Word length: aspects and languages

Ioan-Iovitz Popescu, Bucharest
Sven Naumann, Trier
Emmerich Kelih, Vienna
Andrij Rovenchak, Lviv
Haruko Sanada, Tokyo
Anja Overbeck, Göttingen
Reginald Smith, Rochester
Radek Čech, Olomouc
Panchanan Mohanty, Hyderabad
Andrew Wilson, Lancaster
Gabriel Altmann, Lüdenscheid

Abstract. The article presents some evaluations of four aspects of word length in different languages and compares both models and data: distribution, smoothness, word length in sentence, word length in text. A general discussion of the theoretical background is offered. As is shown, even these four problems require great teams, in order to bring elementary concepts and decisions. The number of co-authors should not surprise: everybody did what (s)he could.

0. Introduction

The word has as many properties as we are able to establish conceptually. Some of these properties supported the rise of different disciplines like semantics, morphology, dialectology, historical linguistics, etc. Some laws of behaviour of these properties are known, but the number of surprises - possibly concealed behind the concept of boundary conditions - is still greater. Here nothing helps but incessant testing, modelling, different viewing of data, modification of hypotheses, collecting of data from new languages, etc. Every "new" language can falsify a beloved theory or force us to modify it. Here, it is not possible to take into account all known properties, we necessarily must restrict ourselves to some selected aspects.

1. Length distributions

If somebody processes the enormous discipline of word problems, his first sight falls on the word length. This property has been treated since the 19th century and belonged for a long time to the preferred research domain of K.-H. Best who maltreated his students misusing them for evaluating word length in about 50 languages using about 4000 texts. It has been shown that both the models themselves vary and even in one language the parameters of identical distributions are different in different texts. This is caused by the boundary conditions sticking to

every text sort, every author, every age, education, etc. Those who tested the law-like nature of word length distribution used a general model and a software that iteratively computed the parameters and automatically performed the chi-square test for goodness-of-fit. However, there were several critical cases in which one special class of words had to be treated separately; for example in Slavic languages there are words of zero syllabic length (some prepositions) which being synsemantics display an excessive frequency. However, if one considers them as proclitics, the problem of modification of the given distribution disappears - but new problems arise.

For evaluating word length one cannot use a different measurement unit but the number of syllables. It is the only way securing generality, because it can be used in all languages. Other candidates, e.g. number of letters, graphemes, phonemes, moras, morphemes, signs, have all their intrinsic problems: there are no letters in some languages; graphemes would be more appropriate but just as phonemes or letters they omit one level, i.e. they are not immediate constituents of words, hence the distributions may be distorted (rather fractal); morphemes do not measure length but complexity (cf. e.g. introflection). In Japanese one works also with moras but for every mora one can securely state in speech whether it is one or two syllables. Thus if we want to approach the mechanism of word formation from the general point of view, we must use syllables as counting units. Besidess, everything that is written belongs to the secondary language.

There is a long bibliographical list of model fittings (cf. Best 1997, 2001; Grzybek 2006; Schmidt 1996) but no language law will ever be fully corroborated, simply because nobody can evaluate all languages or all necessary texts even in one language. The other problem, viz. the finding of boundary conditions, will for ever incriminate the clear conscience of quantitative linguists.

The following distributions have been tested: binomial, modified binomial, extended positive binomial, geometric, Cohen-negative binomial, Cohen-Poisson, Consul-Jain-Poisson, Conway-Maxwell-Poisson, Dacey-Poisson, Fucks-Poisson, Fucks-Gačečiladze-Poisson, geometric, Hirata-Poisson, hyper-Pascal, hyper-Poisson, lognormal, Meyer-Thomas, mixed Poisson, modified negative binomial, modified Poisson, negative binomial, Palm-Poisson, Pandey-Poisson, Poisson, Poisson-uniform, Pólya, Singh-Poisson. Many of them can be derived from a general model, in turn modified, mixed (or compounded) or (Feller-)generalized but it is evident that one must find the "causes" of modifications or generalizations.

The languages from which samples have been taken up to now are as follows: Arabic, Belorussian, Bulgarian, Cheremis, Chinese, Croatian, Czech, Danish, Dutch, Early English, English, Estonian, Eskimo, Faroese, Finnish, French, Gaelic, German, Hindi, Middle German, Swiss, Early New High German, German dialects, Gothic, Greek, Hindi, Hungarian, Irish, Italian, Japanese, Kechua, Korean, Latin, Latvian, Lituanian, Low German, Lower Sorbian, Malayalam, Maori, Marathi, Mordvinian, New High German, New Icelandic, Nor-

wegian, Old Church Slavic, Old High German, Old Greek, Old Hebraic, Odia (earlier Oriya), Old Icelandic, Persian, Polish, Portuguese, Russian, Saami, Serbocroatian, Slovak, Slovenian, Spanish, Swedish, Tamil, Telugu, Turkish, Ukrainian, Usbek, Vogul, Welsh, Yiddish and yearly new languages are added. A comparative within-language and between-language treatment is still missing; this enormous field cannot be captured completely even in a team-work.

This short survey shows that Indo-European languages are better represented than all the other ones, hence one can be sure that the results are somewhat skewed. Many articles try to master the boundary conditions simply by introducing new distributions - as far as the software at our disposal (*Fitter*) contains them.

In the centre of modelling one finds the Poisson distribution which can be derived in many different ways. It is very simple, and if we believe in pure chance in length ordering, it is sufficient. But if the damned boundary conditions intervene, we must clutch at a straw and search for a remedy. Either the language itself brings about a word structure that deviates from our conjectures or the text sort has its own peculiarities (e.g. poetry) or the author wanted to give his text a special air. As a matter of fact, even if all texts of a language follow the same distribution, each of them must have different parameter values. But if the writers and text sorts strongly deviate, the model must be modified. Thus developing ever new models for word length distributions is a legal, desirable activity. There is no end of this activity.

For the sake of lucidity, we present a simple survey of these distributions. The authors of individual articles used the following procedure: first fit all available theoretical distributions to all data, then take that theoretical distribution which fits well to the majority (or all) data. This is an empirically secure way to obtain good results. But if we take into account that texts are not necessarily written spontaneously and what more, after being ready they are corrected, reduced, enlarged, etc. we must expect exceptions. The only text sort written spontaneously and not corrected any more is private letters (especially those written by hand).

Another problem is the maximal length of words. In some languages they are too short, and the fitting cannot be tested because of too few degrees of freedom. In this case one must take a distribution having only one parameter, e.g. Poisson or geometric. Many times the "longest" classes are not very frequent and the theoretical frequencies must be pooled in order to obtain at least $NP_x > 1$ for all x. The way to definitive decisions is very troublesome and one does not obtain always a satisfactory result.

The Poisson distribution, either in its usual form, or displaced to the right, or truncated above the zero point (positive Poisson d.) should be used at the very beginning of any investigation. If Poisson is adequate, we conjecture that the process of writing is performed randomly or spontaneously, without any binding. If boundary conditions play a role and some classes deviate, one can either

perform class modifications and obtain e.g. the Cohen-Poisson d., the Pandey-Poisson d. or the Singh-Poisson d. If all classes deviate, one generalizes the recurrence function by adding a new parameter and may obtain e.g. the Conway-Maxwell-Poisson d., the hyper-Poisson d. or the Palm-Poisson d.; or one takes a distribution which has Poisson as a limiting case, viz. the binomial d., the hyper-binomial d., the negative binomial d., the hyper-Pascal d. and the Pólya d. Even these can be punctually modified: one already found the Cohen-negative binomial d. and G. Djuraš (2012) found a number of other modifications. However, the situation can get even more complex and one performs the Feller-generalization, namely, one replaces the argument t in the probability generating function of the Poisson distribution, $G(t) = \exp(a(t\text{-}1))$ by another probability generating function. In this way one obtained up to now the Hirata-Poisson d., the Consul-Jain-Poisson d., the Poisson-uniform d., and the Meyer-Thomas d. Now, in texts, we have everywhere the problem of possible non-homogeneity evoking the impression that the text has several strata. In such cases, the data must be captured by mixing of distributions. In this way one obtained the simple mixed Poisson d., the Dacey-Poisson d. and the Fucks-Poisson d. Still another way is considering the parameter of the Poisson distribution a variable with its own distribution. These results can be seen in Table 1.1 All of these distributions may be displaced or truncated. Their interrelations are presented in Figure 1.1

Table 1.1
Word length distributions found and their relations to Poisson
(number of parameters)

Class modifications	(positive) Cohen-Poisson(2); Pandey-Poisson(2); Singh-Poisson(2)
Poisson as special case of	Conway-Maxwell-Poisson(2); hyper-Poisson(2); Palm-Poisson(2)
Poisson as limiting case of	binomial(2); negative binomial(2); hyper-Pascal (3); Pólya(3-4)
Feller-generalization of Poisson	Hirata-Poisson(2); Consul-Jain-Poisson(2); Poisson-uniform(2); Meyer-Thomas (3)
Mixing of Poisson	mixed Poisson(3); Dacey-Poisson(2); Fucks-Poisson(≥2)

Some of the above distributions have been modified, too, and in principle it can be done with every distribution (cf. Wimmer, Witkovský, Altmann 1999). But one strives for a theory in which both the models and the variants are linguistically substantiated. As can be seen, the majority of the models have two parameters. There are still some other models (geometric, lognormal, Merkyte, hyper-binomial, extended positive Poisson, etc.) but they could be corroborated only ad hoc.

Since some languages prefer some distributions, it can be conjectured that the boundary conditions can be found directly in the given language. That means, the present distributions could be used also for typology. But we are still far from hitting such a distant target.

A serious problem is always the text size. Whatever we compute, small sizes are not sufficient. There may be outliers, there may be too few degrees of freedom, etc. But if we take long texts, we must give up any homogeneity, we must not trust the chi-square test for goodness-of-fit because with increasing size the chi-square increases and yields "bad" results. The usual technique was to take a coefficient (Cramer, Pearson, Chuprov) which took sample size (or also the degrees of freedom) into account - but the greater the size, the smaller the coefficient, i.e. it is not reliable. Unfortunately, nobody can estimate the ideal text size. There is, of course, a remedy, namely to test the goodness-of-fit by means of the determination coefficient. But in that case we consider the probability mass function a usual continuous function. This is no overrunning the scientific moral, because there are many ways to truth and each of them is only an approximation. But whatever we do, the criterion of the goodness-of-fit is a subjective decision. Even if in statistics one sets $\alpha = 0.05$ for the chi-square or $R^2 > 0.90$ for a usual function, this has nothing to do with reality or with truth, it is a convention giving us a first look of the reliability of the acceptation or rejection of the hypothesis.

The interrelations of the above distributions are presented in Figure 1.1.

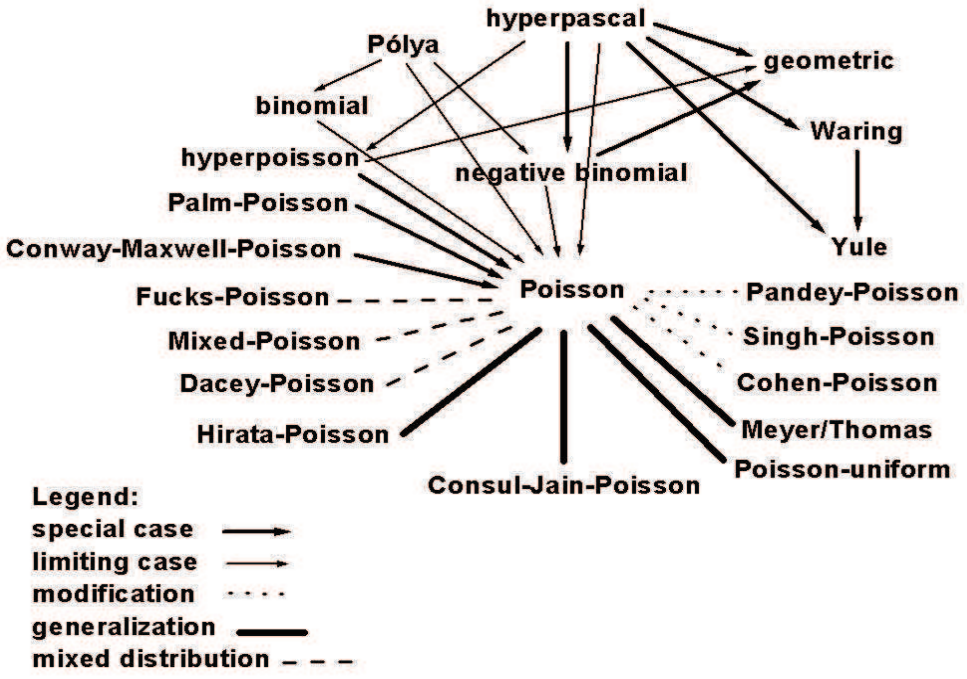


Figure 1.1. Relations of the Poisson distribution to some other ones used in word length research

In order to extend the investigation, we prepared 61 texts in 28 languages and different individual texts and try to show the general trend. The texts we analyzed are rather short hence a number of exceptions can occur. In Table 1.2 we show a survey of texts, the appropriate distributions and their parameters as well as the results of fitting. We shall try to draw some consequences. The testing by means of the chi-square test has been performed mostly with the smallest theoretical class-frequency > 1. But in several cases we pooled some classes in order to obtain better fitting. For the great majority of data we obtained several well fitting distributions. We chose either a common distribution for all texts of a certain language or we took only the "best" distribution from Table 1.1.

Table 1.2
Fitting some distributions to word length data in 61 texts in 28 languages

Language, Text alphabetically	**Distribution**	**Parameters**	X^2	*DF*	*P*	*I*	*S*
Akan Mma Nnsua Ade Bɔne	Positive Poisson	a = 1,0735	2,14	3	0,54	0,4495	1,2804
Akan Agya Yaw Ne Akutu Kwaa	Hirata-Poisson	a = 0,4352 b = 0,3817	6,75	2	0,03	0,4302	0,9519
Bamana Bamakɔ sigicogoya	Mixed Poisson	a = 1,1154 b = 0,0640 α = 0,5409	4,05	2	0,13	0,5469	1,5001
Bamana Masadennin	Mixed Poisson	a = 1,6608 b = 0,1995 α = 0,2662	3,47	2	0,18	0,6609	2,1723
Bamana Namakɔrɔba halakilen	Mixed Poisson	a = 1,8539 b = 0,2960 α = 0,1882	10,27	3	0,02	0,5814	1,9455
Bamana Sonsannin ani Surukuba	Mixed Poisson	a = 1,1755 b = 0,3686 α = 0,1479	0,63	1	0,43	0,3787	1,3181
Bulgarian* Ostrovskij, Kak se kaljavaše stomanata, Chapter 1	Cohen-Poisson	a = 1,2066 α = 0,1692	3,87	3	0,28	0,5737	0,8428
Czech Překvapení v justici	Singh-Poisson	a = 1,4105 α = 0,9131	3,67	3	0,30	0,5901	0,8296
Czech Marek Švehla: Voličův kalkul	Poisson	a = 1,2039	2,06	4	0,73	0,5872	1,4177
Czech Jan Macháček Slovenský (dobrý) příklad	Singh-Poisson	a = 1,3920 α = 0,9124	0,24	3	0,97	0,6158	1,0021
Czech Jan Čulík:O čem	Singh-	a = 1,4912	6,67	5	0,25	0,6652	1,2050

jsou dnešní Spojené státy?	Poisson	$\alpha = 0{,}9070$					
Czech Karel Hvížďala: O předem zpackané prezidentské volbě aneb Jak dlouho budeme bez prezidenta	Hyper-binomial	n = 7 m = 3,6230 q = 0,5693	4,61	3	0,20	0,5753	0,7364
French Dunkerque (Press)	Singh-Poisson	a = 1,1684 $\alpha = 0{,}6933$	0,60	2	0,74	0,6378	1,6376
German Assads Familiendiktatur (Press)	Hyper-Poisson	a = 5,1161 b = 8,1156	6,98	4	0,14	0,8533	2,3159
German ATT0012 (Press)	Hyper-Poisson	a = 6,9874 b = 11,6580	1,89	4	0,76	0,8754	2,1409
German Die Stadt des Schweigens (Press)	Hyper-Poisson	a = 7,0734 b = 12,5014	3,36	4	0,50	0,8756	2,4896
German Terror in Ost Timor (Press)	Hyper-Poisson	a = 2,9826 b = 4,7691	1,25	3	0,74	0,7352	1,9520
German Unter Hackern (Press)	Hyper-Poisson	a = 8,2168 b = 14,0931	10,71	5	0,06	0,8821	2,1106
Hindi Daily Hindi Milap, (31^{st} May, 2012): After the sanction to love marriage, (page 4)	Hyper-poisson	a = 0,5963 b = 0,7097	1,10	2	0,48	0,3907	0,8140
Hindi Swatantra Varta, (31^{st} July, 2012): The Anna Team on a cross-road (page 6)	Hyper-poisson	a = 0,4173 b = 0,5445	0,95	2	0,62	0,3356	0,7173
Hungarian A nominalizmus forradalma **(**Press)	Positive Singh-Poisson	a = 2,7021 $\alpha = 0{,}8788$	16,81	6	0,01	0,9216	1,3449
Hungarian Kunczekolbász **(**Press)	Hyper-Poisson	a = 3,5647 b = 3,5683	4,74	6	0,58	0,8883	1,4403
Indonesian Pengurus PSM terbelah **(**Press)	Conway-Maxwell-Poisson	a = 3,3373 b = 1,9779	6,98	3	0,07	0,3442	0,3323
Indonesian Sekolah ditutup **(**Press)	Cohen-binomial	n = 5 p = 0,3158 $\alpha = 0{,}0080$	4,94	1	0,03	0,4115	0,6787
Italian (Press, Online)	Singh-Poisson	a = 1,5762 $\alpha = 0{,}7773$	2,93	4	0,57	0,7588	1,4915
Japanese Miki, Jinseiron Note	Mixed Poisson	a = 2,6867 b = 0,6312 $\alpha = 0{,}2262$	3,32	4	0,51	0,8672	2,2278

Kikongo Bimpa: Ma Ngo ya Ma Nsiese	Binomial	n = 5 p = 0,1991	2,90	1	0,09	0,4205	0,7749
Kikongo Lumumba speech	Cohen-Poisson	a = 1,1487 α = 0,1540	3,94	3	0,27	0,5395	0,8396
Kikongo Nkongo ye Kisi Kongo	Hyperpa-scal	k = 0,0548 m = 0,0098 q = 0,2113	3,04	1	0,08	0,3392	1,0481
Latin Cicero, In Catilinam I	Extended positive binomial	n = 9 p = 0,1700 α = 0,7227	6,50	3	0,09	0,5666	0,6741
Latin Cicero, In Catilinam 2	Extended positive binomial	n = 9 p = 0,1801 α = 0,7166	8,05	2	0,02	0,6150	0,7940
Macedonian* Ostrovskij, Kako se kaleše čelkiot, Chapter 1	Singh-Poisson	a = 1,6427 α = 0,7696	1,09	3	0,78	0,7498	1,1283
Malayalam 1, Moralistic Hooligans	Positive Cohen-Poisson	a = 3,4255 α = 0,8565	5,01	4	0,54	0,6455	0,8924
Malayalam 2, No one should die	Positive Cohen-Poisson	a = 4,0324 α = 0,5349	4,34	6	0,63	0,7618	1,0833
Maninka Nko Doumbu Kende no. 2	Singh Poisson	a = 0,9601 α = 0,8058	1,95	3	0,58	0,5063	1,1282
Maninka Nko Doumbu Kende no. 7	Singh Poisson	a = 1,0551 α = 0,6788	0,41	3	0,94	0,5551	1,3906
Maninka Sìikán` (Constitution of Guinea, an excerpt)	Singh-Poisson	a = 1,3911 α = 0,6457	5,10	3	0,17	0,6755	1,3162
Maninka Teelen4	Singh-Poisson	a = 0,9500 α = 0,6731	1,73	3	0,63	0,5011	1,3116
Odia The Samaj, Bhubaneshwar (28 June 2012) Title: Who is great? (page 4)	Hyper-poisson	a = 0,9872 b = 0,0986	6,21	3	0,10	0,3580	0,9237
Odia The Dharitri, Balasore (12th February, 2012): Calculation for the District Council President (page 10)	Conway-Maxwell -Poisson	a = 3,7240 b = 1,7405	5,46	4	0,24	0,4056	0,4582
Romanian Paler, Aventuri solitare (excerpt)	1-d. Singh-Poisson	a = 1,5052 α = 0,7221	7,51	4	0,11	0,7235	1,3077
Romanian Popescu D.R.,	1-d.	a = 1,0786	2,02	3	0,56	0,5497	1,1578

Vânătoarea regală, Chapter 2	Singh-Poisson	α = 0,7540					
Romanian Steinhardt, Jurnalul fericirii, Trei soluții	Positive Singh-Poisson	a = 1,9659 α = 0,7901	7,70	4	0,10	0,7424	1,4402
Russian* Ostrovskij, Kak zakaljalas stal', Chapter 1	Singh-Poisson	a = 1,2427 α = 0,9374	1,57	4	0.81	0.5732	1,0362
Serbian* Ostrovskij, Kako se kalio čelik, Chapter 1	Singh-Poisson	a = 1,1926 α = 0,9171	0,28	2	0,87	0,5430	0,8474
Slovak E. Bachletová, Moja Dolná zem	Binomial	n = 11 p = 0,1141	1,01	3	0,80	0,4911	0,7038
Slovak E. Bachletová, Riadok v tlačive: nezamestnaný	Cohen-Poisson	a =1,4859 α = 0,1608	2,74	4	0,60	0,6211	0,7745
Slovenian* Ostrovskij, Kako se je kalilo jeklo, Chapter 1	Cohen-Poisson	a = 0.9856 α = 0.1580	2,51	3	0.47	0,5354	0,9902
Sundanese Agustusan (Salaka Online)	Hyper-poisson	a = 0,7204 b = 0,4222	9,66	3	0,02	0,4054	0,7763
Sundanese Aki Satimi (Salaka Online)	Hyper-poisson	a = 0,6441 b = 0,3345	0,11	2	0,95	0,3779	0,5943
Tagalog Hernandez, Limang Alas: Tatlong Santo	Hyper-poisson	a = 1,9456 b = 2,3391	6,33	5	0,28	0,6493	1,3059
Tagalog Hernandez, Magpisan	Mixed Poisson	a = 1,7016 b = 0,6537 α = 0,5287	8,80	4	0,07	0,6618	1,2883
Tagalog Rosales, Kristal Na Tubig	Mixed Poisson	a = 1,7416 b = 0,5924 α = 0,4877	7,69	4	0,10	0,6794	1,3693
Tamil (Press)	Positive Cohen-Poisson	a = 3,0521 α = 0,9115	5,73	6	0,45	0,6240	1,1262
Telugu Daily Andhra-bhoomi (4th August 2012) Train Journey without safety (page 4)	Hyper-poisson	a = 1,7924 b = 0,2301	4,59	4	0,33	0,5983	1,4391
Telugu Daily Andhra-bhoomi (4th August 2012): Trail-angaswamy: a biography, page10	Positive Cohen-Poisson	a = 3,2346 α = 0,7974	12,04	6	0,05	0,6526	1,0384
Vai Mu ja vaa lɔ	Positive	a = 0,5400	0,35	2	0,84	0,2883	0,4888

(T. Sherman)	Cohen-Poisson	α = 0,5049					
Vai Sabu Mua Ko	Positive Cohen-Poisson	a = 0,3355 α = 0,7572	2,30	1	0,13	0,2571	0,6697
Vai Vande bɛ Wu'u	Poisson	a = 0,4515	0,69	2	0,71	0,2939	0,8527
Welsh T1 Crynodeb Gweithredol	Cohen-binomial	n = 6 p = 0,1949 α = 0,3846	3,35	1	0,07	0,6010	1,0317
Welsh T2 Ffansi camu i esgidiau rhywun enwog?	1-d. Cohen-Poisson	a = 0,6713 α = 0,1270	5,80	1	0,06	0,3797	0,8398

For the analysis of the word length distribution in some Slavic languages marked with asterisk, the first chapter of the Russian novel "How the steel was tempered" (Chapter 1) by N. Ostrovskij and the translations into Slovene, Serbian, Bulgarian and Macedonian were used (for details cf. Kelih 2009). In all Slavic languages several distributions were adequate, we took only the "best" one.

Table 1.3
Some distributions used in word length research

Name	**Definition**
Poisson	$\frac{a^x e^{-a}}{x!},\ x = 0,1,2,...$
Positive Poisson	$\frac{a^x e^{-a}}{x!(1-e^{-a})},\ x = 1,2,3,...$
Pandey-Poisson	$\begin{cases} \frac{\alpha e^{-a} a^x}{x!}, & x = 0,1,2,...,c-1,c+1,c+2,. \\ 1-\alpha+\frac{\alpha e^{-a} a^c}{c!}, & x = c \end{cases}$
Positive Cohen-Poisson	$P_x = \begin{cases} \frac{(1-\alpha)a}{e^a-1-\alpha a}, & x = 1 \\ \frac{a^x}{x!(e^a-1-\alpha a)}, & x = 2,3,4,... \end{cases}$

Cohen-Poisson	$P_x = \begin{cases} e^{-a}(1+a\alpha), & x=0 \\ ae^{-a}(1-\alpha), & x=1 \\ \dfrac{a^x e^{-a}}{x!}, & x=2,3,\ldots \end{cases}$
Singh-Poisson	$P_x = \begin{cases} 1-\alpha+\alpha e^{-a}, & x=1 \\ \dfrac{\alpha a^x e^{-a}}{x!}, & x=2,3,\ldots \end{cases}$
Palm-Poisson	$\dfrac{R_{(x)}a^x}{{}_2F_0(-R,1;-a)}$, x = 0,1,…,$R$
Conway-Maxwell-Poisson	$\dfrac{a^x}{x!^b}P_0$, x = 0,1,2,…
Hyper-Poisson	$\dfrac{a^x}{b^{(x)}{}_1F_1(1;b;a)}$, x = 0,1,2,…
Negative binomial	$\binom{k+x-1}{x}p^k q^x$, x = 0,1,2,…
Hyperpascal	$\dfrac{\binom{k+x-1}{x}}{\binom{m+x-1}{x}}q^x P_0$, $x=0,1,2,\ldots$ $P_0 = ({}_2F_1(k,1;m;q))^{-1}$
Binomial	$\binom{n}{x}p^x q^{n-x}$, x = 0,1,…,n
Pólya (with 3 parameters)	$\dfrac{\binom{-p/s}{x}\binom{-q/s}{n-x}}{\binom{-1/s}{n}}$, $x=0,1,2,\ldots,n$
Hirata-Poisson	$\begin{cases} e^{-a}, & x=0 \\ \sum_{i=0}^{\left[\frac{x}{2}\right]}\binom{x-i}{i}\dfrac{e^{-a}a^{x-i}}{(x-i)!}b^i(1-b)^{x-2i}, & x=1,2,\ldots \end{cases}$

Consul-Jain-Poisson	$\begin{cases} e^{-a}, & x=0 \\ \dfrac{a(a+bx)^{x-1}e^{-(a+bx)}}{x!}, & x=1,2,... \end{cases}$
Poisson-uniform	$(b-a)^{-1}\left[e^{-a}\sum_{j=0}^{x}\frac{a^j}{j!}-e^{-b}\sum_{j=0}^{x}\frac{b^j}{j!}\right], \quad x=0,1,2,...$
Meyer-Thomas	$\frac{e^{-b}}{x!}\sum_{i=1}^{x}\binom{x}{i}ib^{i-1}(im)^{x-i}e^{-im}, \quad x=1,2,3,...$
Mixed Poisson	$\frac{\alpha a^x e^{-a}}{x!}+\frac{(1-\alpha)b^x e^{-b}}{x!}, \quad x=0,1,2,...$
Dacey-Poisson	$\frac{(1-\alpha)a^x e^{-a}}{x!}+\frac{\alpha x a^{x-1}e^{-a}}{x!}, \quad x=0,1,2,...$
Fucks-Poisson	$e^{-a}\sum_{k=0}^{\infty}(\varepsilon_k-\varepsilon_{k+1})\frac{a^{x-k}}{(x-k)!}, \quad x=0,1,2,...$

Discussion. All these distributions may be displaced (by defining the support as x = 1,2,3,… and replacing x by x-1 in the formula if the original support was x = 0,1,2,…); truncated (by setting x = 1,2,3,… and dividing the formula by 1-P_0); modified in different ways; generalized (e.g. by the Feller method), etc. In the literature on word length one finds a number of different adaptations. Unfortunately, they were made ad hoc and did not lead to a theoretical progress. Some of the distributions were ignored simply because they brought an innovation/ deviation. This study is an infinite process. It would be advisable to restrict the investigation to one family of distributions, e.g. that of Wimmer and Altmann (2005) in which one at least knows what kind of force or requirement the parameters may represent. The mixing of distributions is, from the theoretical point of view, no disadvantage, because the text may be stratified (= composed of different classes of elements, cf. Popescu, Altmann, Köhler 2010) but empirically it enlarges the number of parameters and the classical chi-square test does not bring a decisive result. In all cases one can alternatively consider the given distribution as a continuous function and test the deviations (or, decide the relevance of the deviations) by means of the determination coefficient. In most cases not only one of the given distributions is adequate, one always finds several well-fitting ones. The software tests automatically about 200 discrete distributions and it may happen that several of them are at the same time "adequate" for the same data. The curve-software tests about 8000 functions and one can add his own ones. There was only one case (Welsh) in which we were forced to use a more distant relative, the Cohen-binomial distribution, whose limiting cases are both Cohen-

Poisson and Poisson, and it is itself a modification of the binomial. It was not inserted in Figure 1.1.

The fact that one seldom meets the pure Poisson distribution in our data is a sign of special technique both in word formation and in syntax of the given languages. Or simply because languages have their own attractors for word length, and word length is always linked with some other properties.

The fate of this research area is characteristic for social sciences: the more we know, the more chaotically disintegrates the object of research and we cannot even imagine which forces were active.

In order to get a plastic picture of the results, we present the individual data using the criterion of J.K.Ord (1972), applied many times in quantitative studies. It can display groupings, trends, development, etc. One uses the first three moments of the distribution and defines

$$I = \frac{m_2}{m_1'}, \qquad S = \frac{m_3}{m_2}$$

where m'_1 is the mean and m_2, m_3 the second and third central moments of the empirical distribution (cf. Popescu et al. 2009: 154). The computed values are presented in Table 1.2. If we plot the values of I and S in a Carterian coordinate system we obtain the image presented in Figure 1.2. Even if we have a small number of texts form individual languages, one can see that they are fuzzily grouped. The Slavic languages are positioned around a straight line, Indonesian and Sundanese having strong mutual influence are very near to oneanother, Tagalog is more distant from them; Bamana and Vai have their own positions; Romanian occupies a straight line, etc. Thorough investigation of many texts of many text-sorts would furnish us - possibly - a view of text sorts and maybe also a look at the morphological typology of languages but this is not our aim. The straight line through these points shows merely the general trend strengthening our persuasion that there is some background law, a kind of attractor, which makes its way in every text taking into account the given boundary conditions.

If we compare the place of languages in Figure 1.2a, we can see that no language lies in the domain of the negative hypergeometric distribution characterized by $S < 2I - 1$. But this can be considered only the property of the languages studied. A negative m_3 is possible but not very probable. Surely, one can find languages in which S $\rightarrow$ 0 (monosyllabic) but in order to be able to generalize, one must analyze a great number of languages.

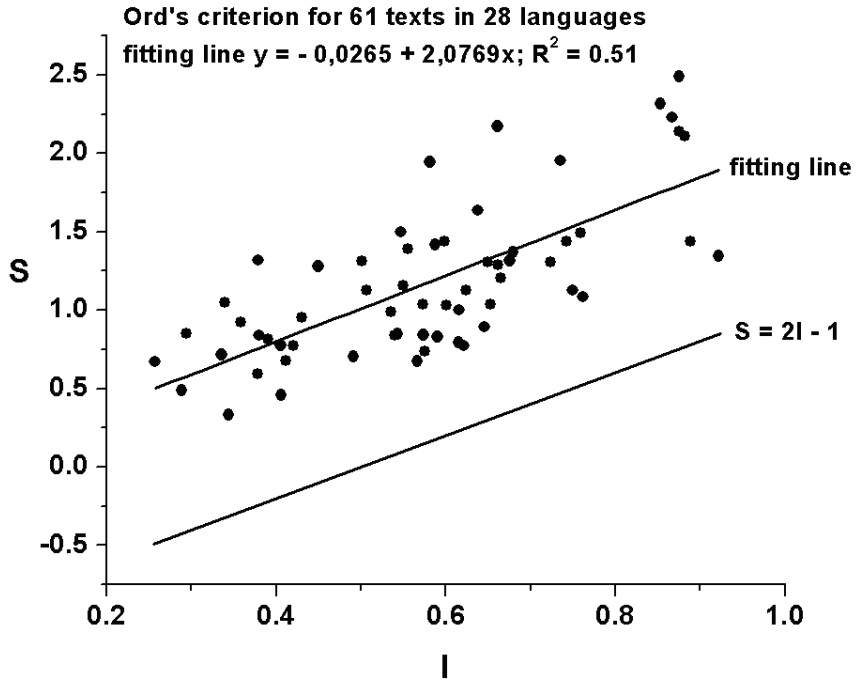


Figure 1.2a. Ord's criterion for 61 texts in 28 languages

Almost all points are placed in a kind of ellipse posited over the S = 2I-1 line. Preliminarily, it can only be said that the points occupy the domain of the hypergeometric and the beta-Pascal distributions. None of the points surpasses I = 1. The preliminary geometry of the ellipse is presented in Figure 1.2b. The ellipse center is defined by the coordinates I_C = mean I = 0,5815 and S_C = mean S = 1,1797 and the major axis by the fitting line y = - 0,0324 + 2,0846x. The distance between the most remote projections on the major axis approximates the major diameter 2*a* = 1,54 (corresponding to the pair Vai, *Sabu Mua Ko* – German, *Die Stadt des Schweigens*). Similarly, the distance between the most remote projections on the minor axis approximates the minor diameter 2*b* = 1,13 (corresponding to the pair Hungarian, *A nominalizmus forradalma* – Bamana, *Masadennin*). It results therefore a *flattening factor* $g = 1 - b/a = 0{,}27$. Though we considered grammatically and phonetically very different languages, the result is not definitive. The existence of outliers is always a reason for checking the complete computation, compare the definition of syllable in the given language with that in other ones, revise the whole theory but expecially to continue analyzing further data. It is also possible that all languages lie on the given straight line within a certain confidence interval.

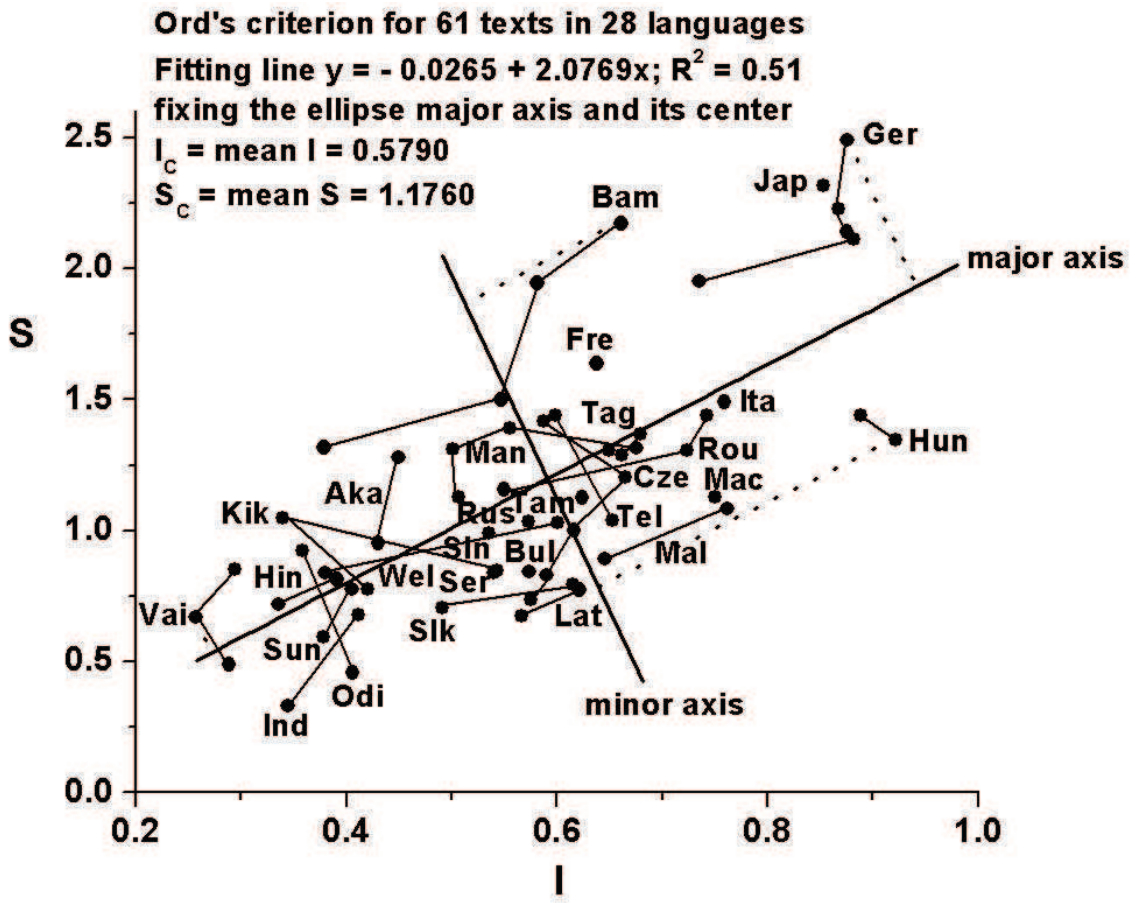


Figure 1.2b. Center and axes of the ellipse enclosing 61 texts in 28 languages (Aka = Akan, Bam = Bamako, Bul = Bulgarian, Cze = Czech, Fre = French, Ger = German, Hin = Hindi, Hun = Hungarian, Ind = Indonesian, Ita = Italian, Jap = Japanese, Kik = Kikongo, Lat = Latin, Mac = Macedonian, Mal = Malayalam, Man = Maninka, Odi = Odia, Rou = Romanian, Rus = Russian, Ser = Serbian, Slk = Slovak, Sln = Slovene, Sun = Sundanese, Tag = Tagalog, Tam = Tamil, Tel = Telugu, Vai = Vai, Wel = Welsh)

2. Smoothness

Length distributions represent an *a posteriori* ordered set of values, i.e., a pattern containing our(!) abstract view. But the sequence from which we obtain the numbers is not ordered. It follows some grammatical, thematic, psychological, text-sort and other rules or customs. These rules join autosemantics mostly by means of synsemantics. Synsemantics are more frequent and *eo ipso* they got historically shorter - unless all words were monosyllabic. In non-monosyllabic languages the sequence of lengths represents an oscillating structure which may be more or less smooth. The extent of oscillation and its (ir)regularity can be measured and expressed in many different ways. In time series analysis one uses mostly autocorrelations which enable us to see the degree of regularity of appear-

ing of the same length in some special distance (lag). In research on fractals one can compute the dimension of the resulting fractal, etc. On the other hand, we may ask about the distribution of distances between equal elements. There is a well developed theory of (random) distances (cf. Zörnig 1987) which may be compared with empirical data. In most cases they will deviate from randomness on different grounds: some grammars do not allow trespassing of special rules; for some entities the Skinner effect of reinforcement holds; some languages or text sorts do not use too long words and emphasize short distances between equal lengths, etc. These all are boundary conditions which should be investigated in individual cases.

For our purposes, we want to express the smoothness of a sequence by considering the local extremes. In the sequence 3,1,2,4,1,2,6,2 there are 6 extremes defined as numbers whose two neighbours are either both greater or both smaller. These are 1,4,1,6. The first and the last numbers are automatically extremes. Thus a sequence has *n* elements and *m* local extremes. The non-weighted non-smoothness can be expressed simply as

$$NS = \frac{m-2}{n-2}. \tag{1}$$

But their simple proportion does not say anything about the strength of the oscillation. This can be computed by taking the arc length (L) between neighbouring elements using the Euclidian distance

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

Here x_i is the value of the i-th element in the sequence. For the above example we obtain

$$L = [(3 - 1)^2 + 1]^{1/2} + [(1 - 2)^2 + 1]^{1/2} + \ldots + [(6 - 2)^2 + 1]^{1/2}.$$

In long texts, this number is usually very great. Instead of taking the length directly we can divide each length by the maximal length present in the text, (in our example it is 6), that is, we can use x_i/x_{max} in the above formula and call it y_i. In this case, however, also the step must be reduced to $1/x_{max}$, otherwise the difference between neighbours almost disappears and the arc almost equals to $(n\text{-}1)^{1/2}$. However, we shall use directly (2). In order to set up a normalized indicator Popescu et al. (2010: 97) defined the roughness as $R = NS(L)/L_{max}$. For our purposes we take into account the omission of "words" of length 0 and obtain

$$L_{\max} = (n-1)[(x_{\max} - 1)^2 + 1]^{1/2} \tag{3}$$

hence the roughness indicator has the form

$$(4) \qquad R = \frac{(m-2)L}{(n-2)L_{\max}}.$$

For the sake of illustration consider the roughness of the above sequence 3,1,2,4,1,2,6,2. Here

n = 8,
m = 6,
x_{max} = 6,
$L = [(3-1)^2+1]^{0.5} + [(1-2)^2+1]^{0.5} + [(2-4)^2+1]^{0.5} + [(4-1)^2+1]^{0.5} +$
$+ [(1-2)^2+1]^{0.5} + [(2-6)^2+1]^{0.5} + [(6-2)^2+1]^{0.5} = 18.7091$
$L_{max} = (8-1)[(6-1)^2+1]^{0.5} = 35.6931.$

Hence

R = (6-2)*18.7091/[(8-2)35.6931)] = 0.3494.

The roughness lies in the interval <0,1>. Computing this indicator for all our data, we obtain the results in Table 2.1, where *n* is the number of words of the considered text.

Table 2.1
Roughness of lengths in 61 texts of 28 languages

Language, Text alphabetically	**n**	**m**	**L**	**R**
Akan Mma Nnsua Ade Bɔne	143	60	218,6213	0,1536
Akan Agya Yaw Ne Akutu Kwaa	201	70	290,4720	0,1569
Bamana Bamakɔ sigicogoya	1138	382	1739,9743	0,1004
Bamana Masadennin	2615	758	4054,0607	0,0635
Bamana Namakɔrɔba halakilen	2392	718	3615,7292	0,0745
Bamana Sonsannin ani Surukuba	1406	356	1890,2265	0,0823
Bulgarian Ostrovskij, Kak se kaljavaše stomanata, Chapter 1	926	464	1644,8522	0,1744
Czech Překvapení v justici	289	116	500,9478	0,1355
Czech Marek Švehla: Voličův kalkul	288	127	482,1356	0,0911
Czech Jan Macháček Slovenský (dobrý) příklad	340	151	599,0170	0,1528
Czech Jan Čulík:O čem jsou dnešní Spojené státy?	2003	868	3633,3438	0,0867
Czech Karel Hvížďala :O předem zpackané prezidentské volbě aneb Jak dlouho budeme bez prezidenta	929	386	1615,0799	0,1185

French Dunkerque (Press)	1532	620	2558,3886	0,0955
German Assads Familiendiktatur (Press)	1415	573	2587,2710	0,0817
German ATT0012 (Press)	1146	478	2153,8356	0,0971
German Die Stadt des Schweigens (Press)	1567	648	2871,0648	0,0835
German Terror in Ost Timor (Press)	1398	591	2476,4021	0,0927
German Unter Hackern (Press)	1363	573	2558,3717	0,0979
Hindi Daily Hindi Milap, (31st May, 2012): After the sanction to love marriage, (page 4)	1106	463	1655,9122	0,1518
Hindi Swatantra Varta,(31st July, 2012): The Anna Team on a cross-road (page 6)	860	321	1212,2651	0,1272
Hungarian A nominalizmus forradalma **(**Press)	1314	666	2841,2580	0,0784
Hungarian Kunczekolbász **(**Press)	458	241	1016,7097	0,1446
Indonesian Pengurus PSM terbelah **(**Press)	345	130	537,8043	0,1144
Indonesian Sekolah ditutup **(**Press)	280	130	456,1638	0,1476
Italian (Press online)	2516	1304	4974,2034	0,1270
Japanese Miki, Jinseiron Note	2043	1099	3951,2210	0,1035
Kikongo Bimpa: Ma Ngo ya Ma Nsiese	956	500	1557,7629	0,1670
Kikongo Lumumba speech	824	375	1397,4363	0,1492
Kikongo Nkongo ye Kisi Kongo	768	360	1139,6341	0,1362
Latin Cicero, In Catilinam I	1064	456	1853,2258	0,1225
Latin Cicero, In Catilinam II	3095	1503	5632,4018	0,1249
Macedonian Ostrovskij, Kako se kaleše čelkiot, Chapter 1	1123	628	2251,3309	0,2197
Malayalam 1, Moralistic Hooligans	282	139	594,3144	0,1284
Malayalam 2, No one should die	288	144	668,4151	0,1434
Maninka Nko Doumbu Kende no. 2	1276	399	1913,3819	0,0917
Maninka Nko Doumbu Kende no. 7	1535	564	2394,5710	0,0941
Maninka Siikán` (Constitution of Guinea, an excerpt)	1662	729	2950,3257	0,1526
Maninka Teelen4	1484	438	2182,7239	0,0849
Odia The Samaj, Bhubaneshwar (28 June 2012): Who is great? (page 4)	348	136	549,2713	0,1008
Odia The Dharitri, Balasore (12th February, 2012): Calculation for the District Council President (page 10)	630	313	1084,2770	0,1403
Romanian Paler, Aventuri solitare (excerpt)	891	456	1681,0150	0,1586

Romanian Popescu D.R., Vânătoarea regală, Chapter 2	1002	437	1658,3234	0,1413
Romanian Steinhardt, Jurnalul fericirii, Trei soluții	1511	708	2718,2766	0,1385
Russian Ostrovskij, Kak zakaljalas stal', Chapter 1	792	327	1316,6654	0,1128
Serbian Ostrovskij, Kako se kalio čelik, Chapter 1	1001	440	1703,1391	0,1811
Slovak E. Bachletová, Moja Dolná zem	872	382	1435,2595	0,1412
Slovak E. Bachletová, Riadok v tlačive: nezamestnaný	924	422	1655,5878	0,1343
Slovenian Ostrovskij, Kako se je kalilo jeklo, Chapter 1	977	376	1556,7049	0,1200
Sundanese Agustusan (Salaka Online)	416	181	664,2653	0,1357
Sundanese Aki Satimi (Salaka Online)	1283	584	2011,2286	0,1729
Tagalog Hernandez, Limang Alas: Tatlong Santo	1738	946	3238,0859	0,1434
Tagalog: Hernandez, Magpisan	1466	870	2838,6311	0,1625
Tagalog: Rosales, Kristal Na Tubig	1958	1201	3794,2703	0,1681
Tamil (Press)	384	167	771,8423	0,1080
Telugu Daily Andhrabhoo mi (4thAugust 2012): Train Journey without safety (p. 4)	665	304	1303,9431	0,0890
Telugu Daily Andhrabhoo mi (4th August 2012): Trailangaswamy:a biography (p. 10)	2295	144	616,9241	0,1261
Vai Mu ja vaa lɔ (T. Sherman)	3140	840	4079,9018	0,0842
Vai Sabu Mua Ko	495	136	631,3964	0,1099
Vai Vande bɛ Wu'u	426	159	571,2930	0,1574
Welsh T1 Crynodeb Gweithredol	985	502	1750,4984	0,1487
Welsh T2 Ffansi camu i esgidiau rhywun enwog?	1002	375	1441,3001	0,1303

As can be seen, the roughness is not very variable. This is based on the fact that x_{max} is used very seldom and R expresses merely the alternation of synsemantics and autosemantics. Hence it expresses formally a background of grammar. A similar indicator would be the fractal dimension of syllabic lengths.

The testing of the difference between two R is lengthy but possible (cf. Popescu, Mačutek, Altmann 2009: 49ff.). Here we shall merely order the languages/texts according to R and delay the problem of finding some variables responsible for the given magnitude of R. Since one finds very different values in the same language, one can conclude that text sort, style, education, etc. play an important role and hinder(!) setting up language typologies. Or, one would be

forced to propose different kinds of normalizations based on boundary conditions.

In any case we see that "great jumps" are not usual in the length sequences. The greatest roughness in our data has Macedonian followed by other Slavic languages. African languages are situated at the bottom of the realized values. This is perhaps the first hint at a possible boundary condition: since in Slavic languages we took the same translated text (Russian, Macedonian, Bulgarian, Serbian) - except for Slovenian - and obtained a great roughness. But Slovenian has a much lower roughness, and in Slovak texts it tends rather to the lower limit of the interval. Hence the above results are merely a first image of a property which can depend on all factors which are effective in a text.

3. Word length in sentence

The fact that the subject of sentence stands mostly at the beginning of sentence leads mechanically to its description, characterisation, circumstances, etc. in the rest of sentence. Here the specifications (attributes, predicates, modificators) of first order, then those of second order, etc. are placed and the longer the sentence, the longer get the specifications. This holds, of course, only in form of averages, not for individual sentences. The idea expressed by L. Uhlířová (1997) has been corroborated in some later investigations (cf. Fenk, Fenk-Oczlon 2006; Kelih 2010; Fan, Grzybek, Altmann 2010). Nevertheless, one can expect text-specific or language-specific exceptions, e.g. in Japanese where a long attributive clause can precede the subject, hence the given explication does not hold. In any case, the hypothesis can be tested using our data on a broad background.

The testing can be performed in two ways:

(1) For each text separately we form groups of sentences having the same number of words; for each group we compute the mean length at each position and investigate the course of length. This way seems to be simple but it can be investigated only using long texts in which every sentence length is repeated several times in order to obtain reliable results. One takes groups containing at least 5 sentences (but 10 is "better").

(2) One considers the text as a whole and computes the mean length of the first word, than that of the second, etc. in all sentences. We obtain the best estimation for the positions at the beginning, but the reliability of the mean decreases. If there are no more than 10 sentences having a certain length, one should cease computing. This approach has a certain disadvantage: short sentences may be constructed differently and destruct the smooth increase of length. Different additional hypotheses are possible: (i) If for different sentence lengths the course of mean word lengths is different, is this difference significant? That is, are the slopes different? (ii) Is it possible that in short sentences the long words stay at the beginning and in long ones (also) at the end? (iii) A quite different possibility

is the hypothesis that the more synthetic a language, the greater is the oscillation of word length in sentence (i.e. the more analytic the language, the smaller the oscillation?).

The views may be combined, different tests can be proposed and if a text does not abide by any of them, literary scientists may search for the causes of this phenomenon.

Evidently, the number of hypotheses can be increased and the way to a theory is still long, even if in some languages one can obtain clearer results. The methodological problem is: when can we say that a hypotheiss of this sort is sufficiently corroborated? Are all our trials membra disiecta? Was our analysis of words "correct" or is a negative result the consequence of a "false" analysis?

As an example of approach (1), we show the results using a Slovak text (E. Bachletová, *Moja Dolná zem*). Since only groups containing at least 5 sentences are relevant, we obtain the results as presented in Table 3.1.

Table 3.1
Mean words lengths in individual positions in sentence groups (Slovak text)
(Approach 1)

Sentence length	**Position in sentence**					
	1	**2**	**3**	**4**	**5**	**6**
2	2,6	2,4				
3	1,6	2	2,6			
5	2,2	2,2	2	2,1	2,6	
6	1,4	2,4	2	1,8	2	2,8

As can be seen, the course is not always monotonously increasing but in each group it is quite different. Nevertheless, the last value is mostly the greatest. It is, of course, possible that the groups of equally long sentences are not sufficiently represented. It is better if one performs this investigation with longer texts. Here, lengths 1, 4, and > 6 could not be taken into account because of too small samples.

Thorough scrutinizing of this hypothesis was performed in Fan, Grzybek, Altmann (2010) who nevertheless stated that there is an increase of mean length in sentence, but the longer the sentence the smaller is the slope *b*. The decrease of the slope is very regular. Unfortunately, for short texts this cannot be shown.

The same text scrutinized under approach (2) (considering all sentences) yields the results presented in Table 3.2 and Figure 3.1.

While the first approach yields not yet interpretable results, the second approach yields an approximately linear dependence presented in Figure 3.1. The strong oscillation is mostly ascribed to the existence of clauses within which the same tendency is repeated. In short texts this can cause a rejection of the hypothesis.

Table 3.2
Mean words lengths in individual positions in sentence groups (Slovak text)
(Approach 2)

Position	1	2	3	4	5	6	7	8	9	10
Mean	2,17	1,88	2,13	2,25	2,21	2,29	2,30	2,25	2,43	2,77

11	12	13	14	15	16	17	18	19	20
2,57	2,32	2,35	2,47	2,50	2,20	2,50	2,27	2,60	2,50

But even if we partition the sentence in clauses and make equal length groups, it is possible that e.g. the third clause in sentence behaves differently and in that case we create again inhomogeneous data.

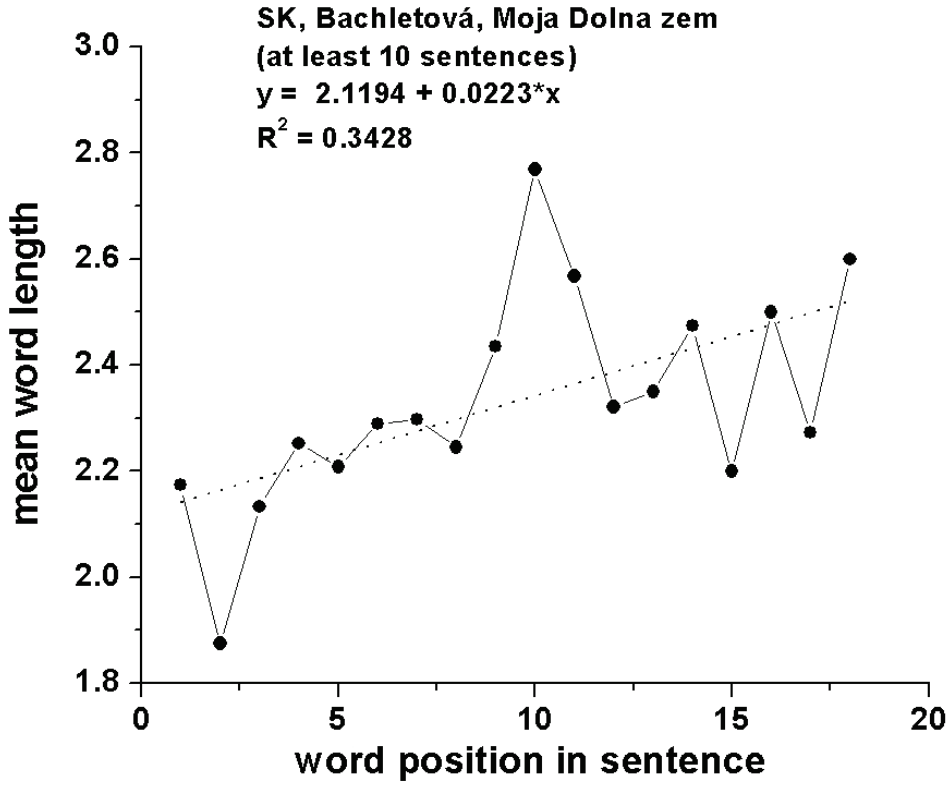


Figure 3.1. Mean word length in positions 1-18 in a Slovak text

For Bachletová's *Moja Dolná zem* we obtain *Mean length* = 2,1194 + 0,0223**Position*, with R^2 = 0,34 which is not sufficiently significant but the trend is visible inspite of great oscillation. The results for linear fitting from other texts are presented in Table 3.3. We notice, however, that power fitting generally

allows better results as, for instance, in this case, *Mean length* = 2,0148**Position*$^{0.0715}$, with R^2 = 0,40.

Table 3.3
The course of mean word lengths in 61 texts and 28 languages
(in contrast to Table 4, here *n* = number of sentences; the mean word length is taken over at least 10 sentences)

Language, Text alphabetically	**n**	**a**	**b**	**R^2**
Akan Agya Yaw Ne Akutu Kwaa	18	1,7453	-0,0327	0,176
Akan Mma Nnsua Ade	15	1,6650	-0,0087	0,007
Bamana Bamakɔ sigicogoya	84	1,6305	-0,0012	0,003
Bamana Masadennin	207	1,5829	0,0038	0,022
Bamana Namakɔrɔba halakilen	248	1,6318	-0,0048	0,053
Bamana Sonsannin ani Surukuba	168	1,5291	-0,0076	0,104
Bulgarian Ostrovskij, Kak se kaljavaše stomanata, Chapter 1	83	2,1471	-0,0008	0,0005
Czech Čulík, O čem jsou dnešní Spojené státy?	114	2,3508	0,0012	0,0023
Czech Hvížďala, O předem zpackané prezidentské volbě	37	2,3372	-0,0008	0,0009
Czech Macháček, Slovenský dobrý příklad	22	1,9034	0,0452	0,314
Czech Spurný, Prekvapení v justici	17	1,8970	0,0421	0,232
Czech Švehla, Editorial, Voličův kalkul	19	1,8672	0,0492	0,236
French Dunkerque (Press)	105	1,7267	0,0082	0,13
German Assads Familiendiktatur	119	2,0271	0,0064	0,028
German ATT0012 (Press)	83	1,9008	0,0210	0,233
German Die Stadt des Schweigens (Press)	120	1,9639	0,0034	0,015
German Terror in Ost Timor (Press)	98	1,9258	0,0054	0,027
German Unter Hackern (Press)	110	1,955	0,0122	0,128
Hindi After the sanction to love marriage	38	1,7734	-0,00003	0,000006
Hindi The Anna Team on a cross-road	42	1,6007	0,0045	0,03
Hungarian A nominalizmus forradalma (Press)	63	2,4654	0,0139	0,09
Hungarian Kunczekolbász (Press)	32	2,4196	0,0286	0,25

Indonesian Pengurus PSM terbelah (Press)	28	2,444	0,0222	0,13
Indonesian Sekolah ditutup (Press)	15	2,7691	-0,0170	0,077
Italian (Press, Online)	92	2.2088	0.0008	0.0014
Japanese Miki, Jinseiron Note, first 100 sentences	100	1,8826	0,0166	0,174
Kikongo Bimpa: Ma Ngo ya Ma Nsiese	79	1,9895	0,0149	0,086
Kikongo Lumumba speech	43	2,0584	-0,0057	0,088
Kikongo Nkongo ye Kisi Kongo	29	1,7176	0,0051	0,052
Latin Cicero, In Catilinam I	80	2,265	0,0095	0,13
Latin Cicero, In Catilinam II	180	2,3819	0,0015	0,004
Macedonian Ostrovskij, Kako se kaleše čelkiot, Chapter 1	89	2,2924	-0,0036	0,043
Malayalam 1, Moralistic Hooligans	32	3,5957	0,0527	0,336
Malayalam 2, No one should die	29	4,0148	0,0259	0,061
Maninka Nko Doumbu Kende no. 2	37	1,7707	-0,0006	0,003
Maninka Nko Doumbu Kende no. 7	34	1,7909	-0,0030	0,04
Maninka Siikán` (Constitution of Guinea, an excerpt)	94	1,9545	-0,0056	0,076
Maninka Teelen4	29	1,6951	-0,0015	0,014
Odia Calculation for the District Council President	28	2,9565	-0,0018	0,0033
Odia Who is great?	36	2,9707	-0,0198	0,08
Romanian Paler, Aventuri solitare (excerpt)	17	1,93	0,0095	0,04
Romanian Popescu D.R., Vânătoarea regală, Chapter 2	61	1,7838	0,0002	0,00008
Romanian Steinhardt, Jurnalul fericirii, Trei soluții	85	1,975	0,0018	0,005
Russian Ostrovskij, Kak zakaljalas stal', Chapter 1	76	2,1320	0,0058	0,02
Serbian Ostrovskij, Kako se kalio čelik, Chapter 1	83	2,1685	-0,0123	0,168
Slovak E. Bachletová, Moja Dolná zem	92	2,1194	0,0223	0,34

Slovak E. Bachletová, Riadok v tlačive: nezamestnaný	78	2,2433	0,0245	0,17
Slovenian Ostrovskij, Kako se je kalilo jeklo, Chapter 1	84	1,9690	-0,0046	0,02
Sundanese Agustusan (Salaka Online)	53	2,158	0,0008	0,0003
Sundanese Aki Satimi (Salaka Online)	147	2,2089	-0,0069	0,09
Tagalog Hernandez, Limang Alas: Tatlong Santo	104	1,9915	0,0121	0,13
Tagalog Hernandez, Magpisan	111	1,9606	0,025	0,13
Tagalog Rosales, Kristal Na Tubig	139	2,1791	-0,0024	0,009
Tamil (Press)	32	3,4551	0,0059	0,0067
Telugu Trailangaswamy	51	3,4916	0,0479	0,131
Telugu Train Journey without safety	61	3,2949	0,0331	0,3049
Vai Mu ja vaa lɔ (T. Sherman)	193	1,4867	-0,0009	0,0060
Vai Sabu Mua Ko	39	1,5749	-0,0130	0,2536
Vai Vande bɛ Wu'u	35	1,4922	-0,0061	0,0270
Welsh T1 Crynodeb Gweithredol	40	2,0535	-0,0040	0,0241
Welsh T2 Ffansi camu i esgidiau rhywun enwog?	34	1,6492	-0,0021	0,0123

(i) As can be seen looking at parameter *b* there is a positive slope, but not with all texts. A Slovenian, Sundanese and Tagalog texts display a slightly decreasing tendency of length which can be considered simply zero. But even the positive slopes are very small, not significant.

(ii) Though the slope is very small, there is, nevertheless, an increasing tendency, but the determination coefficient is in all cases not high enough in order to speak of a real tendency. The oscillation is too great and destructs the linear regression. In all cases, the parameter *b* and the determination coefficient are too small.

Discussion. This tendency - if it exists at all - cannot be presented with persuasion. If we form groups of sentences with equal length, we mix up the text and in any case, we ignore the boundaries of clauses. Whatever we do, the oscillation remains very strong. The most conspicuous case is a stage play in which the speech of persons may be very different. Word length in one- or two-word sentences may strongly differ, e.g. "Yes!", "No!" against "Wonderful!", "Excellent!", G. "Absolut unwahrscheinlich!" That means, the relationship is not quite general and both the texts and the way of constructing data must be selected

according to pre-formulated criteria. Hence, here either there is no law or the boundary conditions necessary for the validity of a law are seldom fulfilled.

On the other hand, word is, perhaps, not the adequate unit for measuring the length rhythm in the sentence. Is it the clause, the phrase, the punctuation or something else?

Hence, the sequence of word lengths in sentence may be random or it may follow some trend which depends on boundary conditions like style, text sort, thematic concentration, text homogeneity and other circumstances which must be scrutinized in detail, in order to find the links of this phenomenon to other phenomena and subsume it under an extensive "word length theory".

4. Word length in increasing text

According to the above hypothesis, word length increases with increasing text. This is in principle impossible and can be realized only in short texts. Word length cannot increase continuously but there may be a limit to which it converges. The test can be performed by evaluating the texts "vertically", i.e. to compute the mean word length of say first 10 sentences, then of the next ten, etc. The grouping can be made differently, e.g. in short texts one takes only groups of 5 sentences, but in long ones one can take 100 or whole chapters.

A second method is to compute stepwise the mean of the first x words and observe the change of the mean. If the hypothesis of increasing length is correct, then the curve must be at least non-decreasing.

There is no "best" method and perhaps the testing must be made differently for every text.

Let us consider first the Slovak text *Moja Dolná zem* by Bachletová containing 92 sentences. We subdivide the text in sentence groups taking 10 sentences together, compute the mean word lengths in the first 10 sentences, then in the next 10 sentences, etc. and for the last group we take the rest. We obtain the result**s** plotted in Figures 4.1. The numerical results of several texts are presented in Table 4.1.

If we compute the linear regression in Bachletová's *Moja Dolná zem*, we even obtain a decreasing function ($R^2 = 0{,}12$)

$$\textit{Mean length} = 2{,}3278 - 0{,}0152\textit{Group},$$

which is not significant but at least shows that the trend is not that simple as conjectured in the hypothesis. Surely, the oscillation may disturb the trend but in any case we have here cases of rejection.

If the texts are short, one may form 5 groups which are enough for studying the tendency - if it is simple. Testing the hypothesis using our texts we obtained the results in Table 4.2. We took means of 10 subsequent sentences.

As can be seen, not only we do not have a significant increase of word length with deployment of text, just on the contrary. In the most cases the coefficient of linear regression *b* is negative and in some cases the regression is even significant (cf. Akan). The tendency may hold for some languages or texts (e.g. French) but it is at least not as general as supposed.

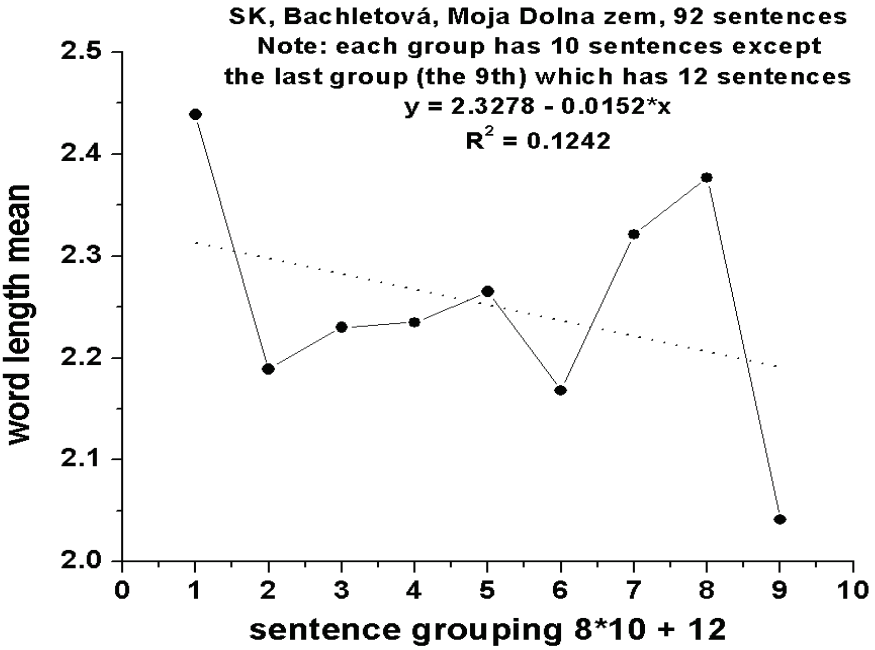


Figure 4.1a. Mean word length in groups of 10 subsequent sentences in *Moja Dolná zem* (Bachletová, Slovak)

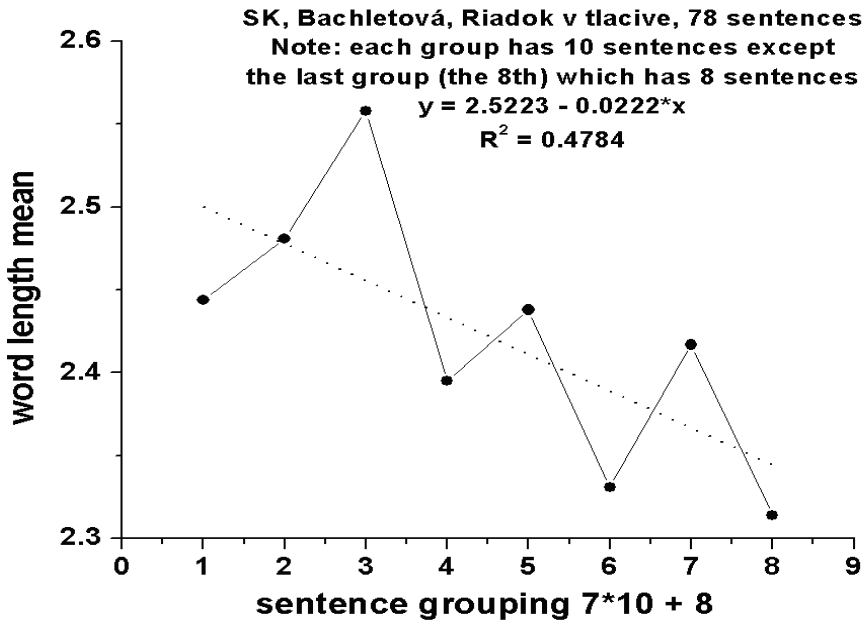


Figure 4.1b. Mean word length in groups of 10 subsequent sentences in *Riadok v tlačive* (Bachletová, Slovak)

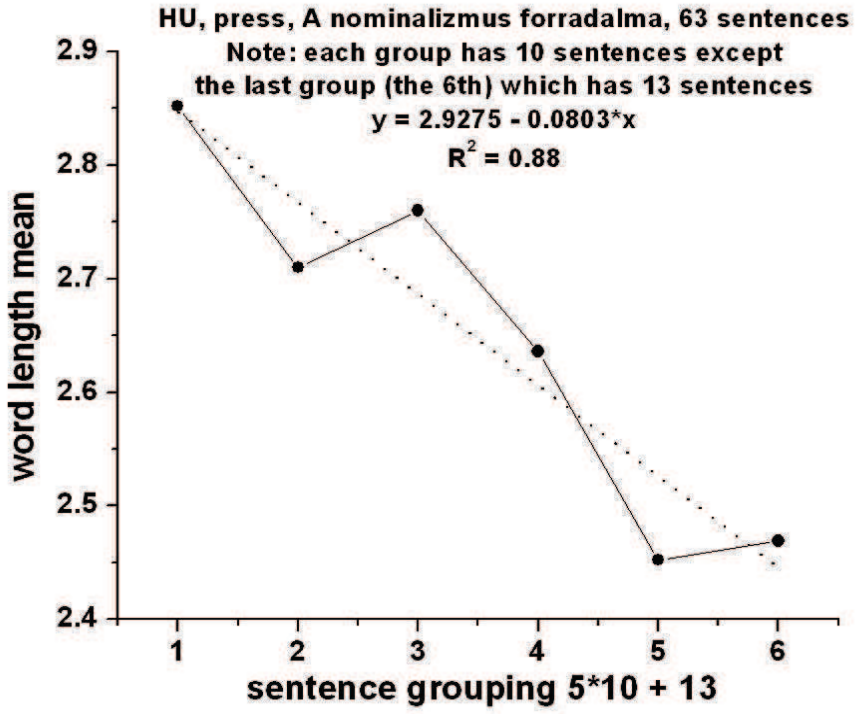


Figure 4.1c. Mean word length in groups of 10 subsequent sentences in *A nominalizmus forradalma* (press, Hungarian)

In spite of strong oscillations, the trend of the mean word length in deploying text can be considered also in detail, by subsequent single (not grouped) sentences, as revealed by comparing Figures 4.1a and 4.2 for Bachletová's *Moja Dolná zem*. In particular, in both plots the slope of the linear fitting is negative, indicating a slight decrease of the mean word length in deploying text. The data of all texts as processed sentence by sentence (that is not grouped as above in Table 4.2) are collected in Table 4.3 below. The only difference in slope *b* sign between single and grouped sentences was found for Latin and Romanian (D.R. Popescu). According to this table, 38 out of 61, that is about 62 % cases have negative *b*.

Both results show that increasing text does not necessarily mean increase of word length. This is perhaps caused by the fact that not only new words appear in the text - which may be longer - but the references to the preceding central words are shorter. As a matter of fact, none of the sequences display a significant increase of word length. Though there are positive results in the literature, we recommend further testing, selection of new texts and searching for boundary conditions.

Table 4.1
Mean word length in increasing texts (by subsequent grouped sentences)

Language, Text alphabetically	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Akan Mma Nnsua Ade Bɔne	1,833	1,688	1,571	1,625	1,429	-	-	-	-	-
Bamana Bamakɔ sigicogoya	1,761	1,541	1,642	1,573	1,512	1,561	1,662	1,686	-	-
Bamana Masadennin	1,701	1,638	1,56	1,637	1,664	1,486	1,512	1,67	1,654	1,556
Bamana Namakɔrɔba halakilen	1,598	1,526	1,594	1,650	1,462	1,475	1,728	1,497	1,703	1,609
Bamana Sonsannin ani Surukuba	1,627	1,625	1,438	1,435	1,436	1,458	1,392	1,514	-	-
Bulgarian Ostrovskij, Kak se kalja-vaše stomanata, Chapter 1	2,385	2,118	2,017	2,146	2,147	1,979	2,057	2,272	-	-
Czech Čulík, O čem jsou dnešní Spojené státy?	2,351	2,434	2,199	2,626	2,356	2,266	2,195	2,244	2,385	2,427
Czech Hvížďala, O předem zpackané prezidentské volbě	2,179	2,436	2,33	2,596	2,306	2,319	2,374	-	-	-
Czech Macháček, Slovenský dobrý příklad	2,175	2,453	2,384	2,151	-	-	-	-	-	-
Czech Spurný, Prekvapení v justici	2,378	2,289	2,271	2,205	2,140	2,306	-	-	-	-
Czech Švehla, Editorial, Voličův kalkul										
French Dunkerque (Press)	1,772	1,758	1,807	1,843	1,873	-	-	-	-	-
German Assads Familiendiktatur (Press)	2,038	2,005	2,109	2,220	2,142	2,036	-	-	-	-
German ATT0012 (Press)	2,113	2,057	2,278	2,208	1,989	1,988	1,935	2,058	-	-
German Die Stadt des Schweigens (Press)	1,811	1,964	1,959	1,966	2,078	1,943	2,108	2,159	2,166	1,931
German Terror in Ost Timor (Press)	1,953	2,005	1,955	1,876	1,85	2,272	2,08	1,896	1,884	1,913

German Unter Hackern (Press)	2,106	2,189	2,275	2,101	2,112	1,722	1,851	2,135	2,048	1,97
Hindi After the sanction to love marriage	1,778	1,698	1,801	1,788	1,764	1,624	1,693	1,862	-	-
Hindi The Anna Team on a cross-road	1,591	1,624	1,566	1,652	1,748	1,727	1,677	1,573	-	-
Hungarian A nominalizmus forradalma (Press)	2,852	2,71	2,76	2,636	2,452	2,469	-	-	-	-
Hungarian Kunczekolbász (Press)	2,737	2,582	2,663	-	-	-	-	-	-	-
Indonesian Pengurus PSM terbelah (Press)	2,626	2,567	2,515	-	-	-	-	-	-	-
Indonesian Sekolah ditutup (Press)	2,644	2,540	2,562	-	-	-	-	-	-	-
Italian (Press, Online)	2,180	2,093	2,102	2,050	2,292	2,286	2,234	2,327	2,354	-
Japanese Miki, Jinseiron Note, first 100 sentences	2,075	2,291	2,071	2,052	1,897	1,995	2,116	2,101	2,094	2,321
Kikongo Bimpa: Ma Ngo ya Ma Nsiese	2,065	2,076	2,062	2,189	2,082	2,101	1,945	2,256	-	-
Kikongo Lumumba speech	1,947	2,032	1,991	2,173	2,000	1,956	1,923	2,095	2,016	-
Kikongo Nkongo ye Kisi Kongo	1,689	1,896	1,738	1,816	1,802	1,758	-	-	-	-
Latin Cicero, In Catilinam I	2,505	2,327	2,484	2,281	2,350	2,204	2,332	2,403	-	-
Latin Cicero, In Catilinam II	2,312	2,403	2,467	2,436	2,395	2,288	2,432	2,278	2,652	2,368
Macedonian Ostrovskij, Kako se kaleše čelkiot, Chapter 1	2,591	2,349	2,112	2,198	2,371	2,017	2,282	2,388	2,260	-
Malayalam 1, Moralistic Hooligans	4,200	4,026	3,760	3,660	3,875	3,472	-	-	-	-
Malayalam 2, No one should die	4,338	4,487	3,962	3,974	4,108	4,268	-	-	-	-
Maninka Nko Doumbu Kende no. 2	1,646	1,799	1,653	2,004	1,657	1,704	1,842	-	-	-
Maninka Nko Doumbu Kende no. 7	1,880	1,659	1,800	1,591	1,656	1,663	1,835	-	-	-
Maninka Siikán` (Constitution of Guinea, an excerpt)	1,901	1,805	2,047	1,906	1,988	1,876	1,821	1,861	1,846	-
Maninka Teelen4	1,745	1,680	1,753	1,667	1,456	1,627	1,559	-	-	-
Odia Calculation for the District Council	3,000	3,180	3,115	2,712	2,750	3,135	-	-	-	-

President										
Odia Who is great?	3,020	2,553	2,772	3,068	2,656	2,606	3,264	-	-	-
Romanian Paler, Aventuri solitare (excerpt)	2,211	1,87	2,117	-	-	-	-	-	-	-
Romanian Popescu D.R., Vânătoarea regală, Chapter 2	1.922	1.667	1.674	1.798	1.839	1.795	-	-	-	-
Romanian Steinhardt, Jurnalul fericirii, Trei soluții	1,972	2,109	1,904	2,159	1,919	2,064	1,97	2,164	1,955	-
Russian Ostrovskij, Kak zakaljalas stal', Chapter 1	2,123	2,115	1,862	1,916	1,924	1,838	1,901	1,995		
Serbian Ostrovskij, Kako se kalio čelik, Chapter 1	2,248	2,297	1,971	1,991	2,090	1,933	2,138	2,042	-	-
Slovak E. Bachletová, Moja Dolná zem	2,439	2,189	2,230	2,235	2,265	2,168	2,321	2,377	2,041	-
Slovak E. Bachletová, Riadok v tlačive: nezamestnaný	2,444	2,481	2,558	2,395	2,438	2,331	2,417	2,314	-	-
Slovenian Ostrovskij, Kako se je kalilo jeklo, Chapter 1	2,000	2,088	1,912	1,706	2,000	2,500	2,059	2,147	-	-
Sundanese Agustusan (Salaka Online)	1,987	2,138	2,295	2,133	2,253	-	-	-	-	-
Sundanese Aki Satimi (Salaka Online)	2,114	2,159	2,137	2,065	2,223	2,085	2,244	-	-	-
Tagalog Hernandez, Limang Alas: Tatlong Santo	2,231	2,035	2,091	2,121	2,277	1,994	1,82	2,018	2,128	2,173
Tagalog Hernandez, Magpisan	2,224	2,264	2,207	2,248	2,159	2,216	2,316	2,154	2,055	2,132
Tagalog Rosales, Kristal Na Tubig	2,102	2,132	2,245	2,096	2,139	2,216	2,133	-	-	-
Tamil (Press)	3,509	3,553	3,436	3,397	3,559	3,734	-	-	-	-
Telugu Trailangaswamy	3,798	4,000	3,784	3,286	3,313	-	-	-	-	-
Telugu Train Journey without safety	3,625	3,495	3,374	3,699	3,681	3,244	-	-	-	-
Vai Mu ja vaa lɔ (T. Sherman)	1,447	1,433	1,466	1,493	1,486	1,478	1,481	1,484	1,500	1,505

Vai Sabu Mua Ko	1,617	1,393	1,486	1,477	-	-	-	-	-	-
Vai Vande bɛ Wu'u	1,467	1,455	1,410	1,500	-	-	-	-	-	-
Welsh T1 Crynodeb Gweithredol	2,144	1,986	1,925	2,014	2,028	1,945	1,943	2,164	-	-
Welsh T2 Ffansi camu i esgidiau rhywun enwog?	1,614	1,698	1,68	1,561	1,582	1,596	1,594	-	-	-

Table 4.2
Mean word length in deploying text by subsequent grouped sentences: all texts

Language, Text alphabetically	n	sentence grouping	a	b	R^2
Akan Mma Nnsua Ade Bɔne	15	5*3	1,8905	-0,0871	0,86
Bamana Bamakɔ sigicogoya	84	7*10 + 14	1,6293	-0,0027	0,006
Bamana Masadennin	207	9*20 + 27	1,6477	-0,0073	0,09
Bamana Namakɔrɔba halakilen	248	9*25 + 23	1,5475	0,0067	0,047
Bamana Sonsannin ani Surukuba	168	7*20 + 28	1,5921	-0,0226	0,376
Bulgarian Ostrovskij, Kak se kalja-vaše stomanata, Chapter 1	83	7*10 + 13	2,2049	-0,0144	0,069
Czech Čulík, O čem jsou dnešní Spojené státy?	114	9*11 + 15	2,3755	-0,005	0,013
Czech Hvížďala, O předem zpackané prezidentské volbě	37	6*5 + 7	2,3161	0,0117	0,038
Czech Macháček, Slovenský dobrý příklad	22	3*5 + 7	2,326	-0,0141	0,015
Czech Spurný, Prekvapení v justici	17	5*3 + 2	2,3521	-0,0249	0,318
Czech Švehla, Editorial, Voličův kalkul	19	3*5 + 4	2,3935	-0,072	0,509
French Dunkerque (Press)	105	4*20 + 25	1,7245	0,0287	0,894
German Assads Familiendiktatur (Press)	119	5*20 + 19	2,0405	0,0146	0,114
German ATT0012 (Press)	83	7*10 + 13	2,1899	-0,0248	0,271
German Die Stadt des Schweigens (Press)	120	12*10	1,8823	0,0229	0,369
German Terror in Ost Timor (Press)	98	9*10 + 8	1,984	-0,0028	0,005
German Unter Hackern (Press)	110	10*11	2,1859	-0,0246	0,207
Hindi After the sanction to love marriage	38	7*5 + 3	1,7506	0,0001	0,00001
Hindi The Anna Team on a cross-road	42	7*5 + 7	1,6063	0,0086	0,093
Hungarian A nominalizmus forradalma **(**Press)	63	5*10 + 13	2,9275	-0,0803	0,88

Hungarian Kunczekolbász **(**Press)	32	2*10 + 12	2,7347	-0,0370	0,23
Indonesian Pengurus PSM terbelah **(**Press)	28	2*10 + 8	2,6377	-0,0235	0,57
Indonesian Sekolah ditutup **(**Press)	15	3*5	2,664	-0,041	0,56
Italian (Press, Online)	92	8*10 + 12	2,0549	0,0316	0,605
Japanese Miki, Jinseiron Note, first 100 sentences	100	10*10	2,0588	0,0077	0,035
Kikongo Bimpa: Ma Ngo ya Ma Nsiese	79	7*10 + 9	2,0599	0,0082	0,048
Kikongo Lumumba speech	43	8*5 + 3	2,0057	0,0019	0,004
Kikongo Nkongo ye Kisi Kongo	29	5*5 +4	1,7691	0,0040	0,011
Latin Cicero, In Catilinam I	80	8*10	2,4390	-0,0174	0,18
Latin Cicero, In Catilinam II	180	10*18	2,3637	0,0072	0,04
Macedonian Ostrovskij, Kako se kaleše čelkiot, Chapter 1	89	8*10 + 9	2,3727	-0,0175	0,081
Malayalam 1, Moralistic Hooligans	32	5*5 + 7	4,2515	-0,1198	0,740
Malayalam 2, No one should die	29	5*5 + 4	4,337	-0,0421	0,1403
Maninka Nko Doumbu Kende no. 2	37	6*5 + 7	1,7004	0,0144	0,055
Maninka Nko Doumbu Kende no. 7	34	6*5 + 4	1,765	-0,0097	0,036
Maninka Sìikán` (Constitution of Guinea, an excerpt)	94	8*10 +14	1,9391	-0,0089	0,096
Maninka Teelen4	29	6*4 + 5	1,7783	-0,0343	0,49
Odia Calculation for the District Council President	28	5*5 + 3	3,0838	-0,0291	0,071
Odia Who is great?	36	6*5 + 6	2,7453	0,0258	0,042
Romanian Paler, Aventuri solitare (excerpt)	17	2*5 + 7	2,16	-0,047	0,07
Romanian Popescu D.R., Vânătoarea regală, Chapter 2	61	1 + 6*10	1,8656	-0,0164	0,16
Romanian Steinhardt, Jurnalul fericirii, Trei soluții	85	8*10 + 5	2,0128	0,0022	0,004
Russian Ostrovskij, Kak zakaljalas stal', Chapter 1	76	7*10 + 6	2,2137	-0,0074	0,014
Serbian Ostrovskij, Kako se kalio čelik, Chapter 1	83	7*10 + 13	2,2094	-0,0268	0,249
Slovak E. Bachletová, Moja Dolná zem	92	8*10 + 12	2,3278	-0,0152	0,12

Slovak E. Bachletová, Riadok v tlačive: nezamestnaný	78	7*10 + 8	2,5223	-0,0222	0,48
Slovenian Ostrovskij, Kako se je kalilo jeklo, Chapter 1	84	7*10 + 14	2,0790	-0,0273	0,41
Sundanese Agustusan (Salaka Online)	53	4*10 + 13	2,0031	0,0527	0,48
Sundanese Aki Satimi (Salaka Online)	147	6*20 + 27	2,0999	0,0117	0,142
Tagalog Hernandez, Limang Alas: Tatlong Santo	104	9*10 + 14	2,1362	-0,0086	0,04
Tagalog Hernandez, Magpisan	111	9*10 + 21	2,274	-0,0139	0,32
Tagalog Rosales, Kristal Na Tubig	139	6*20 + 19	2,1297	0,0055	0,044
Tamil (Press)	32	5*5 + 7	3,4209	0,0315	0,25
Telugu Trailangaswamy	51	4*10 + 11	4,1414	-0,1684	0,7
Telugu Train Journey without safety	61	5*10 + 11	3,6219	-0,0292	0,0893
Vai Mu ja vaa lɔ (T. Sherman)	193	9*20 + 13	1,4427	0,0063	0,7
Vai Sabu Mua Ko	39	3*10 + 9	1,575	-0,0327	0,208
Vai Vande bɛ Wu'u	35	3*10 + 5	1,4445	0,0054	0,035
Welsh T1 Crynodeb Gweithredol	40	8*5	2,0187	-0,00001	0,0000001
Welsh T2 Ffansi camu i esgidiau rhywun enwog?	34	6*5 + 4	1,6696	-0,0129	0,295

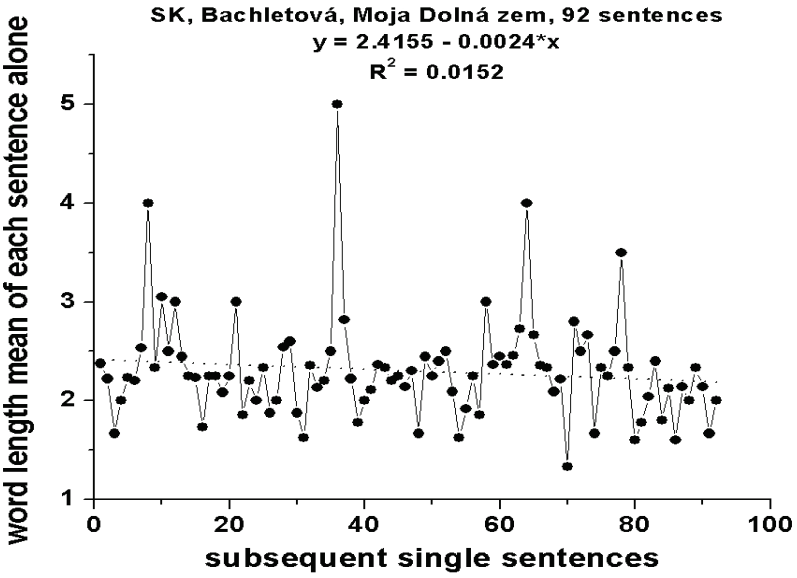


Figure 4.2, Mean word length of single subsequent sentences in *Moja Dolná zem* (Bachletová, Slovak)

Table 4.3
Mean word length in deploying text by subsequent single sentences: all texts

Language, Text alphabetically	**n**	**a**	**b**	**R^2**
Akan Mma Nnsua Ade Bɔne	15	1,9105	-0,0308	0,269
Bamana Bamakɔ sigicogoya	84	1,6354	0,0009	0,0049
Bamana Masadennin	207	1,6034	0,00003	0,00003
Bamana Namakɔrɔba halakilen	248	1,6543	0,0002	0,0008
Bamana Sonsannin ani Surukuba	168	1,6494	-0,0005	0,0021
Bulgarian Ostrovskij, Kak se kaljavaše stomanata, Chapter 1	83	2,2393	-0,0017	0,0095
Czech Čulík, O čem jsou dnešní Spojené státy?	114	2,4454	-0,0016	0,017
Czech Hvížďala, O předem zpackané prezidentské volbě	37	2,3061	0,0022	0,0061
Czech Macháček, Slovenský dobrý příklad	22	2,3762	-0,0113	0,038
Czech Spurný, Prekvapení v justici	17	2,2814	-0,0043	0,0105
Czech Švehla, Editorial, Voličův kalkul	19	2,4003	-0,0189	0,129
French Dunkerque (Press)	105	1,7115	0,0009	0,008
German Assads Familiendiktatur (Press)	119	2,0519	0,0011	0,005
German ATT0012 (Press)	83	2,1897	-0,0033	0,04
German Die Stadt des Schweigens (Press)	120	2,0024	0,0006	0,001
German Terror in Ost Timor (Press)	98	2,0062	-0,0017	0,017
German Unter Hackern (Press)	110	2,2277	-0,0032	0,038
Hungarian A nominalizmus forradalma **(**Press)	63	2,8498	-0,008	0,16
Hungarian Kunczekolbász **(**Press)	32	2,8190	-0,0130	0,09
Indonesian Pengurus PSM terbelah **(**Press)	28	2,7030	-0,0058	0,03
Indonesian Sekolah ditutup **(**Press)	15	2,6414	-0,0010	0,0001
Italian (Press, Online)	92	2,0437	0,0038	0,1708
Japanese Miki, Jinseiron Note, first 100 sentences	100	2,0939	0,0006	0,003

Kikongo Bimpa: Ma Ngo ya Ma Nsiese	79	2,0596	-0,0008	0,0018
Kikongo Lumumba speech	43	2,1644	-0,0041	0,025
Kikongo Nkongo ye Kisi Kongo	29	1,8071	-0,0004	0,0008
Latin Cicero, In Catilinam I	80	2,3208	0,0003	0,0002
Latin Cicero, In Catilinam II	180	2,4055	0,00004	0,00003
Macedonian Ostrovskij, Kako se kaleše čelkiot, Chapter 1	89	2,4005	-0,0023	0,0145
Malayalam 1 (Press)	32	4,1761	-0,0187	0,07
Malayalam 2 (Press)	29	4,5570	-0,0202	0,0835
Maninka Nko Doumbu Kende no. 2	37	1,7203	0,0037	0,013
Maninka Nko Doumbu Kende no. 7	34	1,8257	-0,0039	0,022
Maninka Sìikán` (Constitution of Guinea, an excerpt)	94	1,9605	-0,0008	0,0065
Maninka Teelen4	29	1,7210	-0,0062	0,1526
Odia Calculation for the District Council President	28	3,1279	-0,0102	0,08
Odia Who is great?	36	2,7607	0,0078	0,038
Romanian Paler, Aventuri solitare (excerpt)	17	2,1140	-0,0031	0,005
Romanian Popescu D.R., Vânătoarea regală, Chapter 2	61	1,7864	0,0002	0,0001
Romanian Steinhardt, Jurnalul fericirii, Trei soluții	85	1,9558	0,0017	0,01
Russian Ostrovskij, Kak zakaljalas stal', Chapter 1	76	2,3033	-0,0017	0,004
Serbian Ostrovskij, Kako se kalio čelik, Chapter 1	83	2,2220	-0,0032	0,042
Slovak E. Bachletová, Moja Dolná zem	92	2,4155	-0,0024	0,015
Slovak E. Bachletová, Riadok v tlačive: nezamestnaný	78	2,5138	-0,0042	0,04
Slovenian Ostrovskij, Kako se je kalilo jeklo, Chapter 1	84	2,0996	-0,0033	0,05
Sundanese Agustusan (Salaka Online)	53	2,0697	0,0038	0,037
Sundanese Aki Satimi (Salaka Online)	147	2,1556	0,0007	0,005
Tagalog Hernandez, Limang Alas: Tatlong Santo	104	2,1218	-0,0002	0,0005
Tagalog Hernandez, Magpisan	111	2,2212	-0,0002	0,0004
Tagalog Rosales, Kristal Na Tubig	139	2,1674	0,0002	0,0006
Tamil (Press)	32	3,4603	0,0060	0,0147
Telugu Trailangaswamy	51	4,2411	-0,0184	0,1317

Telugu Train Journey without safety	61	3,6084	-0,002	0,005
Vai Mu ja vaa lɔ (T. Sherman)	193	1,4379	0,0004	0,0140
Vai Sabu Mua Ko	39	1,5679	-0,0045	0,0700
Vai Vande bɛ Wu'u	35	1,4627	-0,0007	0,0014
Welsh T1 Crynodeb Gweithredol	40	2,0516	-0,0004	0,0003
Welsh T2 Ffansi camu i esgidiau rhywun enwog?	34	1,6567	-0,0027	0,0427

For instance Kelih (2012) found a significant interrelation between text length and word length in Russian and Bulgarian texts: the longer the text, the longer the word length. This interrelation has been modelled by a (simple) power function. However, Kelih (2012) used a slightly different approach: firstly, single chapters of a novel have been cumulated stepwise, secondly word length has been measured stepwise in the cumulated chapters, thirdly, the strong and law-like interrelation has been obtained in word-form types only, and fourthly, the length of the used texts was over 100.000 tokens and over 18.000 types respectively. In the case of using word-form tokens, on the contrary, a decrase of word length with growing text length has be obtained (cf. Kelih 2012: 76f.), but this increase is less systematic than the one obtained on the types level. Thus, generally spoken, the analysis of Kelih (2012) is based on rather different "boundary conditions" than the analysis performed above.

The common feature of these sequences is a great oscillation at the beginning and a decrease at the end. Though in the Romanian text by Paler and in some Slavic languages the end slightly increases, the general trend is decreasing. Hence, whatever the way of measurement, the hypothesis that word length increases with increasing text is either falsified or we still neglect some boundary conditions which are not yet known. The sequences cannot be captured by a single function. The texts have their individualities which testify either to non-spontaneous creation or to changes made post-hoc.

As an example consider the course of mean word length in Akan**,** the beginning of the text *Mma Nnsua Ade Bɔne*, as presented in Table 4.4. Here it is sufficient to discard some of the first values and begin with a rather smooth part.

The course of values in individual texts, as shown in Figure 4.3, displays in some cases analogous behaviour (e.g. Vai texts) which does not bring about difficulties for a mathematician, but a linguist stays in front of the door to Dante's hell, reads the famous inscription and if he has courage, he mutters the famous rebellious words of Galilei and continues investigating this infinite domain.

Table 4.4
Cumulative mean word length in Akan

Position in text	Word length	Cumulative mean word length	
	2		
1	1	1.500	discarded
2	3	2.000	discarded
3	3	2.250	
4	1	2.000	
5	1	1.833	
6	1	1.714	
7	1	1.625	
8	4	1.889	
9	1	1.800	
10	3	1.909	
…	…	…	

Since the first word in text may be characteristic of text or language but it does not represent a mean, we propose to omit the first 9 words and begin to count taking the mean of the first 10 values as $x = 1$, then 1-11 as $x = 2$, etc. In this way the sequence changes and displays a more smooth course. If we use this method, i.e. if we smooth the beginning of the sequence, we obtain much more acceptable results. The course of means beginning with a concave course can be captured by the function

$$y = p_1 x^{p_2} (p_3 - \exp(-p_4 x)),$$

where p_i are the parameters. The first part of the function (the power part) represents the decrease of the mean word length, the exponential part represents the increasing beginning of the empirical sequence. The oscillation is not as strong as with the original measurement. However, sometimes even this function cannot capture the course, so that one must choose among three different courses expreesed by the curves

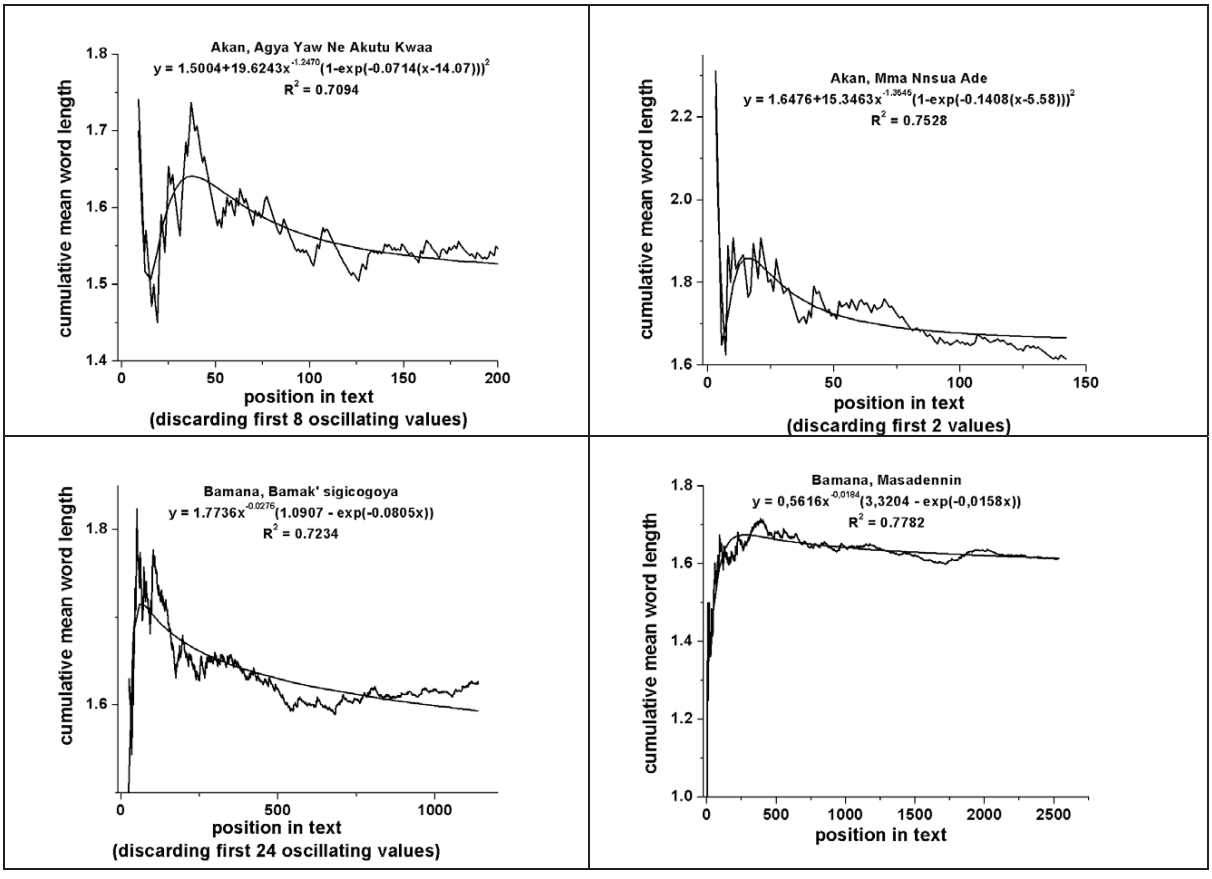
$$y = p_1 x^{-p_2}$$
$$y = p_1 + p_2 (1 - \exp(-p_3 (x - p_4)))^2 .$$

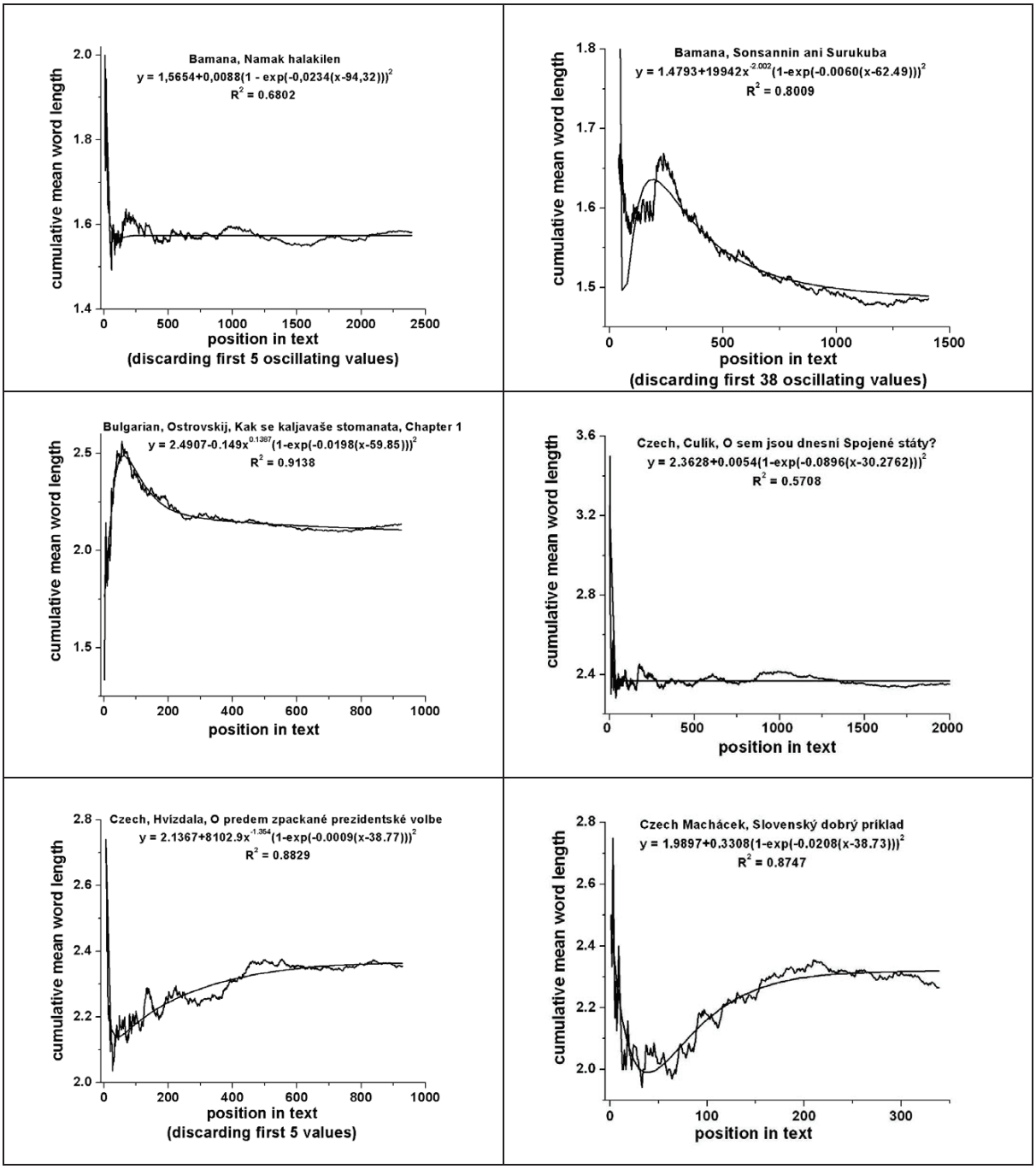
In some cases the first 10 (or more) values must simply be discarded - as has been made in Maninka (Doumbu Kende no 2), Slovak (Moja Dolná zem), or one must perform a further combination of functions, namely to modify the Morse function ad hoc by the power function to obtain

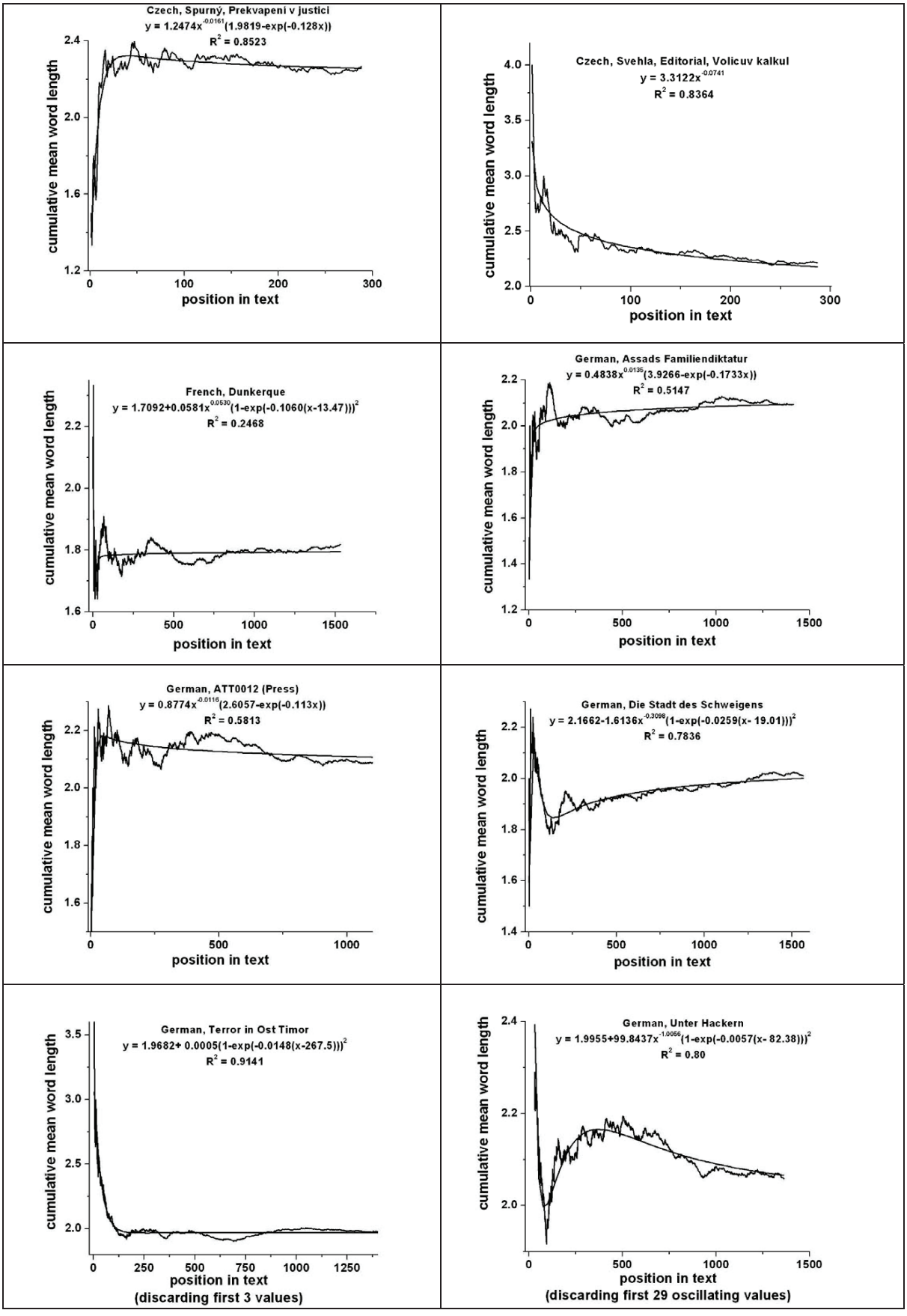
$$y = p_1 + p_2 x^{-p_3}(1 - \exp(-p_4(x - p_5)))^2 .$$

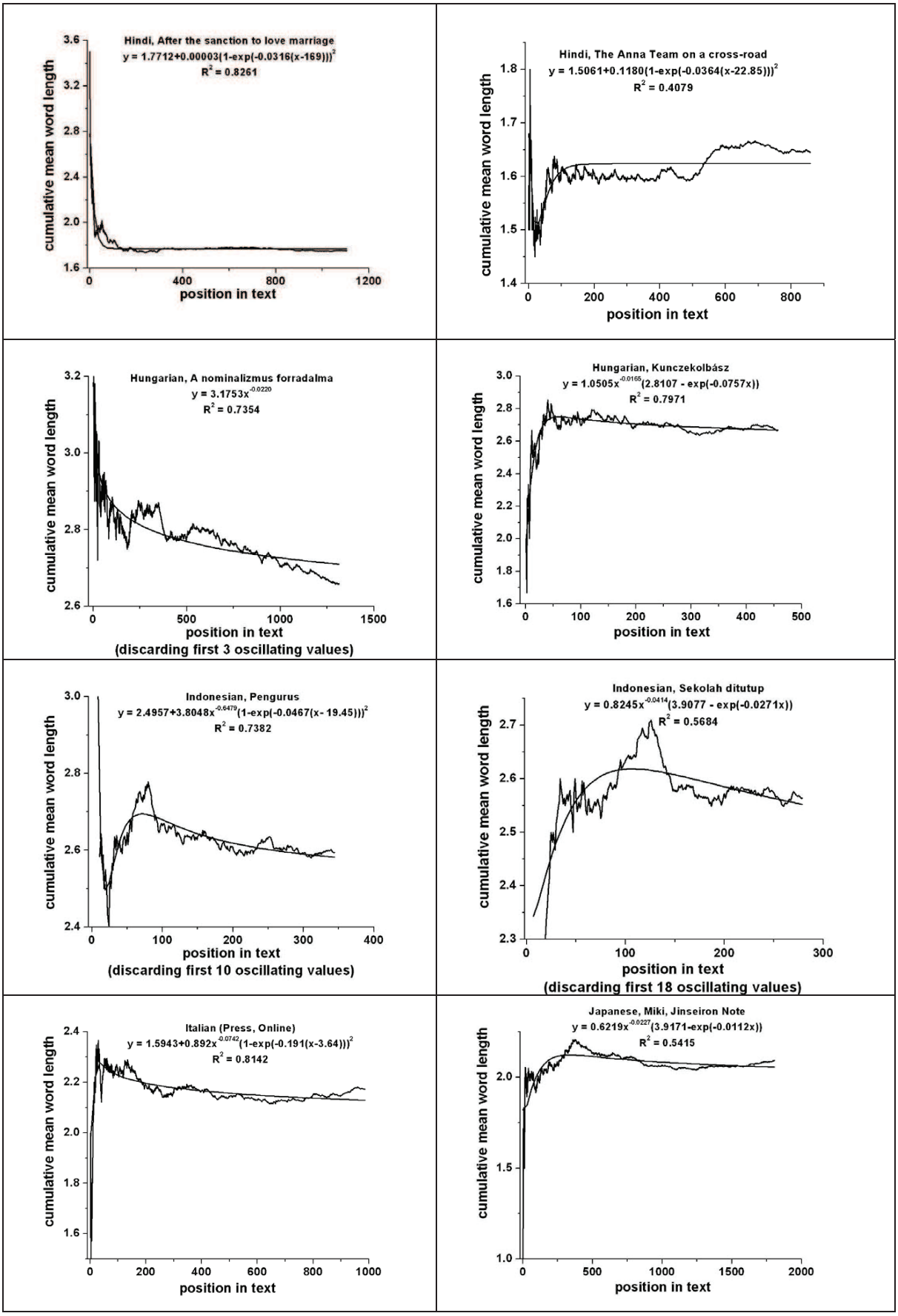
which can, however, roughly fit the general course of the stepwise mean word length of Bachletová's *Moja Dolná zem*. As indicated under the Ox axis title of the plots given in Figure 4.3, the non relevant initial oscillations were circumvented either by counting from the mean of the first 10 values (as in Vai, *Vande*; Maninka, *Sìikán`*; Maninka, *Teleen 4*) or simply discarding them (as in Maninka, *Nko Doumbu Kende no. 2*; Bachletová, *Moja Dolná zem*). Notice that not always these procedures are necessary.

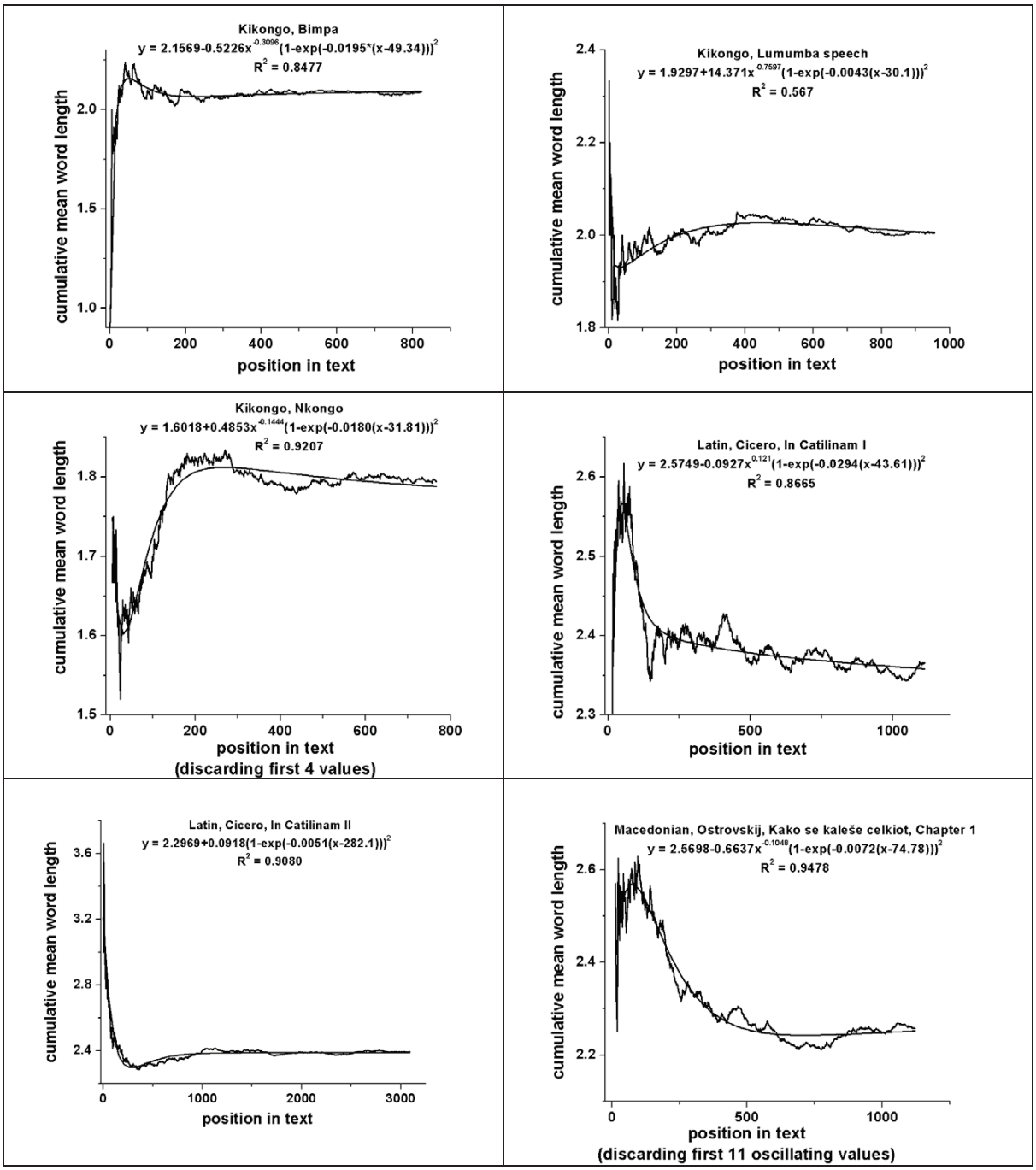
The individual functions are presented in Table 4.5 and Figure 4.4.

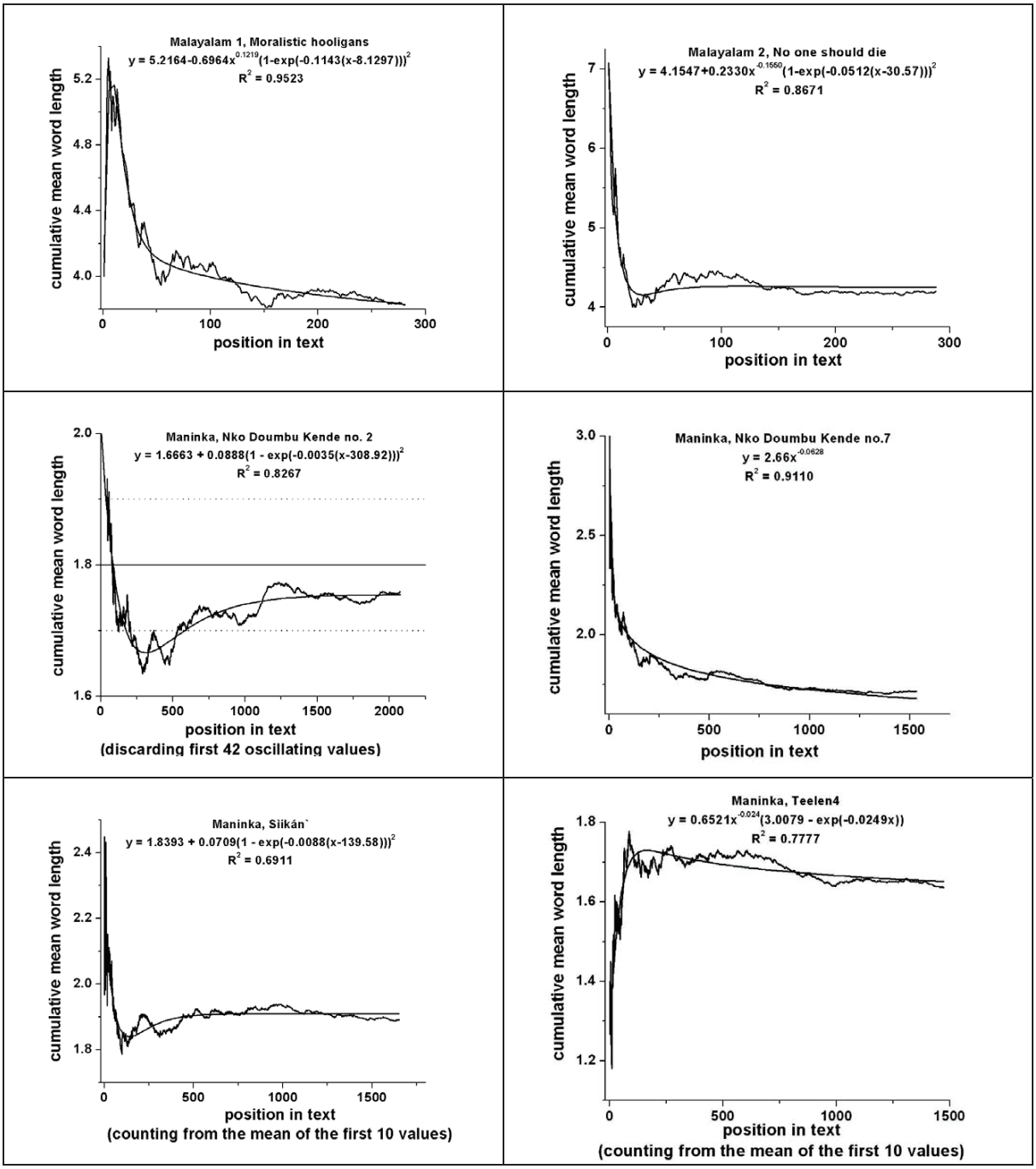


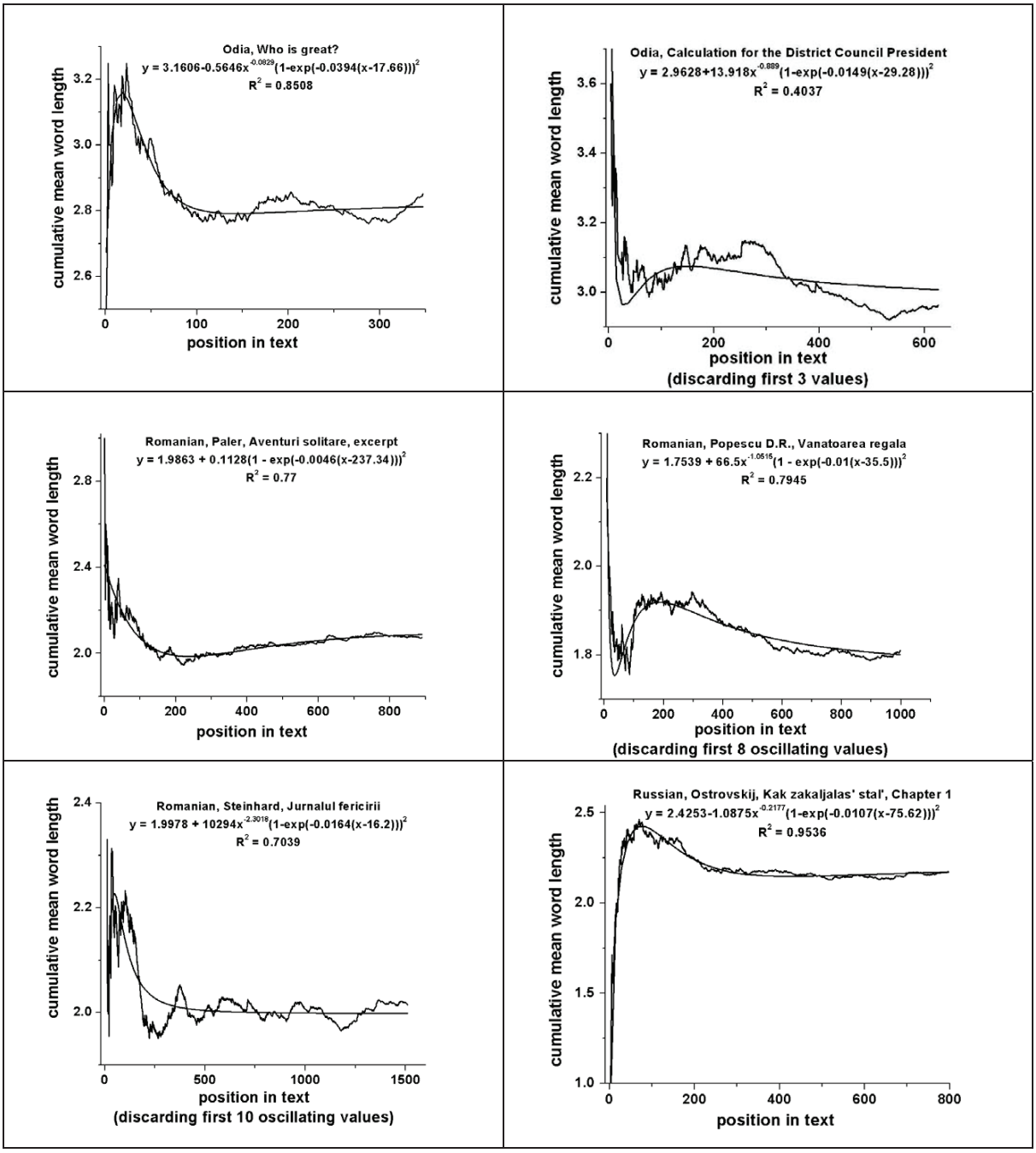

Bamana, Namak halakilen
y = 1,5654+0,0088(1 - exp(-0,0234(x-94,32)))^2
R^2 = 0.6802
cumulative mean word length
position in text
(discarding first 5 oscillating values)
Bamana, Sonsannin ani Surukuba
y = 1.4793+19942x^-2.002(1-exp(-0.0060(x-62.49)))^2
R^2 = 0.8009
cumulative mean word length
position in text
(discarding first 38 oscillating values)
Bulgarian, Ostrovskij, Kak se kaljavaše stomanata, Chapter 1
y = 2.4907-0.149x^0.1387(1-exp(-0.0198(x-59.85)))^2
R^2 = 0.9138
cumulative mean word length
position in text
Czech, Culík, O sem jsou dnesní Spojené státy?
y = 2.3628+0.0054(1-exp(-0.0896(x-30.2762)))^2
R^2 = 0.5708
cumulative mean word length
position in text
Czech, Hvízdala, O predem zpackané prezidentské volbe
y = 2.1367+8102.9x^-1.354(1-exp(-0.0009(x-38.77)))^2
R^2 = 0.8829
cumulative mean word length
position in text
(discarding first 5 values)
Czech Machácek, Slovenský dobrý príklad
y = 1.9897+0.3308(1-exp(-0.0208(x-38.73)))^2
R^2 = 0.8747
cumulative mean word length
position in text

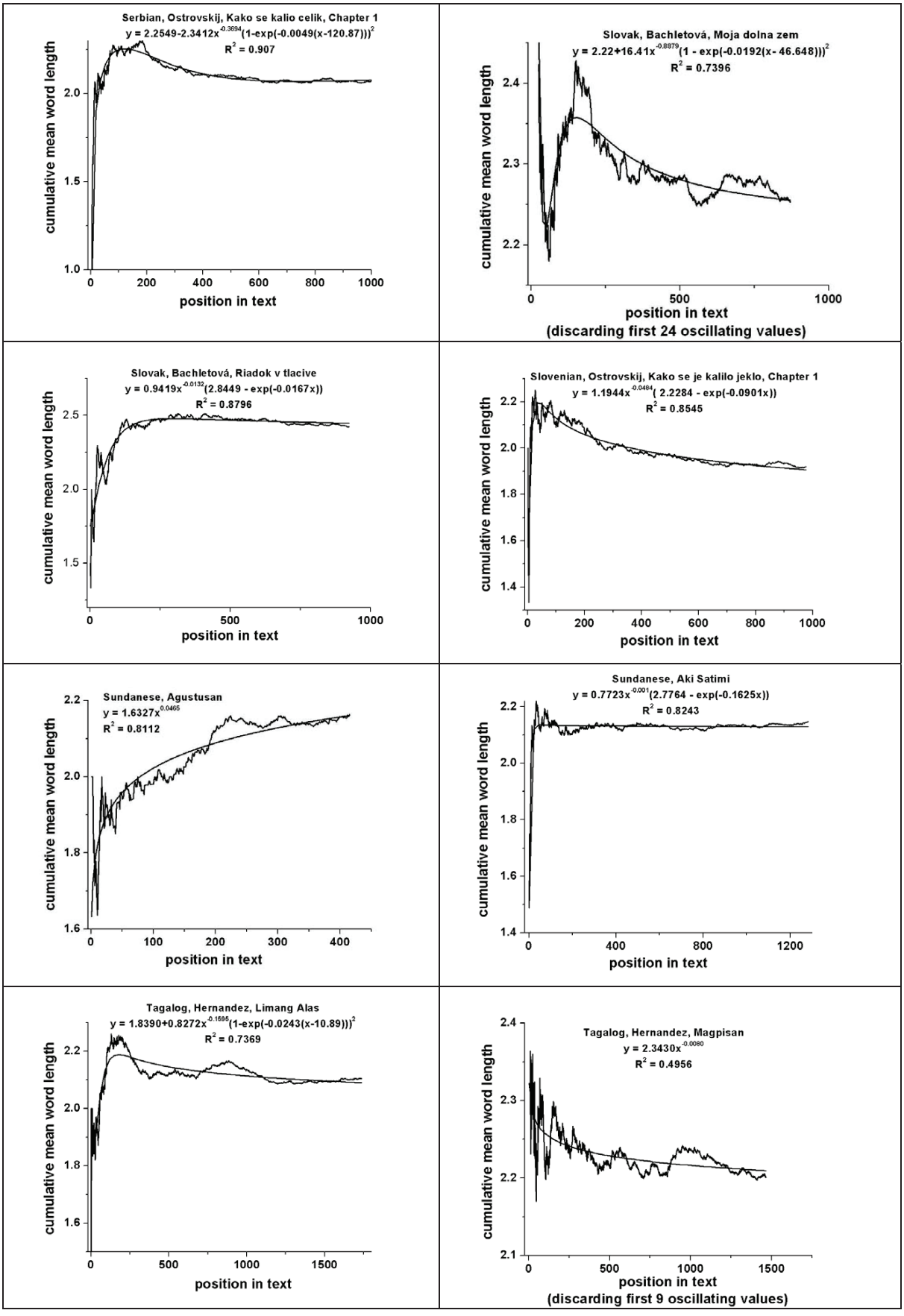

Czech, Spurný, Prekvapení v justici
y = 1.2474x^-0.0161(1.9819-exp(-0.128x))
R^2 = 0.8523
cumulative mean word length
position in text
Czech, Svehla, Editorial, Volicuv kalkul
y = 3.3122x^-0.0741
R^2 = 0.8364
cumulative mean word length
position in text
French, Dunkerque
y = 1.7092+0.0581x^0.0530(1-exp(-0.1060(x-13.47)))^2
R^2 = 0.2468
cumulative mean word length
position in text
German, Assads Familiendiktatur
y = 0.4838x^0.0135(3.9266-exp(-0.1733x))
R^2 = 0.5147
cumulative mean word length
position in text
German, ATT0012 (Press)
y = 0.8774x^-0.0116(2.6057-exp(-0.113x))
R^2 = 0.5813
cumulative mean word length
position in text
German, Die Stadt des Schweigens
y = 2.1662-1.6136x^-0.3098(1-exp(-0.0259(x- 19.01)))^2
R^2 = 0.7836
cumulative mean word length
position in text
German, Terror in Ost Timor
y = 1.9682+ 0.0005(1-exp(-0.0148(x-267.5)))^2
R^2 = 0.9141
cumulative mean word length
position in text
(discarding first 3 values)
German, Unter Hackern
y = 1.9955+99.8437x^-1.0056(1-exp(-0.0057(x- 82.38)))^2
R^2 = 0.80
cumulative mean word length
position in text
(discarding first 29 oscillating values)

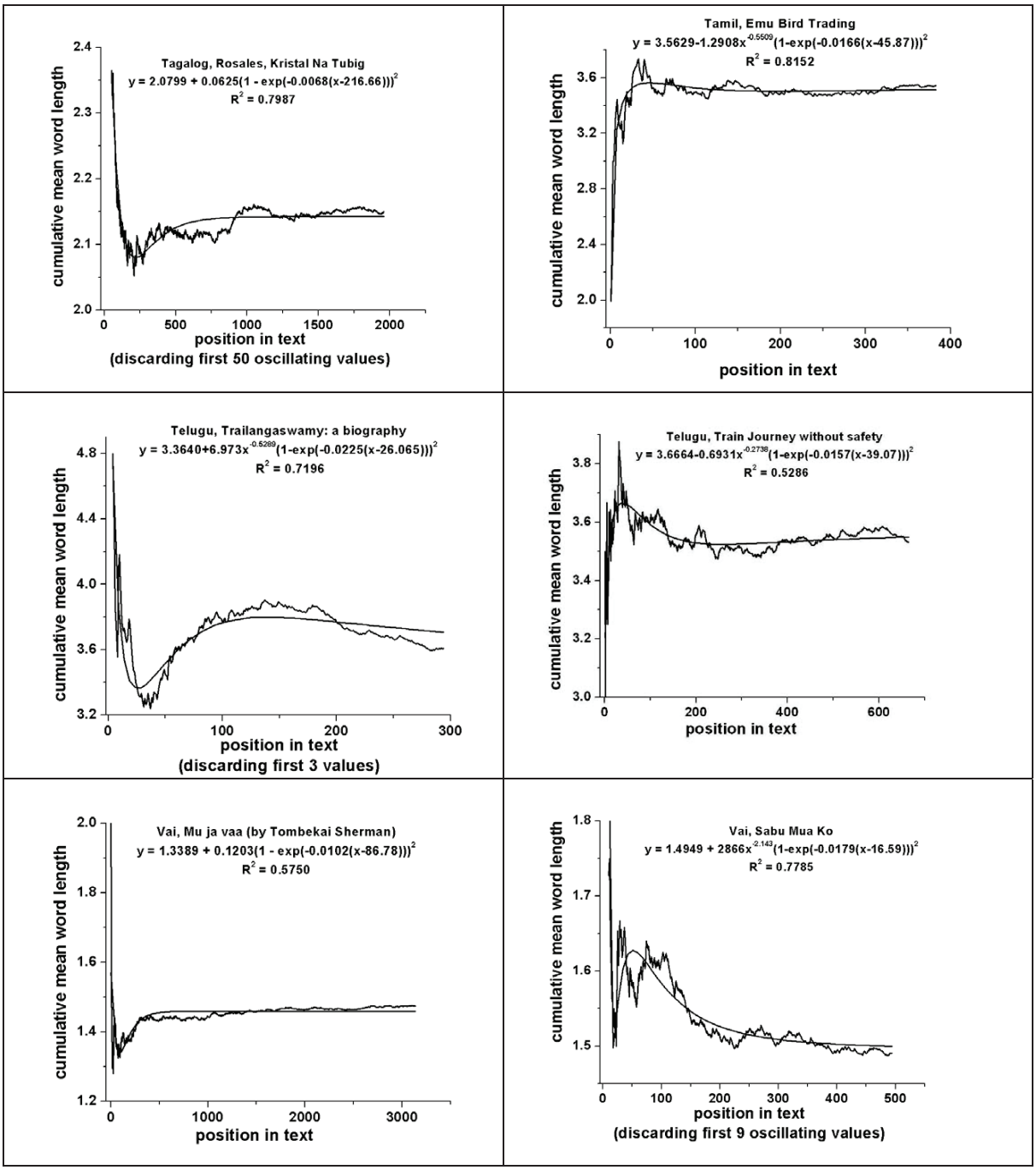

Hindi, After the sanction to love marriage
y = 1.7712+0.00003(1-exp(-0.0316(x-169)))^2
R^2 = 0.8261
cumulative mean word length
position in text
Hindi, The Anna Team on a cross-road
y = 1.5061+0.1180(1-exp(-0.0364(x-22.85)))^2
R^2 = 0.4079
cumulative mean word length
position in text
Hungarian, A nominalizmus forradalma
y = 3.1753x^-0.0220
R^2 = 0.7354
cumulative mean word length
position in text
(discarding first 3 oscillating values)
Hungarian, Kunczekolbász
y = 1.0505x^-0.0165(2.8107 - exp(-0.0757x))
R^2 = 0.7971
cumulative mean word length
position in text
Indonesian, Pengurus
y = 2.4957+3.8048x^-0.6479(1-exp(-0.0467(x- 19.45)))^2
R^2 = 0.7382
cumulative mean word length
position in text
(discarding first 10 oscillating values)
Indonesian, Sekolah ditutup
y = 0.8245x^-0.0414(3.9077 - exp(-0.0271x))
R^2 = 0.5684
cumulative mean word length
position in text
(discarding first 18 oscillating values)
Italian (Press, Online)
y = 1.5943+0.892x^-0.0742(1-exp(-0.191(x-3.64)))^2
R^2 = 0.8142
cumulative mean word length
position in text
Japanese, Miki, Jinseiron Note
y = 0.6219x^-0.0227(3.9171-exp(-0.0112x))
R^2 = 0.5415
cumulative mean word length
position in text


Kikongo, Bimpa
y = 2.1569-0.5226x^-0.3096(1-exp(-0.0195*(x-49.34)))^2
R^2 = 0.8477
cumulative mean word length
position in text
Kikongo, Lumumba speech
y = 1.9297+14.371x^-0.7597(1-exp(-0.0043(x-30.1)))^2
R^2 = 0.567
cumulative mean word length
position in text
Kikongo, Nkongo
y = 1.6018+0.4853x^-0.1444(1-exp(-0.0180(x-31.81)))^2
R^2 = 0.9207
cumulative mean word length
position in text
(discarding first 4 values)
Latin, Cicero, In Catilinam I
y = 2.5749-0.0927x^0.121(1-exp(-0.0294(x-43.61)))^2
R^2 = 0.8665
cumulative mean word length
position in text
Latin, Cicero, In Catilinam II
y = 2.2969+0.0918(1-exp(-0.0051(x-282.1)))^2
R^2 = 0.9080
cumulative mean word length
position in text
Macedonian, Ostrovskij, Kako se kaleše celkiot, Chapter 1
y = 2.5698-0.6637x^-0.1048(1-exp(-0.0072(x-74.78)))^2
R^2 = 0.9478
cumulative mean word length
position in text
(discarding first 11 oscillating values)


Malayalam 1, Moralistic hooligans
y = 5.2164-0.6964x^0.1219(1-exp(-0.1143(x-8.1297)))^2
R^2 = 0.9523
cumulative mean word length
position in text
Malayalam 2, No one should die
y = 4.1547+0.2330x^-0.1550(1-exp(-0.0512(x-30.57)))^2
R^2 = 0.8671
cumulative mean word length
position in text
Maninka, Nko Doumbu Kende no. 2
y = 1.6663 + 0.0888(1 - exp(-0.0035(x-308.92)))^2
R^2 = 0.8267
cumulative mean word length
position in text
(discarding first 42 oscillating values)
Maninka, Nko Doumbu Kende no.7
y = 2.66x^-0.0628
R^2 = 0.9110
cumulative mean word length
position in text
Maninka, Siikán`
y = 1.8393 + 0.0709(1 - exp(-0.0088(x-139.58)))^2
R^2 = 0.6911
cumulative mean word length
position in text
(counting from the mean of the first 10 values)
Maninka, Teelen4
y = 0.6521x^-0.024(3.0079 - exp(-0.0249x))
R^2 = 0.7777
cumulative mean word length
position in text
(counting from the mean of the first 10 values)


Odia, Who is great?
y = 3.1606-0.5646x^-0.0829(1-exp(-0.0394(x-17.66)))^2
R^2 = 0.8508
cumulative mean word length
position in text
Odia, Calculation for the District Council President
y = 2.9628+13.918x^-0.889(1-exp(-0.0149(x-29.28)))^2
R^2 = 0.4037
cumulative mean word length
position in text
(discarding first 3 values)
Romanian, Paler, Aventuri solitare, excerpt
y = 1.9863 + 0.1128(1 - exp(-0.0046(x-237.34)))^2
R^2 = 0.77
cumulative mean word length
position in text
Romanian, Popescu D.R., Vanatoarea regala
y = 1.7539 + 66.5x^-1.0518(1 - exp(-0.01(x-35.5)))^2
R^2 = 0.7945
cumulative mean word length
position in text
(discarding first 8 oscillating values)
Romanian, Steinhard, Jurnalul fericirii
y = 1.9978 + 10294x^-2.3018(1-exp(-0.0164(x-16.2)))^2
R^2 = 0.7039
cumulative mean word length
position in text
(discarding first 10 oscillating values)
Russian, Ostrovskij, Kak zakaljalas' stal', Chapter 1
y = 2.4253-1.0875x^-0.2177(1-exp(-0.0107(x-75.62)))^2
R^2 = 0.9536
cumulative mean word length
position in text


Serbian, Ostrovskij, Kako se kalio celik, Chapter 1
y = 2.2549-2.3412x^-0.3694(1-exp(-0.0049(x-120.87)))^2
R^2 = 0.907
cumulative mean word length
position in text
Slovak, Bachletová, Moja dolna zem
y = 2.22+16.41x^-0.8879(1 - exp(-0.0192(x- 46.648)))^2
R^2 = 0.7396
cumulative mean word length
position in text
(discarding first 24 oscillating values)
Slovak, Bachletová, Riadok v tlacive
y = 0.9419x^-0.0132(2.8449 - exp(-0.0167x))
R^2 = 0.8796
cumulative mean word length
position in text
Slovenian, Ostrovskij, Kako se je kalilo jeklo, Chapter 1
y = 1.1944x^-0.0464(2.2284 - exp(-0.0901x))
R^2 = 0.8545
cumulative mean word length
position in text
Sundanese, Agustusan
y = 1.6327x^0.0465
R^2 = 0.8112
cumulative mean word length
position in text
Sundanese, Aki Satimi
y = 0.7723x^-0.001(2.7764 - exp(-0.1625x))
R^2 = 0.8243
cumulative mean word length
position in text
Tagalog, Hernandez, Limang Alas
y = 1.8390+0.8272x^-0.1595(1-exp(-0.0243(x-10.89)))^2
R^2 = 0.7369
cumulative mean word length
position in text
Tagalog, Hernandez, Magpisan
y = 2.3430x^-0.0080
R^2 = 0.4956
cumulative mean word length
position in text
(discarding first 9 oscillating values)


Tagalog, Rosales, Kristal Na Tubig
y = 2.0799 + 0.0625(1 - exp(-0.0068(x-216.66)))^2
R^2 = 0.7987
cumulative mean word length
position in text
(discarding first 50 oscillating values)
Tamil, Emu Bird Trading
y = 3.5629-1.2908x^-0.5509(1-exp(-0.0166(x-45.87)))^2
R^2 = 0.8152
cumulative mean word length
position in text
Telugu, Trailangaswamy: a biography
y = 3.3640+6.973x^-0.5269(1-exp(-0.0225(x-26.065)))^2
R^2 = 0.7196
cumulative mean word length
position in text
(discarding first 3 values)
Telugu, Train Journey without safety
y = 3.6664-0.6931x^-0.2738(1-exp(-0.0157(x-39.07)))^2
R^2 = 0.5286
cumulative mean word length
position in text
Vai, Mu ja vaa (by Tombekai Sherman)
y = 1.3389 + 0.1203(1 - exp(-0.0102(x-86.78)))^2
R^2 = 0.5750
cumulative mean word length
position in text
Vai, Sabu Mua Ko
y = 1.4949 + 2866x^2.143(1-exp(-0.0179(x-16.59)))^2
R^2 = 0.7785
cumulative mean word length
position in text
(discarding first 9 oscillating values)

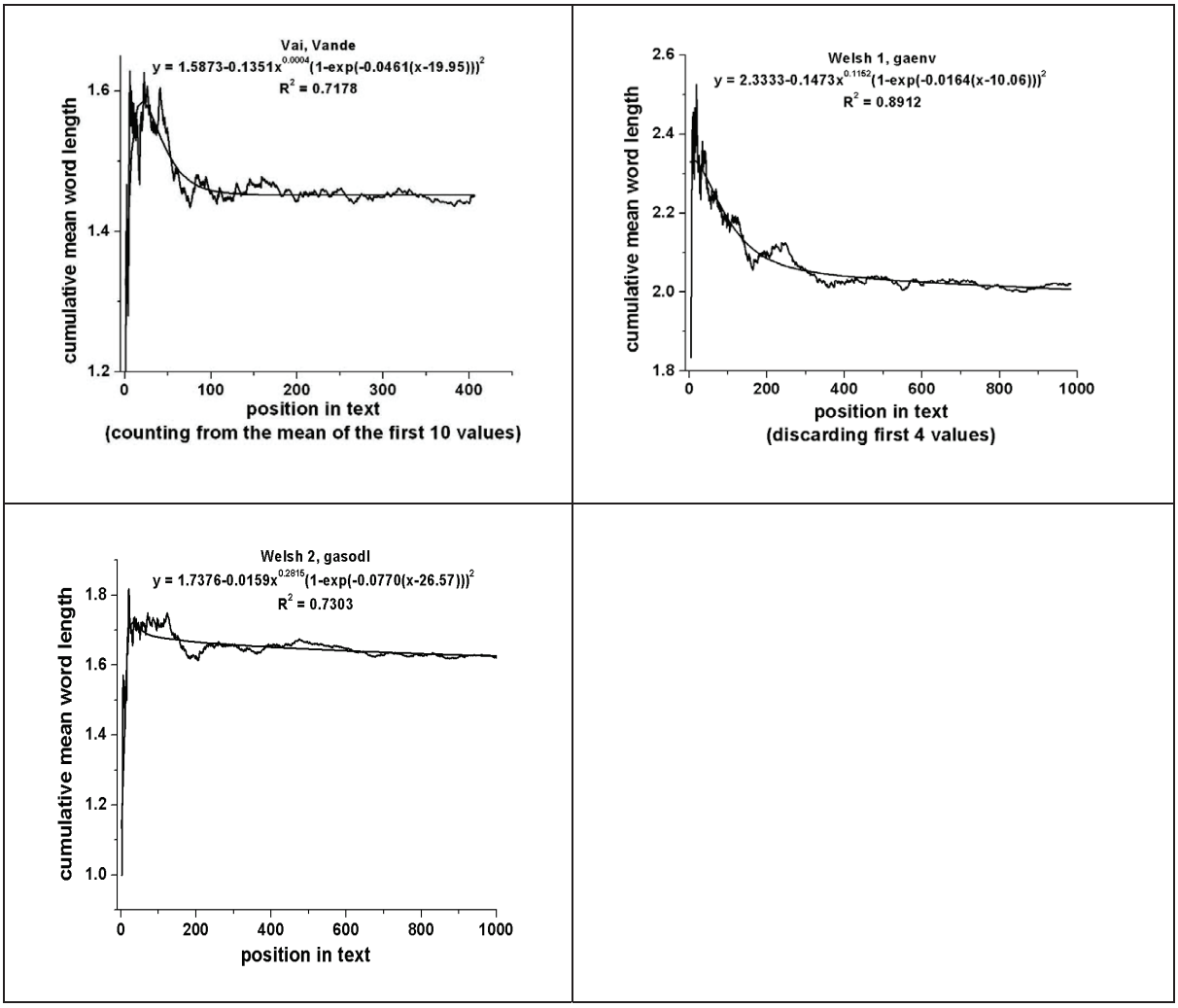


Figure 4.4. Course of mean word lengths in 61 texts of 28 languages fitted by function (5)

This is a purely explorative approach not having preliminarily a sound linguistic substantiation. Nevertheless, the function can be considered a family. The most general function is (J. Mačutek, private communication)

$$y = p_1 + p_2 x^{p_3}\left(p_4 - e^{p_5(x-p_6)}\right)^{p_7}, \tag{5}$$

resulting from the differential equation[1]

[1] Denoting the "shifted" mean word length by $Y = y - p_1$ the following differential equation (6) is of the type Y'/Y = difference between two functions of the position x in text. In other words, we get again the "fight" between two contrary actions along the text, one increasing the word length (autosemantics) and the other one decreasing the word length (synsemantics, auxiliaries).

$$(6) \qquad \frac{y'}{y-p_1} = \frac{p_3}{x} - \frac{p_5 p_7 e^{p_5 |x-p_6}}{p_4 - e^{p_5 (x-p_6)}}$$

where the parameter p_1 on the left hand side shows that the values of y must be greater than a certain constant; usually it is 1. Parameter p_3 is, as usually, the language constant; the functions in the last expression are somehat complex and represent non-linear relations between the force of hearer (numerator) and the force of community (selfregulation in denominator). This is a generalization of the Wimmer-Altmann (2005) model. From formula (5) we obtain all functions used in Table 11 by replacing some parameters by constant values. Hence

$$(7) \qquad y = p_1 + p_2 x^{p_3} \left(1 - e^{p_5 (x-p_6)}\right)^2$$

results if $p_4 = 1$, $p_7 = 2$;

$$(8) \qquad y = p_1 + p_2 \left(1 - e^{p_5 (x-p_6)}\right)^2$$

results if $p_3 = 0$, $p_4 = 1$, $p_7 = 2$;

$$(8) \qquad y = p_2 x^{p_3} \left(p_4 - e^{p_5 x}\right)$$

results if $p_1 = 0$, $p_6 = 0$, $p_7 = 1$;. and

$$(10) \qquad y = p_2 x^{p_3}$$

results if e.g. $p_1 = 0$ and $p_4 - e^{p_5 (x-p_6)} = 1$ or e.g., $p_4 = 2$, $p_5 = 0$, and many other parameter values. This case shows us that some phenomena behave quite differently than expected by the previous theory. Both its extension and inclusion in Köhler's control cycle would be very useful.

Table 4.6
The course of mean word lengths in texts in 61 texts of 28 languages fitted by function (5) (with smoothed or discarded beginning as indicated in the title of the Ox axis of plots given in Figure 5)

Language, Text alphabetically	**Function**	R^2
Akan Agya Yaw Ne Akutu Kwaa	$1,5004+19,6243x^{-1,2470}(1-\exp(-0,0714(x-14,07)))^2$	0,71
Akan Mma Nnsua Ade Bɔne	$1,6476+15,3463x^{-1,3545}(1-\exp(-0,1408(x-5,58)))^2$	0,75
Bamana Bamakɔ sigicogoya	$1,7736x^{-0,0276}(1,0907 - \exp(-0,0805x))$	0,72
Bamana Masadennin	$0,5616x^{-0,0184}(3,3204 - \exp(-0,0158x))$	0,78
Bamana Namakɔrɔba halakilen	$1,5654+0,0088(1 - \exp(-0,0234(x-94,32)))^2$	0,68
Bamana Sonsannin ani Surukuba	$1,4793+19942x^{-2,002}(1-\exp(-0,0060(x-62,49)))^2$	0,80
Bulgarian Ostrovskij, Kak se kaljavaše stomanata, Chapter 1	$2,4907-0,149x^{0,1387}(1-\exp(-0,0198(x-59,85)))^2$	0,91
Czech Čulík, O čem jsou dnešní Spojené státy?	$2,3628+0,0054(1-\exp(-0,0896(x-30,2762)))^2$	0,57
Czech Hvížďala, O předem zpackané prezidentské volbě	$2,1367+8102,9x^{-1,354}(1-\exp(-0,0009(x-38,77)))^2$	0,88
Czech Macháček, Slovenský dobrý příklad	$1,9897+0,3308(1-\exp(-0,0208(x-38,73)))^2$	0,87
Czech Spurný, Prekvapení v justici	$1,2474x^{-0,0161}(1,9819-\exp(-0,128x))$	0,85
Czech Švehla, Editorial, Voličův kalkul	$3,3122x^{-0,0741}$	0,84
French Dunkerque (Press)	$1,7092+0,0581x^{0,0530}(1-\exp(-0,1060(x-13,5)))^2$	0,25
German Assads Familiendiktatur (Press)	$0,4838x^{0,0135}(3,9266-\exp(-0,1733x))$	0,51
German ATT0012 (Press)	$0,8774x^{-0,0116}(2,6057-\exp(-0,113x))$	0,58
German Die Stadt des Schweigens (Press)	$2,1662-1,6136x^{-0,3098}(1-\exp(-0,0259(x- 19,0054)))^2$	0,78
German Terror in Ost Timor (Press)	$1,9682+ 0,0005(1-\exp(-0,0148(x-267,5)))^2$	0,91
German Unter Hackern (Press)	$1,9955+99,8437x^{-1,0056}(1-\exp(-0,0057(x- 82,38)))^2$	0,80
Hindi After the sanction to love marriage	$1,7712+0,00003(1-\exp(-0,0316(x-169)))^2$	0,83
Hindi The Anna Team on a cross-road	$1,5061+0,1180(1-\exp(-0,0364(x-22,85)))^2$	0,41

Hungarian A nominalizmus forradalma **(**Press)	$3{,}1753x^{-0{,}0220}$	0,74
Hungarian Kunczekolbász **(**Press)	$1{,}0505x^{-0{,}0165}(2{,}8107 - \exp(-0{,}0757x))$	0,80
Indonesian Pengurus PSM terbelah **(**Press)	$2{,}4957+3{,}8048x^{-0{,}6479}(1-\exp(-0{,}0467(x- 19{,}45)))^2$	0,74
Indonesian Sekolah ditutup **(**Press)	$0{,}8245x^{-0{,}0414}(3{,}9077 - \exp(-0{,}0271x))$	0,57
Italian (Press, Online)	$1{,}5943+0{,}892x^{-0{,}0742}(1-\exp(-0{,}191(x-3{,}64)))^2$	0,81
Japanese Miki, Jinseiron Note, first 100 sentences	$0{,}6219x^{-0{,}0227}(3{,}9171-\exp(-0{,}0112x))$	0,54
Kikongo Bimpa: Ma Ngo ya Ma Nsiese	$2{,}1569-0{,}5226x^{-0{,}3096}(1-\exp(-0{,}0195(x-49{,}34)))^2$	0,85
Kikongo Lumumba speech	$1{,}9297+14{,}371x^{-0{,}7597}(1-\exp(-0{,}0043(x-30{,}1)))2$	0,57
Kikongo Nkongo ye Kisi Kongo	$1{,}6018+0{,}4853x^{-0{,}1444}(1-\exp(-0{,}0180(x-31{,}81)))^2$	0,92
Latin Cicero, In Catilinam I	$2{,}5749-0{,}0927x^{0{,}121}(1-\exp(-0{,}0294(x-43{,}61)))^2$	0,87
Latin Cicero, In Catilinam II	$2{,}2969+0{,}0918(1-\exp(-0{,}0051(x-282{,}1)))^2$	0,91
Macedonian Ostrovskij, Kako se kaleše čelkiot, Chapter 1	$2{,}5698-0{,}6637x^{-0{,}1048}(1-\exp(-0{,}0072(x-74{,}78)))^2$	0,95
Malayalam 1, Moralistic Hooligans	$5{,}2164-0{,}6964x^{0{,}1219}(1-\exp(-0{,}1143(x-8{,}1297)))^2$	0,95
Malayalam 2, No one should die	$4{,}1547+0{,}2330x^{-0{,}1550}(1-\exp(-0{,}0512(x-30{,}57)))^2$	0,87
Maninka Nko Doumbu Kende no. 2	$1{,}6663 + 0{,}0888(1 - \exp(-0{,}0035(x-308{,}92)))^2$	0,83
Maninka Nko Doumbu Kende no. 7	$2{,}66x^{-0{,}0628}$	0,91
Maninka Siikán` (Constitution of Guinea, an excerpt)	$1{,}8393 + 0{,}0709(1 - \exp(-0{,}0088(x-139{,}58)))^2$	0,69
Maninka Teelen4	$0{,}6521x^{-0{,}024}(3{,}0079 - \exp(-0{,}0249x))$	0,78
Odia Calculation for the District Council President	$2{,}9628+13{,}918x^{-0{,}889}(1-\exp(-0{,}0149(x-29{,}28)))^2$	0,40
Odia Who is great?	$3{,}1606-0{,}5646x^{-0{,}0829}(1-\exp(-0{,}0394(x-17{,}66)))^2$	0,85
Romanian Paler, Aventuri solitare (excerpt)	$1{,}9863 + 0{,}1128(1 - \exp(-0{,}0046(x-237{,}34)))^2$	0,77
Romanian Popescu D.R., Vânătoarea regală, Chapter 2	$1{,}7539 + 66{,}5x^{-1{,}0515}(1 - \exp(-0{,}01(x-35{,}5)))^2$	0,80
Romanian Steinhardt, Jurnalul fericirii, Trei soluții	$1{,}9978 + 10294x^{-2{,}3018}(1-\exp(-0{,}0164(x-16{,}2)))^2$	0,70
Russian Ostrovskij, Kak zakaljalas stal', Chapter 1	$2{,}4253-1{,}0875x^{-0{,}2177}(1-\exp(-0{,}0107(x-75{,}62)))2$	0,95
Serbian Ostrovskij, Kako se kalio čelik, Chapter 1	$2{,}2549-2{,}3412x^{-0{,}3694}(1-\exp(-0{,}0049(x-120{,}87)))^2$	0,91

Slovak Bachletová, Moja Dolná zem	$2,22+16,40x^{-0,8879}(1 - exp(-0,0192(x- 46,648)))^2$	0,74
Slovak Bachletová, Riadok v tlačive: nezamestnaný	$0,9419x^{-0,0132}(2,8449 - exp(-0,0167x))$	0,88
Slovenian Ostrovskij, Kako se je kalilo jeklo, Chapter 1	$1,1944x^{-0,0484}(2,2284 - exp(-0,0901x))$	0,85
Sundanese Agustusan (Salaka Online)	$1,6327x^{0,0465}$	0,81
Sundanese Aki Satimi (Salaka Online)	$0,7723x^{-0,001}(2,7764 - exp(-0,1625x))$	0,82
Tagalog Hernandez, Limang Alas: Tatlong Santo	$1,8390+0,8272x^{-0,1595}(1-exp(-0,0243(x-10,89)))^2$	0,74
Tagalog Hernandez, Magpisan	$2.343x^{-0.008}$	0.50
Tagalog Rosales, Kristal Na Tubig	$2.0799 + 0.0625(1 - exp(-0.0068(x-216.66)))^2$	0.80
Tamil (Press)	$3.5629-1.2908x^{-0.5509}(1-exp(-0.0166(x-45.87)))^2$	0,82
Telugu Trailangaswamy	$3,3640+6,973x^{-0,5289}(1-exp(-0,0225(x-26,065)))^2$	0,72
Telugu Train Journey without safety	$3,6664-0,6931x^{-0,2738}(1-exp(-0,0157(x-39,07)))^2$	0,53
Vai Mu ja vaa lɔ (T. Sherman)	$1.3389 + 0.1203(1 - exp(-0.0102(x-86.78)))^2$	0.58
Vai Sabu Mua Ko	$1.4949 + 2866x^{-2.143}(1-exp(-0.0179(x-16.59)))^2$	0.78
Vai Vande bɛ Wu'u	$1,5873-0,1351x^{0,0004}(1-exp(-0,0461(x-19,95)))^2$	0,72
Welsh T1 Crynodeb Gweithredol	$2,3333-0,1473x^{0,1152}(1-exp(-0,0164(x-10,06)))^2$	0,89
Welsh T2 Ffansi camu i esgidiau rhywun enwog?	$1,7376-0,0159x^{0,2815}(1-exp(-0,0770(x-26,57)))^2$	0,73

In any case, we can state that the development of word length in text may have very different courses. Just as above, it depends on boundary conditions, and the links of this phenomenon to other text properties are a future task for synergetic linguistics.

Resume

Word length is a stochastic phenomenon depending on a great number of factors disturbing the possibility of exact prediction of whatever kind. In long texts, even if their homogeneity cannot not presupposed, there may be a background mechanism arising e.g. from semantic, stylistic, educational, scientific, etc. grounds which get their way and display an observable tendency on some level. As a matter of fact, our results will for ever be marked by some uncertainty because we cannot take all influences into account. Every influencing factor should obtain

its own parameter in our formulas but in that case the number of observations or observation classes may turn out to be smaller than the number of necessary parameters and no test would be possible. The situation would not improve even if we tried to model word length using partial differential equations where the same problem with parameters would arise.

Hence our endeavours are trials and errors. Sometimes the same text sort in two different languages yields very similar results, while two texts of different text sorts in the same language may yield very different results. Since the number of languages is too great and the number of writers and aims multiply this number, we shall be able to set up formulas in explorative or deductive way but we shall never obtain a satisfactory explanation. Nevertheless, the formulas can be derived from a very general background giving space to all possible factors; but here, too, the shortness of words and the number of parameters in the distributions or functions will always collide. Hence, the only remedy is the acceptation of a variety of models and deriving them - in the best case - from a common background.

In order to bring order into this enormous field, individual investigations performed on (supposedly) homogeneous texts are not sufficient. With some groups of texts nice trends may appear and in turn another group displays an opposite trend. The field is far from being systematized.

Here we merely wanted to show the contradictory character of a simple language phenomenon in which linguists are interested since 150 years.

The main results obtained in this article are: (1) The distribution of word lengths in terms of syllable numbers abides by some models related to the family of Poisson distributions. (2) The roughness of the sequence of lengths moves in our data in the interval <0,06; 0,22> but preliminarily, it is not possible to show the "cause", i.e. another property of the text linked with length, which would be responsible for a concrete value. (3) Though in some texts word length increases monotonously from the beginning to the end of sentence, it is not necessarily so. (4) In the same way, the change of word length from the beginning of text to its end is not that smooth as supposed up to now.

Much individual research is necessary to find the control cycle of these four aspects.

References

Best, K.-H. (ed.) (1997). *Glottometrika 16*: *The Distribution of Word and Sentence Length.* Trier: WVT.

Best, K.-H. (ed.) (2001). *Häufigkeitsverteilungen in Texten.* Göttingen: Peust & Gutschmidt Verlag.

Djuraš, G. (2012). *Generalized Poisson Models for Word Length Frequencies in Texts of Slavic Languages.* Diss. University of Graz

Fan, F., Grzybek, P., Altmann, G. (2010). Word length in sentence. *Glottometrics 20, 70-109.*

Fenk, A., Fenk-Oczlon, G. (2006). Within-sentence distribution and retention of content words and function words. In: Grzybek, P. (ed.), *Contributions to the science of text and language. Word length studies and related issues: 157-170*. Dordrecht: Springer.

Grzybek, P. (ed.) (2006). *Word Length Studies and Related Issues*. Dordrecht: Springer.

Kelih, E. (2009): Preliminary analysis of a Slavic parallel corpus. In: Jana Levická und Radovan Garabík (eds.), *NLP, Corpus Linguistics, Corpus Based Grammar Research. Fifth International Conference Smolenice, Slovakia, 25-27 November 2009. Proceedings: 175-183.* Bratislava: Tribun.

Kelih, E. (2012): On the dependency of word length on text length. Empirical results from Russian and Bulgarian parallel texts. In: Naumann, S.; Grzybek, P.; Vulanović, R.; Altmann, G. (eds.), *Synergetic Linguistics. Text and Language as Dynamic Systems: 67-80*. Wien: Praesens.

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

Meyer, P. (1999). Relating word length to morphemic structure: a morphologically motivated class of discrete probability distributions. *Journal of Quantitative Linguistics 6(1), 66–69.*

Ord, J.K. (1972). *Families of frequency distributions*. London: Griffin.

Popescu, I.-I., Altmann, G., Grzybek, P., Jayaram, B.D., Köhler, R., Krupa, V., Mačutek J., Pustet, R., Uhlířová, L., Vidya, M.N. (2009). *Word Frequency Studies*. Berlin-New York: Mouton de Gruyter, XI + 278 pp.

Popescu, I.-I., Altmann, G., Köhler, R. (2010). Zipf's law - another view. *Quality and Quantity 44(4) 713-731.*

Popescu, I.-I., Kelih, E., Mačutek, J., Čech, R., Best, K.-H., Altmann, G. (2010). *Vectors and codes of texts.* Lüdenscheid: RAM.

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

Schmidt, P. (ed.) (1996). *Glottometrika 15*: *Issues in General Linguistic Theory and the Theory of Word Length.* Trier: WVT.

Uhlířová, L. (1997). O vztahu mezi délkou slova a jeho polohou ve větě. *Slovo a slovesnost 57, 174-184.*

Wilson, A. (2012). Word lengths in Welsh: Further investigations on prose and verse. In: Naumann, S., Grzybek, P., Vulanović, R., Altmann, G. (eds.), *Synergetic linguistics. Text and language as dynamic systems: 257-265.* Wien: Praesens

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

Wimmer, G., **Witkovský, V., Altmann, G.** (1999). Modification of probability distributions applied to word length research. *Journal of Quantitative Linguistics 6, 257-268*

Zörnig, P. (1987). A theory of distances between like elements in a sequence. *Glottometrika 8, 1-22.*

Texts

Akan: Agya Yaw Ne Akutu Kwaa, Mma Nnsua Ade Bɔne (short stories, supplied by kasahorow)

Bamana : Bamakɔ sigicogoya, Namakɔrɔba halakilen, Sonsannin ani Surukuba — from the Bamana text corpus (thanks to Valentin Vydrin);

Bamana : Masadennin — the Bamana translation by Bukari Jara of *The Little Prince (Le petit prince)*, a famous 1943 novella by Antoine de Saint-Exupéry (Bamako: Edition Jamana, 1989).

Bulgarian: Ostrovskij, Kak se kaljavaše stomanata, Chapter 1

Czech: Jan Čulík: O čem jsou dnešní Spojené státy? (2. 8. 2012). Britské listy. http://blisty.cz/art/64324.html

Czech: Karel Hvížďala: O předem zpackané prezidentské volbě aneb Jak dlouho budeme bez prezidenta 7. 8. 2012 Blog.aktualne.cz http://blog.aktualne.centrum.cz/blogy/karel-hvizdala.php?itemid=17155

Czech: Jan Macháček Slovenský (dobrý) příklad (4. 8. 2012). Respekt. http://respekt.ihned.cz/audit-jana-machacka/c1-56924030-slovensky-dobry-priklad

Czech: Jaroslav Spurný: Překvapení v justice (2.7. 2012) http://respekt.ihned.cz/komentar/c1-56386120-prekvapeni-v-justici

Czech: Marek Švehla: Voličův kalkul (4. 8.2012). http://respekt.ihned.cz/c1-56902580-editorial-volicuv-kalkul

French: Dunkerque – La route des dunes. Le Blog de François Béguin, fait partie de „Une année en France“, http://dunkerque.blog.lemonde.fr/ 07.05.2012

Hindi: Daily Hindi Milap, (31st May, 2012): After the sanction to love marriage, (page 4)

Hindi: Swatantra Varta,(31st July, 2012): The Anna Team on a cross-road (page 6)

Indonesian: Sekolah ditutup **(**Press online, 01.05.2012)

Indonesian: Pengurus PSM terbelah **(**Press online 01.05.2012)

Italian: [*Il bosone di Higgs scoperto dal Cern potrebbe essere un "impostore"*, http://www.meteoweb.eu/2012/07/il-bosone-di-higgs-scoperto-dal-cern-potrebbe-essere-un-impostore/143186/ http://www.meteoweb.eu/2012/07/fisica-scoperta-la-particella-di-dio-dati-molto-significativi-sul-bosone-di-higgs/142116/; http://www.meteoweb.eu/2012/07/scoperta-la-particella-di-dio-adesso-arrivo-tante-altre-sorprese-peter-higgs-verso-il-premio-nobel/142344/, 11.07.2012]

Japanese: Miki, K. (1941, 1995). Jinseiron Note (Essay on the life). Tokyo: Sôgensha. (CD-ROM edition included in Shinchô Bunko no 100 satsu); Shinchôsha (1995). Shinchô Bunko no 100 satsu (CD-ROM edition of 100 paperbacks extracted from Shinchô Bunko series). Tokyo: Shinchôsha

Kikongo: Independence Day Speech for the Democratic Republic of Congo. Patrice Lumumba, June 30, 1960

Kikongo: Bimpa: Ma Ngo ya Ma Nsiese (Tale: Mr. Leopard and Mr. Antelope) from ngunga.com

Kikongo: Nkongo ye Kisi *K*ongo (The People of Kongo - Customs and Traditions) by the honorable Mr. Ernesto Nzakundomba. Publisher: Imprensa Nacional - E.P. 1st Edition, Luanda, December. 2006.

Latin: Cicero, In Catilinam II, 180sentences, http://www.thelatinlibrary.com/cicero/cat.shtml,

Macedonian: Ostrovskij, Kako se kaleše čelkiot, Chapter 1

Malayalam: Newspaper. Malayala Manorama (21 June 2012). Place of Publication: Cochin. *Moralistic hooligans and their bad activities* (page 10)

Malayalam: Newspaper. Malayala Manorama (21 June 2012). Place of Publication: Cochin. *No one should die in hospital for not getting Oxygen* (page 10)

Maninka: Nko Doumbu Kende no. 2 (press);

Maninka: Nko Doumbu Kende no. 7 (press);

Maninka: Sìikán` (Constitution of Guinea, an excerpt).
Maninka texts (in Nko script) were supplied by Valentin Vydrin

Odia: The Samaj, Bhubaneswar (28 June 2012): Who is great? (page 4)

Odia: The Dharitri, Balasore (12th February, 2012): Calculation for the District Council President (page 10)

Romanian: O. Paler, Aventuri solitare (excerpt)

Romanian: D.R. Popescu, Vânătoarea regală, Chapter 2

Romanian: N. Steinhardt, Jurnalul fericirii, Trei soluții

Serbian: Ostrovskij, Kako se kalio čelik, Chapter 1

Slovenian: Ostrovskij, Kako se je kalilo jeklo, Chapter 1

Sundanese: Agustusan (Salaka Online)**;**

Sundanese: Aki Satimi (Salaka Online)

Tagalog: Hernandez, Limang Alas: Tatlong Santo; Hernandez, Magpisan; Rosales, Kristal Na Tubig

Tamil: Emu Bird Trading– Case filed against Sathiyaraj and Saratkumar http://tamil.oneindia.in/news/2012/08/08/tamilnadu-emu-scam-case-file-against-advertisement-modals-actor-159236.html

Telugu: Daily Andhrabhoo mi (4th August 2012): Train Journey without safety (p. 4)

Telugu: Daily Andhrabhoo mi (4th August 2012): Trailangaswamy: a biography (p. 10)

Vai: Mu ja vaa lɔ (by T. Sherman), supplied by Charles Riley and Tombekai Sherman

Vai: Sabu Mua Ko, Vande bɛ Wu'u — from a Vai book *Kכ' Tíɛ́ Banda Tɛiɛ Nú [Stories We Tell Diring Rice Harvest]*, 2nd edition (The Institute for Liberian Languages: Monrovia, Liberia, 1992; thanks to Valentin Vydrin)

Welsh (gaenv): Welsh-language executive summary from: *Preparing for climate change impacts on freshwater ecosystems (PRINCE).* Science Report SC030300/SR. Bristol: Environment Agency, 2007. http://publications.environment-agency.gov.uk/PDF/SCHO0507BMOJ-E-E.pdf

Welsh (gasodl): Ffansi camu i esgidiau rhywun enwog? *Y Cymro*, 8 July 2011. http://www.y-cymro.com/newyddion/c/44/i/449/desc/ffansi-camu-i-esgidiau-rhywun-enwog/