ 2	0.6090	-1.4207	196.1192	0.9974
MANZ 3	0.5692	-1.3305	112.5002	0.9911

MANZ 4	0.2609	-0.9631	236.4800	0.9815
MANZ 5	0.7312	-1.4874	126.9639	0.9990
MANZ 6	0.3762	-1.2996	429.0543	0.9966
MANZ 7	0.7180	-1.3493	379.4383	0.9935
SABA 1	1.0163	-1.8595	166.8260	0.9982
SABA 2	1.7685	-2.3442	47.0114	0.9991
SABA 3	0.4011	-1.0864	49.0788	0.9948
SABA 4	1.7360	-2.2278	124.6892	0.9959
SABA 5	1.4882	-1.8124	55.5006	0.9753
SERE 1	0.8896	-1.7390	113.7826	0.9943
SERE 2	0.6296	-1.4071	67.7992	0.9886
SERE 3	1.4565	-2.0275	87.8066	0.9968
ZAVA 1	1.0121	-1.3896	128.0355	0.9873
ZAVA 2	1.6640	-2.0518	189.9529	0.9868
ZAVA 3	1.0394	-1.6291	156.4628	0.9902
BOSSI	-0.0307	-0.7887	386.6338	0.9766
PIRA	0.0492	-0.6492	324.5358	0.9621



Portuguese

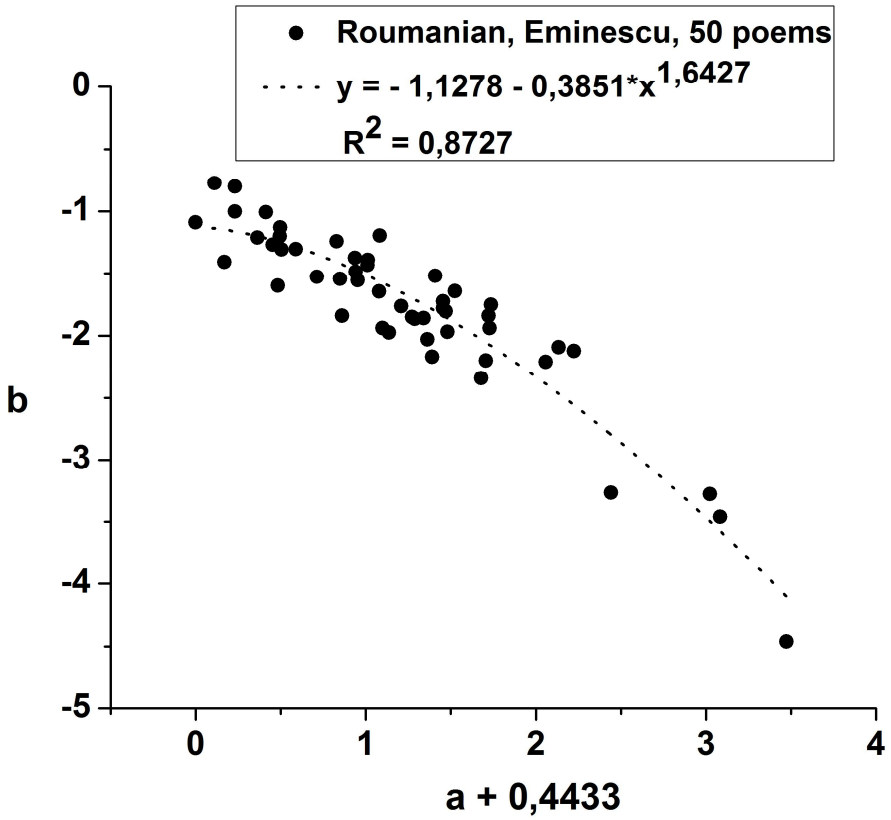
Portuguese (data from Ziegler 1998)				
Text	**a**	**b**	**c**	**R^2**
Text 1	0.1087	-0.9007	159.0536	0.9845
Text 2	-0.3588	-0,9923	178.8420	0.9983
Text 3	-0.0138	-0,7729	141.2437	0.9792
Text 4	-0.4716	-0,4033	145.0756	0.9439
Text 5	-0.3101	-0.6428	206.9095	0.9654
Text 6	-0.5230	-0.5484	218.1370	0.9914
Text 7	-0.7073	-0.3678	182.3211	0.9487
Text 8	-0.1932	-0.5786	179.8537	0.9514
Text 9	0.0631	-0.9060	241.0501	0.9770
Text 10	-0.4604	-0.5433	198.2658	0.9692
Text 11	0.1749	-0.8789	195.3357	0.9825
Text 12	-0.0202	-0.7896	252.0967	0.9750
Text 13	0.2327	-0.9831	193.7607	0.9815
Text 14	0.2232	-0.9903	133.8719	0.9768
Text 15	-0.0788	-0.7133	187.5927	0.9773
Text 16	0.6650	-1.3413	170.8742	0.9932
Text 17	1.1379	-1.8275	141.9166	0.9987
Text 18	0.8148	-1.3149	121.9774	0.9998
Text 19	0.3842	-1.1096	180.8861	0.9880
Text 20	0.6964	-1.2803	179.3583	0.9929



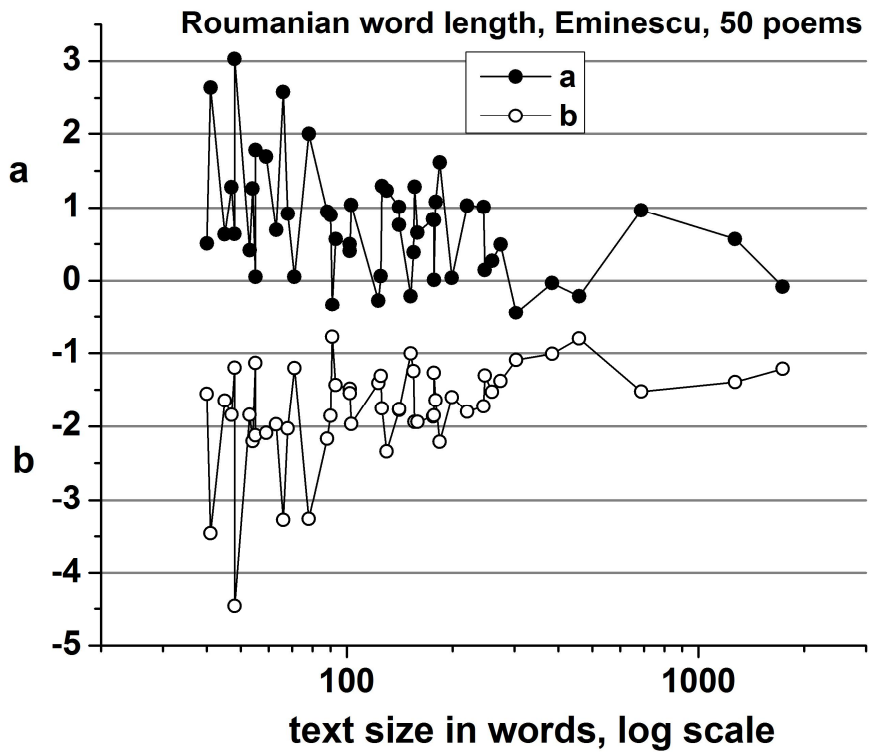
Roumanian

Roumanian (data of 50 poems by Eminescu, 2014)				
Text	**a**	**b**	**c**	**R^2**
T 1	0,6561	-1,9392	84,9177	0,9931
T 2	0,8441	-1,8665	83,9336	0,9970
T 3	1,2928	-1,7545	46,5825	0,7589
T 4	0,4969	-1,4874	47,9784	0,9975
T 5	1,2343	-2,3417	62,9849	0,9993
T 6	1,0115	-1,7808	58,9292	0,9818
T 7	-0,2739	-1,4087	76,0036	0,9998
T 8	0,5681	-1,4328	40,1407	0,9178
T 9	-0,4433	-1,0899	178,9965	0,9991
T 10	0,9184	-2,0300	33,9719	0,9860
T 11	-0,2113	-1,0010	80,9762	0,9986
T 12	0,5093	-1,5568	19,9451	0,9024
T 13	0,4050	-1,5459	50,9889	0,9948
T 14	0,9654	-1,5221	254,9918	0,9934
T 15	0,1458	-1,3055	124,8898	0,9976
T 16	1,0811	-1,6455	67,7849	0,9848
T 17	0,6928	-1,9740	33,9686	0,9743
T 18	-0,2121	-0,7992	213,7626	0,9611
T 19	1,6151	-2,2091	69,9634	0,9979
T 20	0,0532	-1,1313	26,9696	0,9957
T 21	0,0519	-1,2005	36,9444	0,9672
T 22	0,8972	-1,8591	41,9977	1,0000
T 23	2,6392	-3,4600	16,0163	0,9771
T 24	2,5800	-3,2767	25,0063	0,9992
T 25	0,2694	-1,5281	137,9104	0,9969
T 26	-0,0809	-1,2114	938,8828	0,9949
T 27	0,0620	-1,3074	65,9212	0,9941
T 28	0,4941	-1,3775	122,8058	0,9915
T 29	0,3857	-1,2432	67,0711	0,9581
T 30	0,8291	-1,8526	83,9597	0,9979
T 31	1,2627	-2,1989	23,9995	0,9865
T 32	-0,0303	-1,0082	184,1362	0,9613
T 33	0,6392	-1,1967	17,8789	0,9574
T 34	1,9979	-3,2660	37,0317	0,9396
T 35	1,0368	-1,9704	46,0515	0,9939
T 36	0,0400	-1,6013	118,9404	0,9949
T 37	1,2783	-1,8417	17,9966	0,9998
T 38	1,0262	-1,8040	93,7817	0,9878

T 39	3,0292	-4,4622	22,9990	0,9994
T 40	0,7642	-1,7652	65,0851	0,9875
T 41	1,2843	-1,9389	62,6174	0,9513
T 42	0,5694	-1,3928	549,6648	0,9977
T 43	0,6349	-1,6482	21,0279	0,9903
T 44	1,7794	-2,1220	18,1273	0,8544
T 45	1,0111	-1,7253	101,5941	0,9699
T 46	-0,3318	-0,7738	43,8909	0,9806
T 47	0,4170	-1,8420	29,0034	0,9969
T 48	0,9471	-2,1677	44,0092	0,9975
T 49	1,6912	-2,0916	20,8567	0,8661
T 50	0,0102	-1,2695	93,9559	0,9985



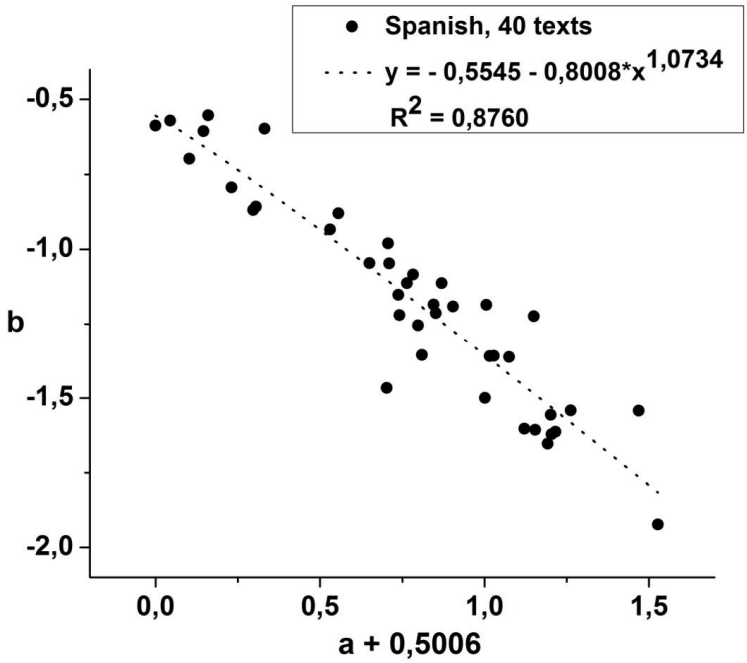
The correlation between the fitting parameters *a* and *b* is illustrated also by their mirror-like dependence on the text size (in words), as illustrated for Roumanian in the following graph below.



Spanish

Spanish (data from Becker 1996)				
Text	**a**	**b**	**c**	**R^2**
Letter 1	0.7031	-1.6209	184.8319	0.9990
Letter 2	0.5051	-1.1856	123.9025	0.9701
Letter 3	0.2372	-1.1520	179.2568	0.9847
Letter 4	0.1498	-1.0465	246.1478	0.9944
Letter 5	-0.5006	-0.5877	170.2027	0.9841
Letter 6	0.2979	-1.2547	359.1728	0.9962
Letter 7	0.5011	-1.4994	269.9825	0.9989
Letter 8	-0.3406	-0.5530	116.2564	0.9801
Letter 9	-0.3547	-0.6056	172.7705	0.9681
Letter 10	0.3449	-1.1844	364.9835	0.9949
Letter 11	0.5277	-1.3585	316.5786	0.9986
Letter 12	0.7613	-1.5419	250.5893	0.9983
Letter 13	0.6209	-1.6028	217.9566	0.9992
Letter 14	0.6540	-1.6067	207.8371	0.9985
Letter 15	0.7156	-1.6132	239.7780	0.9991
Letter 16	0.2018	-1.4662	346.9554	0.9997

Letter 17	0.6912	-1.6531	197.8079	0.9986
Letter 18	1.0268	-1.9235	206.9460	0.9999
Letter 19	0.2410	-1.2197	309.3601	0.9958
Letter 20	0.3091	-1.3561	303.8129	0.9986
Data from Hein 1997				
Letter 1	0.2641	-1.1135	119.7965	0.9976
Letter 2	0.2096	-1.0477	207.4543	0.9934
Letter 3	0.7009	-1.5562	76.9806	0.9913
Letter 4	-0.1953	-0.8578	169.5671	0.9906
Letter 5	0.5744	-1.3623	123.9861	0.9988
Letter 6	0.2826	-1.0843	112.5030	0.9896
Letter 7	-0.2041	-0.8686	156.5496	0.9924
Letter 8	0.0554	-0.8801	90.8494	0.9906
Letter 9	0.4041	-1.1910	55.1414	0.9867
Letter 10	0.2067	-0.9802	227.6703	0.9891
Letter 11	-0.4563	-0.5708	57.9688	0.9926
Letter 12	0.0300	-0.9340	128.6978	0.9966
Letter 13	-0.3977	-0.6982	143.5742	0.9889
Letter 14	0.3514	-1.2131	91.8000	0.9949
Letter 15	0.5154	-1.3598	83.0218	0.9979
Letter 16	0.3689	-1.1133	78.3454	0.9652
Letter 17	0.9683	-1.5423	64.0227	0.9917
Letter 18	-0.1691	-0.5973	65.8213	0.9924
Letter 19	0.6495	-1.2239	70.6934	0.9941
Letter 20	-0.2699	-0.7934	273.2728	0.9942



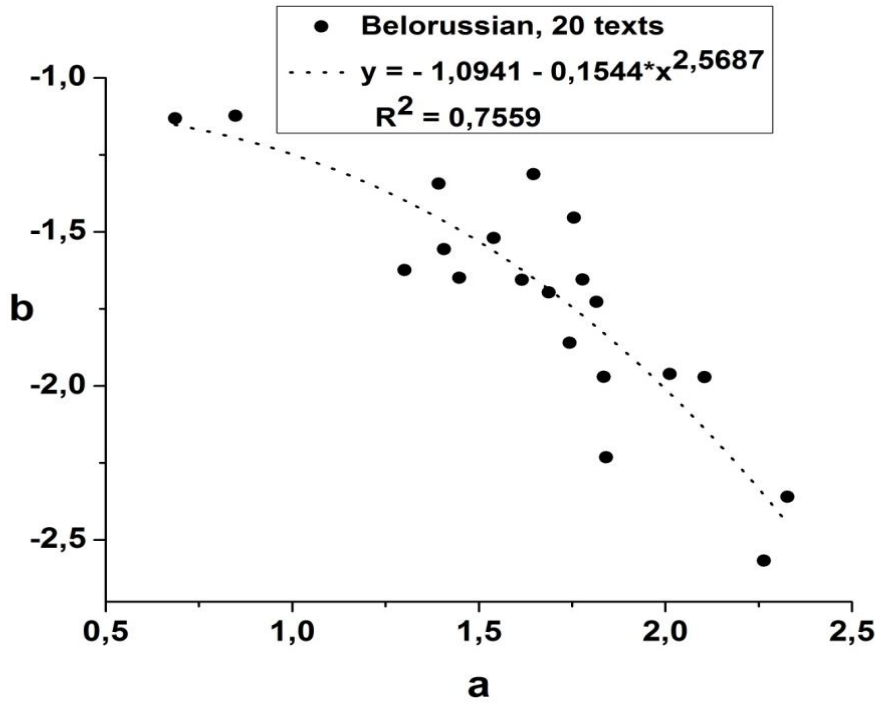
Slavic

In Slavic languages and in Hungarian, non-syllabic words (e.g. prepositions *k, s, v,* z, the Hungarian conjunction *s* < *és*) have been omitted because they represent preclitics. In Polish, Marx (2001) does not show them even in the data. A thorough theoretical analysis can be found in Antić, Kelih, Grzybek (2006).

Too short texts were omitted or a class with zero frequency has been added as the highest class.

Belorussian

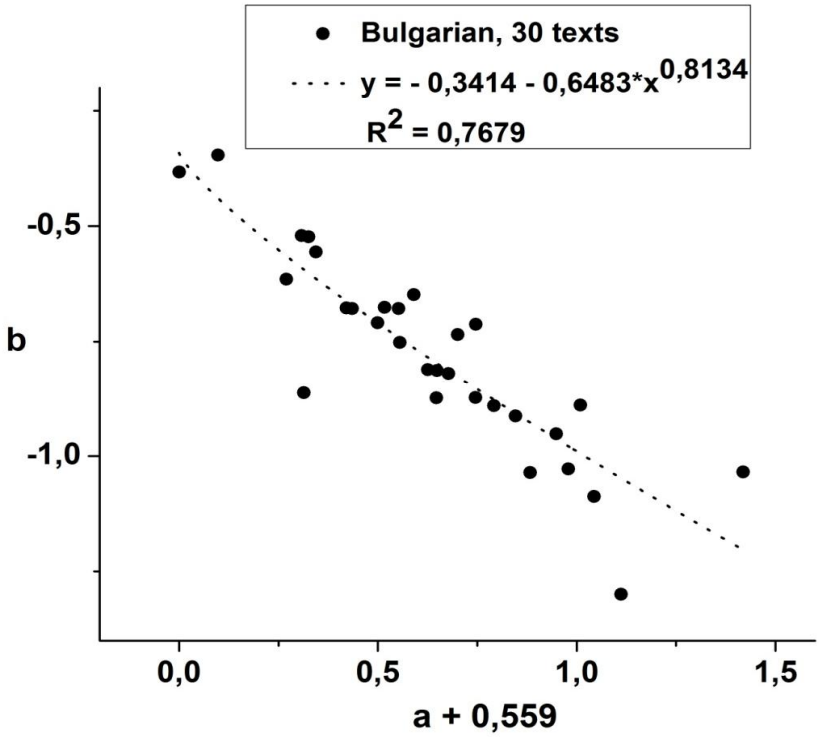
Belorussian (data from Kiyko 2007a)				
Text	**a**	**b**	**c**	**R^2**
T 1	1.3924	-1.3435	66.3758	0.8722
T 2	1.6469	-1.3132	19.8067	0.9284
T 3	1.8415	-2.2316	57.9727	0.9999
T 4	0.8477	-1.1236	66.2774	0.9753
T 5	1.7783	-1.6544	59.9054	0.9426
T 6	2.3270	-2.3600	35.2985	0.9818
T 7	1.6873	-1.6969	59.9945	0.9702
T 8	1.4474	-1.6489	39.3464	0.9820
T 9	1.8348	-1.9705	45.6707	0.9908
T 10	1.8157	-1.7273	60.4649	0.9708
T 11	2.2635	-2.5673	40.2058	0.9879
T 12	1.7431	-1.8604	36.6559	0.9906
T 13	1.5397	-1.5204	86.2119	0.9762
T 14	1.6154	-1.6553	94.6571	0.9755
T 15	2.1049	-1.9715	38.5546	0.9904
T 16	1.4069	-1.5566	57.3726	0.9180
T 17	0.6863	-1.1315	79.4483	0.9871
T 18	2.0115	-1.9617	99.0315	0.9955
T 19	1.3013	-1.6243	147.1718	0.9911
T 20	1.7552	-1.4541	75.6716	0.9432



Bulgarian

Bulgarian (data from Uhlířová 2001)				
Text	**a**	**b**	**c**	**R^2**
Rad 1	0.4503	-0.8899	29.7139	0.9703
Mumi	-0.2505	-0.5202	48.4264	0.9518
Iskra 4	0.1186	-0.8196	48.5726	0.9817
Adam	-0.4608	-0.3459	53.6756	0.9819
Genad 1	0.1420	-0.7349	48.5229	0.9746
Iskra 2	0.3900	-0.9519	54.2807	0.9567
Marg	0.0902	-0.8127	61.0670	0.9364
Iskra 1	-0.0594	-0.7094	64.0625	0.8974
Juri	-0.2449	-0.8630	78.7776	0.9918
Jorn	-0.0423	-0.6757	67.3816	0.9664
Iskra 5	-0.2329	-0.5233	71.3810	0.9480
Dam 1	0.1869	-0.8730	70.8256	0.9959
Kost	0.8594	-1.0348	53.9905	0.8945
Sasa 1	0.4846	-1.0874	93.2363	0.9791
Sasa 2	-0.1377	-0.6769	107.4957	0.9192
Boris 1	0.5526	-1.2997	111.6110	0.9844
Dam 2	0.0884	-0.8739	133.5170	0.9925
Jorn 1	0.0315	-0.6483	118.3234	0.9600
Cen 1	-0.5590	-0.3830	141.0954	0.9843

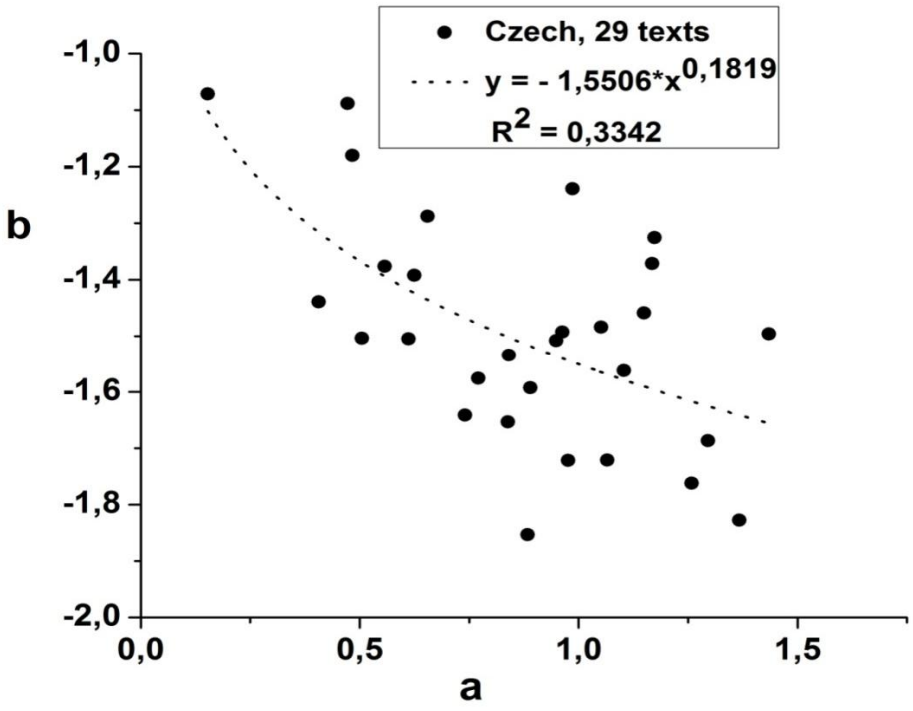
Jan 3	-0.0036	-0.7521	151.7521	0.9438
Jan 1	0.0663	-0.8110	191.7588	0.9653
Alb	0.3243	-1.0357	197.1367	0.9923
Cen 2	0.4198	-1.0278	184.0058	0.9728
Ziv 1	-0.1240	-0.6787	207.5983	0.9866
Jorn 2	0.1876	-0.7127	175.8068	0.9257
Ziv 2	0.2329	-0.8907	201.2096	0.9544
Jan 4	-0.2147	-0.5560	257.8´6366	0.9302
Jan 2	-0.2895	-0.6151	299.3918	0.9746
Boris 2	0.2870	-0.9131	270.9703	0.9446
Bacv 1	-0.0069	-0.6782	291.9774	0.9561



Czech

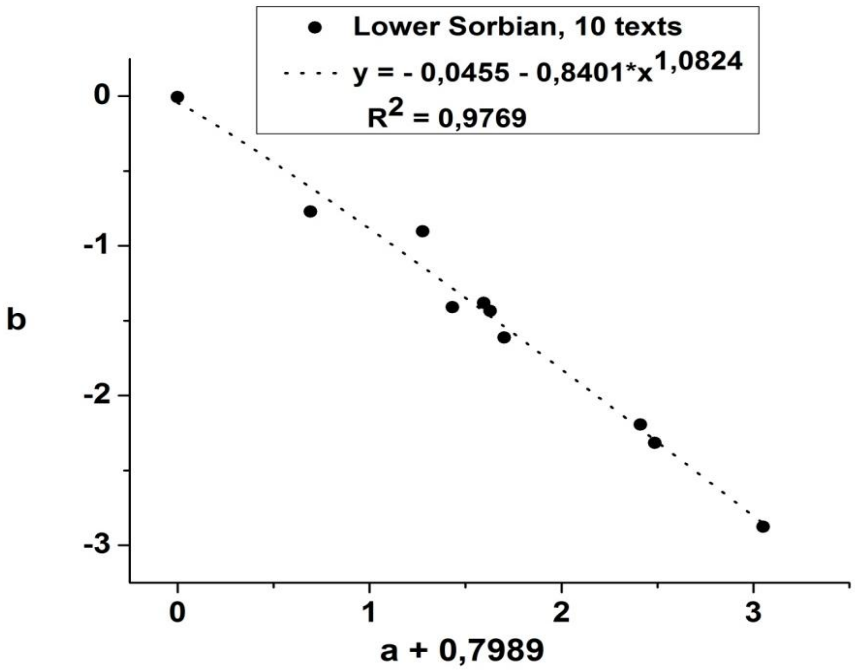
Czech (data from Uhlířová 1996)				
Text	**a**	**b**	**c**	**R^2**
T 310	1.0512	-1.4840	348.8976	0.9824
T 315	0.8897	-1.5929	389.6059	0.9928
T 316	0.5050	-1.5039	1319.1135	0.9976
T 319	1.1037	-1.5621	180.5538	0.9834
T 323	0.4723	-1.0883	410.2530	0.9122
T 327	1.2584	-1.7620	800.7489	0.9852
T 328	0.9759	-1.7218	278.2644	0.9929

T 329	0.5572	-1.3765	198.8570	0.9986
T 330	0.8385	-1.6535	487.8290	0.9952
T 334	0.8409	-1.5343	979.6303	0.9925
T 335	0.7404	-1.6416	823.0826	0.9977
T 338	0.4065	-1.4392	676.1232	0.9977
T 339	0.6118	-1.5049	1252.2630	0.9955
T 82	0.8832	-1.8533	1279.5943	0.9866
T 83	1.0658	-1.7213	1641.9689	0.9915
T 85	0.9632	-1.4926	736.4158	0.9884
T 89	0.7710	-1.5759	1032.6593	0.9921
T 817	0.9492	-1.5081	489.4586	0.9928
T 1	1.1490	-1.4588	154.6072	0.9797
T 2	1.2958	-1.6867	178.8423	0.9941
T 3	1.1738	-1.3258	139.1394	0.9262
T 6	1.1680	-1.3715	146.7677	0.9503
T 81	0.6554	-1.2878	1152.7330	0.9721
T 84	0.6248	-1.3922	757.3974	0.9957
T 325	0.4834	-1.1804	629.5734	0.9803
T 326	0.1532	-1.0713	884.1671	0.9694
T 830	1.4345	-1.4962	239.3593	0.9466
T 4	0.9865	-1.2391	129.0380	0.9516
T 5	1.3670	-1.8278	146.6994	0.9916



Lower Sorbian

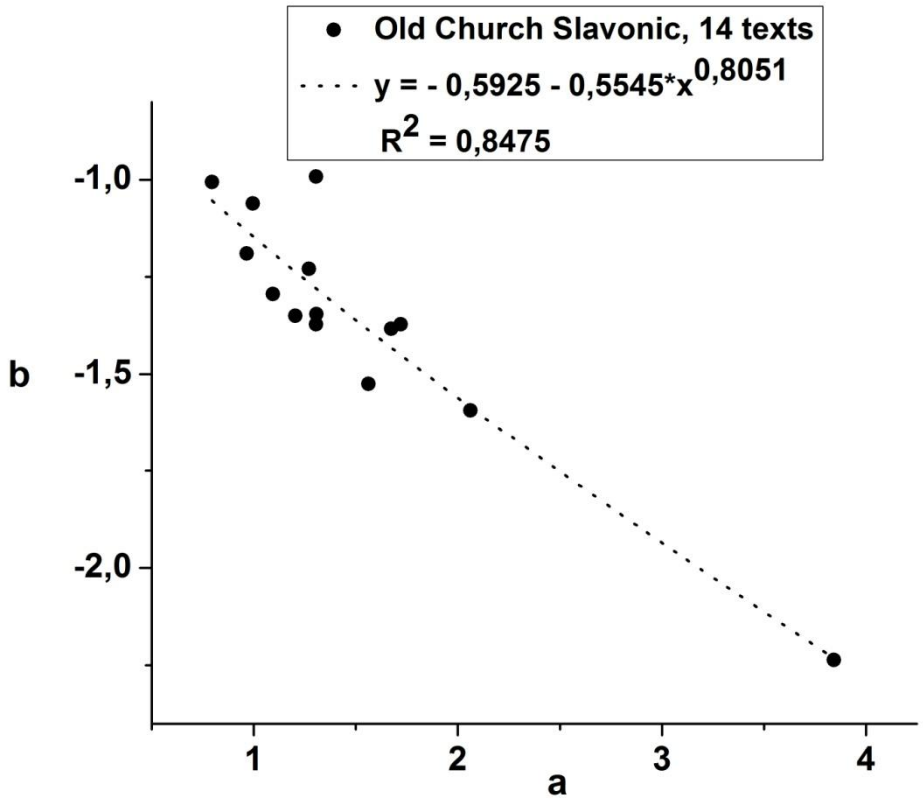
Lower Sorbian (data from Wilson 2006)				
Text	**a**	**b**	**c**	**R^2**
Text 1	1.6121	-2.1934	61.3223	0.9587
Text 2	1.6878	-2.3154	71.1291	0.9957
Text 3	0.6326	-1.4104	68.0488	0.9958
Text 4	0.4779	-0.9023	55.6849	0.9506
Text 5	-0.7989	-0.0065	55.4187	0.8308
Text 6	-0.1073	-0.7701	68.0054	0.9703
Text 7	0.9023	-1.6127	76.6680	0.9803
Text 8	2.2526	-2.8769	51.1729	0.9753
Text 9	0.8290	-1.4338	59.7700	0.9899
Text 10	0.7957	-1.3805	48.1545	0.9822



Old Church Slavonic

Old Church Slavonic (data from Rottmann 1997)				
Text	**a**	**b**	**c**	**R^2**
Text 1	1.5629	-1.5246	197.6037	0.9902
Text 2	0.9662	-1.1895	189.7025	0.9806
Text 3	3.8424	-2.2359	20.4751	0.9847
Text 4	0.9951	-1.0616	206.5180	0.9516

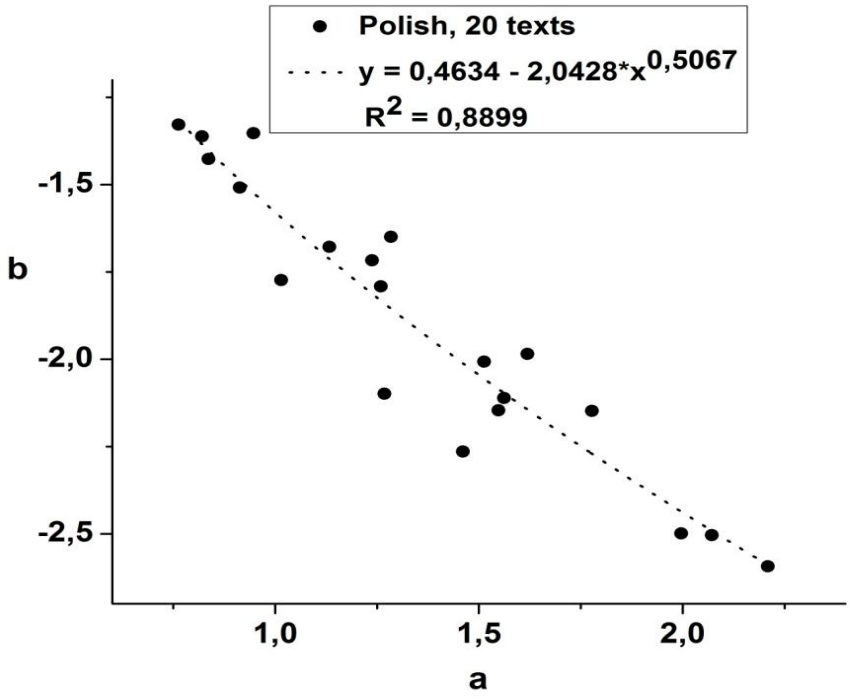
Text 5	1.3063	-0.9917	99.3302	0.8890
Text 6	1.7210	-1.3717	26.5576	0.9252
Text 7	1.3074	-1.3453	339.1914	0.9833
Text 8	2.0613	-1.5929	59.2114	0.9859
Text 9	1.6738	-1.3832	83.4349	0.9751
Text 10	1.2707	-1.2295	191.9095	0.9655
Text 11	1.3057	-1.3717	209.4601	0.9855
Text 12	1.2040	-1.3498	122.2372	0.9857
Text 13	1.0944	-1.2938	281.1007	0.9895
Text 14	0.7963	-1.0062	121.5743	0.9438
Text 15	Text too short			



Polish

Polish (data from Marx 2001)				
Text	**a**	**b**	**c**	**R^2**
Letter 20.09.1901	1.2680	-2.0989	79.0828	0.9959
Letter 02.10.1901	1.4606	-2.2647	91.0922	0.9966
Letter 06.10.1901	1.1328	-1.6782	72.7643	0.9935
Letter 24.04.1903	1.5130	-2.0069	73.0801	0.9947

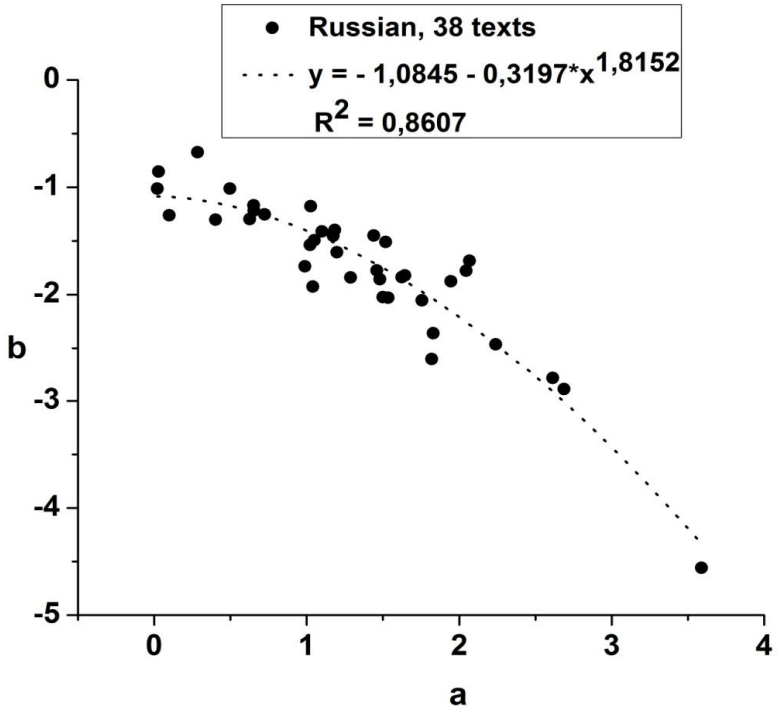
Letter 04.05.1903	2.2093	-2.5936	64.1604	0.9951
Letter 11.05.1903	1.2379	-1.7167	75.8902	0.9967
Letter 12.05.1903	1.6193	-1.9847	67.1067	0.9936
Letter 12.08.1903	1.9967	-2.4990	71.9104	0.9987
Letter 14.08.1905	1.2594	-1.7917	109.1000	0.9984
Letter 18.08.1905	0.9465	-1.3527	73.2497	0.9658
Letter 26.02.1905	1.7777	-2.1485	113.7723	0.9986
Letter 01.03.1905	1.0150	-1.7733	125.9099	0.9993
Letter 23.03.1905	1.5482	-2.1466	100.7785	0.9968
Letter 06.04.1908	0.8206	-1.3621	55.4468	0.9706
Letter 28.09.1908	1.5612	-2.1113	105.9218	0.9993
Letter 29.09.1908	0.7631	-1.3286	63.4868	0.9656
Letter 10.11.1908	2.0720	-2.5039	57.0074	0.9999
Letter 21.11.1908	0.9131	-1.5087	91.6806	0.9894
Letter 24.05.1909	1.2838	-1.6498	62.5699	0.9847
Letter 31.05.1910	0.8361	-1.4265	65.7751	0.9913



Russian

Russian (data from Girzig 1997)				
Text	**a**	**b**	**c**	**R^2**
Text 1	1.8198	-2.6006	90.9922	0.9973
Text 2	1.1741	-1.4536	55.9250	0.9212

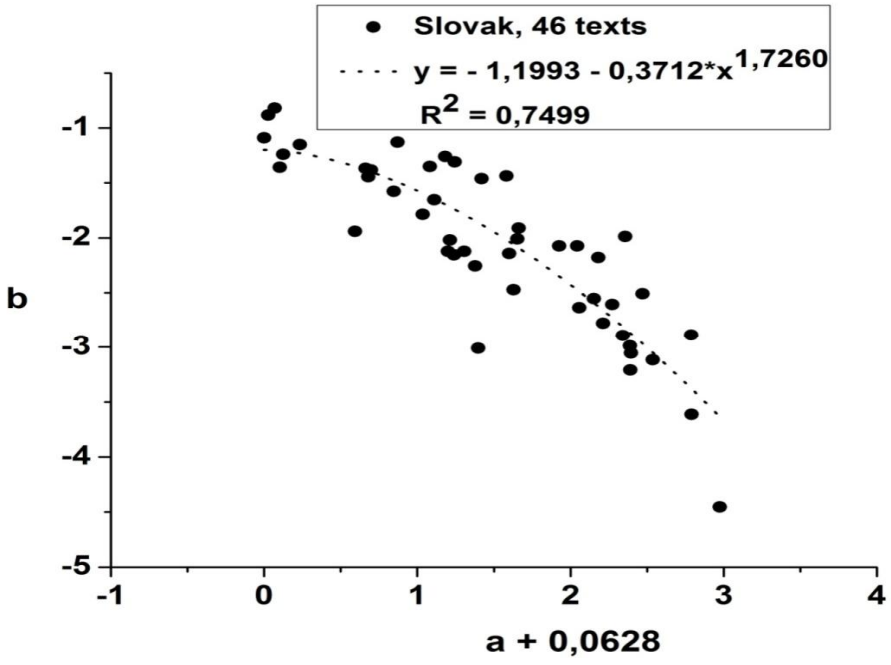
Text 3	1.5001	-2.0247	43.9919	0.9999
Text 4	0.6547	-1.2183	50.4482	0.9361
Text 5	2.0451	-1.7776	16.7882	0.8616
Text 6	1.0500	-1.4938	34.7737	0.9834
Text 7	0.2855	-0.6755	42.8956	0.5487
Text 8	2.6878	-2.8887	44.9380	0.9793
Text 9	0.9877	-1.7377	53.9546	0.9979
Text 10	0.4967	-1.0138	17.9343	0.9805
Text 11	0.0212	-1.0119	52.7775	0.9448
Text 12	1.5346	-2.0291	56.7806	0.9904
Text 13	2.2391	-2.4644	15.9810	0.9925
Text 14	1.8304	-2.3608	50.9387	0.9988
Text 15	1.2889	-1.8414	40.7837	0.9772
Text 16	0.7251	-1.2526	142.6674	0.9625
Text 17	1.0268	-1.1770	18.6350	0.9330
Text 18	1.1850	-1.3994	61.9884	0.9704
Text 19	1.9461	-1.8763	36.4135	0.9820
Text 20	0.6521	-1.1684	12.9059	0.9329
Text 21	0.0287	-0.8552	64.3479	0.9525
Text 22	1.4597	-1.7764	24.1545	0.9861
Text 23	1.1003	-1.4114	17.8963	0.9907
Text 24	1.0398	-1.9243	36.9761	0.9994
Text 25	0.6270	-1.2976	50.6740	0.9736
Text 26	1.5173	-1.5104	52.5809	0.8184
Text 27	0.4020	-1.3024	55.7960	0.9884
Text 28	1.4799	-1.8580	23.9290	0.9913
Text 29	0.0984	-1.2616	55.9290	0.9914
Text 30	2.6143	-2.7873	11.9484	0.9897
Text 31	3.5881	-4.5605	29.0020	0.9999
Text 32	1.0230	-1.5375	260.8585	0.9947
Text 33	1.7551	-2.0534	234.6309	0.9988
Text 34	1.6437	-1.8230	340.6770	0.9908
Text 35	1.1975	-1.6047	451.8179	0.9962
Text 36	1.4393	-1.4508	286.8780	0.9674
Text 37	1.6243	-1.8372	109.6030	0.9915
Text 38	2.0679	-1.6849	122.9063	0.9700



Slovak

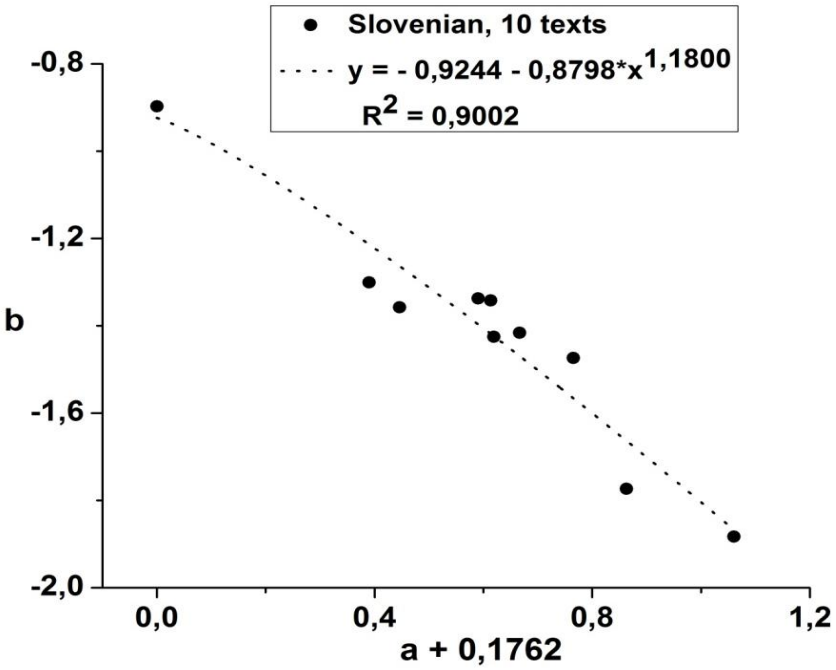
Slovak (data from Nemcová, Altmann 1994)				
Text	**a**	**b**	**c**	**R^2**
R1	2.1179	-2.1789	32.7476	0.9788
R2	2.7251	-2.8859	75.2078	0.9944
R3	2.4078	-2.5086	28.0965	0.9945
R4	2.7268	-3.6114	45.0787	0.9652
R5	1.2442	-2.1224	70.9071	0.9931
R6	2.0907	-2.5537	53.9218	0.9970
R7	0.6351	-1.3811	60.6317	0.9269
R8	2.3250	-2.9868	41.7734	0.9996
R9	1.5907	-2.0082	47.6438	0.9455
R10	2.1504	-2.7795	24.9665	0.9953
B1	0.7845	-1.5764	26.8927	0.9611
B2	2.3276	-3.2089	25.0232	0.9929
B3	2.4738	-3.1169	28.0371	0.9937
B4	1.0486	-1.6535	31.8582	0.9866
B5	1.9950	-2.6368	30.0522	0.9913
B6	2.2947	-1.9876	11.4413	0.8083
B7	2.3320	-3.0543	31.9758	0.9977
B8	1.1507	-2.0185	36.9795	0.9988
B9	2.2109	-2.6069	210.0262	1.0000
B10	1.1763	-2.1540	1032.8061	0.9981

M1	2.9125	-4.4529	44.0000	1.0000
M2	0.9734	-1.7844	37.9105	0.9840
M3	0.5309	-1.9397	58.0000	1.0000
M4	1.3145	-2.2556	51.0132	0.9996
M5	1.5368	-2.1428	32.9667	0.9991
M6	1.3358	-3.0091	47.0055	0.9990
M7	1.5651	-2.4717	74.9004	0.9940
M8	1.1373	-2.1234	34.0047	0.9999
C1	2.2780	-2.8920	464.6398	0.9984
C2	1.5985	-1.9111	260.0237	0.9738
C3	1.9815	-2.0720	460.2925	0.9775
J1	1.0190	-1.3513	170.3784	0.9847
J2	0.8080	-1.1297	109.1389	0.9310
J3	1.3574	-1.4610	151.3128	0.9579
J4	1.1191	-1.2588	140.9890	0.9654
J5	1.5183	-1.4356	138.9120	0.9745
J6	1.8640	-2.0742	117.6928	0.9989
J7	1.1825	-1.3075	127.7935	0.9923
J8	0.0062	-0.8182	155.0248	0.9693
BED	0.0620	-1.2392	543.2900	0.9960
DUS	0.6002	-1.3674	419.7150	0.9907
KAP	0.0394	-1.3572	775.4323	0.9988
ZEL	0.1718	-1.1517	1357.3317	0.9970
LAS	-0.0628	-1.0890	842.2996	0.9957
BAL	0.6168	-1.4431	2174.3386	0.9966
MA	-0.0351	-0.8839	196.3447	0.9942



Slovenian

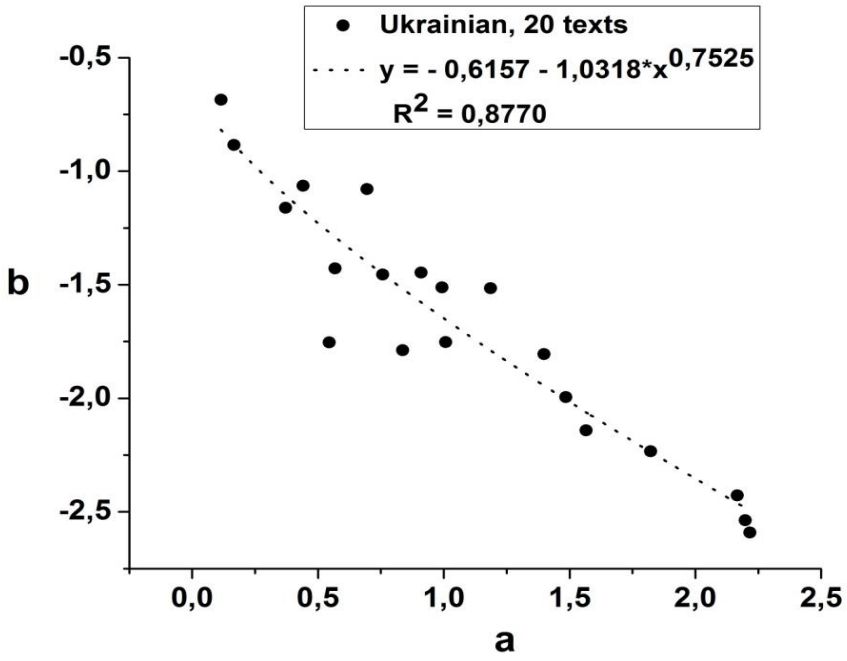
Slovenian (data from Antić, Kelih, Grzybek 2006)				
Text	**a**	**b**	**c**	**R^2**
T 1	0.5894	-1.4741	265.5665	0.9957
T 2	0.6869	-1.7736	477.6039	0.9981
T 3	0.2696	-1.3576	505.9051	0.9910
T 4	0.4433	-1.4254	375.0351	0.9896
T 5	0.8840	-1.8831	380.4095	0.9954
T 6	-0.1762	-0.8975	432.6304	0.9856
T 7	0.4902	-1.4165	440.0140	0.9931
T 8	0.4144	-1.3376	669.2370	0.9778
T 9	0.4376	-1.3420	421.3910	0.9818
T 10	0.2136	-1.3008	558.6851	0.9891



Ukrainian

Ukrainian (data from Best, Zinenko 1999) (Poetry)				
Text	**a**	**b**	**c**	**R^2**
T 1	0.9105	-1.4461	38.9118	0.9925
T 2	0.9932	-1.5116	41.9260	0.9983
T 3	0.3712	-1.1616	60.4434	0.9505
T 4	0.8366	-1.7881	38.9358	0.9935
T 5	0.1662	-0.8841	38.5355	0.8891

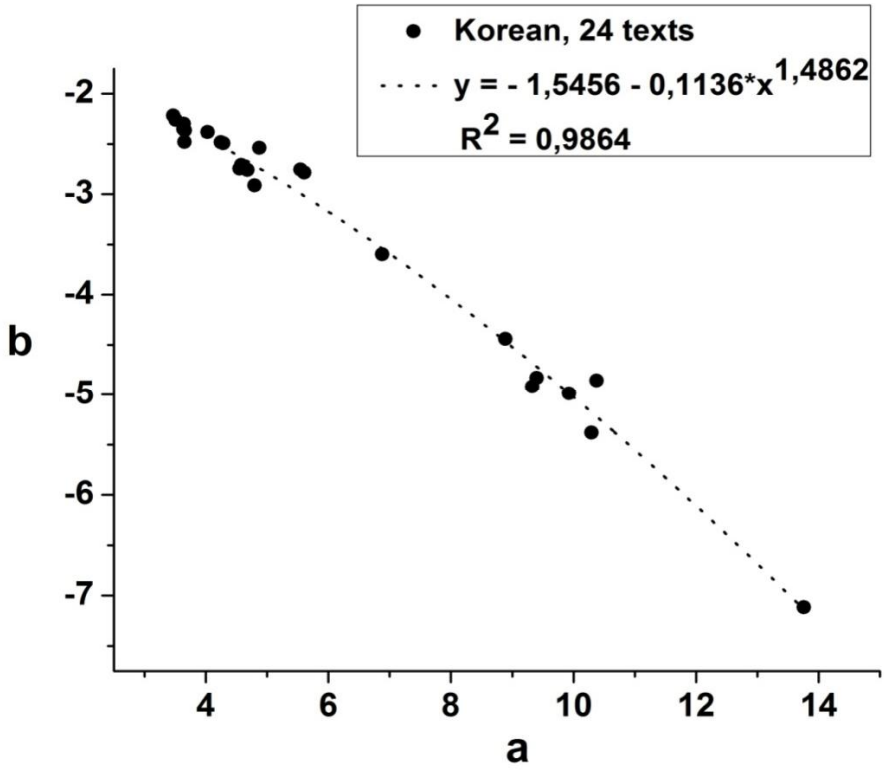
T 6	2.2166	-2.5911	29.2201	0.9278
T 7	0.5682	-1.4279	67.9038	0.9941
T 8	1.3980	-1.8058	28.8176	0.9608
T 9	2.1672	-2.4278	27.0430	0.9980
T 10	0.6948	-1.0790	30.5629	0.8244
T 11	2.1983	-2.5368	43.8715	0.9920
T 12	0.5453	-1.7535	93.9934	0.9996
T 13	1.5664	-2.1413	29.8771	0.9720
T 14	0.4414	-1.0646	61.0543	0.9991
T 15	1.0079	-1.7524	55.0485	0.9979
T 16	0.1147	-0.6855	32.6504	0.8453
T 17	0.7576	-1.4551	101.3175	0.9657
T 18	1.8226	-2.2331	47.8124	0.9834
T 19	1.1856	-1.5155	42.2784	0.9081
T 20	1.4852	-1.9946	41.1575	0.9860



Korean

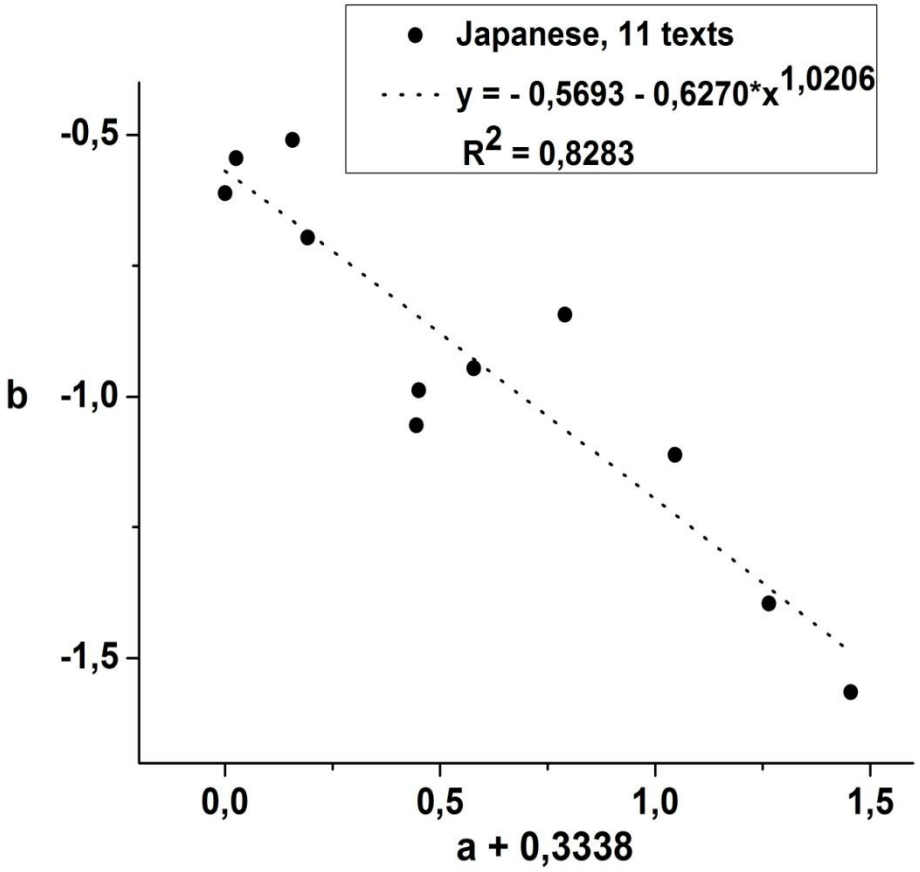
Korean (data from Kim, Altmann 1996)				
Text	**a**	**b**	**c**	**R^2**
T1	4.2450	-2.4847	33.3197	0.9657
T2	4.7998	-2.9133	37.3874	0.9704
T3	4.0320	-2.3826	42.9835	0.8786
T4	3.6531	-2.4816	96.5581	0.9920
T5	3.4729	-2.2170	71.5419	0.9737

T6	9.9271	-4.9871	2.4023	0.9039
T7	10.2943	-5.3822	2.8634	0.9776
T8	13.7563	-7.1149	0.5351	0.9469
T9	3.5075	-2.2610	116.7914	0.9083
T10	5.5447	-2.7567	8.3129	0.9362
T11	8.8874	-4.4400	2.3627	0.9784
T12	9.3271	-4.9199	8.6625	0.9539
T13	3.6426	-2.2990	76.5288	0.9681
T14	4.5522	-2.7470	104.7626	0.9667
T15	4.6789	-2.7594	23.2045	0.9668
T16	9.3979	-4.8299	4.7295	0.9445
T17	3.6376	-2.3538	108.6116	0.9904
T18	4.5805	-2.7091	107.8526	0.9535
T19	3.6604	-2.3664	207.4840	0.9817
T20	4.2839	-2.4947	96.3604	0.9554
T21	5.6094	-2.7843	10.1605	0.9904
T22	4.8734	-2.5405	25.6925	0.9848
T23	10.3744	-4.8582	1.2302	0.9616
T24	6.8793	-3.5973	23.8554	0.9857



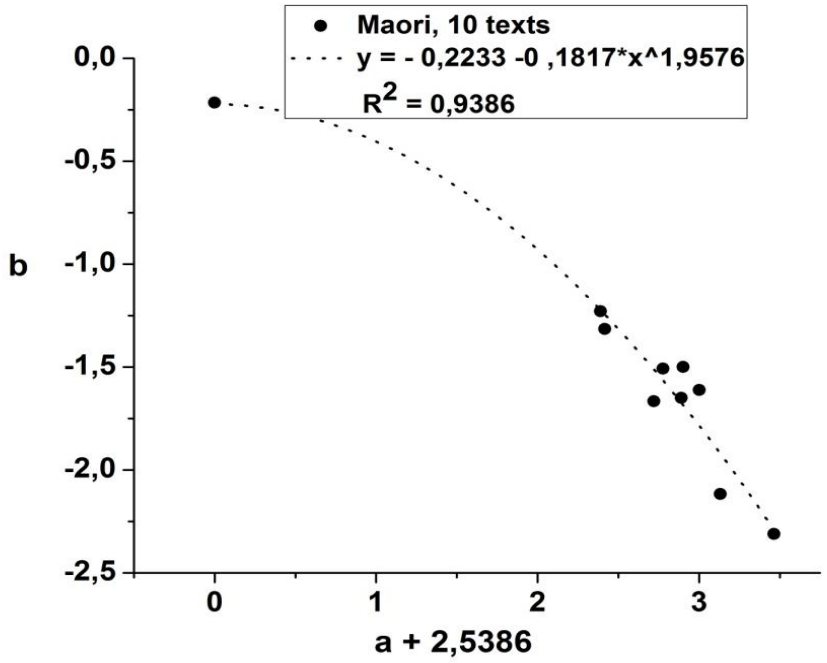
Japanese

Japanese (data from Riedemann 1997)				
Text	**a**	**b**	**c**	**R^2**
Text 1	0.7125	-1.1120	100.9124	0.9737
Text 2	0.2446	-0.9451	57.4642	0.9694
Text 3	0.9305	-1.3957	80.9576	0.9925
Text 4	-0.1417	-0.6957	160.8355	0.9377
Text 5	0.1112	-1.0544	146.6624	0.9930
Text 6	-0.3338	-0.6117	161.4454	0.9638
Text 7	1.1206	-1.5647	67.8309	0.9879
Text 8	-0.1766	-0.5100	86.3359	0.8781
Text 9	-0.3080	-0.5449	76.3266	0.9736
Text 10	0.4564	-0.8431	36.1676	0.9219
Text 11	0.1164	-0.9871	130.4897	0.9886



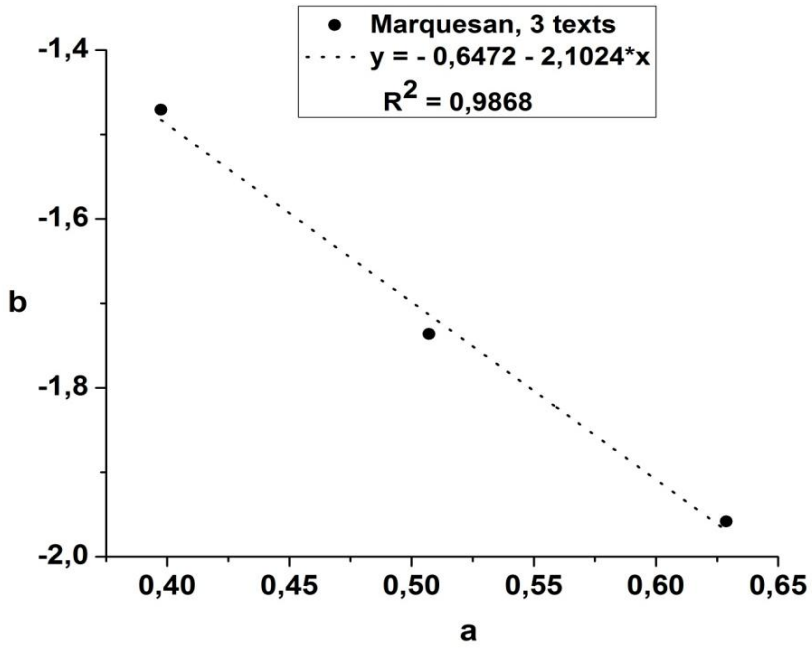
Polynesian: Maori

Maori (data from Krupa 1994)				
Text	**a**	**b**	**c**	**R^2**
T 1	0.9250	-2.3111	766.9139	0.9999
T 2	-0.1218	-1.3149	576.6504	0.9983
T 3	0.5921	-2.1171	607.0137	1.0000
T 4	-2.5386	-0.2149	545.0149	0.9999
T 5	0.2379	-1.5082	592.4472	0.9947
T 6	0.3618	-1.4991	548.8161	0.9995
T 7	0.3515	-1.6500	394.9810	0.9998
T 8	0.4632	-1.6117	374.6147	0.9970
T 9	-0.1493	-1.2296	598.8756	0.9998
T 10	0.1819	-1.6666	582.0558	0.9999



Marquesan

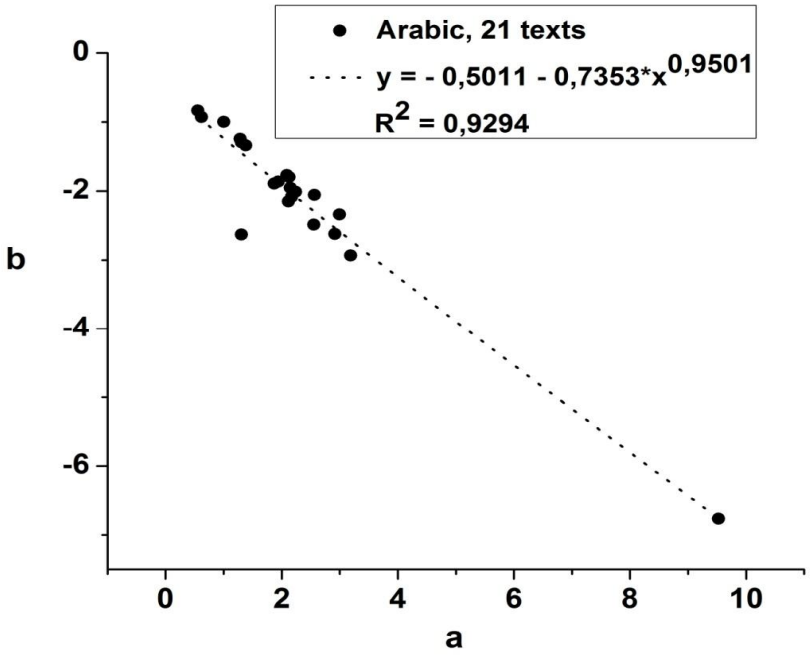
Marquesan (data from Krupa 1993)				
Text	**a**	**b**	**c**	**R^2**
T 1	0.5071	-1.7363	1025.9471	0.9952
T 2	0.6288	-1.9584	606.6434	0.9977
T 3	0.3975	-1.4709	194.8115	0.9954



Semitic: Arabic

Arabic (data from Abbe 2000)					
Text	**a**	**b**	**c**	**k**	**R^2**
Letter 1	1.0008	-0.9981	46.5048	-	0.8777
Letter 2	2.1185	-2.1504	62.9398	2	0.9811
Letter 3	1.8706	-1.8912	51.9788	2	0.9950
Letter 4	2.5539	-2.4852	81.0123	2	0.9975
Letter 5	1.3050	-2.6293	39.9445	2	0.9709
Letter 6	2.1705	-2.0864	52.9800	2	0.9959
Letter 7	1.2842	-1.2467	21.4827	-	0.9307
Letter 8	2.2418	-2.0122	32.9905	2	0.9987
Letter 9	1.3080	-1.2926	25.6495	-	0.9636
Letter 10	2.0892	-1.7688	35.9719	2	0.9935
Letter 11	9.5255	-6.7668	18.0000	2, 3	1.0000
Letter 12	1.3823	-1.3388	38.9977	-	0.9362
Letter 13	1.9419	-1.8635	38.8975	2	0.9428
Letter 14	2.1477	-1.9543	30.9367	2	0.9397
Letter 15	2.9942	-2.3375	31.8908	2	0.9841
Letter 16	2.5653	-2.0570	24.9208	2	0.9503
Letter 17	2.1251	-1.7971	21.8513	2	0.8789
Letter 18	0.5525	-0.8357	48.9235	2	0.9479
Letter 19	3.1876	-2.9340	30.0093	2	0.9948
Letter 20	0.6203	-0.9273	31.3548	-	0.8665

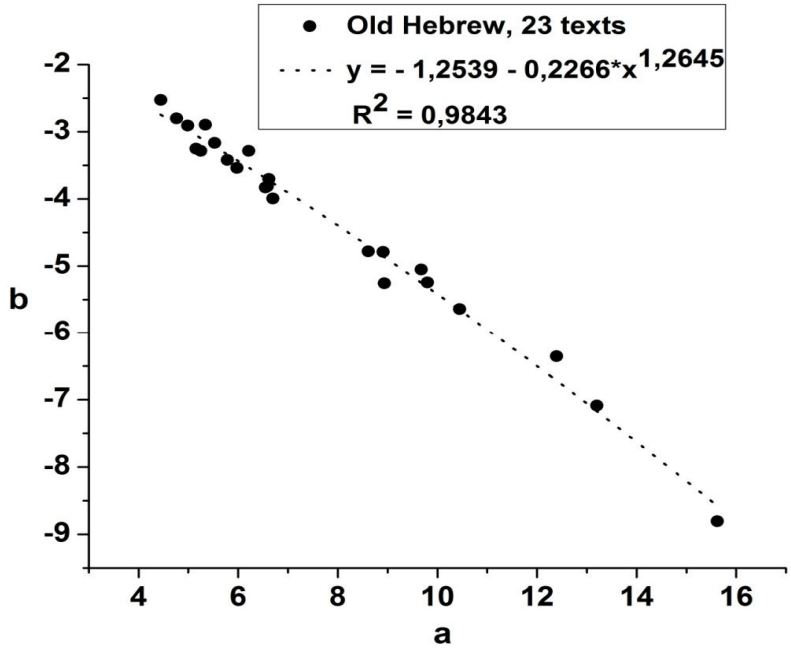
Letter 21	2.9187	-2.6220	34.0351	2	0.9683
(Letter 11 of Arabic is very irregular)					



Old Hebrew

Old Hebrew (data from Balschun 1997)				
Text	**a**	**b**	**c**	**R^2**
Psalm 5	8.9138	-4.7914	0.6618	0.9391
Psalm 7	8.9388	-5.2587	1.4348	0.9935
Psalm 10	15.6237	-8.8084	0.0901	0.9915
Psalm 16	9.6745	-5.0510	0.3051	0.9689
Psalm 18	5.5312	-3.1680	12.2110	0.9881
Psalm 19	5.2503	-3.2841	6.1019	0.9999
Psalm 21	4.4464	-2.5276	4.7202	0.9380
Psalm 22	6.6184	-3.7030	4.8485	0.9776
Psalm 25	4.9892	-2.9079	6.0595	0.9854
Psalm 26	4.7626	-2.8006	3.7175	0.9797
Psalm 31	9.7980	-5.2430	0.9472	0.9590
Psalm 33	5.7819	-3.4202	5.0054	0.9583
Psalm 34	5.9776	-3.5397	4.8014	0.9829
Psalm 35	6.5881	-3.8135	5.2228	0.9977
Psalm 38	12.3915	-6.3548	0.1799	0.9909
Psalm 44	6.2107	-3.2874	3.5050	0.9516
Psalm 49	5.1555	-3.2519	8.2443	0.9246
Psalm 55	8.6160	-4.7837	1.6531	0.9776

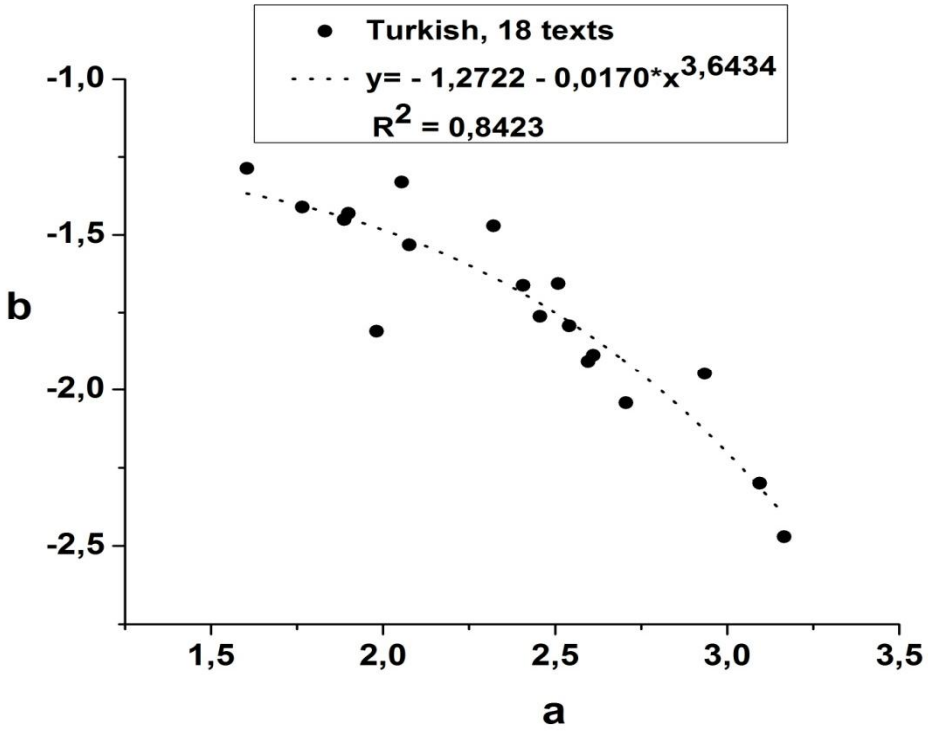
Psalm 57	6.5511	-3.8307	2.5217	0.9962
Psalm 68	13.2006	-7.0888	0.3219	0.9323
Psalm 105	6.6977	-3.9941	7.0625	0.9977
Psalm 107	10.4507	-5.6423	0.9628	0.9933
Psalm 145	5.3398	-2.8959	3.9312	0.9788



Turkic: Turkish

Turkish (data from Altmann, Erat, Hřebíček 1996)				
Text	**a**	**b**	**c**	**R^2**
T1	2.5956	-1.9068	120.8349	0.8985
T2	2.4068	-1.6625	79.3682	0.9241
T3	1.9816	-1.8098	164.3958	0.9831
T4	1.6045	-1.2865	96.6657	0.8952
T5	1.8877	-1.4510	143.6048	0.8960
T6	2.6105	-1.8873	58.1740	0.9357
T7	2.7062	-2.0423	113.6451	0.9863
T8	3.0942	-2.2991	48.2113	0.9863
T9	2.9349	-1.9479	40.5279	0.9479
T10	2.5408	-1.7921	73.9569	0.9848
T11	3.1656	-2.4709	90.8757	0.9957
T12	2.0763	-1.5322	110.6910	0.9541
T13	1.8998	-1.4306	172.8509	0.9686
T14	2.3211	-1.4714	61.7202	0.9041
T15	2.5087	-1.6571	148.6440	0.9713
T16	2.0539	-1.3309	63.2873	0.8496

T17	1.7656	-1.4107	111.5246	0.9705
T18	2.4556	-1.7617	88.8462	0.9275



Uzbek

Uzbek (data from Kaydanova 2004/5)				
Text	**a**	**b**	**c**	**R^2**
T 1	2.8665	-1.9061	43.6548	0.9501
T 2	4.6788	-3.3063	17.5198	0.9079
T 3	1.9445	-1.5967	42.5662	0.9142
T 4	2.7611	-2.0535	42.5291	0.9737
T 5	3.1030	-2.0623	56.2442	0.9594
T 6	3.8419	-2.0332	77.2856	0.9966
T 7	2.9784	-1.9797	22.5047	0.9835
T 8	2.5406	-1.8743	61.7870	0.9208
T 9	4.0695	-2.5100	14.5458	0.9874
T 10	3.6507	-2.7791	34.9762	0.9955
T 11	3.1049	-2.2465	28.8502	0.9818
T 12	1.8981	-1.5275	72.7949	0.9214
T 13	3.3195	-2.1380	15.2216	0.9362
T 14	3.1216	-2.3595	43.8737	0.9964
T 15	2.7960	-2.2184	62.3033	0.9953
T 16	2.8656	-2.0068	41.2379	0.9437

T 17	4.1307	-3.0425	23.4256	0.9885
T 18	19.2809	-9.8014	0.0180	0.9045
T 19	4.6588	-2.9496	24.9927	0.9809
T 20	2.6065	-1.9151	58.8040	0.9829

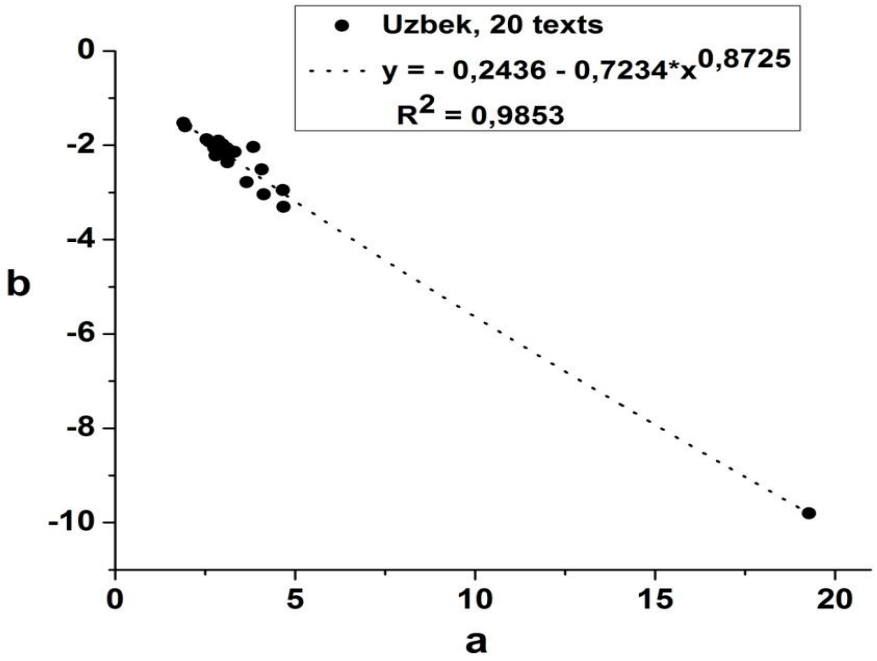


Figure 4.1. The series of b = f(a) graphs and fittings

Discussion

Some of the results are not satisfactory but we do not want to search for individual boundary conditions – such an enterprise would be impossible. The overwhelming majority of the data conforms to the model; a determination coefficient smaller than 0.8 is rather an exception. In some languages one finds a great variation of the parameter *b*. This can be considered the result of the effect of writer, style, text-sort, historical development, etc. i.e. a case of emergence in language.

One can state that an older stage of a language has in general a greater *a* than a younger one. Though it cannot be decided for all languages analyzed, we find Latin at a higher position than Italian, Spanish, Portuguese and French; Old Icelandic is "over" Modern Icelandic; Early Modern English is over English; Old Church Slavonic "is over" Slovak, Polish, Czech, Bulgarian, Russian, Slovenian, Ukrainian, Lower Sorbian, but not over Belorussian. In German, we have a sequence: Old High German – Middle High German – Early New High German (Luther) – Modern German – Early New High German (Dürer). Nevertheless, many other analyses of the historical development of individual languages would be necessary in order to state some developmental regularity. In Slavic languages the non-syllabic words (preposition *k, s, v, z,*…) developed after the family

disintegrated and the "jer" had been eliminated in many words. It would be premature to set up typological hypotheses but the highest place of Old Hebrew is a hint. The last item contains texts encompassing 500 years and a special text sort (letters), hence the hypothesis of correlations between age and parameter *a* must be enriched by boundary conditions or specialized to individual text sorts and at least the individual centuries must be considered separately.

Further, one can see that Indo-European languages have in general a smaller parameter *a* than the languages of other genetic groups. However, Chinese is an exception. The Polynesian languages are at low positions.

Some few fittings are not quite good but this could be expected in a system in which emergence of new features – mostly local ones – is a quite normal phenomenon. Usually one calls them outliers but only specialists for the given language, historical epoch and text sort could find the cause of the deviation. This is the normal state of research which must consider boundary conditions.

Now, the capturing of the relationship between parameter *a* and *b* in the Zipf-Alekseev formula yields always a very regular function that can be expressed by the power function $y = k + mx^b$ or linearly, i.e. with $b = 1$. In some languages we displaced the function slightly to the right as signalized by (*a* + number) below the figures, in order to plot the values in the positive domain. Of course, the area filled by the dots of $b = f(a)$ could be captured also by an ellipse but preliminarily one can dispense with this procedure. Nevertheless, the above function opens further vistas and can lead to a look at the internal life of language. The simple question: which are the properties with which the parameters *k*, *m* and *b* in the above function are associated? They are elements of a system whose depth cannot be even conjectured. Perhaps, later on, one will be able to obtain a survey but first many more languages and texts must be analyzed and many more properties must be measured. If one sets up a differential equation like (2), then parameter *a* is given by some properties of language, parameter *b* is given by the requirements of the speaker/writer which may be of different origin: style, age, text-sort, etc. Though we consider *b* as a function of *a*, the reverse influence is always present because speakers are the cause of changes in language. Hence a thorough investigation of texts in one language would be appropriate in order to track down the background of the individual parameters (cf. Köhler 2005). Thus word length is not simply a surface property, it depends both on language, its development and its use. What is more, the first impression says that if a language diversifies in dialects, parameter *a* decreases, cf.. German vs. Low German and Palatine.

Tables 4.3 and 4.4. display the relationships of *a* and *b* in all the data analyzed.

Table 4.3
Linear relation between *a* and *b*

		Linear fitting		
Language	**# of texts**	**Intercept**	**Slope ascending**	**R^2**
German	26	-1,1365	-1,3829	0,7392
Icelandic	20	-0,6246	-1,3786	0,9253
Chinese 2	21	-3,5423	-1,1300	0,6751
Gaelic	31	-1,8008	-1,1290	0,8280
Faeroese	21	-1,6349	-1,0480	0,8174
Old Icelandic	20	-1,7110	-0,9982	0,9164
Low German	27	-2,1622	-0,9883	0,9025
Chinese 1	19	-1,5725	-0,9702	0,9275
Lower Sorbian	10	-0,7221	-0,9272	0,9782
French	16	-1,5272	-0,9261	0,8219
Spanish	20	-0,9730	-0,9012	0,9060
Early New High German: Luther's letters	21	-0,8884	-0,8933	0,9008
Palatine	16	-1,3967	-0,8896	0,8644
Polish	20	-0,6958	-0,8824	0,8906
Middle High German	18	-1,8385	-0,8797	0,7610
Old High German	21	-0,8300	-0,8267	0,8654
Swedish	18	-1,0908	-0,8262	0,8609
Slovak	46	-0,9046	-0,8217	0,7192
Roumanian	50	-1,1376	-0,8134	0,8345
Estonian	19	-0,5766	-0,8107	0,9048
Early New High German: Dürer's letters	20	-0,0399	-0,8084	0,9255
Russian	38	-0,6786	-0,8066	0,7957
Italian	54	-0,6580	-0,7870	0,8539
Koine Greek	10	-0,4326	-0,7104	0,9556
Sami	25	-0,8101	-0,6979	0,7761
Portuguese	20	-0,8470	-0,6923	0,8868
Dutch	16	-1,4525	-0,6868	0,8424
Hungarian	22	-0,7088	-0,6786	0,8387
Turkish	18	-0,1871	-0,6520	0,8031
Arabic	21	-0,6127	-0,6498	0,9327
English	21	-1,6382	-0,6415	0,4955
Japanese	11	-0,5650	-0,6324	0,8473
Bulgarian	30	-0,7349	-0,5892	0,7690
Finnish	20	-0,7084	-0,5707	0,7484

Old Hebrew	23	-0,3615	-0,5134	0,9831
Latin	18	-0,7127	-0,4691	0,8183
Korean	24	-0,6060	-0,4480	0,9798
Quechua	24	-0,5520	-0,4111	0,9765
Old Church Slavonic	14	-0,7733	-0,3847	0,8582
Latvian	12	-1,3069	-0,3668	0,2643
Czech	29	-1,1835	-0,3560	0,2835
Inuktitut	19	-0,2293	-0,3510	0,9901

Table 4.4
Power relation between parameters *a* and *b*

Language	**Power law fitting function**	**R^2**
German	$b = - 0,6817 - 0,7655*(a + 0,7096)^{1,6867}$	0,7758
Icelandic	$b = - 2,7711 - 0,1193*a^{2,5705}$	0,9305
Chinese 2	$b = 0,2171 - 0,7518*(a + 3,4498)^{1,3116}$	0,7651
Gaelic	$b = - 1,0873 - 0,9861*(a + 0,7400)^{1,1476}$	0,8051
Faeroese	$b = - 1,8560 - 0,7942*b^{1,3220}$	0,8101
Old Icelandic	$b = - 2,0238 - 0,7456*a^{1,1586}$	0,9134
Low German	$b = - 0,4129 - 0,4226*(a + 2,3531)^{1,6023}$	0,9073
Chinese 1	$b = - 0,2962 - 1,8461*a^{0,7486}$	0,9253
Lower Sorbian	$b = - 0,3861 - 0,5425*(a + 0,7989)^{1,3744}$	0,9666
French	$b = 0,1047 - 0,8730*(a + 1,7913)^{1,1270}$	0,7785
Spanish	$b = - 0,3340 - 1,0986*(a + 0,5006)^{0,7741}$	0,8837
Early New High German: Luther's letters	$b = 2,2773 - 4,0354*(a + 0,6225)^{0,3009}$	0,8830
Palatine	$b = - 0,3757 - 0,9753*(a + 1,0653)^{0,8672}$	0,8374
Polish	$b = 0,4634 - 2,0428*a^{0,5067}$	0,8899
Middle High German	$b = - 1,7788 - 0,9088*(a + 0,2414)^{1,1686}$	0,9368
Old High German	$b = - 0,4756 - 1,1678*(a + 0,5222)^{0,8125}$	0,8548
Swedish	$b = - 0,1099 - 0,5462*(a + 1,4422)^{1,4830}$	0,8638
Slovak	$b = - 1,1993 - 0,3712*(a + 0,0628)^{1,7260}$	0,7499
Roumanian	$b = - 1,1278 - 0,3851*(a + 0,4433)^{1,6427}$	0,8727
Estonian	$b = 0,1118 - 1,4133*a^{0,7183}$	0,9018
Early New High German: Dürer's letters	$b = - 0,4009 - 0,5857*(a + 2,1703)^{1,1334}$	0,8985
Russian	$b = - 1,0845 - 0,3197*a^{1,8152}$	0,8607
Italian	$b = - 0,6226 - 0,8232*(a + 0,1336)^{0,9803}$	0,8600
Koine Greek	$b = 0,8543 - 2,0826b*(a + 0,4983)^{0,3234}$	0,9217
Sami	$b = 0,4750 - 1,8427*a^{0,5669}$	0,7767
Portuguese	$b = -0,5314 - 0,4727*(a + 0,0206)^{1,5537}$	0,8848

Dutch	$b = 1{,}1092 - 1{,}9079*(a + 1{,}9736)^{0{,}4443}$	0,8417
Hungarian	$b = - 0{,}2457 - 1{,}1583*(a + 0{,}0206)^{0{,}5657}$	0,8415
Turkish	$b = - 1{,}2722 - 0{,}0170*a^{3{,}6434}$	0,8423
Arabic	$b = - 0{,}5011 - 0{,}7353*a^{0{,}9501}$	0,9294
English	$b = 0{,}3517 - 1{,}0876*(a + 2{,}4821)^{0{,}5788}$	0,3972
Japanese	$b = - 0{,}5177 - 0{,}6856*(a + 0{,}1336)^{0{,}8944}$	0,8090
Bulgarian	$b = - 0{,}1790 - 0{,}8102*(a + 0{,}559)^{0{,}5965}$	0,7401
Finnish	$b = - 1{,}1382*a^{0{,}6906}$	0,7630
Old Hebrew	$b = - 1{,}2539 - 0{,}2266*a^{1{,}2645}$	0,9843
Latin	$b = 190{,}0505 - 191{,}2309*x^{0{,}0033}$	0,8456
Korean	$b = - 1{,}5456 - 0{,}1136*a^{1{,}4862}$	0,9864
Quechua	$b = - 1{,}0688 - 0{,}2057*a^{1{,}2242}$	0,9793
Old Church Slavonic	$b = - 0{,}5925 - 0{,}5545*a^{0{,}8051}$	0,8475
Latvian	$b = 2106{,}3639 - 2108{,}0144*a^{2{,}7967E\text{-}4}$	0,2656
Czech	$b = - 1{,}5506*a^{0{,}1819}$	0,3342
Inuktitut	$b = - 0{,}2185 - 0{,}3455*(a + 1{,}8461)^{1{,}0111}$	0,9853

In general, we consider the model as a sufficient representation of this property in language. Of course, one cannot predict further development of the research but a "better" approach should have a care for at least the following issues:

(1) The function must be as simple as possible, i e. it should contain the minimal number of parameters.

(2) The parameters should represent some linguistically supposed forces, e.g. those presented in Köhler (2005).

(3) One should try to find other collateral properties of the word and study their link to word length.

(4) In general, the result should represent a synergetic image of length.

(5) A further step into the depth would be the consideration of the lengths of individual parts of speech (systemic and pragmatic view).

(6) The establishment of "mixed classes" consisting of words that can belong to several classes simultaneously, e.g. all German verbs can be used as nouns, many adjectives are simultaneously adverbs, etc. This could, perhaps, shed light also on some grammatical and typological matters. We are still accustomed to use the classical Latin classification of parts of speech which is not sufficient in many languages. Even some strongly synthetic languages use the possibility of changing the main part of speech of a word to another, e.g. using analytic means.

We conjecture that parameters *a* and *b* are linked also with the morphological complexity of words, with the proneness of language to form compounds, with the number of grammatical categories in language expressed by means of agglutination or (internal) inflection, with some syntactic properties, with the number of homonyms, etc. That means, after having measured some other prop-

erty, one can try to set up a control cycle using the parameters *a, b, c, k, m* and link them with other aspects of language, as proposed above.

But not only the given language alone is responsible for length: all languages having contact (either geographic or cultural) with other languages are influenced by them, both in the vocabulary and the syntax. If one models the length distribution by a number of discrete probability distributions, one cannot use all parameters and search for their links with other properties; this is possible only if one has a unique model.

Though any conjectures about the grouping of languages according to the mean parameter *a* are premature, we may risk a trial to set up a graph capturing the similarities between languages concerning word length. To this end we shall compute the simplified criterion

$$(5) \qquad u = \frac{\left|\bar{a}_1 - \bar{a}_2\right|}{\sqrt{Var(\bar{a}_1) + Var(\bar{a}_2)}}$$

where $Var(\bar{a}) = Var(a)/n$, representing the quantile of the normal distribution. We use all data in Table 4.2. For example, the difference between Old Church Slavonic and Bulgarian is

$$u = \frac{\left|1.5077 - 0.0762\right|}{\sqrt{0.5643/14 + 0.0991/30}} = 6.85.$$

This value is highly significant (= greater than 1.96) and says that the given texts do not belong to the same word-length group. In Table 4.5 one finds the necessary numbers

Table 4.5
Means and variance of *a* in syllable lengths

Language	**Average *a***	**Variance**	**n**
English	-1,3736	0,3625	21
Low German	-0,8957	0,6663	27
Dutch	-0,8658	0,3748	16
French	-0,612	0,1801	38
Swedish	-0,418	0,3962	18
Palatine	-0,4166	0,2489	16
Welsh	-0,0698	0,2995	12
Maori	0,0304	0,9153	10
Early New High German	0,0582	0,9586	20

(Dürer)			
Portuguese	0,0682	0,2404	20
Bulgarian	0,0762	0,0991	30
Gaelic	0,1477	0,2875	31
Japanese	0,2484	0,2502	11
Spanish	0,2795	0,1595	40
German	0,2899	0,2114	26
Slovenian	0,4253	0,0824	10
Marquesan	0,5111	0,0134	3
Italian	0,6464	0,18	54
Hungarian	0,7007	0,1852	22
Roumanian	0.7707	0.5876	50
Greek Koine	0,7732	0,3693	10
Early New High German (Luther)	0,8048	0,3711	21
Lower Sorbian	0,8284	0,7858	10
Czech	0,875	0,0997	29
Faeroese	1,0225	0,2001	21
Early Modern English	1,0295	0,6532	18
Ukrainian	1,0724	0,4403	20
Middle High German	1,0896	1,0921	19
Gothic	1,1931	0,5022	20
Latin	1,3009	0,1422	19
Russian	1,312	0,5887	38
Chinese	1,314	8,1738	40
Polish	1,3617	0,1835	20
Old High German	1,3958	1,5033	21
Slovak	1,4182	0,684	46
Latvian	1,4453	0,2427	12
Old Church Slavonic	1,5077	0,5643	14
Belorussian	1,6523	0,1675	20
Vogul/Mansi	1,9949	0,3873	20
Arabic	2,2811	3,2931	21
Estonian	2,3136	0,3679	19
Turkish	2,3672	0,2059	18
Icelandic	2,5723	0,1179	20
Old Icelandic	2,7232	0,8397	20
Sami	2,7333	1,1548	25
Finnish	2,9894	0,338	20

Erzja-Mordvin	3,1187	0,2819	15
Uzbek	4,0109	13,5117	20
Cheremis	4,2941	0,2266	14
Inuktitut	4,8892	8,1674	19
Quechua	6,1321	19,4963	24
Korean	6,1507	8,6036	24
Old Hebrew	7,7177	8,9848	23

Performing all tests between pairs of languages we may present the results either in the form of a matrix in which similarities (= not significant differences) are presented by an X, or by a graph with as many vertices as there are languages (texts). Graphs of this dimension usually get obscure, and the same holds for too large tables which must be presented in parts. Hence we would rather present the respective results in the form of vectors whose elements are the numbers designating the individual languages (cf. Table 4.6). For example Low German is "similar" to Dutch(3), French(4) and Greek Koine(20).

Table 4.6
Vectors of "similar" languages: word length (in terms of mean *a*)

Language (No)	**„Similar" languages**	#
1 English	-	0
2 Low German	3,4,20	3
3 Dutch	2,4	2
4 French	2,3,5,6	4
5 Swedish	4,6,7,8,9,18	6
6 Palatine	4,5,7,8,9	5
7 Welsh	5,6,8,9,10,11,12,13,17	9
8 Maori	5,6,7,9,10,11,12,13,14,15,16,17,22	13
9 Early New High German (Dürer)	5,6,7,8,10,11,12,13,14,15,16	11
10 Portuguese	7,8,9,11,12,13,14,15,31	9
11 Bulgarian	7,8,9,10,12,13	6
12 Gaelic	7,8,9,10,11,13,14,15	8
13 Japanese	7,8,9,10,11,12,14,15,16,22	10
14 Spanish	8,9,10,12,13,15,16,22	8
15 German	8,9,10,12,13,14,16,22	8
16 Slovenian	8,9,13,14,15,17,20,22,31	9

17 Marquesan	7,8,13,16,18,19,20,22,31	8
18 Italian	5,17,19,20,21,22,25,27,31,53	10
19 Hungarian	17,18,20,21,22,23,25,27,31,53	10
20 Greek Koine	2,16,17,18,19,21,22,23,24,25,26,28,31,33,53	15
21 Early New High German (Luther)	18,19,20,22,23,24,25,26,27,28,31,53	12
22 Low Sorbian	8,13,14,15,16,17,18,19,20,21,22,23,24,25,26, 27,28,29,30,31,32,33,53	23
23 Czech	19,20,21,22,24,25,26,27,28,31,33,53	12
24 Faeroese	20,21,22,23,25,26,27,28,30,31,33	11
25 Early Modern English	18,19,20,21,22,23,24,26,27,28,29,30,31,32, 33,34,35,36,53	19
26 Ukrainian	20,21,22,23,24,25,27,28,29,30,31,32,33,34,35,36,53	17
27 Middle High German	18,19,20,21,22,23,24,25,26,28,29,30,31,32,33,34, 35,36,53	19
28 Gothic	20,21,22,23,24,25,26,27,29,30,31,32,33,34,35,36	16
29 Latin	22,25,26,27,28,30,31,32,33,34,35,36	12
30 Russian	22,24,25,26,27,28,29,31,32,33,34,35,36	13
31 Chinese	10,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30, 32,33,34,35,36,37,38,39,53	25
32 Polish	22,25,26,27,28,29,30,31,33,34,35,36	12
33 Old High German	20,22,23,24,25,26,27,28,29,30,31,32,34,35,36,37,39	17
34 Slovak	25,26,27,28,29,30,31,32,33,35,36,37	12
35 Latvian	25,26,27,28,29,30,31,32,33,34,36,37	12
36 Old Church Slavonic	25,26,27,28,29,30,31,32,33,34,35,37,38	13
37 Belorussian	31,33,34,35,36,39	6
38 Vogul/Mansi	31,39,40	3
39 Arabic	31,33,36,37,38,40,41,42,43,44,45,47	12
40 Estonian	38,39,41,42,43,44	6
41 Turkish	39,40,42,43,44	5
42 Icelandic	39,40,41,43,44,47	6
43 Old Icelandic	39,40,41,42,44,45,46,47	8
44 Sami	39,40,41,42,43,45,46,47	8
45 Finnish	39,43,44,46,47	5
46 Erzja-Mordvin	43,44,45,47	4
47 Uzbek	39,42,43,44,45,46,48,49,50	9
48 Cheremis	47,49	2
49 Inuktitut	47,48,50,51	4
50 Quechua	47,49,51,52	4

51 Korean	49,50,52	3
52 Old Hebrew	50.51	2
53 Roumanian	18,19,20,21,22,23,25,26,27,31	10

Ordered according to the size of the similarity set

English 1	-	0
Dutch 3	2,4	2
Cheremis 48	47,49	2
Old Hebrew 52	50.51	2
Low German 2	3,4,20	3
Vogul/Mansi 38	31,39,40	3
Korean 51	49,50,52	3
French 4	2,3,5,6	4
Erzja-Mordvin 46	43,44,45,47	4
Inuktitut 49	47,48,50,51	4
Quechua 50	47,49,51,52	4
Palatine 6	4,5,7,8,9	5
Turkish 41	39,40,42,43,44	5
Finnish 45	39,43,44,46,47	5
Swedish 5	4,6,7,8,9,18	6
Bulgarian 11	7,8,9,10,12,13	6
Belorussian 37	31,33,34,35,36,39	6
Estonian 40	38,39,41,42,43,44	6
Icelandic 42	39,40,41,43,44,47	6
Gaelic 12	7,8,9,10,11,13,14,15	8
Spanish 14	8,9,10,12,13,15,16,22	8
German 15	8,9,10,12,13,14,16,22	8
Marquesan 17	7,8,13,16,18,19,20,22,31	8
Old Icelandic 43	39,40,41,42,44,45,46,47	8
Sami 44	39,40,41,42,43,45,46,47	8
Welsh 7	5,6,8,9,10,11,12,13,17	9
Portuguese 10	7,8,9,11,12,13,14,15,31	9
Slovenian 16	8,9,13,14,15,17,20,22,31	9
Italian 18	5,17,19,20,21,22,25,27,31,53	10
Hungarian 19	17,18,20,21,22,23,25,27,31,53	10
Uzbek 47	39,42,43,44,45,46,48,49,50	9

Japanese 13	7,8,9,10,11,12,14,15,16,22	10
Roumanian 53	18,19,20,21,22,23,25,26,27,31	10
Early New High German (Dürer) 9	5,6,7,8,10,11,12,13,14,15,16	11
Early New High German (Luther) 21	18,19,20,22,23,24,25,26,27,28,31,53	12
Czech 23	19,20,21,22,24,25,26,27,28,31,33,53	12
Faeroese 24	20,21,22,23,25,26,27,28,30,31,33	11
Latin 29	22,25,26,27,28,30,31,32,33,34,35,36	12
Polish 32	22,25,26,27,28,29,30,31,33,34,35,36	12
Slovak 34	25,26,27,28,29,30,31,32,33,35,36,37	12
Latvian 35	25,26,27,28,29,30,31,32,33,34,36,37	12
Arabic 39	31,33,36,37,38,40,41,42,43,44,45,47	12
Maori 8	5,6,7,9,10,11,12,13,14,15,16,17,22	13
Russian 30	22,24,25,26,27,28,29,31,32,33,34,35,36	13
Old Church Slavonic 36	25,26,27,28,29,30,31,32,33,34,35,37,38	13
Greek Koine 20	2,16,17,18,19,21,22,23,24,25,26,28,31,33,53	15
Ukrainian 26	20,21,22,23,24,25,27,28,29,30,31,32,33,53 34,35,36	17
Gothic 28	20,21,22,23,24,25,26,27,29,30,31,32,33, 34,35,36	16
Old High German 33	20,22,23,24,25,26,27,28,29,30,31,32,34, 35,36,37,39	17
Early Modern English 25	18,19,20,21,22,23,24,26,27,28,29,30,31, 32,33,34,35,36,53	19
Middle High German 27	18,19,20,21,22,23,24,25,26,28,29,30,31, 32,33,34,35,36,53	19
Lower Sorbian 22	8,13,14,15,16,17,18,19,20,21,22,23,24, 25,26,27,28,29,30,31,32,33,53	23
Chinese 31	10,16,17,18,19,20,21,22,23,24,25,26,27, 28,29,30,32,33,34,35,36,37,38,39,53	25

As can be seen, English differs from all languages. This is caused most probably by its very mixed character. Chinese, on the contrary, displays similarity with almost the half of the languages studied. It is to be remarked that here merely a formal property – word length – is concerned, no other typological, semantic or syntactic properties.

5. Length of compounds

Compounds are special kinds of words or phrases. Since they represent semantic units with specialized meaning, one of the possible measurements of their length is in terms of component numbers. As a matter of fact, the immediate components are stems; the next level consists of morphemes but we shall not skip a level. We consider German press texts and use the data presented in Poppe (2007). The results of fitting are presented in Table 5.1. As can be seen, the model fits the data excellently. In some cases marked with "*" we added a class with frequency 0, otherwise the program could not compute all the parameters.

Table 5.1
Compounds in component numbers

Compound length (data from Poppe 2007; German press texts)				
Text	**a**	**b**	**c**	$\mathbf{R^2}$
T 1	-3.6658	1.3102	91.0787	0.9936
T 2	-4.8926	1.9975	101.0455	0.9965
T 3	-6.6935	3.3610	66.9937	0.9801
T 4	-7.1723	3.5048	79.0021	0.9995
T 5	-4.5179	1.8660	81.0388	0.9959
T 6	-4.1073	1.4473	86.0559	0.9937
T 7	-5.9516	2.6320	87.0026	0.9993
T 8	-4.2558	1.3303	91.9926	0.9967
T 9	0.3019	-2.9435	80.0014	0.9978
T 10	-2.8015	-0.6093	176.0036	0.9996
T 11	-8.0721	3.9775	180.0154	0.9993
T 12	-4.2495	1.6038	130.0055	1.0000
T 13*	-0.5900	-1.7483	27.9983	0.9995
T 14	-3.0958	0.9990	18.0254	0.9924
T 15*	-1.4860	-1.5815	125.9988	0.9999
T 16	-6.1992	2.5949	101.0138	0.9989
T 17	-1.1635	-2.2719	60.0003	0.9996
T 18	-3.1060	-0.4973	76.0008	0.9999
T 19	0.2607	-3.6936	64.0000	1.0000
T 20*	-2.2942	-0.1278	41.9971	0.9992
T 21*	-1.5347	-1.0998	59.0000	1.0000

*Last class added (= 0)

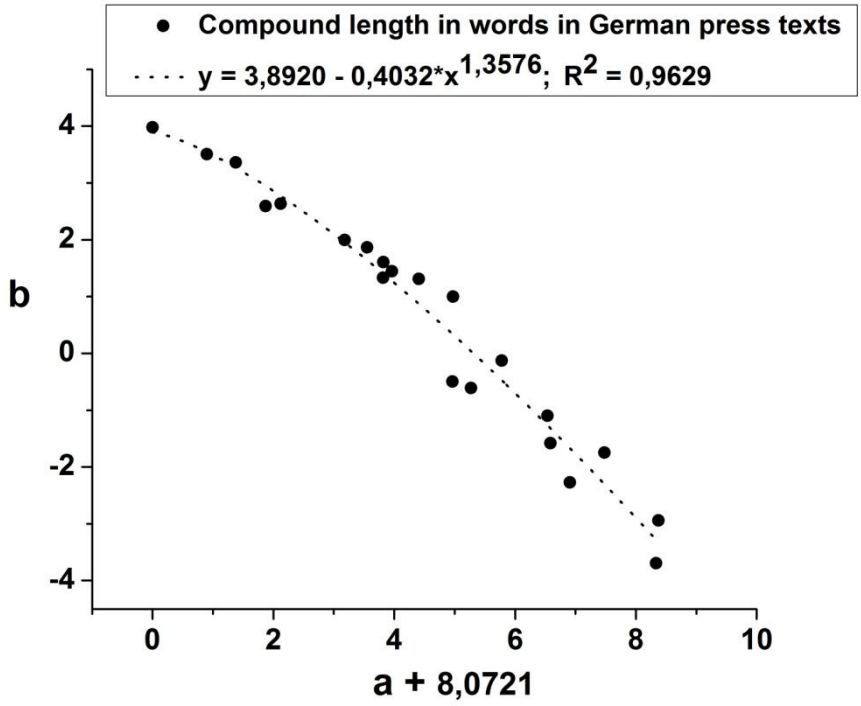


Figure 7.1. Compound length in words b = f(a)

Compounds are, so to say, not phonetic but rather semantic units. They arise on the basis of the requirement for specification of meaning, i.e. on purely semantic grounds. One can see that except for one text, all values of a are smaller than zero.

6. Rhythmic units

On the other hand, some units (e.g. syllables or morphemes) are adequate also for the measurement of other units or their properties, e.g. the length of rhythmic units, the morphological complexity of words, etc. Best (2002) considers length of rhythmic units as the number of steps necessary to reach the next accented syllable, practically the number of unaccented and the accented syllables between the accented syllables. Considering his data in German short prose, we obtain the results presented in Table 6.1. The study of rhythmic poetry would bring stricter distributions. It can be conjectured that the distribution would have very large excess.

Table 6.1
Length of rhythmic units in German prose

Length of rhythmic units (data from Best 2002; German prose)				
Text	**a**	**b**	**c**	**R^2**
T 1	8.3381	-3.8399	0.4386	0.9717
T 2	4.1193	-2.1022	5.7223	0.9860
T 3	4.6690	-2.2847	3.7950	0.9885
T 4	5.6900	-2.7908	2.3622	0.9791
T 5	6.1201	-3.0286	2.1879	0.9927
T 6	3.0582	-1.7409	10.2393	0.9442
T 7	6.2344	-2.7543	1.0294	0.9552
T 8	3.0813	-1.6254	10.1122	0.9088
T 9	3.6984	-1.9710	6.7919	0.9810
T 10	5.2800	-2.8000	3.3366	0.9544
T 11	4.2063	-1.9000	2.6481	0.9391
T 12	4.1681	-2.0464	3.2375	0.9380
T 13	6.9355	-3.3382	1.4100	0.9809
T 14	4.4029	-2.2078	4.3146	0.9192
T 15	6.1465	-3.1243	2.2762	0.9474
T 16	7.4454	-3.4071	0.7799	0.9312

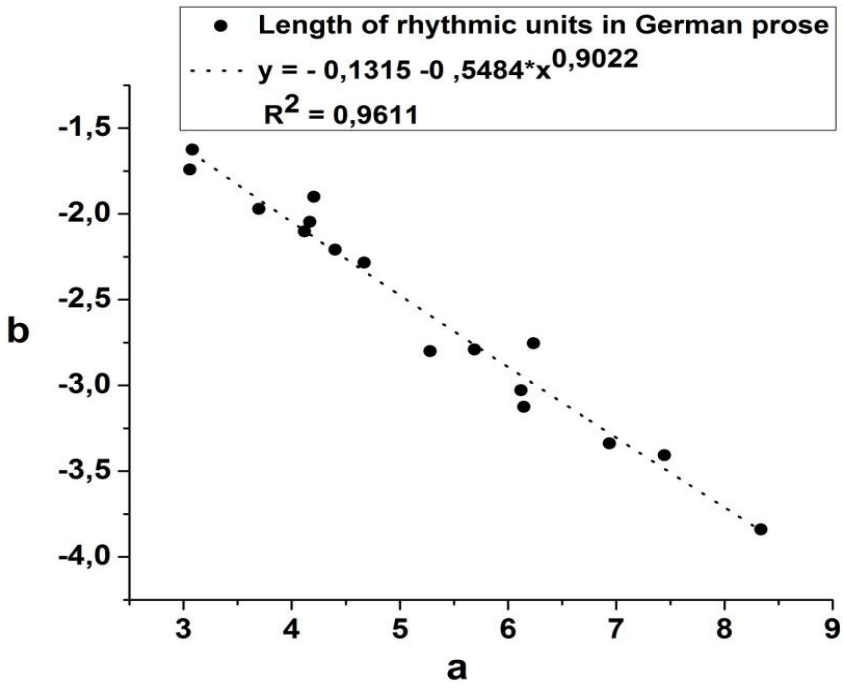


Figure 6.1. Length of rhythmic units in German prose

Though the number of data is very small, one can see that – as compared with non-phonetic features – the parameter *a* is relatively great. Again a hint at the possible stratification of language.

7. Verse length in words

Even if the verse need not have a syllabic background, it is clearly demarcated by the end of line. This boundary need not coincide with any grammatical phenomena, hence the number of words in the verse can be considered its length. This holds also in the case that the rhythm is based on a regular succession of accented and unaccented syllables.

Let us fit the given function to the length of German poems measured in terms of numbers of words. Best (2012, 2012b) used the binomial distribution for the measurement of verse length in terms of words and obtained satisfactory results. The fitting by the above function is presented in Table 7.1.

Table 7.1
Verse length in terms of word numbers

Verse length (data from Best 2012; German poems by G.A.Bürger)				
Text	**a**	**b**	**c**	**R^2**
P 1	22.5728	-7.6166	2.0047E-006	0.9351
P 2	20.7564	-8.0145	7.6662E-005	0.9822
P 3	44.4356	-13.8878	8.7858E-015	0.7932
P 4	69.2491	-22.4537	1.9254E-022	0.9852
P 5	50.8425	-15.5780	2.6936E-017	0.9320
P 6	18.8174	-6.3562	1.7465E-005	0.9272
P 7	52.6671	-15.4202	1.7212E-018	0.9359
P 8	28.0658	-8.7891	4.1342E-009	0.8551
P 9	56.8804	-16.6649	2.3083E-020	0.9893
P 10	33.1999	-9.8891	2.0232E-011	0.9232
P 11	78.0405	-19.5588	2.9877E-033	0.7857
P 12	33.0376	-10.1957	2.4029E-010	0.9486
P 13	49.6374	-14.1329	5.3175E-0.18	0.9947
P 14	42.3667	-13.3961	8.1206E-013	0.9601
P 15	14.8095	-4.8321	0.0005	0.8897
P 16	24.7155	-7.7195	7.5274E-008	0.9402
P 17	29.6206	-9.3052	5.0712E-009	0.9767
P 18	20.2039	-6.2312	3.1676E-006	0.8750
P 19	16.4733	-5.1313	4.0671E-005	0.9726
P 20	33.2372	-9.2080	2.2019E-012	0.9458

All distributions are concave (bell-shaped). The fact that the extreme classes contain merely a small number of cases may evoke slightly worse fittings. In that case one can use the given function with an additive parameter (+ 1). The relationship between the parameters *a* and *b* is visualized in Figure 7.1. As can be

seen, even with this “indirect” measurement there is a well expressed relationship.

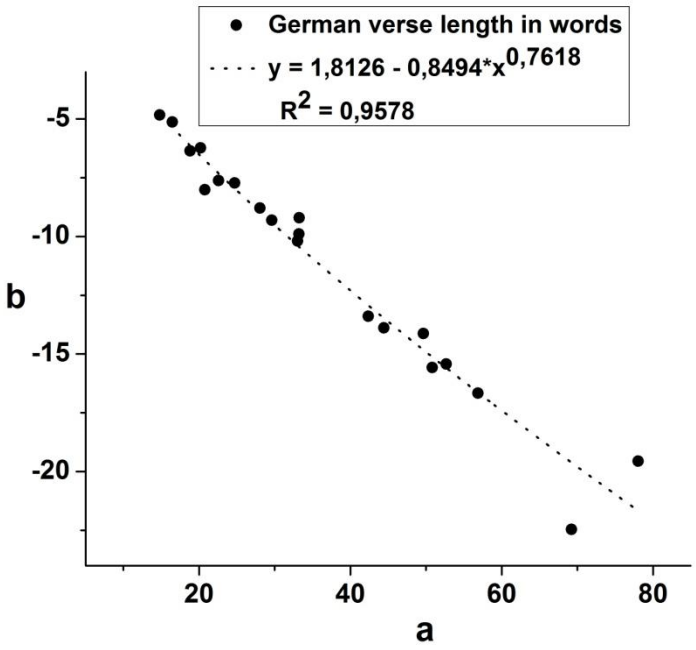


Figure 7.1. German verse length in words

Another investigation by Best (2012b) yields the verse length in the work *Atta Troll* by H. Heine. There are 27 chapters and some other texts (Goethe: *Der Erlkönig*, and *Totentanz*, and *Rodogune* by Corneille). The results of fitting are presented in Table 7.2 and $b = f(a)$ is plotted in Figure 7.2.

Table 7.2
Verse length in terms of word numbers in German

Verse length (data from Best 2012b; German poem by H. Heine)				
Text	**a**	**b**	**c**	**R^2**
C 1	33.0524	-11.1111	7.7382E-010	0.9633
C 2	39.3791	-12.6165	1.0960E-012	0.9903
C 3	15.7245	-5.3674	0.0001	0.9998
C 4	23.0693	-7.5223	6.7687E-007	0.9299
C 5	65.1031	-20.7396	3.4091E-021	0.9722
C 6	18.6692	-6.3670	2.5267E-005	0.9742
C 7	45.8480	-14.6403	5.2266E-015	0.9504
C 8	26.9974	-8.9864	4.7094E-008	0.9705
C 9	11.7514	-4.1508	0.0022	0.7240
C 10	57.8208	-18.2345	4.2092E-0.19	0.9678
C 11	61.6552	-19.5676	3.1400E-020	0.9758
C 12	42.8363	-13.3904	5.9699E-014	0.9443

C 13	28.2363	-9.3826	1.0779E-008	0.9684
C 14	39.0545	-12.2380	6.9913E-013	0.7966
C 15	31.6895	-10.4903	9.2557E-010	0.9380
C 16	46.9673	-14.2144	3.2401E-016	0.9149
C 17	27.6262	-9.5100	6.8478E-008	0.9759
C 18	35.0088	-11.3570	7.6520E-011	0.9153
C 19	61.2142	-19.2130	4.2660E-020	0.9612
C 20	26.8250	-8.9630	8.8772E-008	0.9871
C 21	25.8084	-8.6037	2.0072E-007	0.9535
C 22	29.3066	-9.5129	8.9100E-009	0.9150
C 23	27.7304	-9.0820	2.8900E-008	0.9719
C 24	78.7207	-25.5902	1.3895E-025	0.9895
C 25	25.4028	-8.5330	1.5979E-007	0.9727
C 26	57.8767	-18.2079	4.0779E-019	0.9743
C 27	27.2602	-9.6659	1.1945E-007	0.9899
Complete text	32.9003	-10.7102	9.1654E-009	0.9697

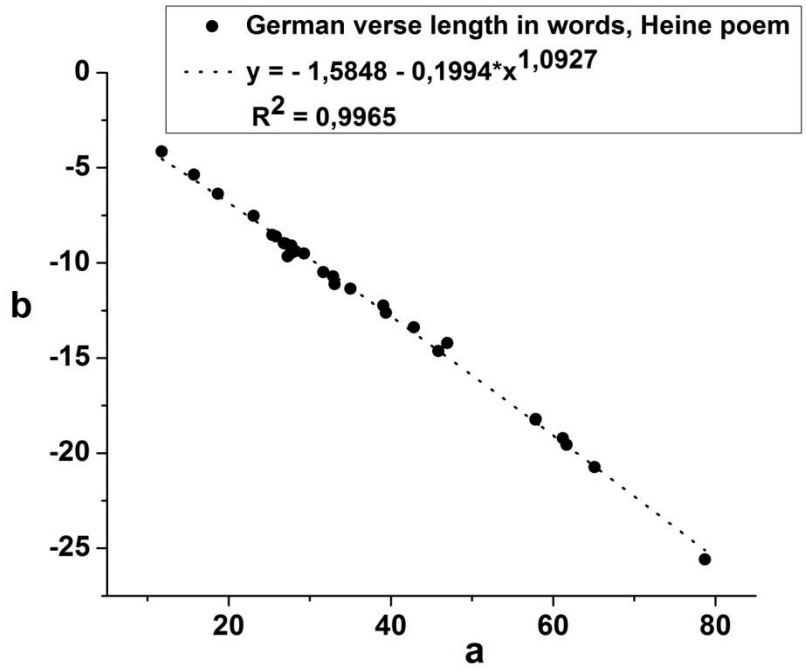


Figure 7.2. German verse length in words, Heine poem

One can observe the extremely high values of the parameter *a.*

8. Sentence length

Though there are many investigations analyzing sentence length in number of words, here we adhere to the principle of measuring it in terms of immediate constituents, i.e. clauses. Hence our view here is grammatical. The number of investigations is, unfortunately, very small. We found the work by Wittek (2001) who analyzed 80 German texts taken from scientific journals on geography and the analysis by Niehaus (1997) concerning different text sorts. There are no data for other languages known to us but since our aim is merely the testing of the given hypothesis we shall fit the model to all data. The results of fitting concerning the data of Wittek are presented in Table 8.1. Since Wittek studied the development of sentence, we can compare the development of the parameters.

Table 8.1
Sentence length in clause numbers in German

Sentence length (data from Wittek 2001: years 1896-1905)				
Text	**a**	**b**	**c**	$\mathbf{R^2}$
T 1	3.2287	-1.6678	7.0799	0.8857
T 2	3.0787	-2.5590	14.7667	0.9929
T 3	1.4781	-1.8619	75.8878	0.9991
T 4	0.6467	-1.1474	58.7368	0.9872
T 5	1.0075	-1.3836	39.4284	0.9602
T 6	1.9850	-2.2673	56.2743	0.9914
T 7	0.3196	-1.3633	34.9264	0.9844
T 8	2.4036	-1.7921	27.6908	0.9791
T 9	1.9431	-1.8805	59.3776	0.9860
T 10	2.5014	-2.0973	37.9483	0.9872
T 11	1.6671	-1.3288	21.0603	0.8835
T 12	1.4828	-1.9361	48.0272	0.9955
T 13	2.0417	-1.8134	44.4296	0.9964
T 14	0.8883	-1.9843	50.9548	0.9940
T 15	2.8735	-1.8166	10.4861	0.9126
T 16	0.8952	-1.4345	62.8573	0.9956
T 17	1.2163	-1.5296	34.6233	0.9790
T 18	1.7943	-2.0724	43.1592	0.9938
T 19	1.1143	-1.1571	46.7478	0.9080
T 20	2.5373	-2.6835	50.2276	0.9885
Years 1929-1933				
T 21	1.2990	-2.2543	36.0041	0.9978

T 22	1.0171	-1.2152	59.1148	0.9241
T 23	-0.8380	-0.2342	33.8059	0.9620
T 24	-0.0787	-1.2074	109.8744	0.9944
T 25	0.7242	-1.4347	105.5709	0.9916
T 26	1.1829	-1.7769	40.0682	0.9918
T 27	1.5542	-1.5981	29.8124	0.9975
T 28	1.1404	-1.6279	20.9788	0.9883
T 29	1.6659	-1.5369	43.3898	0.9901
T 31	1.8006	-1.6378	33.4183	0.9244
T 32	0.8244	-1.6153	71.6294	0.9683
T 33	1.7970	-2.3563	82.0280	0.9995
T 34	0.4228	-1.3985	74.8359	0.9998
T 35	1.2517	-2.4062	113.0395	0.9987
T 36	1.7051	-2.0364	57.6145	0.9884
T 37	0.9929	-1.6962	54.8378	0.9881
T 38	1.3621	-2.1298	53.0045	1.0000
T 39	2.3974	-3.0492	18.0157	0.9965
T 40	0.4256	-0.9378	47.3543	0.9579
Years 1959-1960				
T 41	0.4585	-1.7927	84.9680	0.9894
T 42	0.0415	-1.9446	153.9806	0.9994
T 43	0.1425	-0.9260	42.8077	0.9839
T 44	3.0556	-3.2707	38.0756	0.9976
T 45	2.0880	-3.4623	72.0004	0.9992
T 46	2.4793	-3.3600	81.0192	0.9986
T 47	1.6510	-2.4845	70.0564	0.9986
T 48	0.4765	-2.1517	115.0089	0.9998
T 49	1.6961	-2.2016	29.0625	0.9925
T 50	1.3570	-2.6221	83.9946	0.9999
T 51	-0.1745	-0.9118	79.8082	0.9913
T 52	1.6045	-3.7157	96.0025	0.9997
T 53	1.2978	-1.7940	41.9108	0.9930
T 54	0.7155	-1.7112	95.8408	0.9934
T 55	0.0315	-1.2780	44.9119	0.9855
T 56	1.7289	-2.2430	46.9944	1.0000
T 57	0.2447	-1.7537	64.9854	0.9994
T 58	-0.5510	-1.0540	145.9902	0.9995
T 59	1.5424	-2.7854	96.9863	0.9993
T 60	1.1345	-2.1613	73.0401	0.9975
Years 1993-1994				
T 61	1.4176	-3.2323	76.0034	0.9997
T 62	0.7949	-1.6694	56.9500	0.9987
T 63	0.7152	-2.0040	97.9530	0.9984

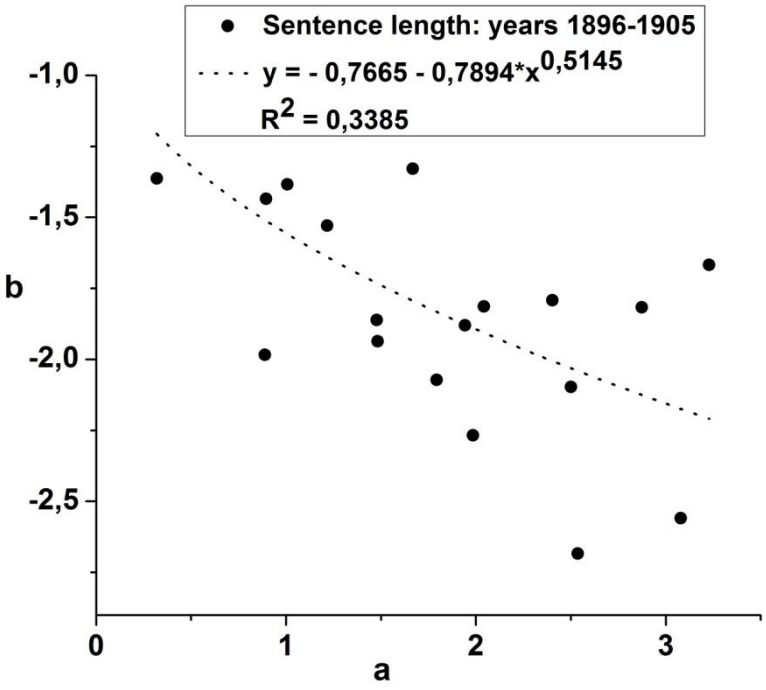
T 64	1.1960	-2.8953	57.9939	0.9993
T 65	0.9705	-2.0554	85.9042	0.9955
T 66	-0.0003	-1.9285	90.9985	1.0000
T 67	-0.1047	-1.1756	61.9266	0.9942
T 68	0.5273	-2.4944	98.9968	0.9999
T 69	0.2075	-2.2636	66.9948	0.9994
T 70	1.6795	-2.6705	36.0061	0.9987
T 71	-0.0427	-1.7565	83.9951	0.9999
T 72	1.5602	-2.5767	63.9181	0.9993
T 73	1.7596	-3.1357	55.9960	0.9998
T 74	1.4119	-2.2627	47.9906	0.9998
T 75	1.5347	-3.2421	99.9972	0.9998
T 76	-0.7083	-0.7919	76.1677	0.9744
T 77	-0.5941	-1.2063	114.9551	0.9677
T 78	0.9797	-2.0196	56.9251	0.9900
T 79	-1.0817	-0.7438	73.9726	0.9976
T 80	0.2554	-1.9483	121.9910	0.9998

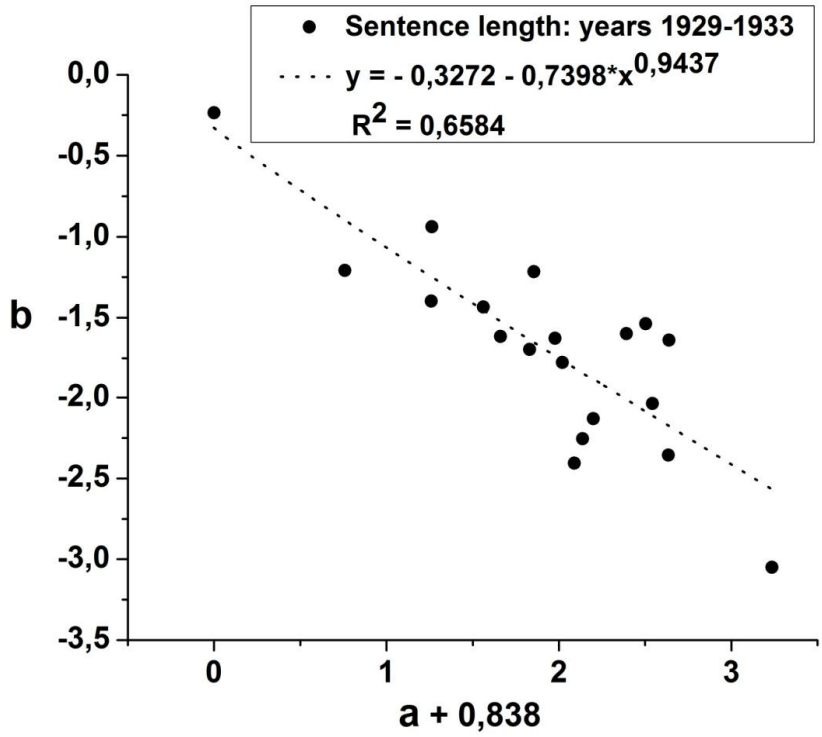
The dependence of *b* on *a* is visualized in Figure 8.1. The theoretical lines are presented separately for the four time intervals. The fifth figure is for all together. As can be seen, the general trend is present but for each time period it is different. Taking all the data together, the oscillation of the values is considerable.

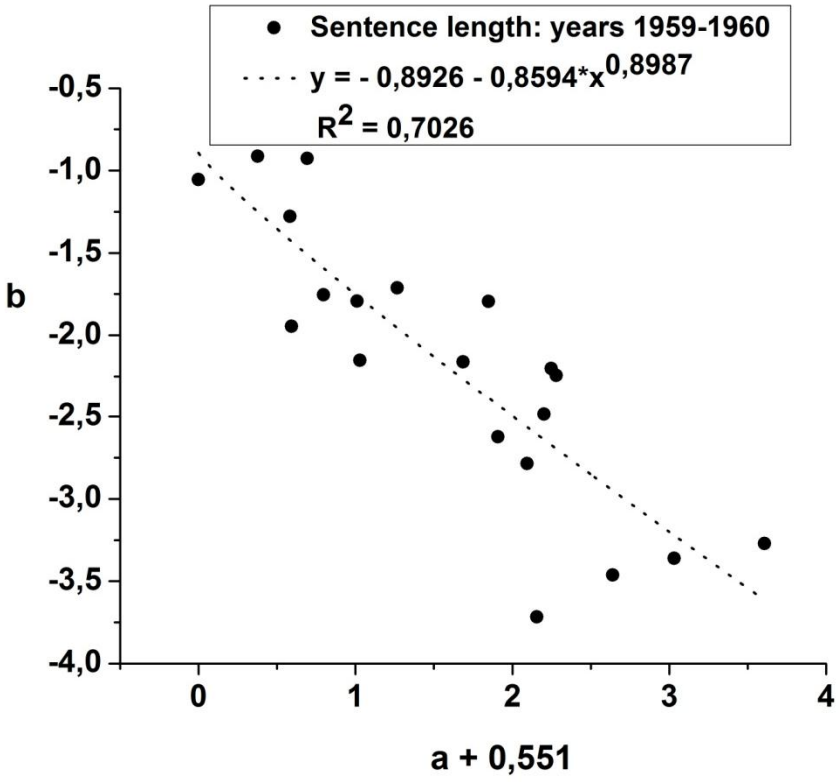
Computing the mean parameter *a* in the individual time periods we obtain an interesting result:

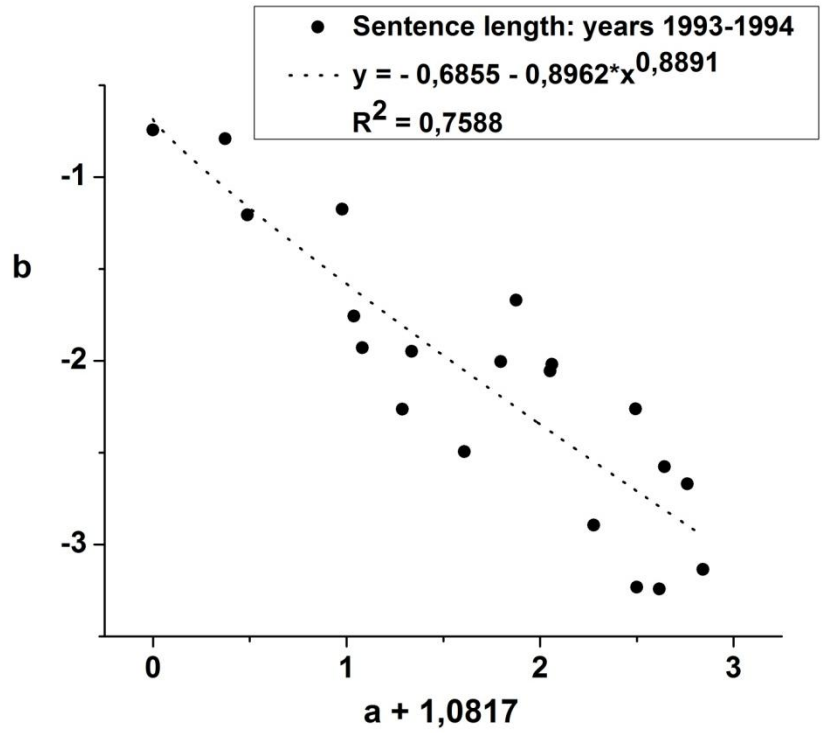
1896-1905: 1.7552
1929-1933: 1.0867
1959-1960: 1.0510
1993-1004: 0.6239

The development is evident but we cannot conjecture its cause. Further research both in German and other languages is necessary.


Sentence length: years 1896-1905
y = - 0,7665 - 0,7894*x^0,5145
R^2 = 0,3385
-1,0
-1,5
-2,0
-2,5
b
0
1
2
3
a


Sentence length: years 1929-1933
y = - 0,3272 - 0,7398*x^0,9437
R^2 = 0,6584
0,0
-0,5
-1,0
-1,5
-2,0
-2,5
-3,0
-3,5
b
0
1
2
3
a + 0,838


Sentence length: years 1959-1960
y = - 0,8926 - 0,8594*x^0,8987
R^2 = 0,7026
b
-0,5
-1,0
-1,5
-2,0
-2,5
-3,0
-3,5
-4,0
0
1
2
3
4
a + 0,551


Sentence length: years 1993-1994
y = - 0,6855 - 0,8962*x^0,8891
R^2 = 0,7588
b
-1
-2
-3
0
1
2
3
a + 1,0817

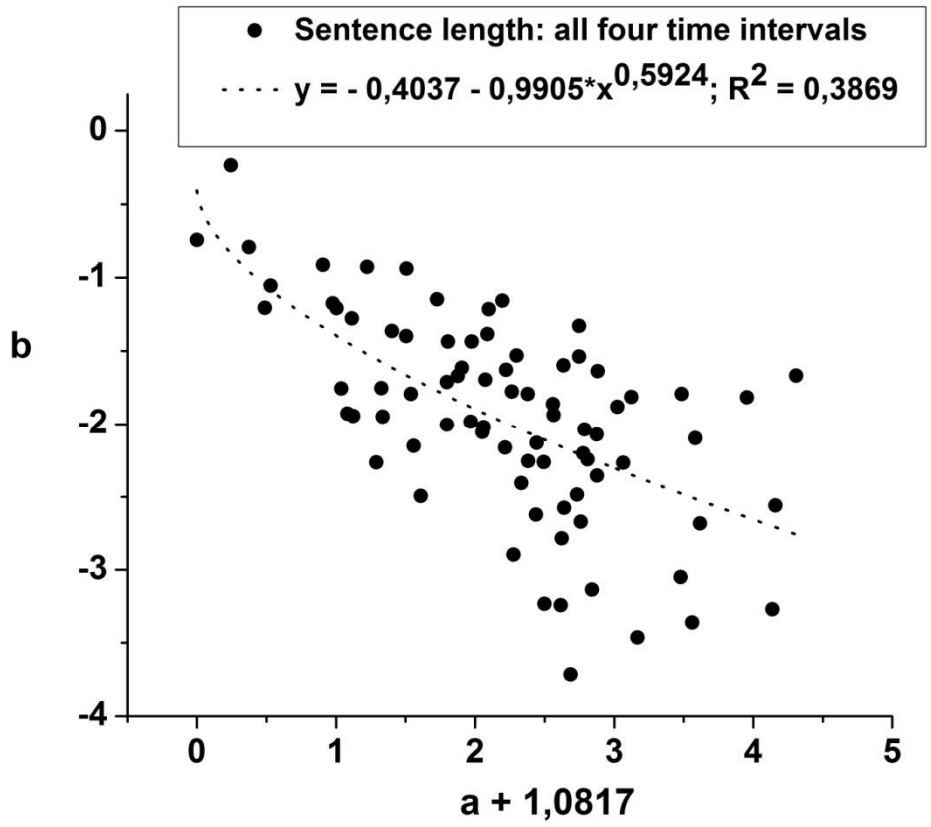


Figure 8.1. The dependence of a on b in different periods

As can be seen, the link is stronger if the time periods are scrutinized separately. Taking into account a period of 100 years may change the numerical form of the relationship. Taking all the data together yields a very great dispersion of values.

The data collected by Niehaus (1997) are presented in Table 8.2. Niehaus used different text sorts which can be summarized as follows:

Text No	**Text sort**
1 - 17	Prose for children
18 - 34	Prose for adults
35 - 51	“Der Spiegel”
52 - 57	Jurisprudence
58 – 63	Economic texts
64 – 68	History
69 – 85	Philosophy

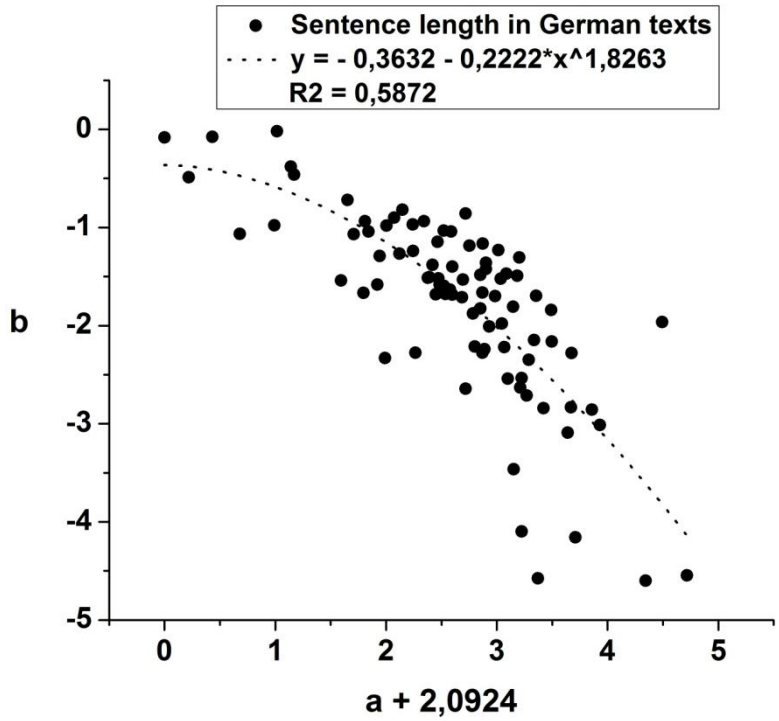
Table 8.2
Sentence length in German texts

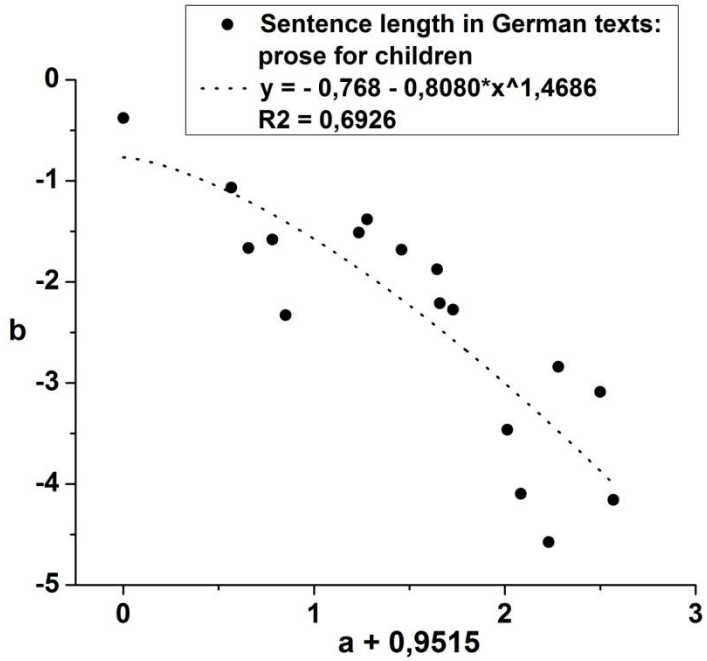
Sentence length (data from Niehaus 1997)				
Text	**a**	**b**	**c**	**R^2**
T 1	1.1329	-4.0973	159.9999	1.0000
T 2	0.7075	-2.2128	102.9910	0.9998
T 3	0.5074	-1.6837	78.8904	0.9871
T 4	0.7771	-2.2752	92.0161	0.9995
T 5	1.5482	-3.0905	104.1002	0.9996
T 6	1,3292	-2.8415	84.0113	0.9992
T 7	1.2781	-4.5754	130.0002	0.9999
T 8	-0.1696	-1.5815	93.9930	0.9999
T 9	1.0614	-3.4642	118.9990	0.9999
T 10	1.6168	-4.1586	100.9997	1.0000
T 11	0.3272	-1.3819	125.0586	0.9975
T 12	-0.9515	-0.3796	74.7802	0.9933
T 13	-0.1003	-2.3302	151.0013	0.9999
T 14	0.6933	-1.8769	84.9107	0.9948
T 15	-0.3850	-1.0666	113.8898	0.9956
T 16	-0.2956	-1.6667	144.9904	0.9996
T 17	0.2838	-1.5121	79.9897	0.9999
T 18	0.9742	-2.2198	85.9829	0.9993
T 19	0.4444	-1.6785	115.0127	0.9996
T 20	0.2493	-0.9372	71.9064	0.9988
T 21	0.7777	-1.6633	85.8239	0.9950
T 22	0.6612	-1.1840	56.7085	09848
T 23	0.3723	-1.1475	69.8596	0.9932
T 24	-0.0187	-0.9002	68.7433	0.9872
T 25	-0.4398	-0.7190	107.6458	0.9819
T 26	0.0288	-1.2668	75.8825	0.9912
T 27	-1.8730	-0.4869	153.9885	0.9998
T 28	-2.0924	-0.0824	106.9939	1.0000
T 29	-1.4132	-1.0661	173.9939	0.9999
T 30	-0.9225	-0.4620	90.7939	0.9915
T 31	-0.2821	-0.9350	97.7620	0.9868
T 32	-1.6597	-0.0765	102.8888	0.9984
T 33	-1.0763	-0.0196	56.3746	0.9715
T 34	0.1484	-0.9685	99.6866	0.9924
T 35	2.2507	-4.5983	111.0007	0.9999
T 36	0.5064	-1.3998	101.9928	0.9971
T 37	0.6010	-1.5306	114.9428	0.9993

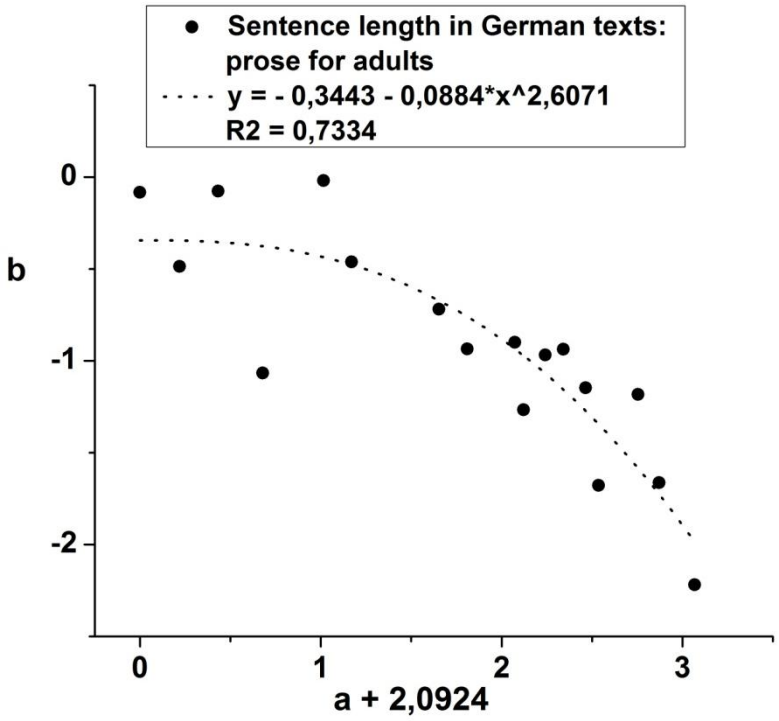
T 38	0.3559	-1.6802	75.0204	0.9995
T 39	1.1780	-2.7127	109.0013	1.0000
T 40	0.6253	-2.6423	173.0017	1.0000
T 41	0.4282	-1.5936	73.8564	0.9863
T 42	2.6240	-4.5445	108.0001	0.9999
T 43	1.8398	-3.0138	68.9929	0.9998
T 44	1.1967	-2.3479	95.9720	0.9995
T 45	1.5767	-2.8328	120.0291	0.9993
T 46	0.4846	-1.6332	124.7959	0.9914
T 47	0.1729	-2.2769	143.0011	1.0000
T 48	-0.4985	-1.5389	136.9819	0.9993
T 49	0.4945	-1.0418	73.5171	0.9800
T 50	1.1316	-2.5334	90.9966	1.0000
T 51	-0.2507	-1.0418	109.8070	0.9908
T 52	0.3808	-1.5180	106.8673	0.9955
T 52a	-0.1489	-1.2912	117.9804	0.9990
T 53	1.2448	-2.1470	87.9505	0.9993
T 54	0.8103	-1.4221	68.0939	0.9935
T 55	1.0540	-1.8081	98.6727	0.9900
T 56	2.4006	-1.9628	33.7504	0.9541
T 57	1.5836	-2.2806	83.1270	0.9938
T 58	-1.0997	-0.9782	120.0001	1.0000
T 59	1.1190	-2.6319	132.0065	0.9999
T 60	0.8103	-1.3606	52.7709	0.9912
T 61	0.9445	-1.5211	71.8843	0.9981
T 62	1.1119	-1.3055	67.8886	0.9728
T 63	1.7662	-2.8549	107.9718	0.9994
T 64	0.9548	-1.9773	109.0446	0.9984
T 65	1.4035	-2.1627	111.8709	0.9981
T 66	0.8924	-1.6997	104.6972	0.9900
T 67	0.1521	-1.2380	116.9170	0.9991
T 68	0.7954	-2.2415	121.0480	0.9978
T 69	0.5944	-1.7127	108.0469	0.9994
T 70	0.3915	-1.5829	86.9619	0.9993
T 71	0.2986	-1.5054	112.9537	0.9994
T 72	0.0552	-0.8177	71.6676	0.9899
T 73	1.0074	-2.5422	129.0575	0.9963
T 74	-0.0879	-0.9810	73.8354	0.9945
T 75	0.7603	-1.8255	106.0403	0.9993
T 76	0.6270	-0.8571	46.9620	0.9086
T 77	0.9946	-1.4697	84.2369	0.9964
T 78	1.2620	-1.6967	60.7212	0.9883
T 79	0.9220	-1.2312	60.4589	0.9891

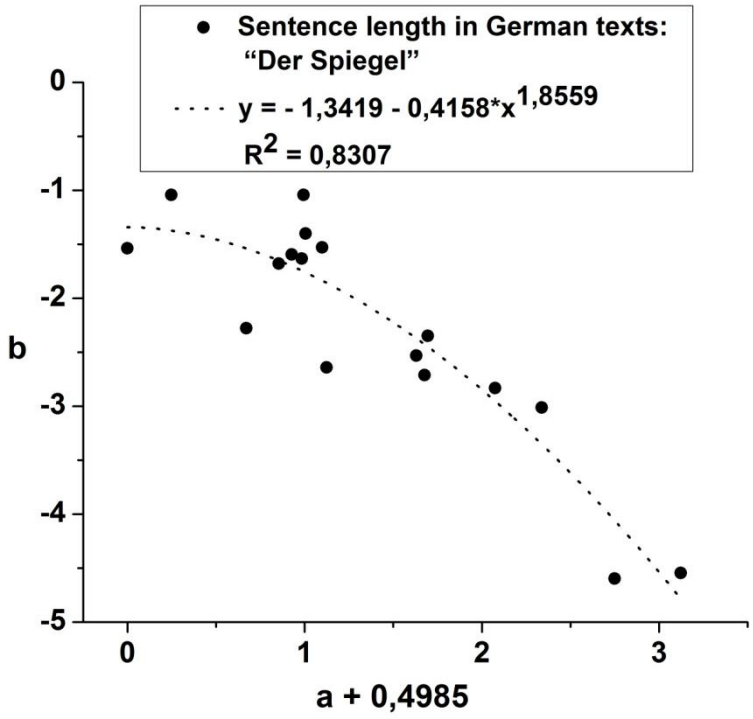
T 80	0.8393	-2.0086	79.0189	0.9997
T 81	0.7597	-1.4834	73.0548	0.9994
T 82	0.4289	-1.0310	73.4202	0.9819
T 83	0.7800	-1.1656	51.7763	0.9902
T 84	1.3991	-1.8420	66.9996	0.9976
T 85	1.0935	-1.4917	51.4631	0.9768

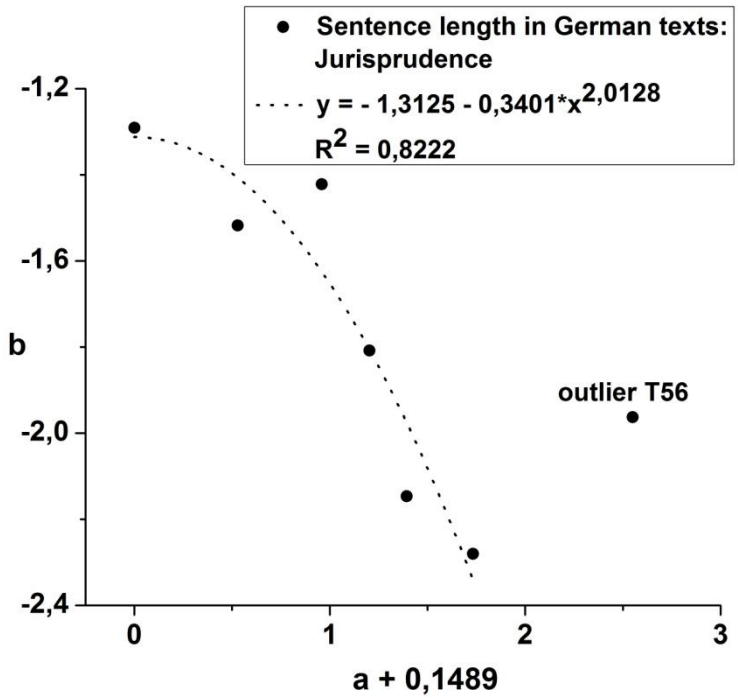
If one puts all texts together, one can see that the oscillation of $b = f(a)$ is very strong. This is merely the sign of differences between text sorts. Taking them separately we obtain smoother relations.

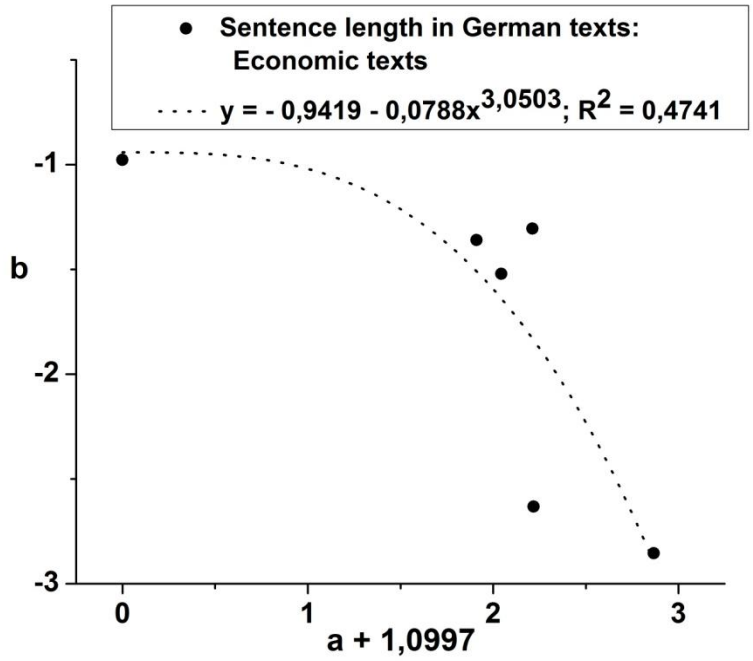


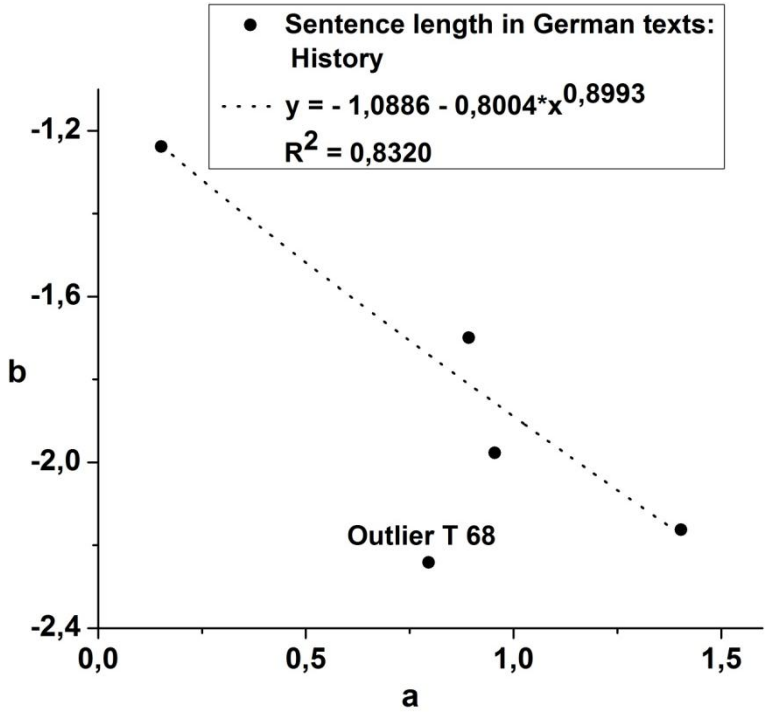

Sentence length in German texts:
prose for children
y = - 0,768 - 0,8080*x^1,4686
R2 = 0,6926
b
a + 0,9515

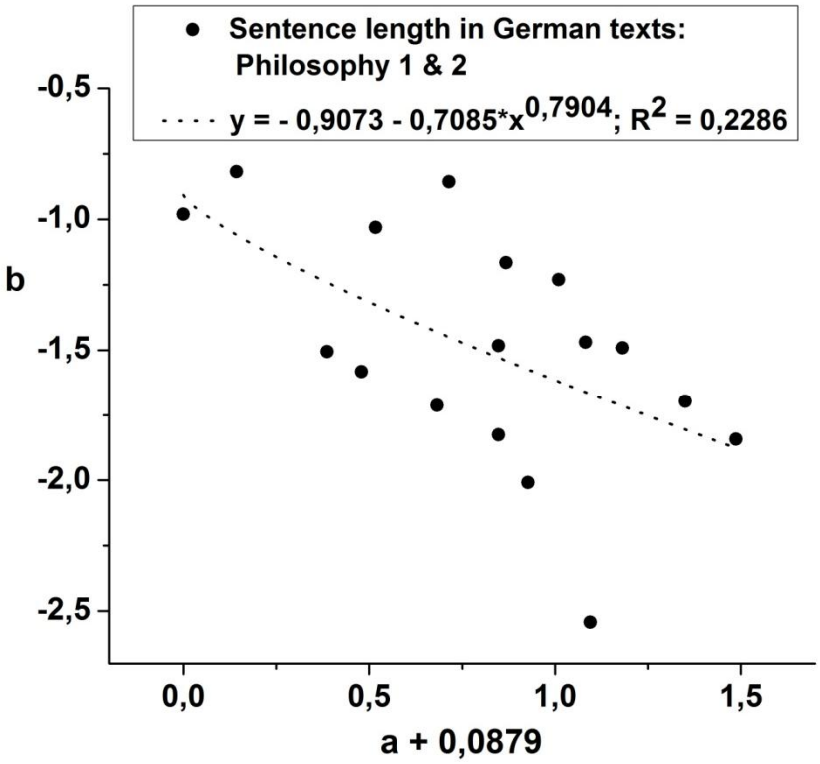

Sentence length in German texts:
prose for adults
y = - 0,3443 - 0,0884*x^2,6071
R2 = 0,7334
b
a + 2,0924


Sentence length in German texts:
"Der Spiegel"
y = - 1,3419 - 0,4158*x^1,8559
R^2 = 0,8307
b
a + 0,4985

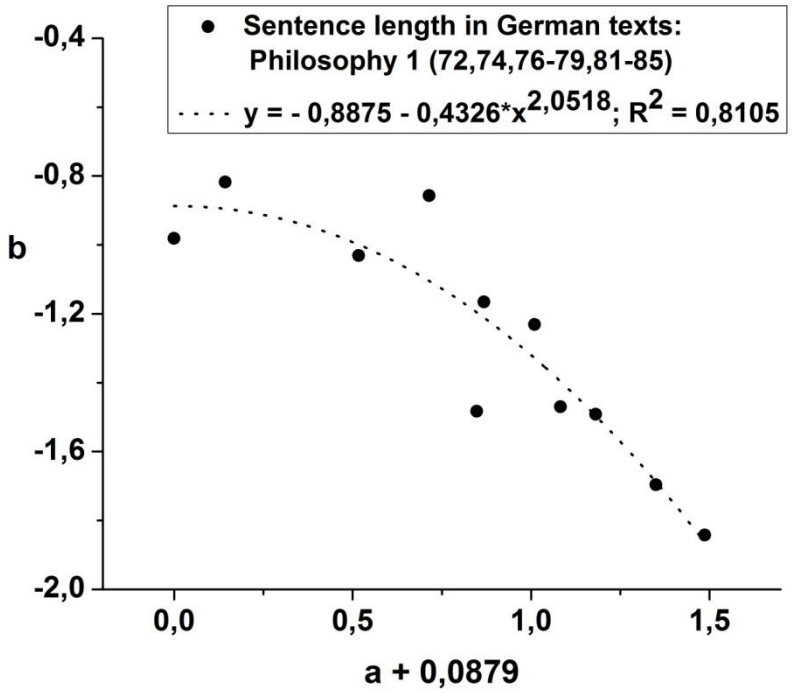

Sentence length in German texts:
Jurisprudence
y = - 1,3125 - 0,3401*x^2,0128
R^2 = 0,8222
outlier T56
b
a + 0,1489


Sentence length in German texts:
Economic texts
y = - 0,9419 - 0,0788x^3,0503; R^2 = 0,4741
-1
-2
-3
b
0
1
2
3
a + 1,0997


Sentence length in German texts:
History
y = - 1,0886 - 0,8004*x^0,8993
R^2 = 0,8320
-1,2
-1,6
-2,0
-2,4
b
Outlier T 68
0,0
0,5
1,0
1,5
a



A better fitting can be found in the next figures where the philosophical texts are divided in two groups. This fact shows that texts should be classified according to selected criteria and the homogeneity should be as strong as possible.



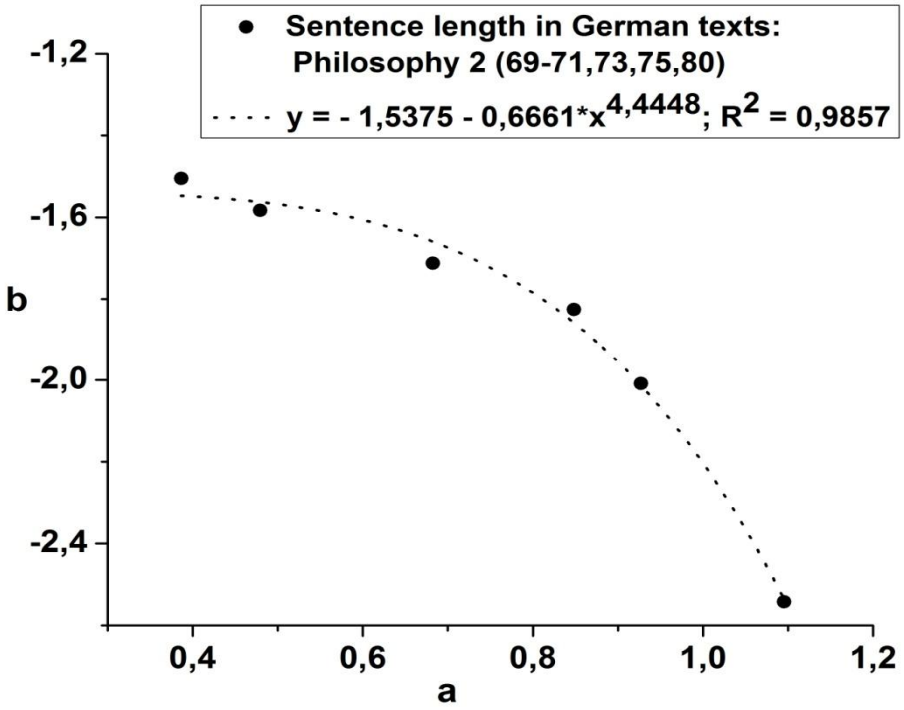


Figure 8.2. Sentence length in German texts, presented together or separately

The above case shows not only the importance of homogeneity for modeling and testing but also the possibility of finding quite formal differences between texts and text sorts.

9. Speech act chains

Sentence is not the "highest" level in the linguistic hierarchy. There is a number of further levels that can be analyzed (cf. Altmann 2014). Some of them are not necessarily situated in the general formal hierarchy, they represent a different view of the text. But again, each new level is a source of further hierarchies and there is no end. At these levels, "length" can be considered a quite complex phenomenon that can be defined in terms of sequence length, runs, distances, hreb sizes, etc. These domains are not studied frequently because they become ever more abstract and the definition of units and kinds of length increases.

Here we shall merely touch on the level of speech acts which can be studied either in stage-plays or in registered conversations. As an example we show the lengths of chains of illocutionary speech acts as produced by two children and two adults in a family conversation. The data are from Rothe, Altmann, Wagner (1992). For each person separately the number of different acts in uninterrupted sequence is registered, and the lengths establish a distribution. The speech acts were classified exactly before the analysis, and to be sure there are many possibilities of how to do it. Hence further qualitative and quantitative research would be necessary.

Since there are merely four data points, we present them in Table 9.1.

Table 9.1
Length of speech acts in a German conversation
of two children (G and C) and two adults (N and M)

Length	G	C	N	M
1	360	40	366	231
2	169	13	111	59
3	111	7	28	14
4	56	1	12	2
5	28		4	1
6	25		1	
7	12		2	
8	5		1	
9	8			
11	3			
12	2			
13	2			
14	1			
a	-0,6794	-1,1634	-0,8000	-0,8815
b	-0,4812	-0,5938	-1,3375	-1,5653
c	358,9512	39,9741	366,0260	230,9948
R^2	0,9958	0,9831	0,9998	0,9999

The values of parameters and the resulting R^2 are presented in the last row of the table. The link between *a* and *b* is presented in Figure 9.1. The resulting parameters *a* can be the source of further hypotheses concerning age, dominance of person in the conversation, etc.

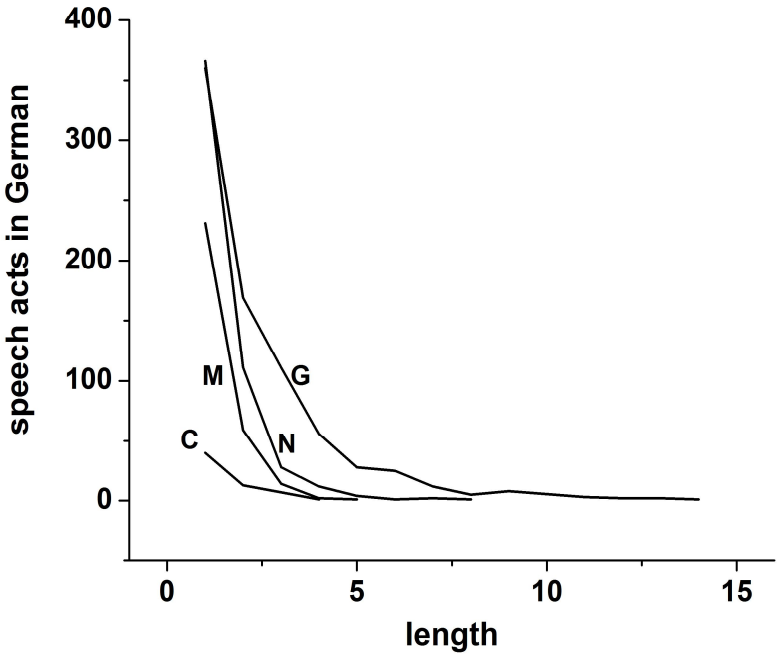


Figure 9.1. Speech acts in German

10. The levels

Though a comparison of levels should be performed only using homogeneous data in one language, e.g. taking the same texts and perform the investigations of all entities in the same texts, we restrict ourselves to intuitive comparisons which may be helpful for further research. For this purpose we can use only German analyzed already on different levels.

We shall consider modern German texts and compute the mean of the parameter *a* for more or less homogeneous classes of texts and different units. We shall not perform tests. The survey is presented in Table 10.1

Table 10.1
Mean values of parameter *a* in German texts

Unit	**Text type**	**Average a**	**N**
Verse in syllables	Bürger (Best 2012)	36.9815	20
	Heine (Best 2012)	37.4309	27
Syllable in phonemes	Press (Cassier 2001)	13.4704	21
	Prose (Best 2010)	15.1987	20
Rhythmic unit in syllables	Prose (Best 2002)	5.2246	16
Morph in phonemes	Press (Best 2000)	1.9281	21
	Prose (Best 2000)	1.6739	18
Sentence in clauses	Geography (Wittek 2001)	1.1297	80
	Mixed (Niehaus 1997)	0.5379	85
Word in syllables	Mixed (Altmann, Best 1996)	0.2899	26
Speech acts	Conversation (Rothe, Altmann, Wagner 1992)	-0.8832	4
Compound in stems	Press texts (Poppe 2007)	-3.5851	21

For the time being we do not risk even a conjecture. For more thorough analysis one would be forced to take individual texts and perform their complete length analysis at all levels. But a preliminary look at Table 10.1 shows that the more one moves away from the phonetic level, the smaller the mean value of the parameter *a* becomes. This is merely a first impression which must be tested on many texts in many languages.

11. Conclusion

The main aim of the investigation has been fulfilled. It has been stated that the length of any unit in language abides by the same regularity which can be considered now a law. The results can be statistically compared, the languages can be classified – if necessary – and the relationship between length and other properties can be put to use. Not only word length linked with other properties but an extensive net of links may be constructed. Though the basic behavior of length is known (cf. Köhler 2005), we are far from attaining the deeper levels of relations in language. Up to now not all properties have been scrutinized and not all possible and reasonable units have been established. The length research can be extended to special units, e.g. speech acts of different kinds, individual parts of speech, adnominal constructs, distances between dependent entities, hrebs, etc. The model itself seems to be no problem, the topical questions are now the parameters, and their change in time, in text sort and in the degree of the applicable linguistic level.

Though a great number of data points have been used for testing, in linguistics it is never enough. Not only do the languages differ but even the individual texts of a unique speaker display different parameters. The only possibility to find a common and more stable background is extensive testing. But even if we find a common background and its links to other phenomena, for each text the boundary conditions must be traced back. Which requirements were active at the creation of the text? And even if we find them, why were they realized in the given form? But "why"-questions cannot be answered by simple statements, only laws can give an answer. One sees that the way into the inner life of language does not have an end. Wherever one begins to set up hypotheses and test them, one necessarily comes to a point at which one must make a big step in the hierarchy of explanations.

The proposal of a unified model does not exclude the possibility of applying in the future stochastic processes, urn models etc. in order to obtain explanatory approaches of other kinds.

Appendix

Title and size of Roumanian poems by M. Eminescu
used in the present book

Text	Title	Size
	alphabetically	in words
T 1	Adio	159
T 2	Atât de fragedă…	176
T 3	Când	126
T 4	Ce te legeni…	102
T 5	Criticilor mei	130
T 6	Cu mâne zilele-ți adaogi…	141
T 7	De ce nu-mi vii	123
T 8	De-aş avea	93
T 9	Despărțire	303
T 10	Din noaptea	68
T 11	Din valurile vremii…	152
T 12	Doi aştri	40
T 13	Dorința	102
T 14	Egipetul	688
T 15	Floare-albastră	247
T 16	Freamăt de codru	179
T 17	Îngere palid...	63
T 18	Junii corupți	458
T 19	La Bucovina	184
T 20	La mijloc de codru…	55
T 21	La steaua	71
T 22	Lacul	90
T 23	Lebăda	41
T 24	Lida	66
T 25	Locul aripelor	259
T 26	Luceafărul	1737
T 27	Mai am un singur dor	125
T 28	Melancolie	274
T 29	Misterele nopții	155
T 30	Noaptea…	177

T 31	Nu e steluță	54
T 32	Nu mă-nțelegi	384
T 33	Numai poetul	48
T 34	O stea pin ceruri	78
T 35	Odă in metru antic	103
T 36	Pe lângă plopii fara soți	199
T 37	Peste vârfuri	47
T 38	Povestea codrului	220
T 39	Prin nopți tăcute	48
T 40	Revedere	141
T 41	Sara pe deal	156
T 42	Scrisoarea I	1272
T 43	Se bate miezul nopții…	45
T 44	Somnoroase păsărele…	55
T 45	Speranța	245
T 46	Stelele-n cer	91
T 47	Şi dacă…	53
T 48	Trecut-au anii	88
T 49	Unda spumă	59
T 50	Vis	177

References

Abbe, S. (2000). Word length distribution in Arabic letters. *Journal of Quantitative Linguistics 7(2), 121-127*

Ahlers, A. (2001). The distribution of word length in different types of Low German texts. In: Best, K.-H. (ed.), *Häufigkeitsverteilungen in Texten: 43-58.* Göttingen: Peust und Gutschmidt.

Altmann, G., Best, K.-H. (1996). Zur Länge der Wörter in deutschen Texten**.** In: P. Schmidt (ed.), *Glottometrika 15, 166-180.*Trier: WVT.

Altmann, G., Erat, E., Hřebíček, L. (1996). Word length distribution in Turkish texts. In: P. Schmidt (ed.), *Glottometrika 15: 195-204.* Trier: WVT.

Ammermann, S. (2001). Zur Wortlängenverteilung in deutschen Briefen über einen Zeitraum von 500 Jahren. In: Best, K.-H. (ed.), *Häufigkeitsverteilungen in Texten: 59-91*. Göttingen: Peust & Gutschmidt.

Antić, G., Kelih, E., Grzybek, P. (2006). Zero-syllable words in determining word length. In: Grzybek, P. (ed.), *Contributions to the science of text and language. Word length studies and related issues: 117-116.* Dordrecht: Springer.

Arlt, I. (2006). Zur Wortlängenverteilung in SMS-Texten. *Göttinger Beiträge zur Sprachwissenschaft 13, 9-21*.

Balschun, C. (1997). Wortlängenhäufigkeiten in althebräischen Texten. In: Best, K.H. (ed.), *Glottometrika 16: 174-179.* Trier: WVT.

Bartens, H.-H., Best, K.-H (1996). Wortlängen in estnischen Texten. *Ural-altaische Jahrbücher N.F. 14, 112-128.*

Bartens, H.-H., Best, K.-H. (1997). Wortlängen in erzamordwinischen Texten. *Linguistica Uralica XXIII, 5-13* .

Bartens, H.-H., **Best, K.-H.** (1997). Word-length distribution in Sámi texts. *Journal of Quantitative Linguistics 4(1-3), 45-52.*

Bartens, H.-H., Best, K.-H. (1997a). Wortlängen im Tscheremissischen (Mari). *Finnisch-Ugrische Mitteilungen 20, 1-20.*

Bartens, H.-H., Zöbelin, T. (1997). Wortlängenhäufigkeiten im Ungarischen. In: Best, K.H. (ed.), *Glottometrika 16: 195-203.* Trier: WVT

Becker, C. (1996). Word lengths in the letters of the Chilean author Gabriela Mistral. *Journal of Quantitative Linguistics 3(2), 128-131.*

Best, K.H. (1996). Zur Wortlängenhäufigkeit in schwedischen Pressetexten. In: In: P. Schmidt (ed.), *Glottometrika 15: 147-157.* Trier: WVT.

Best, K.-H. (1996). Word length in Old Icelandic songs and prose texts. *Journal of Quantitative Linguistics 3(2), 97-105.*

Best, K.-H. (1996). Zur Bedeutung von Wortlängen, am Beispiel althochdeutscher Texte. *Papiere zur Linguistik 55, 141-152.*

Best, K.-H. (1996). Wortlänge in mittelhochdeutschen Texten. In: Best, Karl-Heinz (ed.), *Glottometrika 16, 40-54.*Trier: WVT.

Best, K.-H. (1997). Zur Wortlängenhäufigkeit in deutschsprachigen Pressetexten. In: Best, Karl-Heinz (Hrsg.), *Glottometrika 16, 1-15*. Trier: Wissenschaftlicher Verlag Trier.

Best, K.-H. (2000). Morphlängen in Fabeln von Pestalozzi. *Göttinger Beiträge zur Sprachwissenschaft 3, 19-30.*

Best, K-H. (2001). Wortlängen in Texten gesprochener Sprache. *Göttinger Beiträge zur Sprachwissenschaft 6, 31-42.*

Best, K.-H. (2002). The distribution of rhythmic units in German short prose. *Glottometrics 3, 136-142.*

Best, K-.H. (2010). Silben-, Wort- und Morphlängen bei Lichtenberg. *Glottometrics 21, 2010, 1-13*

Best, K.-H. (2012). Zur Verslänge bei G. A. Bürger. *Glottometrics 23, 56-61.*

Best, K.-H. (2012b). How many words are in a verse? An exploration. In: Naumann, S., Grzybek, P., Vulanović, R., Altmann, G. (eds.), *Synergetic linguistics. Text and language as dynamic systems: 13-22.* Wien: Praesens.

Best, K.-H., Brynjólfsson, E. (1997). Wortlängen in isländischen Briefen und Pressetexten. *Skandinavistik 27, 24-40.*

Best, K.-H., Kaspar, I. (1997). Wortlängen in Faröischen Briefen. *Naukovyj Visnik Černivec'koho Universytetu, Vypusk 41, Hermans'ka Filologija 3-14.*

Best, K.-H., Kaspar, I. (2001). Wortlängen in Färöischen. In: Best, K.-H. (ed.), *Häufigkeitsverteilungen in Texten: 92-100.* Göttingen: Peust und Gutschmidt.

Best, K.-H., Medrano, P. (1997). Wortlängen in Ketschua-Texten. In: Best, K.-H. (ed.), *Glottometrika 16: 204-212.* Trier: WVT

Best, K.-H., Zinenko, S. (1999). Wortlängen in Gedichten des ukrainischen Autors Ivan Franko. In: Genzor, J., Ondrejovič, S. (eds.), *Pange lingua. Zborník na počesť Viktora Krupu: 201-213.* Bratislava: Veda.

Cassier, F.-U. (2001). Silbenlängen in Meldungen der deutschen Tagespresse. In: Best, K.-H. (ed.), *Häufigkeitsverteilungen in Texten: 33-42.* Göttingen: Peust & Gutschmidt.

Dieckmann, S., Judt, B. (1966). Untersuchung zur Wortlängenverteilung in französischen Pressetexten und Erzählungen**.** In: P. Schmidt (ed.), *Glottometrika 15: 158-165.* Trier: WVT.

Drechsler, J. (2001). Häufigkeitsverteilung von Wortlängen in gälischen Texten. In: Best, K.-H. (ed.), *Häufigkeitsverteilungen in Texten: 115-123.* Göttingen: Peust und Gutschmidt.

Egbers, J., Groen, C., Podehl, R., Rauhaus, E. (1997). Zur Wortlängenhäufigkeit in griechischen Koine-Texten. In: Best, K.H. (ed.), *Glottometrika 16: 108-120.* Trier: WVT

Feldt, S., Janssen, M., Kuleisa, S. (1997). Untersuchung zur Gesetzmäßigkeit von Wortlängenhäufigkeiten in französischen Briefen und Pressetexten. In: Best, K.H. (ed.), *Glottometrika 16: 145-151.* Trier: WVT

Gaeta, L. (1994). Wortlängenverteilung in italienischen Texten. *ZET – Zeitschrift für empirische Textforschung 1, 44-48.*

Girzig, P. (1997). Untersuchung zur Häufigkeit von Wortlängen in russischen Texten. In: Best, K.H. (ed.), *Glottometrika 16: 152-162.* Trier: WVT.

Grzybek, P. (1998). Explorative Untersuchungen zur Wort- und Satzlänge kroatischer Sprichwörter. (Am Beispiel der ‚Poslovice' von Đuro Daničić, 1871) In: Nikolaeva, T.M. (ed.), *Polytropon. K 70-letiju Vladimira Nikolaeviča Toporova: 447-465.* Moskva: Indrik.

Grzybek, P. (1999). Wie lang sind slowenische Sprichwörter? Zur Häufigkeitsverteilung von (in Worten berechneten) Satzlängen slowenischer Sprichwörter. *Anzeiger für Slavische Philologie 27, 87-108.*

Grzybek, P. (2000). Zum Status der Untersuchung von Satzlängen in der Sprichwortforschung. Methodologische Vor-Bemerkungen. In: Lilič, G.A., Birih, A.K., Nikolaeva, E.K. (eds.), *Slovo vo vremeni i prostranstve. K 60-letiju professora V.M. Mokienko: 430-457.* Sankt Peterburg: Folio-Press.

Grzybek, P. (2006). History and methodology of word length studies. In: Grzybek, P. (ed.), *Contributions to the science of text and language. Word length studies and related issues: 15-90.* Dordrecht: Springer.

Grzybek, P. (2010). Text difficulty and the Arens-Altmann law In: P. Grzybek, E. Kelih, J. Mačutek (eds.), *Text and Language. Structures · Functions · Interrelations. Quantitative Perspectives: 57-70.* Wien: Praesens.

Grzybek, P. (2011). Der Satz und seine Beziehungen. I: Satzlänge und Wortlänge im Russischen (Am Beispiel von L.N. Tolstojs «Анна Каренина»). *Anzeiger für Slavische Philologie 39; 39-74.*

Grzybek, P. Kelih, E., Stadlober, E. (2008). The relation between word length and sentence length. An intra-systemic perspective in the core data structure. *Glottometrics 16, 111-121.*

Grzybek, P., Schlatte, R. (2002). Zur Satzlänge deutscher Sprichwörter. Ein Neu-Ansatz. In: Piirainen, E., Piirainen, I. (eds.), *Phraseologie in Raum und Zeit. Akten der 10. Tagung des Westfälischen Arbeitskreises ‚Phraseologie/Parömiologie' (Münster 2001): 287-305.* Hohengehren: Schneider.

Grzybek, P., Stadlober, E. (2007). Do we have problems with Arens' law? A new look at the sentence-word relation In: Grzybek, P., Köhler, R. (eds.), *Exact Methods in the Study of Language and Text. Dedicated to Gabriel Altmann on the Occasion of his 75th Birthday: 205-217.* Berlin, New York: Mouton de Gruyter, 205-217. [Quantitative Linguistics; 62]

Grzybek, P., Stadlober, E., Kelih, E. (2007). The relationship of word length and sentence length. The inter-textual perspective. In: Decker, R., Lenz, H.-J. (eds.), *Advances in Data Analysis. Proceedings of the 30th Annual Conference of the Gesellschaft für Klassifikation e.V., Freie Universität Berlin, March 8-10, 2006: 611-618.* Berlin, Heidelberg: Springer, [Studies in classification, data analysis and knowledge organization].

Hasse, A., **Weinbrenner, M.** (1997). Zur Häufigkeit von Wortlängen in englischen Texten. In: Best, K.H. (ed.), *Glottometrika 16: 98-107.* Trier: WVT

Hein, M. (1997). Wortlängen in Briefen des spanischen Dichters Federico García Lorca. Best, K.H. (ed.), *Glottometrika 16: 138-144.* Trier: WVT

Hollberg, C. (1997). Wortlängenverteilungen in italienischen Pressetexten. In: Best, K.H. (ed.), *Glottometrika 16: 127-137.* Trier: WVT.

Kahl, S. (2002). Wortlängenverteilungen in wogulischen Texten. *Göttinger Beiträge zur Sprachwissenschaft 7, 51-63.*

Kaydanova, L. (2004/5). Zur Wortlängenhäufigkeit in usbekischen Texten. *Göttinger Beiträge zur Sprachwissenschaft 10/11, 57-66.*

Kelih, E., Grzybek, P. (2004). Häufigkeiten von Satzlängen. Zum Faktor der Intervallgröße als Einflussvariable (am Beispiel slowenischer Texte). *Glottometrics 8, 23-41*.

Kelih, E., Grzybek, P, (2005). Satzlänge: Definitionen, Häufigkeiten, Modelle. (Am Beispiel slowenischer Prosatexte). In: *Quantitative Methoden in Computerlinguistik und Sprachtechnologie 20, 31-35.* [Special Issue of: LDV-Forum. Zeitschrift für Computerlinguistik und Sprachtechnologie//Journal for Computational Linguistics and Language Technology].

Kelih, E., Grzybek, P., Antić, G., Stadlober, E. (2006). Quantitative text typology. The impact of sentence length. In: Spiliopoulou, M., Kruse, R., Nürnberger, A., Borgelt, Ch., Gaul, W. (eds.), *From Data and Information Analysis to Knowledge Engineering: 382-389.*. Heidelberg, Berlin: Springer.

Kiefer, A. (2001). Wortlängenverteilung im Pfälzischen. In: Best, K.-H. (ed.), *Häufigkeitsverteilungen in Texten: 124-131*. Göttingen: Peust und Gutschmidt.

Kim, I., Altmann, G. (1996). Zur Wortlänge in koreanischen Texten. *Glottometrika 15, 1996, 205-213.*

Kiyko, S. (2007). Wortlängen im Gotischen. *Glottometrics 13, 47-58.*

Kiyko, S. (2007a). Wortlängen im Weißrussischen. *Glottometrics 14, 46-57.*

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds). *Quantitative Linguistics. An International Handbook: 760-774.* Berlin: de Gruyter.

Köhler, R. (1986). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik.* Bochum: Brockmeyer.

Krupa, V. (1993). Wortlängen in der Sprache der Marquesas. Private communication.

Krupa, V. (1994). Wortlängen in der Sprache der Maori. Private communication.

Mačutek, J., Altmann, G. (2007). Discrete and continuous modeling in quantitative linguistics. *Journal of Quantitative Linguistics 14(1), 81-94.*

Marx, M. (2001). Zu den Wortlängen in polnischen Briefen. *Glottometrics 1, 52-62.*

Meyer, P. (1997). Word-length distribution in Inuktitut narratives: Empirical and theoretical findings. *Journal of Quantitative Linguistics 4(1-3), 143-155*

Müller, F. (2003). Wortlängen in finnischen E-Mails und Briefen. *Göttinger Beiträge zur Sprachwissenschaft 8, 71-85.*

Nemcová, E., Altmann, E. (1994) Zur Wortlänge in slowakischen Texten. *Zeitschrift für empirische Textforschung 1, 1994, 40-43*

Niehaus, B. (1997). Untersuchung zur Satzlängenhäufigkeit im Deutschen**.** In: Best, K.H. (ed.), *Glottometrika 16: 213-275.* Trier: WVT

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

Pande, H., Dhami, H.S. (2012). Model generation for word frequencies in texts with the application of Zipf's order approach. *Journal of Quantitative Linguistics 19(4), 249-261.*

Popescu, I.-I. et al. (2013) Word length: aspects and languages. In: Köhler, R., Altmann, G. (eds.), *Issues in Quantitative Linguistics Vol. 3: 224-281.* Lüdenscheid: RAM-Verlag.

Poppe, S. (2007). Die Verteilung von Kompositalängen in deutschen journalistischen Texten. *Göttinger Beiträge zur Sprachwissenschaft 15, 79-85.*

Pustet, R., Altmann, G. (2005). Morpheme length distribution in Lakota. *Journal of Quantitative Linguistics 12(1), 53-63.*

Rheinländer, N. (2001). Die Wortlängenhäufigkeit im Niederländischen. In: Best, K.-H. (ed.), *Häufigkeitsverteilungen in Texten: 142-152.* Göttingen: Peust und Gutschmidt.

Riedemann. G. (1997). Wortlängenhäufigkeiten in japanischen Pressetextne. In: Best, K.H. (ed.), *Glottometrika 16: 180-184.* Trier: WVT.

Rothe, U., Altmann, G., Wagner, K.R. (2014). Verteilung der Länge von Sprechakten in der Kindersprache. In: Wagner, K.R. (ed.), *Kindersprachstatistik: 47-56.* Essen: Die Blaue Eule.

Röttger, W., Schweers, A. (1997). Wortlängehäufigkeit in Plinius-Briefen. In: Best, K.H. (ed.), *Glottometrika 16: 121-126.* Trier: WVT.

Rottmann, O.A. (1977). Word-length counting in Old Church Slavonic. *Jorunal of Quantitative Linguistics 4, 252-256.*

Rottmann, O.A. (2003). Word lengths in the Baltic languages - are they of the same type as the word lengths in the Slavic languages? *Glottometrics 6, 52-60.*

Strobel, H. (1996). *Wortlängen in Briefen und Erzählungen von Böll und Hemingway.* Staatsexamensarbeit, Göttingen.

Uhlířová, L. (1996). How long are words in Czech? In: P. Schmidt (ed.), *Glottometrika 15: 134-146.* Trier: WVT.

Uhlířová, L. (2001). On word length, clause length and sentence length in Bulgarian. In: Uhlířová, L. et al. (eds.), *Text as a linguistic paradigm: Levels, constituents, constructs: 266-282.* Trier: WVT.

Wilson, A. (2006). Word-length duistribution in present-day Lower-Sorbian newspaper letters. In: Grzybek, P. (ed.), *Contributions to the science of text and language. Word length studies and related issues: 319-327.* Dordrecht: Springer.

Wilson, A. (2003). Word-Length distribution in modern Welsh prose texts. *Glottometrics 6, 35-39.*

Wimmer, G., Altmann, G. (2005). Unified derivation of some linguistic laws. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 791-807*. Berlin: de Gruyter.

Wittek, M. (2001). Zur Entwicklung der Satzlänge im gegenwärtigen Deutschen. In: Best, K.-H. (ed.), *Häufigkeitsverteilungen in Texten: 219-247.* Göttingen: Peust & Gutschmidt.

Zhu, J., Best, K.-H. (1997). Zur Modellierung der Wortlängen im Chinesischen. In: Best, K.H. (ed.), *Glottometrika 16: 185-194.* Trier: WVT.

Ziegler, A. (1998). Word length in Portuguese texts. *Journal of Quantitative Linguistics 5(1-2), 115-120.*

Zuse, M. (1996). The distribution of word length in Early Modern English letters of Sir Philip Sidney. *Journal of Quantitative Linguistics 3, 272-276.*

Author index

Subject index