Glottometrics 20
2010

RAM-Verlag

ISSN 2625-8226

Glottometrics

Glottometrics ist eine unregelmäßig erscheinende Zeitdchrift (2-3 Ausgaben pro Jahr) für die quantitative Erforschung von Sprache und Text.
Beiträge in Deutsch oder Englisch sollten an einen der Herausgeber in einem gängigen Textverarbeitungssystem (vorrangig WORD) geschickt werden.
Glottometrics kann aus dem **Internet** heruntergeladen werden (**Open Access**), auf **CD-ROM** (PDF-Format) oder als **Druckversion** bestellt werden.

Glottometrics is a scientific journal for the quantitative research on language and text published at irregular intervals (2-3 times a year).
Contributions in English or German written with a common text processing system (preferably WORD) should be sent to one of the editors.
Glottometrics can be downloaded from the **Internet (Open Access)**, obtained on **CD-ROM** (as PDF-file) or in form of **printed copies.**

Herausgeber – Editors

Bestellungen der CD-ROM oder der gedruckten Form sind zu richten an

Orders for CD-ROM or printed copies to RAM-Verlag RAM-Verlag@t-online.de

Herunterladen/ Downloading: https://www.ram-verlag.eu/journals-e-journals/glottometrics/

Die Deutsche Bibliothek – CIP-Einheitsaufnahme
Glottometrics. 20 (2010), Lüdenscheid: RAM-Verlag, 2010. Erscheint unregelmäßig.
Diese elektronische Ressource ist im Internet (Open Access) unter der Adresse
https://www.ram-verlag.eu/journals-e-journals/glottometrics/ verfügbar.
Bibliographische Deskription nach 20 (2010) **ISSN 2625-8226**

Contents

Emmerich Kelih
The type-token relationship in Slavic parallel texts 1-11

Anja Overbeck, Arjuna Tuzzi, Ioan-Iovitz Popescu, Gabriel Altmann
Analysis of Italian word classes 12-28

Karl-Heinz Best, Jinyang Zhu
Ein Modell für die Zunahme chinesischer Schriftzeichen 29-33

Karl-Heinz Best
Zur Entwicklung des Wortschatzes der deutschen Umgangssprache 34-37

Haitao Liu, Yiyi Zhao, Wei Huang
How do Local Syntactic Structures Influence
Global Properties in Language Networks? 38-58

Andrei Beliankou, Reinhard Köhler
The distribution of parts-of-speech in Russian texts 59-69

Fengxiang Fan, Peter Grzybek, Gabriel Altmann
Dynamics of word length in sentence 70-109

History of Quantitative Linguistics

Karl-Heinz Best
Laut- und Buchstabenzählungen im frühen 19. Jahrhundert 110-11**4**

Glottometrics 20, 2010, 1-11

The type-token relationship in Slavic parallel texts

Emmerich Kelih[1]

Abstract. The aim of the paper is to analyse the statistical regulation of the type token relationship in Slavic parallel texts. Furthermore it is shown that this relationship in parallel texts can be explained due to morphological and typological characteristics.

Keyords: type-token relationship, Slavic languages, corpus, parallel texts

0. Introduction

Parallel texts are a reliable empirical resource for the cross-linguistic analyses of typological and morphological features of languages. One seemingly trivial characteristic of parallel texts is text length. The main purpose of this paper is to show that text length of parallel texts is a property that can be used for the measurement of the typological "closeness" of languages. Furthermore, an attempt to model the relation between text length (number of tokens) and vocabulary size (number of types), including a typological interpretation of parameters derived from the model used, is offered. The empirical data base consists of the translations of the Russian novel "How the steel was tempered" (N.A. Ostrovskij) into eleven Slavic Standard Languages (Slovenian, Croatian, Serbian, Bulgarian, Macedonian, Slovak, Czech, Polish, Upper-Sorbian, Belarusian and Ukrainian).

1. Parallel Text Corpora and the Type Token Ratio

1.1. Text material and linguistic input

For the examination of the relationship between text length and vocabulary size, a multi-parallel corpus was built, which contains ten chapters of the Russian novel "Kak zakaljalas' stal'/How the steel was tempered" in twelve Slavic Standard Languages, i.e., Russian, Slovenian, Croatian, Serbian, Bulgarian, Macedonian, Ukrainian, Belarusian, Polish, Upper-Sorbian, Czech and Slovak. The novel was written by N.A. Ostrovskij in the years 1932-1934. For details about the translations used and general linguistic problems of parallel corpora research cf. Kelih (2009a, 2009b).

The calculation of text length in terms of the number of tokens (*N*) and the number of types (V) was performed by means of specific software (Wordsmith 5.0). Linguistically, the word form is defined using purely orthographical criteria, e.g. every sequence of letters/ graphemes between two blanks[2] (for further details cf. Antić/Kelih/Grzybek 2006) is treated as

[1] Address correspondence to: emmerich.kelih@uni-graz.at .

[2] The orthographical criterion is considered to be the simplest definition of the word and word form respectively. or a detailed and illuminating discussion of the "graphematical" word in German cf. Fuhrhop (2008). In Kelih (2007) it has been shown that the use of different word definitions (based on

one word form. The hyphen is understood as punctuation mark which has a delimitative function. The number of tokens (*N*) is the total number of word forms occurring in the whole text. Not taking into account the frequency of word forms in the text, we obtained the number of word form types (*V*), which are sometimes called tokemes (cf. Andersen, Altmann 2006). The texts are thus analysed without any lemmatisation, i.e. the analysis focuses on the lexical level, including morphological and morphosyntactical information.

1.2. Description: Number of Tokens (N) vs. number of Types (V) in translated texts

The analysis of the text length, i.e. the number of tokens and types respectively, of parallel corpora has so far not been in the focus of quantitative and synergetic linguistics. Leaving aside practical problems concerning the calculation of a translator's remuneration based on text length, it is also of interest from a theoretical point of view. Without a full list of text length related linguistic problems, one has to consider at least two different influence factors:

1. According to contemporary translational studies, one of the main strategies of translation is the simplification of the translated text. Empirically, this can be seen not only at the syntactical level (e.g. shorter sentence length in the target text than in the source text), but on the lexical level too. In this regard, a smaller number of tokens and types of the target texts can be understood as a lexical simplification (cf. Baker 1995, 1996). Furthermore, the well-known translators' strategy of explicitation, e.g. the process of rendering information that is only implicit in the source text explicit in the target text (cf. Frankenberg-Garcia 2009: 48), plays a crucial role in text length. It is reasonable to assume that explicitation produces the effect that target texts are longer than the source text (cf. Teich 2003). Therefore in parallel text research a priori simplification as well as explicitation has to be taken into account.
2. For a comparison of the text length in source and target texts, morphological and morphosyntactical characteristics of the languages involved nevertheless have to be taken into consideration. Theoretically, no absolute 1:1 correspondence of the types and tokens level between two languages, but rather deviations from this correspondence, have to be assumed. This will now be demonstrated on the basis of the available translations of the Russian source text in 11 Slavic languages.

First, a brief analysis of the text length of our parallel texts is offered, measured in terms of the token numbers (*N*). Cf. Table 1 with an overview of the different text lengths in translated Slavic texts.

Let us start with a comparison of text length (*N*) between the source text and the languages that are traditionally treated as the members of the South Slavic group. While the Russian source text contains 49672 tokens, the Slovenian translations – the language with the highest number of tokens – has over 62600 tokens, i.e. the Slovenian text has approximately 12000 tokens (= 26%) more than the Russian source text. A relatively similar picture is to be obtained if one compares the Russian source text with another south Slavic language, such as, Macedonian, in which the translations consist of 58819 tokens; similarly large differences can be obtained in a comparison of the Russian texts with the Serbian and Croatian translations, which have approximately a 12% higher text length than the source text; for Bulgarian it is even slightly more (15%).

---

orthographical, orthographical-phonetical and phonological criteria) leads to a systematic shift of the calculated word length and related empirical parameters.

Table 1
Types (*V*) and Tokens (*N*) of „How the steel was tempered“

Language	**N**	**V**
Russian	49672	15053
Ukrainian	49612	14645
Belarusian	49874	14814
Slovak	52093	14025
Czech	52180	14136
Polish	52735	14979
Upper-Sorbian	58480	14465
Slovenian	62646	13940
Serbian	56227	13637
Croatian	56415	13830
Bulgarian	57165	12303
Macedonian	58819	11437

The same, i.e. translated texts being longer than the source, holds true for the Western Slavic languages: All of these (Slovak, Czech, Polish, Upper-Sorbian) are longer than the original Russian text, e.g. the Slovak, Czech and Polish translations of ”How the steel was tempered“ have approximately 2600 tokens more than the Russian one; only the Upper-Sorbian translation with its 58480 tokens has a behaviour similar to that of South Slavic languages in this regard.

Furthermore a phenomenon was observed that is even more important than these cross-linguistic comparisons, namely that the text length of genetically close languages is almost the same: For instance, the Eastern Slavic languages (Russian, Ukrainian, and Belarusian) share approximately the same text length, both in respect to the number of types and tokens. The difference in relation to Russian of 63 tokens (Ukrainian) and 335 tokens (Belarussian) is relatively marginal, i.e. no striking differences in text length are to be obtained. The analysis of text length of Western Slavic languages yields a relatively similar picture: For Slovak, Polish and Czech a more or less similar text length (≈ 52300 Tokens) was obtained. This especially holds true for Slovak and Czech, with a difference of only 82 tokens. The genetically “very close” languages Croatian (56424 tokens) and Serbian (56230 tokens) again do not show any notable differences whatsoever; similar small differences regarding the number of tokens can be obtained for the Bulgarian and Macedonian translations.

Generally, it can be seen that the text length (*N*) appears to be a rather robust characteristic of parallel texts (in Slavic languages). Interestingly enough, it roughly coincides with the areal and geographical affiliation of the languages under examination. Only the Slovenian and the Upper-Sorbian translations show a somehow different behaviour in this respect: The Slovenian text – the longest text with 62646 tokens – is slightly above the south Slavic average (Croatian, Serbian, Bulgarian, Macedonian) with a mean text length of approximately 58000 tokens. The same holds true for the Upper-Sorbian text (58480 tokens), which is “above” the West Slavic average of approximately 53000 tokens.

The differences obtained in text length are not caused by stylistic preferences of the translators or by an over-explicitation of the Russian original text (cf. Kelih 2009a, 2009b), but by the typological, especially morphological or morphosyntactical, features of the languages under examination. This will be discussed in more detail in 1.3.

Similar differences in text length are obtained at the type-level (*V*) too, albeit to a lesser extent than at the token level. In this respect, the Russian text has the largest number of types (= 15053), whereas the Slovak and Slovenian translations have approximately the same *V* (approx. 14000 types), i.e. 1000 types less than the Russian source text. Only the Macedonian and the Bulgarian texts are, with respect to the number of types in relation to other Slavic languages, clearly outliers, inasmuch as they have "only" 11437 and 12327 types respectively. This can in particular be reasonably explained by the commonly known extremely morphologically limited case flexion system of these two south Slavic languages.

Finally, the relation between types and tokens has to be analysed. As can be seen from figure 1, there is no particular (strong) statistical correlation between the number of tokens (*N*) and types (*V*) in Slavic languages. Evidently, there are no mechanisms of compensation in the way that a small vocabulary (*V*) of one language is accompanied by a long text length in the number of tokens.

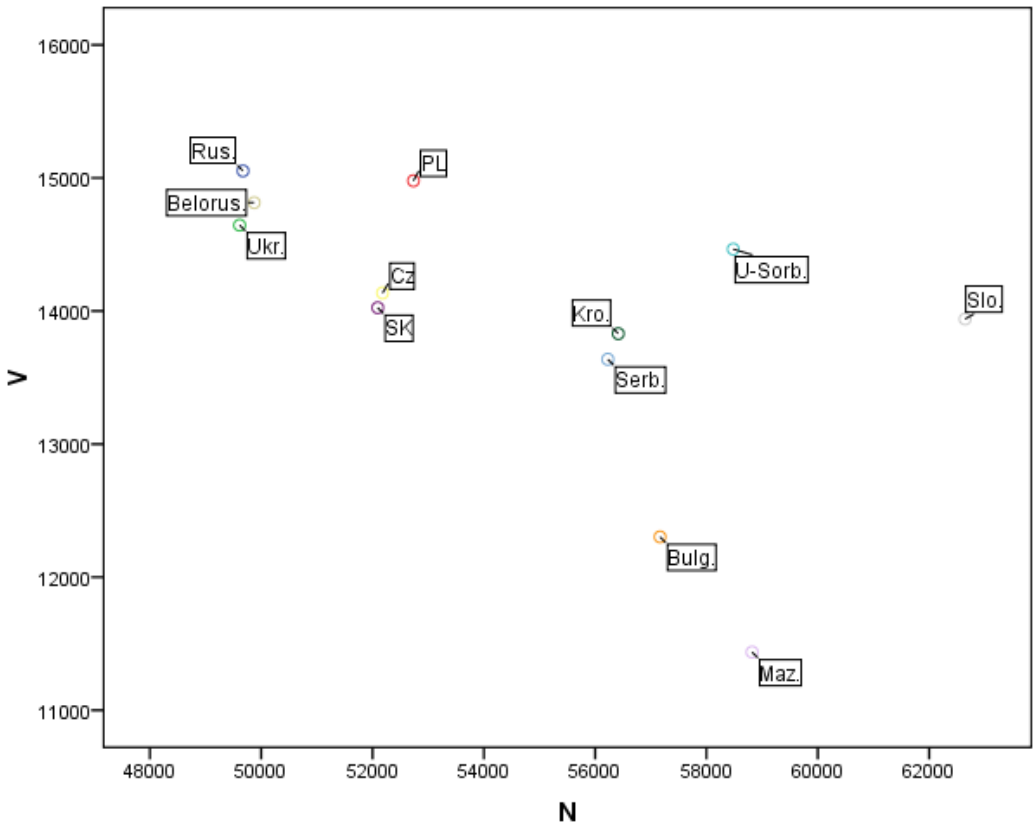


Figure 1: Relation between number of tokens (*N*) and types (*V*) in KZS

Although there are no systematic inter-lingual relations at text length level, one more analysis has to be performed: The relation between the number of types (*V*) and tokens (*N*) within the individual chapters of the translated novel, i.e. within the different languages. It is well known from quantitative linguistics that the relation between the number of tokens (*N*) and number of types (*V*) is systematically organised and controlled by a statistical mechanism, which has to be analysed in two respects: On the one hand in regard to the intra-textual regulation in the various chapters of the novel, and on the other hand again cross-linguistically, i.e. the control of types and tokens, in dependency of the morphological characteristics (productivity of the inflection system, morphological expression of the reflexivity, analytic or synthetic expression of temporal forms etc.) of the languages.

1.3. Modelling Text length (*N*) and Vocabulary Size (*V*): Intra-textual perspective

For the modelling of the relation between text length (*N*) and vocabulary size (*V*) many different continuous and discrete models have been proposed in the past. It is beyond the scope of this paper to present and discuss these different models in detail (cf. Altmann/ Altmann 2008: 107ff. and Fan (2008) for an extensive overview). Generally, not only is the relation between text length (*N*) and vocabulary size (*V*) discussed, but different indices and coefficients are derived from these two text characteristics such as the Type Token Ratio (TTR), calculated as *V*/*N* or *N*/*V*. These indices are usually understood as measurements of the vocabulary richness and style (cf. Herdan 1966: 77) of a text. However, the TTR clearly depends on the type of the analysed texts and the text length (cf. Tuldava 1974, 1993, 1995), and thus an interpretation of the TTR as a stylistic feature must be approached with caution.

A slightly different interpretation of the TTR – which is indirectly related to the absolute number of types and tokens too – is offered by Altmann/Altmann (2008: 107), who understands the TTR as a measurement of the information flow within one particular text. The more often one token is repeated within one text/chapter, the slower the spread of new information. It goes without saying that again this information regulation is text type specific and clearly dependent on the length[3] of the texts under examination, but nevertheless this interpretation is much more reasonable than the stylistic one offered above.

However, both interpretations of TTR, i.e. as the control of the information flow within one text and as the measurement of the vocabulary/stylistic richness of one text, seem to fail in the analysis of parallel texts. Differences in the number of types and tokens between the source text and the target text can, at least for the KZS corpus, be explained mainly[4] by morphological and morphosyntactical features of the translated languages, as shown below.

Thus, from our point of view the number of types and tokens in parallel-text research represents the morphological productivity of the languages under examination: The more productive the inflection system (especially of verbs and nouns), the greater the number of different morphological forms that are generated by the language system. For a language that has a highly productive flexion system, the probability of the appearance of the same word form type seems to be lower than for a language with a less productive flexion system. Thus, there must be a direct impact on the number of types and tokens of translated texts.

Furthermore, differences in text length at the type and token level in parallel-text research can, as already pointed out by Popescu/Altmann (2008: 371) and Popescu/ Mačutek/Altmann (2009: 99), be connected with the discussion of the significance of the number of Hapax Legomena for language typology issues, used as an indicator for the degree of analyticity/ syntheticity of the languages under examination. Thus – stylistic variation by the translators has to be excluded for the sake of simplicity – it can be stated that the longer the translated text in relation to the source text, the higher the analytic expression of the morphological and morphosyntactical categories of the target languages. To at least illustrate this with one example, the expression of reflexivity in Slavic languages, for instance, has to be kept in mind: In Russian it is part of an (orthographically defined) word form (e.g. podnjat'sja – to stand up), whereas in Slovenian the reflexive marker appears as a separate word form (vzdigniti se – to stand up); similar phenomena can also be obtained for the analytic and synthetic morphological expression of the temporal system. Without going into much more

[3] A quite similar problem has been discussed in quantitative linguistics in connection with Zipf's law and Zipf's size *U*, a hypothetical length of a text constructed by a speaker. Cf. Orlov (1982) and Orlov/Boroda/Nadarejšvili (1982).

[4] For the sake of simplicity, the well known phenomena of simplification and explication of translated texts are not taken into consideration here.

detail, it is now quite clear that the analysis of the number of types and tokens comes far behind the interpretation of the information flow and the problem of vocabulary richness of texts.

To give a more elaborate insight into the statistical behaviour of the relation between the number of types and tokens (and not of the TTR in general), the intra-textual behaviour now has to be analysed and modelled. For all the translations, the dynamic relationship between *V* and *N* of the entire novel in twelve languages was computed at a chapter-spaced interval. The results are in Table 2.

Table 2
Type and Token in cumulated chapters: 12 Slavic languages

	Russian		**Ukrainian**		**Belarusian**		**Slovak**	
chapter	**N**	**V**	**N**	**V**	**N**	**V**	**N**	**V**
1	4107	1907	4119	1895	4145	1916	4275	1895
1-2	8243	3538	8279	3491	8322	3524	8600	3472
1-3	14567	5591	14561	5498	14689	5536	15096	5405
1-4	18300	7003	18325	6865	18480	6890	18981	6701
1-5	22070	7956	22080	7751	22271	7802	22843	7568
1-6	29604	9822	29622	9533	29818	9646	30864	9260
1-7	35624	11388	35621	11061	35881	11217	37201	10677
1-8	40976	12864	40983	12534	41243	12674	42982	12020
1-9	44270	13635	44261	13309	44555	13447	46394	12735
1-10	49672	15053	49612	14645	49874	14814	52093	14025
	Czech		**Polish**		**Upper-Sorbian**		**Slovenian**	
chapter	**N**	**V**	**N**	**V**	**N**	**V**	**N**	**V**
1	3925	1773	4348	1970	4851	1976	5209	1955
1-2	8306	3375	8716	3625	9663	3547	10408	3490
1-3	14976	5360	15410	5668	17085	5496	18379	5420
1-4	18896	6693	19413	7084	21541	6884	23166	6737
1-5	22748	7559	23410	8010	25813	7768	27886	7565
1-6	30865	9295	31347	9849	34608	9506	37432	9188
1-7	37255	10717	37695	11399	41666	10992	44952	10624
1-8	42993	12086	43448	12839	47982	12406	51744	11990
1-9	46444	12840	46949	13600	51832	13157	55849	12696
1-10	52180	14136	52735	14979	58480	14465	62646	13940
	Serbian		**Croatian**		**Bulgarian**		**Macedonian**	
chapter	**N**	**V**	**N**	**V**	**N**	**V**	**N**	**V**
1	4579	1899	4582	1900	4653	1709	4810	1636
1-2	9235	3435	9271	3458	9387	3127	9708	2944
1-3	16328	5330	16431	5386	16611	4807	17178	4509
1-4	20618	6639	20747	6718	20916	5991	21602	5619
1-5	24859	7494	25002	7586	25193	6710	26027	6295
1-6	33425	9085	33555	9169	33866	8170	34941	7624
1-7	40241	10436	40396	10574	40858	9421	42094	8788
1-8	46270	11746	46471	11913	47100	10605	48508	9873
1-9	50019	12427	50231	12613	50887	11202	52358	10421
1-10	56227	13637	56415	13830	57165	12303	58819	11437

For the statistical modelling of the relation between types and tokens, the power model $V = aN^b$ was used. As a result, this model is appropriate without any exception for all languages analysed here, attaining in all cases a $R^2 > 0.99$. A graphical representation of the modelled interrelation of four selected Slavic languages is given in Figures 1a-1d. Table 3 shows the goodness-of-fit results and the parameters *a* and *b* of the model used.

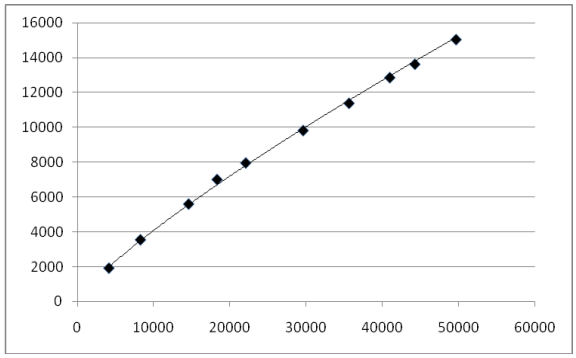


Figure 1a.: Russian: *N* vs. *V*

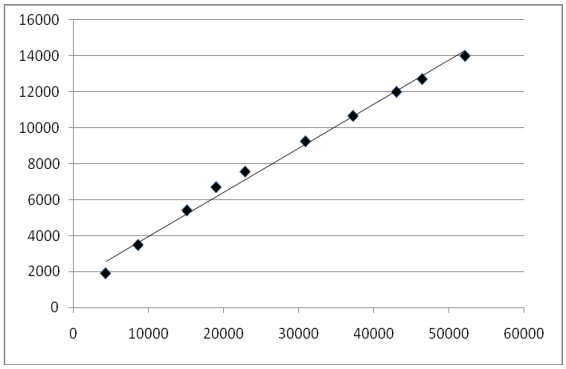


Figure 1b: Slovak: *N* vs. *V*

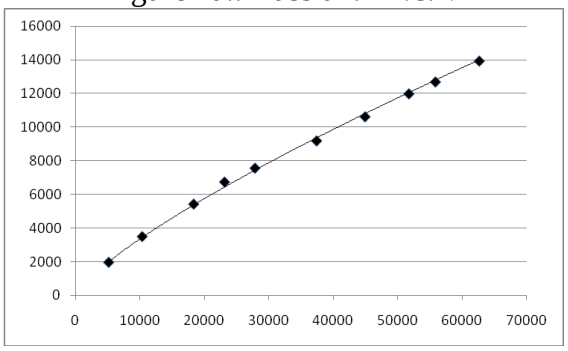


Figure 1c: Slovenian: *N* vs. *V*

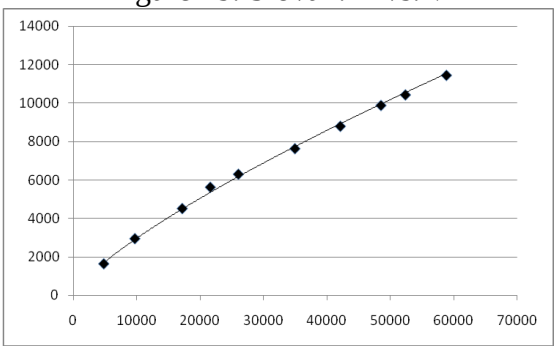


Figure. 1d: Macedonian: *N* vs. *V*

Table 3
Goodness of fit and parameters

Language	**Parameter b**	**Parameter a**	**R^2**
Russian	0.7974	2.6990	0.9993
Slovak	0.7630	3.5115	0.9991
Slovenian	0.7636	3.0114	0.999
Macedonian	0.7454	3.1736	0.999
Czech	0.7678	3.3560	0.9991
Serbian	0.7520	3.6417	0.999
Croatian	0.7569	3.4909	0.9989
Bulgarian	0.7514	3.2640	0.9990
Belarusian	0.7980	2.6274	0.9994
Ukrainian	0.7932	2.7479	0.9992
Polish	0.7810	3.0616	0.9993
Upper-Sorbian	0.7753	2.9102	0.9993

As can be seen in Figure 2, which displays the dependence of the combined theoretical values of *V* on the number of tokens (*N*) in five Slavic languages (Russian, Slovak, Slovene, Macedonian and Upper-Sorbian), control of text length (*N*) and the vocabulary size (*V*) are evidently strongly determined by the morphological status of the language. There is no

overlap of the analysed languages, and the supposed typological relevance in the comparison of text length of translated texts/languages can be shown empirically. A striking characteristic of the relation of types and tokens is the different increase of the fitting curves, which again can be interpreted as important typological information: The slower the increase of the curve, the more analytic the language.

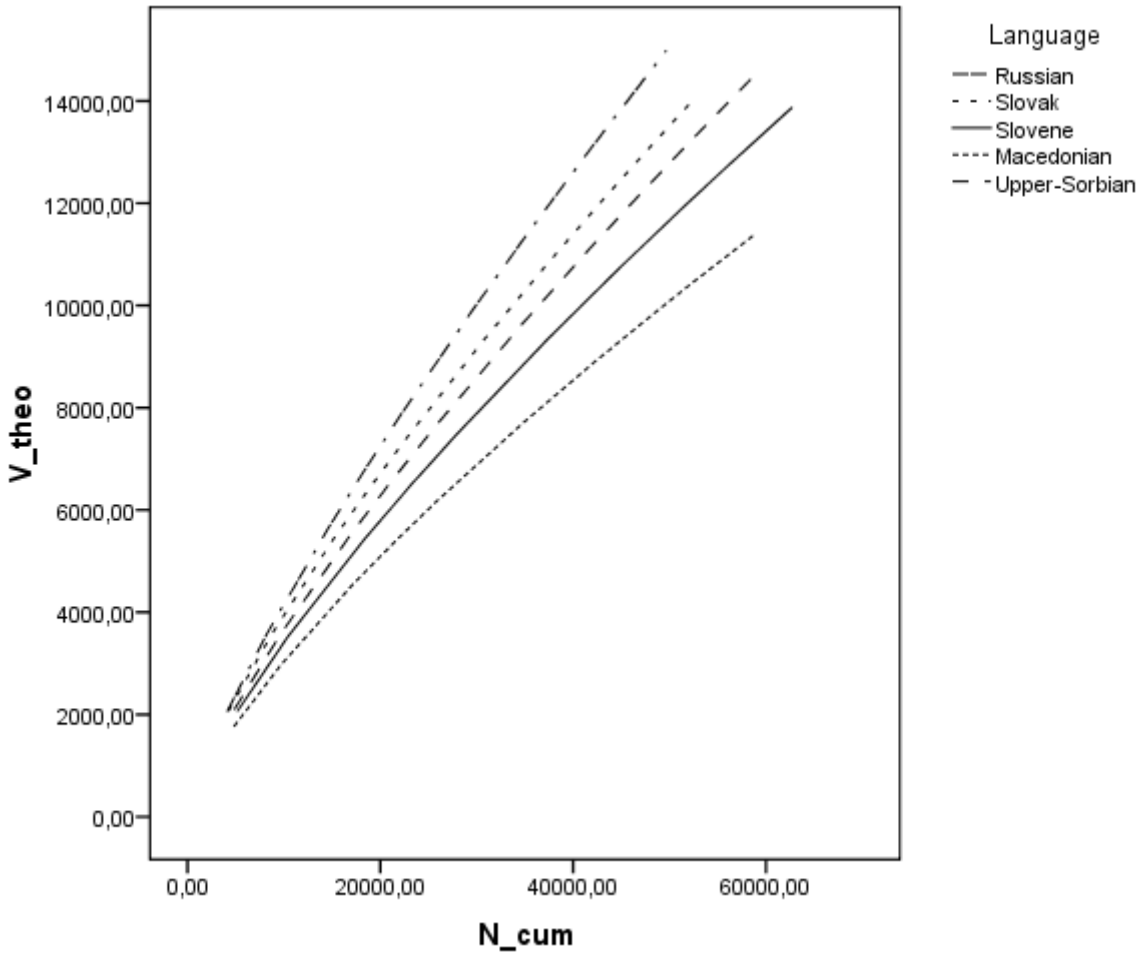


Figure. 2. Interrelation between *N* und *V* (theoretical values) in five Slavic languages

Furthermore, the conclusion can be drawn from Figure 2 that with increasing text length the discrepancy, i.e. the distance between the analysed languages (= texts) becomes wider and wider. In other words, to reach maximal comparability in parallel corpora research, the texts that are analysed should not be too short.

1.3.1. Significant difference in the increase?

Having found a simple and appropriate model for the relation between number of tokens (*N*) and number of types (*V*) for different Slavic parallel texts, it can finally be tested whether there is a significant difference in the increase of the fitting curves, i.e. of the regression coefficients. To do so, the abovementioned model is linearised by logarithmisation and trans-

formed into *log(V) = log(a) + blog(N)*, and it is tested whether there are significant differences in the regression coefficients. Cf. Zöfel (2002: 146) for details of the procedure used.

In Figure 3 the calculated regression coefficients of the analysed texts (= languages) are plotted in ranked order. Russian is at the upper most level and Macedonian at the lower most level. In particular, there is no need for complicated and systematic comparisons between the texts, but rather in a first step it is sufficient to compare the languages, with the maximal and minimal regression coefficients only.

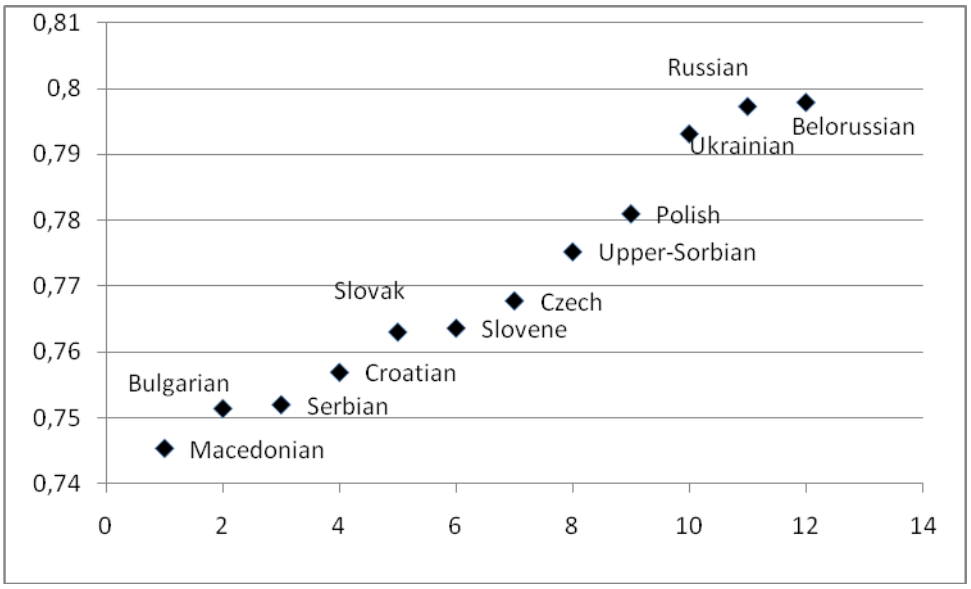


Figure 3. Parameter *b* (ranked values)

Indeed, the comparison of the Russian and Macedonian coefficients shows no significant differences ($p > 0.05$). As long as there are no statistical significant differences between these two turning points, the same results are naturally to be obtained for other languages. Cf. Table 3 with some selected comparisons, for instance Russian-Bulgarian with a $p = 0.9961$, Russian-Slovak with a $p = 0.9973$, etc.

Table 4
Results of *t*-distributed test statistics (selected languages)

pairs of comparison		t-value	DF	p
Russian	Macedonian	0.0008	26486	0.9993
	Bulgarian	0.0048	27352	0.9961
	Slovak	0.0033	29074	0.9973
	Slovene	0.0043	28989	0.9965
	Belarusian	0.0007	29863	0.9994

Interestingly enough, there are, in general, no significant differences in the increase of the relation between text length (*N*) and vocabulary size (*V*) obtained above. Nevertheless, the order of parameter *b* ranked above is particularly interesting for language typology: Russian, Belarusian, Polish, Upper-Sorbian, Czech, Slovak, Slovene, Croatian, Serbian, Bulgarian and Macedonian. This rank order roughly represents the degree of syntheticity/ analycity of the Slavic texts/languages under examination. It starts with synthetic Russian and ends with the

analytic south-eastern Slavic languages, i.e. Macedonian and Bulgarian. For the time being, it is not known whether or not this typological order can also be obtained for other Slavic parallel text corpora. For future research it is necessary to take also other parameters into consideration, such as, the number of Hapax Legomena, the productivity of the nominal flexion, etc.

2. Summary

Let us summarize the results of the present study. The relation between text length (*N*) and vocabulary size (*V*) in parallel text research is considered to be an efficient tool for the analysis of morphological features of the examined texts from 12 Slavic languages. Furthermore, it has been shown that selective parameters of appropriate models capturing the relation between *V* and *N* of the parallel Slavic texts can be interpreted as the degree of analyticity/ syntheticity of the analysed texts/languages. Nevertheless, further parallel texts have to be analysed in this regard.

References

Altmann, V.; Altmann, G. (2008). *Anleitungen zu quantitativen Textanalysen. Methoden und Anwendungen.* Lüdenscheid: RAM-Verlag..

Andersen, S.; Altmann, G. (2006). Information content of words in text. In: Grzybek, P. (ed.), *Word length studies and related issues: 93-117*. Boston: Kluwer.

Baker, Mona (1995). Corpora in translation studies: An overview and some suggestions for future research. In: *Target 7(2), 223–245.*

Baker, Mona (1996). Corpus-based translations studies: The challenge that lie ahead. In: Somers, Harold (ed.), *Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager; 175-186*. Amsterdam: Benjamins.

Fan, Fengxiang (2008). A Corpus-Based Study on Random Textual Vocabulary Coverage. *Corpus Linguistic and Linguistic Theory 4(1), 1–17.*

Fuhrhop, Nanna (2008). Das graphematische Wort (im Deutschen): Eine erste Annäherung. *Zeitschrift für Sprachwissenschaft 27, 198-228.*

Frankenberg-Garcia, Ana (2009). Are Translations Longer Than Source Texts? A Corpus-Based Study of Explicitation. In: Beeby, Allison; Rodríguez Inés, Patricia; Sánchez-Gijón, Pilar (eds.), *Corpus Use and Translating: Corpus Use for Learning to Translate and Learning Corpus Use to Translate: 47-58*. Amsterdam: Benjamins.

Kelih, E. (2007). Zur Frage der Wortdefinitionen in Wortlängenuntersuchungen. In: Kaliuščenko, Volodymir; Köhler, Reinhard, Levickij, Viktor (eds.), *Problems of Typological and Quantitative Lexicology: 91-105*. Chernivtsi: Ruta..

Kelih, E. (2009). Slawisches Parellel-Textkorpus: Projektvorstellung von „Kak zakaljalas' stal' (KZS)". In: Kelih, E.; Levickij, V.V.; Altmann, G. (eds.), *Methods in Text Analysis:* 106-124.. Chernivtsi: Ruta.

Orlov, Ju.K. (1982). Ein Modell der Häufigkeitsstruktur des Vokabulars. In: Guiter, H.; Arapov, M.V. (eds.*), Studies on Zipf's law*: *154-233.* Bochum: Brockmeyer.

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

Popescu, Ioan-Iovitz; Altmann, Gabriel (2008). Hapax legomena and language typology. *Journal of Quantitative Linguistics 15(4), 370-378.*

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

Teich, Elke (2003). Cross-linguistic Variation in System and Text. A Methodology for the Investigation of Translations and Comparable Texts. Berlin, New York: Mouton de Gruyter.

Tuldava, Ju.A. (1974). O statističeskoj strukture teksta. In: *Sovetskaja pedagogika i škola* 9, *5-33*. [English translation in: Tuldava, Ju. (1995): *Methods in Quantitative Linguistics*. Trier: WVT.]

Tuldava, J. (1993). The statistical structure of a text and its readability. In: Hřebíček, L., Altmann, G. (eds.), *Quantitative text analysis: 215-227*. Trier: WVT.

Tuldava, Ju.A. (1995). The ratio of words forms and lexemes in texts. In: Tuldava, Ju. (1995), *Methods in Quantitative Linguistics*. Trier: WVT.

Glottometrics 20, 2010, 12-28

Analysis of Italian word classes

Anja Overbeck, Göttingen
Arjuna Tuzzi, Padova
Ioan-Iovitz Popescu, Bucharest
Gabriel Altmann, Lüdenscheid

Abstract. In the article, the development and the stabilization of the occurrence of parts of speech in Italian texts are analyzed. The data were won from opera librettos since 1607 and from all end-of-year speeches of Italian presidents (1949-2008). The results display both developmental, text-sort and content-dependent differences. In order to obtain a more complete image, the analysis should be applied to many other Italian texts.

Keywords: Italian, parts of speech, Zipf´s law, ternary plot, golden section

0. Introduction

Any class of linguistic entities is a conceptual construction. We conjecture its existence because some entities behave similarly if considered from a certain view, e.g. grammatical, semantic, lexicological, historical, phonetic, etc. However, in the reality only individuals exist. The establishment of classes is a gnoseological bridge allowing us to order the reality, recognize some background mechanisms that affect the behaviour of entities if they get into the gravitation of an attractor, and last but not least, to control the elements of the class.

The classes of linguistic objects are mostly fuzzy but usually we decide about their membership categorically, according to the aim of the study. As to the word classes (parts of speech), the Latin tradition still survives and in many languages it is useful for numerous purposes. But in many cases a classification-driven disambiguation of words is possible only within the sentence with the use of some criteria. However, the criteria are not contained in the sentence, they are established by us, i.e., they are conceptual constructions, too. In this situation we cannot say that a classification is true or not true but only that it is more or less productive (prolific) for a certain purpose.

There are, of course, criteria which are so to say external and do not have any influence on the affiliation to one of the members of a class but their use is helpful in deciding about the reasonableness of a class. This is, for example, the frequency of occurrence. We suppose that a class is established rationally if the rank-frequency of the members abides by a known regularity. Since ranking presupposes ordering, the resulting distribution cannot be either uniform or normal, it is a convex hyperbolic function. The whole domain is called "Zipf´s law" but the resulting formulas have many different forms (cf. http://www.nslij-genetics.org/wli/zipf).

In order to study the behaviour of Italian word classes (parts of speech, POS) we shall consider two kinds of texts: (a) nine librettos of Italian operas (1607-1887) furnishing us with some historical data, and (b) sixty end-of-year speeches of Italian presidents (1949-2008). First we shall study the homogeneity of texts from the viewpoint of POS, then we shall study the distribution of POS and the properties of some parameters. If there is a historical change, it will automatically follow from the results.

The data are presented in Table 1 and 2. The librettos are as follows

L1 Alessandro Striggio the Younger (1573-1630), *La favola d'Orfeo* 1607
L2 Aurelio Aureli (ca. 1630- ca. 1709), *L'Orfeo* 1672
L3 Apostolo Zeno (1668-1750), *Teuzzone* 1719
L4 Pietro Metastasio (1698-1782), *L'Olimpiade* 1733
L5 Giambattista Varesco (1736-1805), *Idomeneo, re di Creta* 1781
L6 Gaetano Rossi (1774-1855), *Semiramide* 1823
L7 Felice Romani (1788-1865), *Norma* 1831
L8 Francesco Maria Piave (1810-1876), *Ernani* 1844
L9 Arrigo Boito (1842-1918), *Otello* 1887

The Italian presidents are as follows: Luigi Einaudi 1949-1954, Giovanni Gronchi 1955-1961, Antonio Segni 1962-1963, Giuseppe Saragat 1964-1970, Giovanni Leone 1971-1977, Sandro Pertini 1978-1984, Francesco Cossiga 1985-1991, Oscar Luigi Scalfaro 1992-1998, Carlo Azeglio Ciampi 1999-2005, Giorgio Napolitano 2006-2008.

Table 1
Frequencies of POS in Italian librettos
(Adj = adjective, Adv = adverb, C = conjunction, D = article, I = interjection, N = noun, PN = personal name, Pp = preposition, Pn = pronoun, V = verb)

	Adj	**Adv**	**C**	**D**	**I**	**N**	**PN**	**Pp**	**Pn**	**V**	**Sum**
L1 Striggio	516	349	255	311	79	847	72	441	498	598	**3966**
L2 Aureli	870	574	480	705	107	1737	253	931	1210	1740	**8607**
L3 Zeno	792	547	508	695	116	1632	123	821	1215	1561	**8010**
L4 Metastasio	900	768	458	676	216	1560	234	921	1403	1959	**9095**
L5 Varesco	463	309	234	338	159	957	121	413	651	775	**4420**
L6 Rossi	480	333	256	393	97	991	192	598	845	934	**5119**
L7 Romani	367	372	193	318	199	846	88	434	773	1035	**4694**
L8 Piave	391	304	152	403	77	942	99	446	615	903	**4332**
L9 Boito	643	401	293	369	114	1342	128	540	968	1209	**6007**
Sum	**5422**	**3957**	**2829**	**4208**	**1164**	**10854**	**1310**	**5545**	**8178**	**10814**	**54181**

1. Non-homogeneity

As can be seen the numbers are not equal and transforming them in proportions would display some differences. Nevertheless, the distributions of frequencies can be homogeneous between the authors. The overall homogeneity can be tested by a simple chi-square test using the formula

$$X^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(x_{ij} - E_{ij})^2}{E_{ij}}, \quad (1)$$

where m = number of POS classes (here 10), n = number of authors (here 9), further let

$s_{i.} = \sum_{j=1}^{m} x_{ij}$, representing the sums of rows, $c_{.j} = \sum_{i=1}^{n} x_{ij}$ representing the sums of columns, and $E_{ij} = \frac{s_{i.} c_{.j}}{N}$, N being the sum of all frequencies. The chi-square has (m-1)(n-1) degrees of freedom (DF).

Another approximation is the use of the information statistics. In this case the test criterion is

$$2I = 2\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij} \ln x_{ij} + 2N \ln N - 2\sum_{i=1}^{m} s_{i.} \ln s_{i.} - 2\sum_{j=1}^{n} c_{.j} \ln c_{.j} \quad (2)$$

which is distributed approximately as chi-square with (m-1)(n-1) degrees of freedom. Inserting the numbers from the table in these formulas we obtain X^2 = 776.62 and 2I = 759.07 with 72 degrees of freedom, indicating that the distributions are far from being homogeneous. That means that either there are stylistic differences – a result that can be accepted without objections – or there is a development in Italian. It has been shown that both possibilities apply in any manner (Overbeck 2009: chapter 3.1.1.). For testing the latter case one would be forced to collect data from different texts for regular time intervals. Nevertheless, the development of non-homogeneity can be tested by stepwise comparison of authors. In the first step one compares Striggio (1607) with Aureli (1672) using the numbers in Table 2 representing the first two rows of Table 1.

Table 2
Comparison of the first two data

	Adj	**Adv**	**C**	**D**	**I**	**N**	**PN**	**Pp**	**Pn**	**V**	**Sum**
L1 Striggio	516	349	255	311	79	847	72	441	498	598	**3966**
L2 Aureli	870	574	480	705	107	1737	253	931	1210	1740	**8607**

The information statistics yields 2I = 110.75 and the chi-square 109.99 with 9 DF showing that Striggio and Aureli significantly differ in the use of POS.

Since we compare the bulk of previous data with the new ones, the next step is adding L1 + L2 and comparing it with L3. In that case we obtain the data in Table 3.

Table 3
Comparison of L1 + L2 with L3

L1 +L2	1386	923	735	1016	186	2584	325	1372	1708	2338
L3 Zeno	792	547	508	695	116	1632	123	821	1215	1561

Now 2I = 50.60, X^2 = 49.43 (DF 9) which is significant but smaller, telling that Zeno lies between the two authors. If we continue in this way, viz. comparing L1-L3 with L4, then L1-L4 with L5 we obtain the results in Table 4.

Table 4
The development of homogeneity in librettos

Texts	2I	X^2	DF
L1 vs L2	110.75	109.99	9
L1-L2 vs L3	50.60	49.43	9
L1-L3 vs L4	132.84	133.95	9
L1-L4 vs. L5	86.33	97.17	9
L1-L5 vs L6	66.99	70.80	9
L1-L6 vs L7	176.34	193.04	9
L1-L7 vs L8	64.77	61.83	9
L1-L8 vs L9	70.11	68.61	9

As can be seen, there is no regular development. The greatest deviation can be found with L7 Romani. It has already been shown that this special libretto takes an exceptional position within the libretto corpus concerning the frequency of verbs (especially imperatives) and interjections which may be explained by its exemplary character for the romantic *melo-dramma* (Overbeck 2009: chapters 3.1.1, 4.1.2. and 4.2.2.). The sums of the 2I or X^2 are equal to the overall 2I and X^2 respectively, computed for Table 1.

Our next question is the state in the end-of-year speeches of Italian presidents.

The data are presented in Table 5. The first column contains the name of the president and the year of his speech, the other columns contain the numbers of individual POS. In this table the numerals (Nm) are considered a separate word class which must be eliminated if we compare the presidential speeches with librettos.

Using the above formulas we obtain a very clear result. The criteria yield 2I = 7.08 and X^2 = 6.98, i.e. almost identical and very small results. The number of degrees of freedom is DF = 590, hence the speeches are homogeneous. Since the speeches belong to the same genre and underlie the same laws of text linguistics and stylistics, the significant heterogeneity of librettos must be caused both by stylistic factors and historical development.

Table 5
POS distribution in presidential speeches

PoS	**Adj**	**Adv**	**C**	**D**	**I**	**N**	**PN**	**Nm**	**Pp**	**Pr**	**V**
1949Einaudi	30	6	15	14	0	41	0	1	37	17	33
1950Einaudi	15	4	9	15	0	42	0	1	36	8	20
1951Einaudi	41	15	11	18	0	50	0	0	40	21	34
1952Einaudi	27	7	11	12	0	46	0	0	35	13	28
1953Einaudi	34	6	12	9	0	47	0	1	42	15	24
1954Einaudi	43	14	18	17	0	57	1	0	54	20	36
1955Gronchi	64	21	29	31	0	83	0	1	78	30	51
1956Gronchi	88	26	58	71	0	180	0	0	121	34	87
1957Gronchi	170	59	79	93	0	267	6	5	241	84	126
1958Gronchi	131	42	74	82	0	201	3	1	162	63	127
1959Gronchi	92	36	72	71	0	181	0	1	135	29	80
1960Gronchi	112	38	63	78	0	196	3	2	161	45	106
1961Gronchi	184	75	111	105	0	304	0	2	244	65	162
1962Segni	120	29	54	73	0	196	0	0	147	36	83

1963Segni	170	45	68	93	0	257	14	8	219	52	131
1964Saragat	85	17	42	40	0	102	0	3	84	28	64
1965Saragat	141	45	78	85	0	267	6	3	211	79	138
1966Saragat	185	50	75	109	0	324	2	5	239	66	144
1967Saragat	167	36	59	96	2	263	14	3	207	64	145
1968Saragat	176	56	86	95	0	304	2	8	243	70	134
1969Saragat	222	72	103	165	2	394	3	8	284	99	232
1970Saragat	272	86	112	186	2	490	5	17	389	113	257
1971Leone	37	6	11	30	0	70	2	3	51	17	35
1972Leone	134	24	45	69	0	182	3	5	149	45	111
1973Leone	174	63	103	97	0	298	1	1	232	76	205
1974Leone	120	35	59	66	0	197	1	1	141	42	139
1975Leone	200	69	122	97	0	312	0	2	244	91	191
1976Leone	196	73	113	112	0	321	3	1	239	97	211
1977Leone	216	113	142	115	0	358	2	4	270	122	262
1978Pertini	156	86	125	106	0	332	10	17	248	130	283
1979Pertini	279	115	201	184	2	499	8	8	345	219	442
1980Pertini	164	61	101	104	2	316	10	9	228	121	244
1981Pertini	331	196	227	231	1	571	38	14	377	261	571
1982Pertini	322	139	172	233	2	509	62	19	332	202	495
1983Pertini	452	206	275	360	3	786	55	33	510	308	760
1984Pertini	163	51	97	129	0	302	20	17	197	95	269
1985Cossiga	404	93	192	207	2	612	10	3	427	120	289
1986Cossiga	215	77	106	130	1	321	0	1	232	79	187
1987Cossiga	349	107	163	184	0	501	11	11	414	103	248
1988Cossiga	369	134	183	199	0	557	5	14	467	146	311
1989Cossiga	302	101	145	154	0	441	31	6	399	102	231
1990Cossiga	534	173	305	277	0	800	35	18	646	163	396
1991Cossiga	64	29	48	26	2	95	0	4	71	22	57
1992Scalfaro	360	151	250	231	2	656	3	4	435	208	472
1993Scalfaro	387	168	247	236	1	684	22	8	501	218	469
1994Scalfaro	482	207	267	284	1	866	12	15	633	248	590
1995Scalfaro	523	246	357	332	3	994	22	38	682	290	741
1996Scalfaro	326	110	128	183	0	535	11	16	348	115	313
1997Scalfaro	522	329	368	429	10	1113	54	33	712	397	1048
1998Scalfaro	415	254	289	399	5	972	23	35	577	251	775
1999Ciampi	278	82	89	206	1	504	9	24	347	110	291
2000Ciampi	273	88	95	168	0	432	23	12	338	124	291
2001Ciampi	262	96	89	224	2	549	18	15	395	109	338
2002Ciampi	304	98	132	209	0	556	10	7	389	112	312
2003Ciampi	214	79	75	142	1	408	4	2	297	112	231
2004Ciampi	265	88	93	147	0	455	19	8	353	111	268
2005Ciampi	166	40	55	114	1	290	10	12	235	89	181
2006Napolitano	286	146	191	169	1	502	10	7	377	159	356
2007Napolitano	242	115	144	123	3	419	13	5	352	104	274
2008Napolitano	220	86	127	135	1	409	4	2	328	120	281

If we compare the librettos as a whole with the speeches as a whole (skipping the numerals) we obtain the data in Table 6 and the results are 2I = 7948.33 and X^2 = 7910.06 with 9 DF signalizing a significant heterogeneity between these two groups. It is to be noted that the chi-square increases with increasing sample size and one frequently uses modifications or normalizations in form of different coefficients (Pearson, Cramér, Tschouproff, etc.) which are then interpreted rather intuitively. Here we dispense with this possibility.

Table 6
Comparison of presidential speeches and librettos

Presidents	13275	5119	7170	8399	53	23016	633	16927	6489	15410
Librettos	5422	3957	2829	4208	1164	10854	1310	5545	8178	10814

Further comparisons, e.g. president-wise partitionings are not necessary because the whole domain is homogeneous. In order to make some substantiated statements about the style, genre or development, more different texts should be analyzed.

2. Rank-frequency analysis

In order to study whether the given POS classification abides by an external criterion, we consider the rank-frequency distribution of word classes in individual texts. The empirical results are presented in Table 5. Usually one tests at the first place the famous Zipf´s law being a power function. One can consider the data either as frequency distribution or simply a sequence that can be captured by a simple function. Since both approaches are conceptual approximations to reality, we follow the second way and test two functions, namely

(3) $y = c/x^a$ (Zipf) and

(4) $y = 1 + A \exp(-x/a)$ (Popescu).

Both functions have been tested on Italian presidential texts and function (4) yielded in all cases a better fit.

Table 5
Rank-frequency sequences of POS in librettos

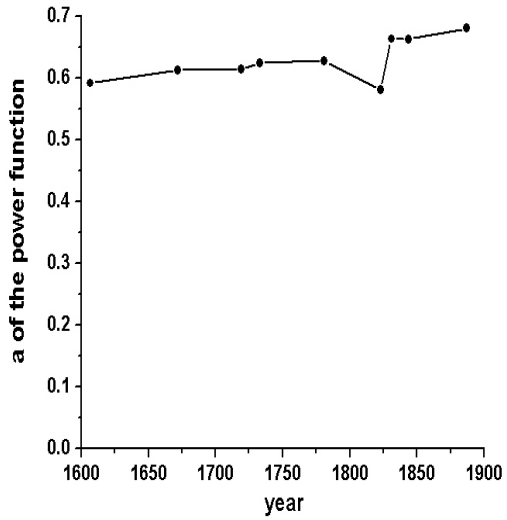
Rank	L1	L2	L3	L4	L5	L6	L7	L8	L9
1	847	1740	1632	1959	957	991	1035	942	1342
2	598	1737	1561	1560	775	934	846	903	1209
3	516	1210	1215	1403	651	845	773	615	968
4	498	931	821	921	463	598	434	446	643
5	441	870	792	900	413	480	372	403	540
6	349	705	695	768	338	393	367	391	401
7	311	574	547	676	309	333	318	304	369
8	255	480	508	458	234	256	199	152	293
9	79	253	123	234	159	192	193	99	128
10	72	107	116	216	121	97	88	77	114

The fitting of (3) and (4) can be performed iteratively. As can be seen in Table 5 the exponential fitting is in each case better than the power function fitting. At the same time it expresses the fact that parts of speech form an acceptable classification consisting of one stratum only (cf. Tuzzi, Popescu, Altmann 2010: 112 ff).

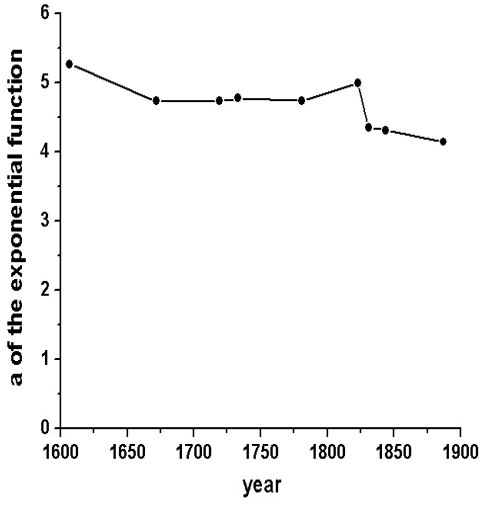
Table 5
Fitting (3) and (4) to librettos (1607-1887)

	Zipf fitting: $y = c/x^a$			**Exp. fitting: $y = 1 + A \cdot \exp(-x/a)$**		
Libretto	**c**	**a**	**R^2**	**A**	**a**	**R^2**
L1 Striggio	894.0	0.5914	0.8441	981.3	5.2629	0.9322
L2 Aureli	2005.8	0.6128	0.8130	2325.1	4.7316	0.9576
L3 Zeno	1870.8	0.6143	0.8016	2166.9	4.7309	0.9462
L4 Metastasio	2139.0	0.6243	0.8709	2429.4	4.7704	0.9743
L5 Varesco	1041.8	0.6277	0.9051	1181.7	4.7314	0.9923
L6 Rossi	1146.7	0.5804	0.7977	1323.1	4.9957	0.9642
L7 Romani	1139.9	0.6628	0.8697	1328.2	4.3508	0.9664
L8 Piave	1072.3	0.6622	0.8304	1265.6	4.3052	0.9587
L9 Boito	1519.6	0.6798	0.8474	1815.9	4.1411	0.9787

From the historical point of view a development of the rank-frequency sequence can be observed: The parameter *a* in both fittings displays an almost linear trend with the same outlier (L6 Rossi, which is a short libretto featuring a high frequency of verbs, concurrently with a low frequency of types, Overbeck 2009: chapter 3.1.1. and 3.2.1.) in both cases: a positive one with the power function and a negative one with the exponential function as can be seen in Figure 1.



1a. Parameter *a* in the power function



1b. Parameter *a* in the exponential function

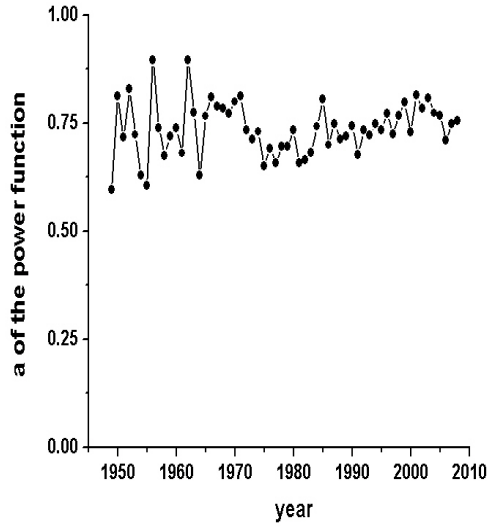
Figure 1. Historical trend of the parameter *a* in librettos (1607-1887)

In order to follow this trend we compare it with the results won from the analysis of presidential speeches (1949-2008). Using the data presented in Tuzzi, Popescu, Altmann (2010: 118-120) and shown in Table 6 we can state that the developmental trend stabilized and forms a horizontal straight line as can be seen in Figure 2.

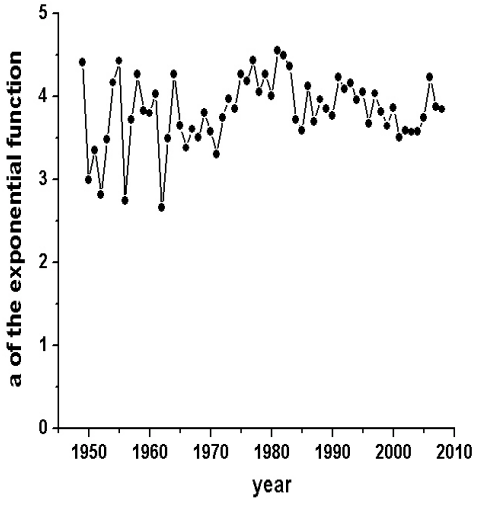
Table 6
The fitting of (3) and (4) to presidential speeches (1949-2008)

	Zipf fitting: $y = c/x^a$			**Exp fitting: $y = 1 + A \cdot \exp(-x/a)$**		
Text	*c*	*a*	R^2	*A*	*a*	R^2
1949Einaudi	47.07	0.5966	0.7315	55.4131	4.4061	0.9054
1950Einaudi	46.08	0.8131	0.8788	59.6584	2.9894	0.9664
1951Einaudi	44.26	0.7185	0.7681	56.1011	3.3544	0.9106
1952Einaudi	43.05	0.8300	0.8511	57.2212	2.8153	0.9477
1953Einaudi	53.39	0.7232	0.8121	66.9476	3.4773	0.9595
1954Einaudi	65.39	0.6288	0.7780	78.2885	4.1636	0.9330
1955Gronchi	94.85	0.6055	0.7765	112.0836	4.4270	0.9300
1956Gronchi	160.67	0.8950	0.9203	214.7349	2.7377	0.9597
1957Gronchi	297.24	0.7397	0.8308	364.2917	3.7208	0.9584
1958Gronchi	223.42	0.6749	0.8013	262.1934	4.2630	0.9408
1959Gronchi	192.28	0.7195	0.8828	228.9571	3.8226	0.9481
1960Gronchi	215.74	0.7381	0.8441	260.2692	3.7961	0.9645
1961Gronchi	329.68	0.6805	0.8636	393.3201	4.0293	0.9669
1962Segni	180.13	0.8950	0.8879	248.7756	2.6638	0.9816
1963Segni	281.31	0.7741	0.8393	352.5834	3.4962	0.9769
1964Saragat	112.73	0.6301	0.7636	134.6237	4.2649	0.9270
1965Saragat	289.04	0.7664	0.8653	350.8234	3.6439	0.9615
1966Saragat	350.29	0.8100	0.8867	436.2819	3.3792	0.9840
1967Saragat	292.45	0.7891	0.8406	358.7014	3.6043	0.9747
1968Saragat	332.6	0.7834	0.8779	410.9684	3.5014	0.9798
1969Saragat	429.42	0.7727	0.8389	512.3810	3.8062	0.9575
1970Saragat	541.52	0.7995	0.8561	662.1603	3.5796	0.9738
1971Leone	74.66	0.8131	0.8649	92.6163	3.3028	0.9644
1972Leone	203.65	0.7339	0.7917	250.6557	3.7467	0.9483
1973Leone	330.42	0.7133	0.8157	396.2084	3.9666	0.9510
1974Leone	216.28	0.7305	0.8063	260.6625	3.8525	0.9394
1975Leone	338.22	0.6500	0.8354	398.4730	4.2682	0.9506
1976Leone	350.97	0.6922	0.8137	412.1600	4.1778	0.9371
1977Leone	392.82	0.6576	0.7886	456.3595	4.4352	0.9205

1978Pertini	363.17	0.6967	0.8141	434.2952	4.0518	0.9481
1979Pertini	570.71	0.6963	0.7846	671.3088	4.2701	0.9402
1980Pertini	351.56	0.7338	0.8197	417.2399	4.0059	0.9573
1981Pertini	667.3	0.6584	0.7659	774.9393	4.5525	0.9208
1982Pertini	591.23	0.6646	0.7873	689.4811	4.4881	0.9476
1983Pertini	909.84	0.6817	0.7916	1068.0115	4.3566	0.9469
1984Pertini	323.95	0.7425	0.8219	398.3592	3.7149	0.9640
1985Cossiga	667.32	0.8046	0.8502	812.9602	3.5882	0.9670
1986Cossiga	349.87	0.6997	0.8180	412.0190	4.1275	0.9426
1987Cossiga	550.19	0.7489	0.8294	676.8828	3.6957	0.9666
1988Cossiga	618.77	0.7124	0.8265	744.3419	3.9614	0.9606
1989Cossiga	496.24	0.7193	0.8255	605.9822	3.8465	0.9673
1990Cossiga	869.28	0.7428	0.8439	1058.5020	3.7696	0.9722
1991Cossiga	103.02	0.6770	0.8091	120.0938	4.2254	0.9410
1992Scalfaro	717.94	0.7338	0.8168	840.8369	4.0915	0.9418
1993Scalfaro	753.09	0.7218	0.8171	880.3619	4.1586	0.9453
1994Scalfaro	953.87	0.7475	0.8247	1131.8520	3.9566	0.9498
1995Scalfaro	1093.9	0.7346	0.8327	1288.6011	4.0523	0.9575
1996Scalfaro	559.62	0.7727	0.8465	675.5472	3.6738	0.9426
1997Scalfaro	1265.5	0.7248	0.8116	1508.0736	4.0325	0.9437
1998Scalfaro	1067.4	0.7673	0.8594	1273.9222	3.8173	0.9653
1999Ciampi	546.17	0.7988	0.8544	659.4942	3.6409	0.9634
2000Ciampi	467.53	0.7286	0.8146	566.8884	3.8555	0.9542
2001Ciampi	599.44	0.8146	0.8610	736.5069	3.5044	0.9746
2002Ciampi	591.54	0.7840	0.8629	720.7311	3.5887	0.9622
2003Ciampi	445.42	0.8073	0.8603	541.8572	3.5698	0.9703
2004Ciampi	492.24	0.7735	0.8489	606.5151	3.5775	0.9653
2005Ciampi	325.46	0.7683	0.8320	395.4018	3.7380	0.9719
2006Napolitano	556.45	0.7099	0.8040	649.5857	4.2250	0.9381
2007Napolitano	471.12	0.7486	0.8191	567.3769	3.8705	0.9584
2008Napolitano	459.23	0.7550	0.8194	553.2141	3.8427	0.9571



2a. Parameter *a* of the power function in presidential speeches



2b. Parameter *a* of the exponential function in presidential speeches

Figure 2. Historical trend of the parameter *a* in presidential speeches

As can be seen in Figure 2 even the oscillation decreases and the parameter *a* flows in a narrow interval. However, even under these advantageous conditions no prediction of the future development can be made since language steadily rearranges its means. Besides, the classification of words in POS is a matter of definition and the choice of the criterion (syntactic or semantic).

3. The ternary plot

Every rank-frequency sequence can be characterised by a three-component vector $U(V,f_1,L)$ containing the three basic quantities, viz. V – the inventory of pertinent units being identical with the highest rank of the sequence, f_1 – the frequency of the most frequent unit in the sample, and L – the length of the arc joining f_1 and V consisting of the sum of Euclidian distances between neighbouring frequencies, i.e.

$$L = \sum_{r=1}^{V-1} D_r = \sum_{r=1}^{V-1} [(f_r - f_{r+1})^2 + 1]^{1/2}, \tag{5}$$

where D_r are the Euclidian distances between the individual frequencies ranked ordinally (cf. Popescu et al. 2010). There are also other possibilities of characterising, e.g. using the so-called *h*-point instead of L. Since the presentation of this vector in a three-dimensional plot would be rather obscure, one usually tries to set up a normalized vector

$$U(x,y,z), \tag{6}$$

such that $x + y + z = 1$. To this end one uses the data of the sample and performs a series of transformations. The first transformation concerns the frequencies themselves. If the rank-frequency sequence is too short – and we have this case with the POS where $V_{max} = 11$ – then

one of the data in the sample having great *N* can distort the whole picture. Thus it is advisable to transform all frequencies in the following way

(7) $f^*_r = f_r - f_V + 1$

i.e. to subtract from each frequency that of the smallest frequency (that of the last class) and add 1. In this way the new $f^*_V = 1$ and the influence of *N* is eliminated. This transformation is not necessary e.g. with data concerning word frequencies where there are always words having frequency 1.

The second transformation is performed in the following way

(8) $$X = \frac{V - V_{min}}{V_{\max} - V_{min}}, \quad Y = \frac{f_1^* - f_{1,min}^*}{f_{1,\max}^* - f_{1,min}^*}, \quad Z = \frac{L - L_{min}}{L_{max} - L_{min}}.$$

taking the minima and the maxima from the sample because there is e.g. no general maximum of f_1. In our joined data of librettos and presidential speeches we have

$V_{min} = 8$ $V_{max} = 11$ (i.e. at least 8 POS classes and at most 11)
$f^*_{1,\min} = 40$ $f^*_{1,\max} = 1744$
$L_{\min} = 40.04$ $L_{\max} = 1743.071$

that is, we obtain the values of *X, Y* and *Z* as

$$X = \frac{V - 8}{11 - 8} = \frac{V - 8}{3}, \qquad Y = \frac{f_r^* - 40}{1744 - 40} = \frac{f_r^* - 40}{1704},$$
$$Z = \frac{L - 40.04}{1743.071 - 40.04} = \frac{L - 40.04}{1702.67}.$$

The last transformation yields a vector whose sum of components is equal to 1, namely

$$x = \frac{X}{X + Y + Z}, \quad y = \frac{Y}{X + Y + Z}, \quad z = \frac{Z}{X + Y + Z}.$$

The results of computation of *x,y,z* for the librettos and the presidential speeches are presented in Table 7 and ordered according to *x*. It is to be remarked that if new data will be added and some minima or maxima are more extreme than here, the whole computation must be made anew.

Table 7
The computation of the normalized vector U

Text	V	f_1*	L	X	Y	Z	Sum	x	y	z
1956Gronchi	8	155	154.59	0.0000	0.0675	0.0673	0.1348	0.0000	0.5008	0.4992
1962Segni	8	168	167.22	0.0000	0.0751	0.0747	0.1498	0.0000	0.5015	0.4985
L4 Metastasio	10	1744	1743.07	0.6667	1.0000	1.0000	2.6667	0.2500	0.3750	0.3750

L2 Aureli	10	1634	1633.19	0.6667	0.9354	0.9355	2.5376	0.2627	0.3686	0.3686
L3 Zeno	10	1517	1516.12	0.6667	0.8668	0.8667	2.4002	0.2778	0.3611	0.3611
L9 Boito	10	1229	1228.08	0.6667	0.6978	0.6976	2.0620	0.3233	0.3384	0.3383
L7 Romani	10	948	947.20	0.6667	0.5329	0.5327	1.7322	0.3849	0.3076	0.3075
L6 Rossi	10	895	894.05	0.6667	0.5018	0.5015	1.6699	0.3992	0.3005	0.3003
L8 Piave	10	866	865.11	0.6667	0.4847	0.4845	1.6359	0.4075	0.2963	0.2962
L5 Varesco	10	837	836.07	0.6667	0.4677	0.4674	1.6018	0.4162	0.2920	0.2918
1990Cossiga	10	783	782.12	0.6667	0.4360	0.4357	1.5384	0.4333	0.2834	0.2832
L1 Striggio	10	776	775.15	0.6667	0.4319	0.4316	1.5302	0.4357	0.2823	0.2821
1997Scalfaro	11	1104	1103.11	1.0000	0.6244	0.6242	2.2486	0.4447	0.2777	0.2776
1995Scalfaro	11	992	991.12	1.0000	0.5587	0.5585	2.1171	0.4723	0.2639	0.2638
1998Scalfaro	11	968	967.29	1.0000	0.5446	0.5445	2.0891	0.4787	0.2607	0.2606
1994Scalfaro	11	866	865.30	1.0000	0.4847	0.4846	1.9693	0.5078	0.2461	0.2461
1975Leone	9	311	310.21	0.3333	0.1590	0.1586	0.6510	0.5120	0.2443	0.2437
1961Gronchi	9	303	302.21	0.3333	0.1543	0.1539	0.6416	0.5195	0.2406	0.2399
1988Cossiga	10	553	552.17	0.6667	0.3011	0.3007	1.2684	0.5256	0.2373	0.2371
2002Ciampi	10	550	549.31	0.6667	0.2993	0.2990	1.2650	0.5270	0.2366	0.2364
1983Pertini	11	784	783.11	1.0000	0.4366	0.4363	1.8729	0.5339	0.2331	0.2330
1996Scalfaro	10	525	524.32	0.6667	0.2846	0.2844	1.2357	0.5395	0.2303	0.2301
1987Cossiga	10	491	491.19	0.6667	0.2647	0.2649	1.1962	0.5573	0.2213	0.2214
1993Scalfaro	11	684	683.22	1.0000	0.3779	0.3777	1.7556	0.5696	0.2153	0.2151
1992Scalfaro	11	655	654.92	1.0000	0.3609	0.3610	1.7220	0.5807	0.2096	0.2097
2004Ciampi	10	448	447.37	0.6667	0.2394	0.2392	1.1453	0.5821	0.2091	0.2088
1989Cossiga	10	436	435.54	0.6667	0.2324	0.2322	1.1313	0.5893	0.2054	0.2053
1985Cossiga	11	611	610.58	1.0000	0.3351	0.3350	1.6701	0.5988	0.2006	0.2006
1981Pertini	11	571	571.24	1.0000	0.3116	0.3119	1.6235	0.6159	0.1919	0.1921
2001Ciampi	11	548	547.35	1.0000	0.2981	0.2979	1.5960	0.6266	0.1868	0.1866
1977Leone	10	357	356.66	0.6667	0.1860	0.1859	1.0386	0.6419	0.1791	0.1790
1982Pertini	11	508	507.19	1.0000	0.2746	0.2743	1.5490	0.6456	0.1773	0.1771
1999Ciampi	11	504	503.26	1.0000	0.2723	0.2720	1.5443	0.6475	0.1763	0.1761
2006Napolitano	11	502	501.40	1.0000	0.2711	0.2709	1.5420	0.6485	0.1758	0.1757
1979Pertini	11	498	498.18	1.0000	0.2688	0.2690	1.5378	0.6503	0.1748	0.1749
1970Saragat	11	489	488.70	1.0000	0.2635	0.2634	1.5269	0.6549	0.1726	0.1725
1966Saragat	10	323	322.32	0.6667	0.1661	0.1657	0.9985	0.6677	0.1663	0.1660
1978Pertini	10	323	322.28	0.6667	0.1661	0.1657	0.9985	0.6677	0.1663	0.1660
1959Gronchi	9	181	180.64	0.3333	0.0827	0.0826	0.4986	0.6685	0.1659	0.1656
1986Cossiga	10	321	321.34	0.6667	0.1649	0.1652	0.9968	0.6688	0.1654	0.1657
1976Leone	10	321	320.77	0.6667	0.1649	0.1648	0.9964	0.6691	0.1655	0.1654
1968Saragat	10	303	302.26	0.6667	0.1543	0.1540	0.9750	0.6838	0.1583	0.1579
1973Leone	10	298	298.20	0.6667	0.1514	0.1516	0.9697	0.6875	0.1561	0.1563

2000Ciampi	11	421	420.20	1.0000	0.2236	0.2232	1.4468	0.6912	0.1545	0.1543
2007Napolitano	11	417	416.47	1.0000	0.2212	0.2210	1.4423	0.6933	0.1534	0.1533
2008Napolitano	11	409	408.84	1.0000	0.2165	0.2166	1.4331	0.6978	0.1511	0.1511
1984Pertini	10	286	285.49	0.6667	0.1444	0.1441	0.9552	0.6980	0.1511	0.1509
2003Ciampi	11	408	407.86	1.0000	0.2160	0.2160	1.4319	0.6984	0.1508	0.1508
1969Saragat	11	393	392.74	1.0000	0.2072	0.2071	1.4143	0.7071	0.1465	0.1464
1965Saragat	10	265	264.88	0.6667	0.1320	0.1320	0.9307	0.7163	0.1419	0.1418
1957Gronchi	10	263	262.66	0.6667	0.1309	0.1307	0.9283	0.7182	0.1410	0.1408
1963Segni	10	250	249.27	0.6667	0.1232	0.1229	0.9128	0.7304	0.1350	0.1346
1980Pertini	11	315	314.76	1.0000	0.1614	0.1613	1.3227	0.7560	0.1220	0.1220
2005Ciampi	11	290	289.44	1.0000	0.1467	0.1464	1.2932	0.7733	0.1135	0.1132
1958Gronchi	10	201	200.54	0.6667	0.0945	0.0942	0.8554	0.7794	0.1105	0.1102
1974Leone	10	197	197.47	0.6667	0.0921	0.0924	0.8512	0.7832	0.1082	0.1086
1960Gronchi	10	195	194.69	0.6667	0.0910	0.0908	0.8484	0.7858	0.1072	0.1070
1967Saragat	11	262	261.67	1.0000	0.1303	0.1301	1.2604	0.7934	0.1034	0.1033
1972Leone	10	180	180.39	0.6667	0.0822	0.0824	0.8312	0.8020	0.0988	0.0991
1964Saragat	9	100	99.85	0.3333	0.0352	0.0351	0.4037	0.8258	0.0872	0.0870
1955Gronchi	9	83	83.11	0.3333	0.0252	0.0253	0.3839	0.8684	0.0657	0.0659
1991Cossiga	10	94	93.79	0.6667	0.0317	0.0316	0.7299	0.9133	0.0434	0.0432
1954Einaudi	9	57	57.16	0.3333	0.0100	0.0101	0.3534	0.9433	0.0282	0.0285
1971Leone	10	69	69.19	0.6667	0.0170	0.0171	0.7008	0.9513	0.0243	0.0244
1953Einaudi	9	47	46.85	0.3333	0.0041	0.0040	0.3414	0.9763	0.0120	0.0117
1950Einaudi	9	42	43.00	0.3333	0.0012	0.0017	0.3362	0.9913	0.0035	0.0052
1949Einaudi	9	41	41.26	0.3333	0.0006	0.0007	0.3346	0.9961	0.0018	0.0021

In Table 7 1951Einaudi and 1952Einaudi are missing because they have extreme values, namely $U(x,y,z) = U(0,0,1)$. The last three columns in Table 7 can be presented graphically in two dimensions as is done in Figure 3.

As can be seen both in Table 7 and in Figure 3 the librettos occupy a special domain. Though the form of the ternary plot is very characteristic for classes with a small number of elements, the position of librettos is, perhaps, a sign of difference in style. A comparison with other texts could contribute to verify this.

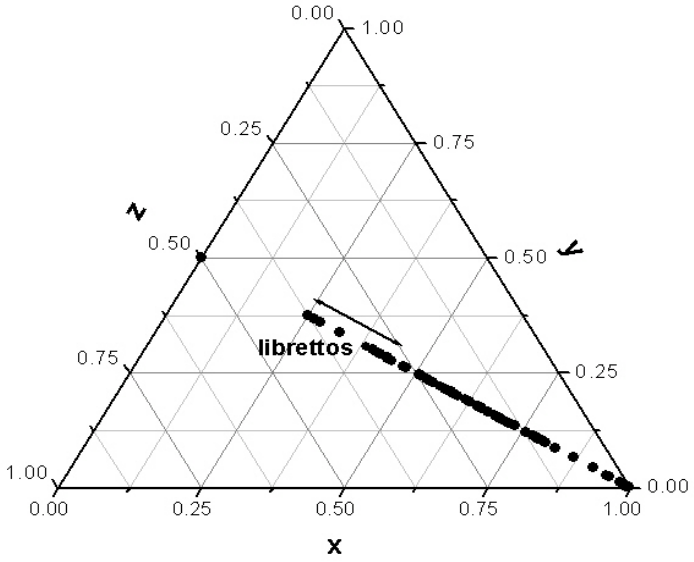


Figure 3. Ternary plot of POS characteristics

If one takes the normalized components in pairs, it can easily be shown that they are in a perfect linear relationship. Since *y* and *z* are almost identical, it is sufficient to show the relationship $y = f(x)$. In that case we obtain the result as presented in Figure 4.

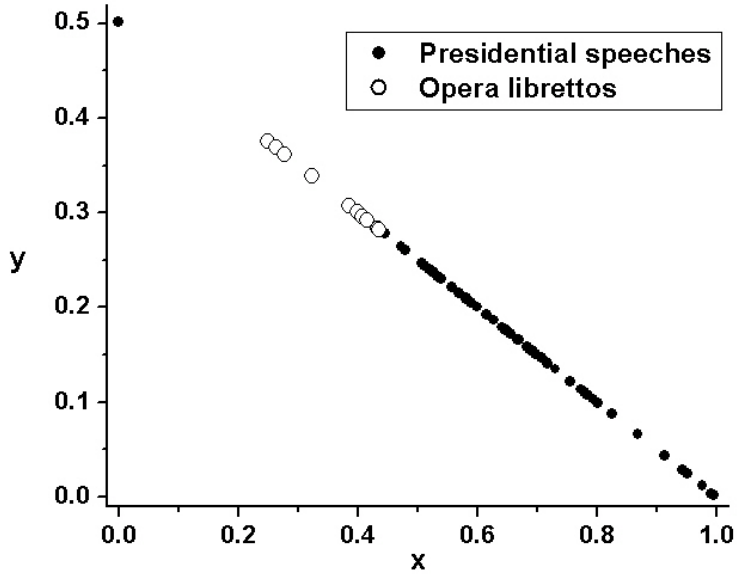


Figure 4. The relationship $y = f(x)$

4. The golden section

In the rank-frequency sequence another entity can be found which is mostly called *the golden section*. In order to show it one must first compute the so-called *h*-point separating in a fuzzy way the autosemantics from synsemantics. It is defined as

$$h = \begin{cases} r, & \text{if there is an } r = f(r) \\ \dfrac{f(i)r_j - f(j)r_i}{r_j - r_i + f(i) - f(j)}, & \text{if there is no } r = f(r) \end{cases} \quad (9)$$

that is, either it can be read directly from the rank-frequency sequence (if there is an $r = f(r)$) or it is computed from two (mostly) neighbouring ranks and frequencies. It is to be noted that the frequencies are transformed according to (7). For different uses of the *h*-point cf. Popescu et al. (2009), Popescu, Mačutek, Altmann (2009), Tuzzi, Popescu, Altmann (2010). Let us illustrate the procedure by an example. We take libretto L1 Striggio in which we have the individual frequencies

516,349,255,311,79,847,72,441,498,598, N = 3966.

Reordering them we obtain

1	2	3	4	5	6	7	8	9	10
847	598	516	498	441	349	311	255	79	72

Performing the transformation (7) $f(r)^* = f(r) - f(V) + 1$, i.e. subtracting 72 and adding 1 to each frequency we obtain

1	2	3	4	5	6	7	8	9	10
776	527	442	427	370	278	218	184	8	1

Evidently, we must compute the second part of formula (9). Taking $r_i = 8$, $r_j = 9$, $f(8) = 184$, $f(9) = 8$ and inserting them in the second part of formula (9) we obtain

$h = [184(9)-8(8)]/[9-8+184-8] = 8.99 \approx 9$

Since *h, N, L* are interrelated in a special way described e.g. in Popescu, Mačutek, Altmann (2009: 144ff), Tuzzi, Popescu, Altmann (210: 103ff), it has been shown that the indicator

$$GS = (L_{\max} - L)\left(\frac{1}{\sqrt{N}} + \frac{1}{h-1}\right) \quad (10)$$

takes on values in the vicinity of the golden section 1.618… The results of the computation for the librettos are presented in Table 8. Here the maximal arc length is computed separately for each text as

$$L_{max} = V - 1 + f^*(1) - 1. \quad (11)$$

As can be seen, in librettos this ideal is not yet fulfilled. The difference to presidential speeches which can be found in Tuzzi, Popescu, Altmann (2010: 104) is very great. A comparison of both classes is shown in Figure 5.

Table 8
The indicator GS for the librettos

ID	N	V	f*(1)	h	L	L_{max}	GS
L1 Striggio	3966	10	776	9.00	775	784	1.2473
L2 Aureli	8607	10	1634	9.94	1633	1642	1.0803
L3 Zeno	8010	10	1517	9.00	1516	1525	1.2091
L4 Metastasio	9095	10	1744	9.53	1743	1752	1.1404
L5 Varesco	4420	10	837	9.77	836	845	1.1526
L6 Rossi	5119	10	895	9.91	894	903	1.1291
L7 Romani	4625	10	948	9.91	947	956	1.1171
L8 Piave	4332	10	866	9.61	865	874	1.1673
L9 Boito	6007	10	1229	9.41	1228	1237	1.1761

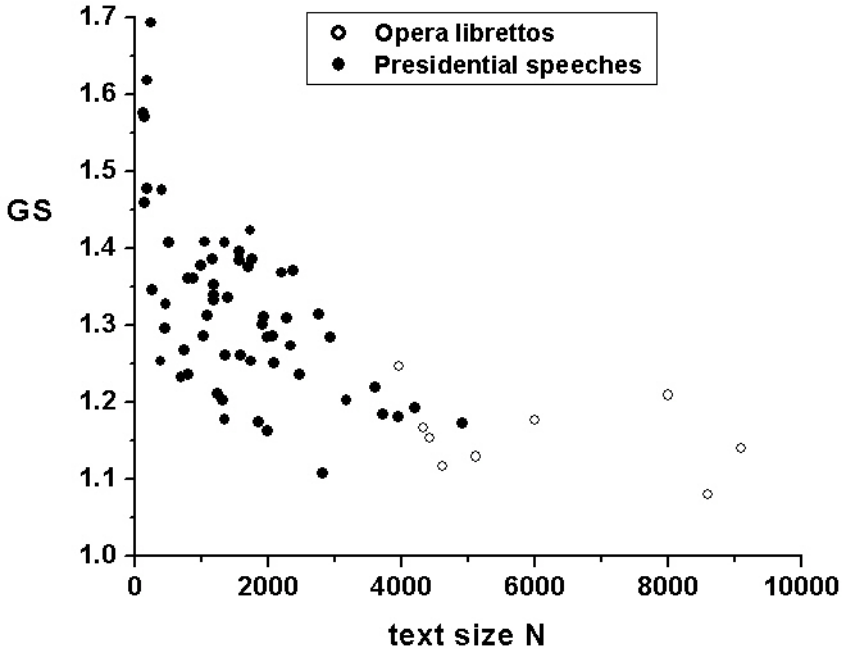


Figure 5. The indicator GS in librettos and in presidential speeches

The result can preliminarily be viewed from two perspectives: GS may be considered a property which either strongly develops or strongly distinguishes genres. If we consider the texts regardless of difference in time and genre, we see that the longer the text, the more it moves away from the golden section. This hypothesis is plausible because by enlarging the text, the authors stepwise lose the intuitive idea of golden section or it becomes effaced by stratification represented by individual chapters (acts, persons). Thus in future investigations it would be more appropriate to consider separately the individual parts of the text.

5. Conclusions

The results show that parts of speech were in Italian a developing class that stabilized at least in a closed genre represented by texts having the same destination (presidential speeches). In librettos dispersed in three centuries and thus underlying many cultural and esthetical changes, there is much less homogeneity regarding parts of speech. Besides, the use of certain parts of speech is often depending on matters and contents of the respective libretto. Each literary period has its own thematical and stylistical focuses, influencing a relatively open genre like that of the opera libretto. The first examples of the emerging genre (beginning with Ottavio Rinuccini's *La Dafne*, 1598) show a very different linguistic organisation than e.g. the librettos of the famous operas by Verdi. Thus parts of speech are not only syntactic or semantic but also stylistic means.

The good fitting of the exponential to the rank-frequency sequence shows that POS is a prolific class whose members are in mutual equilibrium state, i.e. they build a class with a Zipfian background mechanism controlling their occurrence. They are part of self-regulation. The ternary plot shows a very expressed mechanism of control.

Futher, the POS show the influence of sample size *N* on the disturbing of the golden section.

Needless to say, two sets of coherent data are not enough to show the structuring of POS even in one language in all of its aspects. Further examinations both in Italian and in other languages are necessary to corroborate or reject this conjectured phenomenon.

References

Overbeck, A. (2009), *Con lingua più gentile / qui si parla d'amor». Quantitative und qualitative Studien zu Lexik, Syntax und Stil italienischer Opernlibretti*. University of Göttingen, Professoral Dissertation (printing in preparation).

Popescu, I.-I. et al. (2009). *Word frequency studies*. Berlin-New York: Mouton de Gruyter.

Popescu, I.-I. et al. (2010). *Vectors and codes of texts* (submitted).

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

Tuzzi, A., Popescu, I.-I., Altmann, G. (2010), *Quantitative analysis of Italian texts*. Lüdenscheid: RAM-Verlag.

Glottometrics 20, 2010, 29-33

Ein Modell für die Zunahme chinesischer Schriftzeichen

Karl-Heinz Best[1]
Jinyang Zhu

Abstract. The purpose of this paper is to present some further evidence for the validity of the Piotrowski law (logistic law) in language change. To this end we test some data concerning the increase of the number of Chinese characters.

Keywords: Chinese, script evolution

1. Verlauf von Sprachwandel

Ein Themenkomplex, dem sich die Quantitative Linguistik in den letzten 3 Jahrzehnten immer wieder gewidmet hat, ist die Frage danach, wie Sprachwandel verläuft und ob er, wenn er denn einmal stattfindet, irgendwelchen Gesetzmäßigkeiten folgt. Eine Theorie dazu, welche Gesetzmäßigkeiten infrage kommen, haben Altmann (1983) sowie Altmann, von Buttlar, Rott & Strauß (1983) vorgelegt, die mit drei unterschiedlichen Typen von Sprachwandel rechnen: dem vollständigen Sprachwandel, bei dem alle alten Formen durch neue ersetzt werden, dem unvollständigen Sprachwandel, bei dem sich sprachliche Neuerungen bis zu einem gewissen Grad durchsetzen, und dem reversiblen Sprachwandel, bei dem sich eine einmal eingeschlagene Entwicklung irgendwann wieder umkehrt. Untersuchungen zu verschiedenen Phänomenen (Wandel von Sprachsystem und –verwendung, Entlehnungen, Spracherwerb) in unterschiedlichen Sprachen haben diese Annahmen immer wieder bestätigt (Best 2003). Diese Gesetzmäßigkeit ist in der Quantitativen Linguistik zu Ehren der St. Petersburger Forscher Piotrovskaja und Piotrovski (Piotrovskaja & Piotrovski 1974), die wohl als erste ein mathematisches Modell für Sprachwandel entwickelten, als „Piotrowski-Gesetz“ bekannt; es reiht sich ein in die Wachstumsgesetze, die in den verschiedensten Wissenschaften seit fast 200 Jahren bekannt sind und auf Modelle zur Bevölkerungsentwicklung zurückgehen (Verhulst 1838/45).

Zum Sprachwandel des Chinesischen wurden bisher drei Phänomene untersucht: die Ausbreitung anlautender stimmloser Obstruenten, der Zuwachs an Schriftzeichen und die Zunahme der durchschnittlichen Wortlänge (Best & Zhu 2006). In allen drei Fällen ließ sich zeigen, dass die Prozesse sich gesetzmäßig entwickeln. In diesem Beitrag geht es noch einmal um die Frage, ob auch der Zuwachs an Schriftzeichen im Chinesischen, der sich über mehr als 2 Jahrtausende verfolgen lässt. Der Grund für diese thematische Wiederaufnahme besteht darin, dass uns jetzt mehr Daten bekannt sind und wir einen etwas anderen Ansatz bei der Aufnahme von Daten wählen wollen.

2. Zunahme chinesischer Schriftzeichen

Angaben zur Zahl chinesischer Schriftzeichen in den verschiedenen Entwicklungsstadien findet man bei mehreren Autoren (Coulmas 1982: 85, Dürscheid 2006: 70; Fazzioli 2003:

[1] Address correspondence to: kbest@gwdg.de

17/19, Kaeding 1897/98: S. 651f., Li (1996: 1409), Schindelin 2005a: 96f., Schindelin 2005b: 948).

Die Daten fallen recht unterschiedlich aus. So berichtet Kaeding von Zählungen zum Chinesischen, die bereits im 19. Jahrhundert von anderen Autoren veröffentlicht wurden, darunter zur Bibel, die insgesamt einen Umfang von 676827 Schriftzeichen habe, von denen 4182 verschiedene Schriftzeichen seien und macht darauf aufmerksam, dass die letztere Zahl bei verschiedenen Übersetzungen unterschiedlich ausfalle. In Werken der klassischen chinesischen Literatur seien 6544 verschiedene Schriftzeichen enthalten, von denen jedoch viele selten und veraltet seien, so dass „zwischen 5000 und 6000 verschiedene Schriftzeichen" (Kaeding 1897/98: 952) übrig blieben.

Die Angaben der übrigen Autoren stimmen nicht immer in den absoluten Zahlen, wohl aber in der Dimension einigermaßen überein. Wir stützen uns hier auf Daten zu Wörterbüchern, die in der Fachliteratur zu finden sind. Dabei wurde darauf geachtet, solche Angaben zu berücksichtigen, bei denen eine Aussicht besteht, die Gesamtmenge der Schriftzeichen zur jeweiligen Zeit zu erfassen.

Diskrepanzen zwischen den verschiedenen Quellen kommen besonders bei den Angaben für das 20. Jahrhundert in den Blick. So werden für drei Wörterbücher dieses Jahrhunderts zwischen 60000 und über 80000 Schriftzeichen angegeben. Diese hohen Zahlen für die Gegenwart erklären sich nach Schindelin (2005b: 948) dadurch, dass darin noch jede „peripherste Variante" aufgelistet wurde. Nach Beseitigung von Allographen und Fehlschreibungen komme man auf etwa 27000 Zeichen, von denen ein großer Teil nicht gebräuchlich sei. Schindelin zitiert darauf bezogen, „dass der maximale Umfang des Schriftzeicheninventars zwischen 9.742 und 20.107 liegen müsste."

Die beschriebene Auswahl der Daten hat also einen Nachteil: Sie überzeichnet die Zahl der Schriftzeichen. Das bereits von Kaeding angeschnittene Problem, dass nämlich ein beträchtlicher Teil der Zeichen veraltet ist, charakterisiert auch Tabelle 1. Es wäre wünschenswert, eine ähnliche Zusammenstellung zu haben, in der nur die tatsächlich noch gebräuchlichen Zeichen der jeweiligen Zeitabschnitte enthalten wären. Leider sind solche Angaben in gleicher Ausführlichkeit derzeit nicht verfügbar.

Nimmt man nun die Daten zur Zahl verschiedener Schriftzeichen in den unterschiedlichen Zeitabschnitten, so wie sie anhand entsprechender Wörterbücher ermittelt wurden, ergibt sich eine sehr gute Grundlage dafür, die Hypothese zu prüfen, dass die Bildung immer neuer Schriftzeichen dem *Piotrowski-Gesetz* in seiner Form für den unvollständigen Sprachwandel folgt.

Dieses Modell lautet nach Altmann (1983: 60f.):

$$(1) \qquad n = \frac{c}{1 + ae^{-kt}}$$

wobei n die Zahl der Schriftzeichen bezeichnet.

Die Tabelle ist aus den Angaben von Fazzioli (2003), He (2005), Li (1996) und Schindelin (2005a,b) zusammengestellt. Die Angaben für das 20. Jahrhundert, für das drei verschiedene Schätzwerte aufgenommen wurden, sind zeitlich differenziert; in allen anderen Fällen wurde auf solch eine Differenzierung verzichtet, da auch jeweils nur eine Angabe für den betreffenden Zeitraum vorliegt und nicht alle Daten als gesichert angesehen werden können. Die starke Streuung der Zahlenwerte wird in Kauf genommen. Im Grunde genommen gibt es nur einen Ausreißer, nämlich bei $t = 13$, der bereits das Niveau des $t = 20$ erreicht. Ohne diesen wäre die Kurve recht glatt, wie man in Abbildung 1 sehen kann.

Tabelle 1
Zunahme chinesischer Schriftzeichen

t	Datierung	Verfasser	Titel	Schriftzeichen beobachtet	Schriftzeichen berechnet
1	ca. 200 v. Chr.	Lī Sī	Sān Cāng	3300	6038.86
4	um 100 oder 121 n. Chr.	Xu Shen	Shuo wen jie zi	9353	9152.17
8	543	Gu Yewang	Yupian	12158	15558.12
9	601	Lu Fayan	Qieyun	16917	17662.58
13	1039	Ding Du usw.	Jiyun	53525	28425.58
14	12. Jahr-hundert 1161 vollendet	Zheng Qiao	Liu shu, in Tongzhi	24235	31715.43
19	1615	Mei Yingzuo	Zihui	33179	51081.56
20	1717	Zhang Yushu usw.	Kangxi zidian	48641	55318.09
22.38	1938	Wang Yunwu usw.	Zhongshan Dacidian	ca. 60000	65410.91
22.94	1994	Leng Yulong	Zhonghua zihai	85568	67738.01
22.95	1995		Quan Hanzi Shu	ca. 70000	67779.27
a = 20.9268 c = 115000 k = 0.1483 D = 0.81					

Legende zur Tabelle 1:
a, *c* und *k* sind die Parameter des Modells; *c* gibt an, auf welchen Zielwert der beobachtete Prozess bei gleichbleibender Tendenz hinsteuert. Die Anpassung des Modells an die beobachteten Daten ist mit *D* = 0.81 recht gut. (Der Determinationskoeffizient *D* soll mindestens 0.80 erreichen, um eine gute Übereinstimmung zwischen Modell und Beobachtungswerten anzuzeigen; er kann aber nicht größer als *D* = 1.00 werden.)

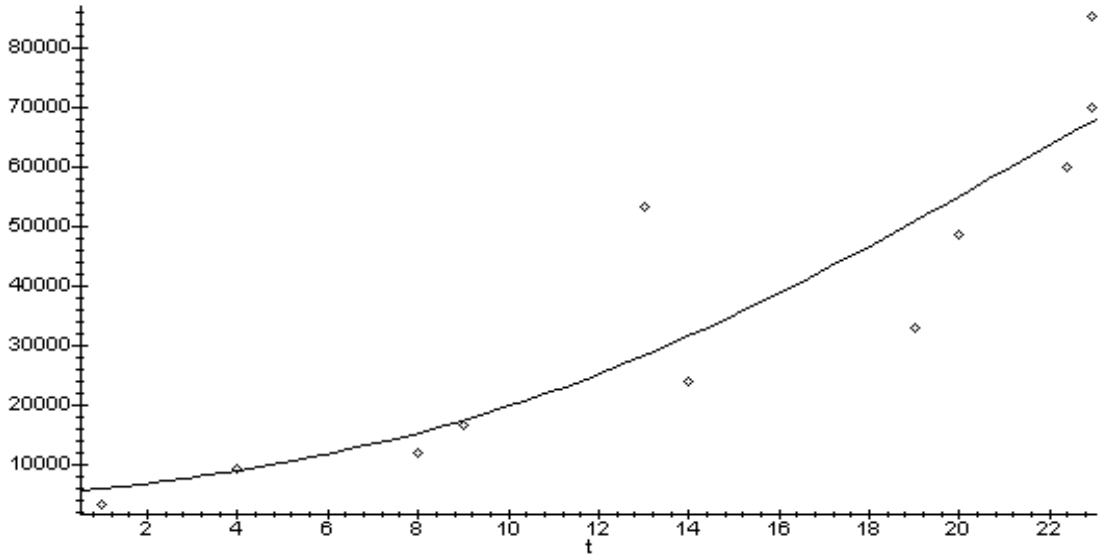


Abbildung 1. Zunahme chinesischer Schriftzeichen

3. Zusammenfassung

Zunächst kann festgestellt werden, dass das Piotrowski-Gesetz geeignet ist, den Zuwachs an unterschiedlichen Schriftzeichen im Chinesischen mit gutem Ergebnis zu modellieren. Der Trend der Entwicklung, die sich über 2300 Jahre erstreckt, wird sehr deutlich. Die Problematik der Daten wurde bereits angesprochen; deshalb hier nur ein ergänzender Hinweis: Dürscheid (2006: 70) gibt zur Zahl der Schriftzeichen an: „Heute besteht das gesamte Inventar an Schriftzeichen aus ca. 70000 Zeichen." Die Berechnung deutet an, dass sie damit vermutlich recht hat.

Der Parameter *c* ist, wie bereits oben gesagt, ein Schätzwert dafür, auf welche Gesamtzahl der beobachtete Prozess letztlich hinausläuft. Diesen Wert darf man nicht übertrieben interpretieren, vor allem aus folgenden Gründen: Bei der Anpassung des Modells musste der Wert für *c* versuchsweise eingesetzt werden; die Software NLREG hat ihn nicht aufgrund der Daten berechnet. Außerdem muss man sich dessen bewusst sein, dass die Zahl der tatsächlich gebräuchlichen Schriftzeichen wesentlich geringer ist als die Zahl der in der Literatur identifizierten Zeichen. Als drittes ist zu beachten, dass die chinesische Sprachpolitik darauf ausgerichtet ist, die Zahl der Schriftzeichen zu verringern (Coulmas 1982: 87f.). Und viertens, das Ansetzen eines endlichen Parameters *c* ist realistisch, da es keine unendlichen linguistischen Inventare gibt. Parameter *c* repräsentiert den status quo.

Abschließend sei darauf hingewiesen, dass nicht nur die Menge der chinesischen Schriftzeichen und ihre Zunahme die Forscher interessiert. Auch andere Aspekte der Schrift sind für die Quantitative Linguistik von Bedeutung. So hat Bohn (1998) untersucht, wie sich das Menzerath-Altmann-Gesetz („Je größer das Ganze, desto kleiner seine direkten Konstituenten.") bei chinesischem Sprachmaterial bewährt, darunter auch bei chinesischen Schriftzeichen. Yu (2001) wiederum hat an Texten untersucht, ob die Komplexität chinesischer Schriftzeichen sich auf die Häufigkeit ihrer Verwendung auswirkt und konnte die Wirksamkeit eines Verteilungsgesetzes nachweisen. Schindelin (2005b: 966) schließlich hat Köhlers Regelkreis, der anhand der Lexik des Deutschen entwickelt wurde (Köhler 1986: 74), so überarbeitet, dass er auf die Schriftzeichen angewendet werden kann. Eine Weiterentwicklung dieses Modells schlägt Altmann (2008: 160) vor, die aber noch nicht an chinesischem Material erprobt ist.

Literatur

Altmann, Gabriel (1983). Das Piotrowski-Gesetz und seine Verallgemeinerungen. In: Best, Karl-Heinz, & Kohlhase, Jörg (Hrsg.) (1983). *Exakte Sprachwandelforschung:* 54-90. Göttingen: edition herodot.

Altmann, Gabriel (2008). Towards a theory of script. In: Altmann, Gabriel, & Fan, Fengxiang (eds.), *Analyses of Script. Properties of Characters and Writing Systems*: *149-164.*. Berlin/New York: Mouton de Gruyter.

Altmann, Gabriel, von Buttlar, H., Rott, W., & Strauß, U. (1983). A law of change in language. In: Brainerd, B. (ed.), *Historical linguistics: 104-115*. Bochum: Brockmeyer.

Best, Karl-Heinz (2003). Spracherwerb, Sprachwandel und Wortschatzwachstum in Texten. Zur Reichweite des Piotrowski-Gesetzes. *Glottometrics 6, 9-34.*

Best, Karl-Heinz, & Zhu, Jinyang (2006). Sprachwandel im Chinesischen. *Archív Orientální 74, 203-214.*

Bohn, Hartmut (1998). *Quantitative Untersuchungen der modernen chinesischen Sprache und Schrift*. Hamburg: Kovač.

Coulmas, Florian (1982). *Über Schrift.* Frankfurt: Suhrkamp.

Dürscheid, Christa (2006). *Einführung in die Schriftlinguistik.* Göttingen: Vandenhoeck & Ruprecht

Fazzioli, Edoardo (2003). *Gemalte Wörter*. Wiesbaden: Fourier Verlag.

He, Jiuying (2005). Zhongguo gudai yuyanxue shi (= Geschichte der chinesischen Sprachwissenschaft im alten China). Guangzhou: Guangdong Jiaoyu.

Kaeding, F.W. (1897). *Häufigkeitswörterbuch der deutschen Sprache. 1. Teil: 37-42.* Steglitz bei Berlin: Selbstverlag des Herausgebers. (Auch: *Grundlagenstudien aus Kybernetik und Geisteswissenschaft 4/ 196, Beiheft.)*

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

Piotrovskaja, A.A., & Piotrovskij, R.G. (1974). Matematičeskie modeli diachronii i tekstoobrazovanija. In: *Statistika reči i avtomatičeskij analiz teksta: 361-400*. Leningrad: Nauka.

Schindelin, Cornelia (2005a). Zur Geschichte quantitativ-linguistischer Forschungen in China. In: Köhler, Reinhard, Altmann, Gabriel, & Piotrowski, Raimund (Hrsg.), *Quantitative Linguistik - Quantitative Linguistics. Ein internationales Handbuch: 96-115.* Berlin/ NewYork: de Gruyter.

Schindelin, Cornelia (2005b). Die quantitative Erforschung der chinesischen Sprache und Schrift. In: Köhler, Reinhard, Altmann, Gabriel, & Piotrowski, Raimund (Hrsg.), *Quantitative Linguistik - Quantitative Linguistics. Ein internationales Handbuch: 947-970.* Berlin/ NewYork: de Gruyter.

Verhulst, P.-F. (1838). Notice sur la loi que la population suit dans son accroissement. *Correspondance Mathématique et Physique, Tome X, 3-21*.

Verhulst, P.-F. (1845). Recherches mathématiques sur la loi d'accroissement de la population. *Nouveaux Mémoires de l'Académie Royale des Sciences et Belles-Lettres de Bruxelles, Tome XVIII, 5-38.*

Yu, Xiaoli (2001). Zur Komplexität chinesischer Schriftzeichen. *Göttinger Beiträge zur Sprachwissenschaft 5, 121-129.*

Verwendete Software

MAPLE V Release 4 (1996). Berlin u.a.: Springer.

NLREG. *Nonlinear Regression Analysis Program*. Ph. H. Sherrod. Copyright (c) 1991 - 2001.

Adresse des „Göttinger Projekts" im Internet (mit ausführlicher Bibliographie): http://wwwuser.gwdg.de/~kbest

Glottometrics 20, 2010, 34-37

Zur Entwicklung des Wortschatzes der deutschen Umgangssprache

Karl-Heinz Best, Göttingen

Abstract. The purpose of this paper is to present some further evidence for the validity of the Piotrowski law (logistic law) in language change. In this paper the domain of its application is the development of Colloquial German.

Keywords: German spoken lanbguage, word stock, Piotrowski law,

1. Gesetzmäßigkeit von Sprachwandel

Lebende Sprache verändern sich ständig; dieser Wandel vollzieht sich, soweit das bisher zu beobachten ist, immer gesetzmäßig. Seit den Arbeiten von Altmann (1983) und Altmann u.a. (1983) ist das zugrundeliegende Gesetz des Sprachwandels bekannt als „Piotrowski-Gesetz", das in drei Varianten den vollständigen, den unvollständigen und den reversiblen Sprachwandel modelliert. Der vollständige Sprachwandel kann ebenso wie der unvollständige als Zuwachs-, aber auch als Zerfallsprozess vorkommen; der reversible Wandel kann zuerst als Zuwachs, dann als Abnahme und auch umgekehrt beobachtet werden. Betroffen sind sowohl einzelne Erscheinungen der Sprache als auch ganze Teilsysteme, sowohl der Wortschatz als auch die Grammatik, die Schrift, die Orthographie und das phonetische bzw. phonologische System (zu etlichen dieser Phänomene vgl. Best 2003). Es geht in dieser vorliegenden Untersuchung nun darum, ein weiteres Sprachwandelphänomen in diese Betrachtungen einzubeziehen.

.

2. Zum Wortschatz der deutschen Umgangssprache

Ein spezieller Prozess, der bisher noch nicht darauf hin überprüft wurde, ob auch er dem Piotrowski-Gesetz unterliegt, ist die Entwicklung des umgangssprachlichen Wortschatzes im Deutschen. Es konnten bereits mehrere Untersuchungen zu anderen Aspekten des Wortschatzes durchgeführt werden: zur Ausbreitung von Entlehnungen und zur Entwicklung des Erbwortschatzes im Deutschen (Körner 2004) sowie zur Ausbreitung von gebundenen lexikalischen Morphemen und Affixen (u.a. Best 2007). Dabei handelt es sich immer um standardsprachliche Phänomene. Hier soll nun der umgangssprachliche Wortschatz anhand der Übersicht, die Meier (1967: 47) zu diesem Thema erarbeitet hat, auf seinen gesetzmäßigen Verlauf hin überprüft werden.

Meier (1967: 47) stellt in einer Graphik dar, aus welchem Jahrhundert ein wie hoher Anteil der „Begriffswörter" der gegenwärtigen deutschen Umgangssprache stammt; „Formwörter" wurden bei der Erhebung ausgeschlossen. Grundlage dieser Zusammenstellung ist Küppers *Wörterbuch der deutschen Umgangssprache* (1955, 1963), der den Versuch unternommen hatte herauszufinden, in welchem Jahrhundert die Wörter der Umgangssprache erstmals nachweisbar sind. Dieses Wörterbuch hat Meier ausgewertet und in Form einer Graphik

dargestellt. Meier weist auf eine Besonderheit des umgangssprachlichen Wortschatzes hin: Während im Häufigkeitswörterbuch der Standardsprache vier Fünftel der Wortformen aus alt- und mittelhochdeutscher Zeit stammt, ist in der Umgangssprache lediglich ein Prozent so alt (Meier 1967: 47). Der Wortschatz der Umgangssprache ist also verglichen mit dem der Standardsprache sehr jung.

3. Das mathematische Modell

Es geht nun darum, Meiers Befunde darauf hin zu prüfen, ob auch die Entwicklung des umgangssprachlichen Wortschatzes dem Sprachwandelgesetz folgt. Da seine Angaben in Prozent (und nicht in absoluten Zahlen) erfolgen, wurde das Piotrowski-Gesetz für den unvollständigen Sprachwandel in leicht abgewandelter Form als

$$(1) \qquad p = \frac{100}{1+ae^{-bt}}.$$

Altmann (1983: 60) getestet: Im Zähler steht eine 100 für 100%, da ja der zur Zeit der Untersuchung von Meier beobachtete umgangssprachliche Wortschatz mit 100% angesetzt werden muss.

4. Die gesetzmäßige Entwicklung der „Begriffswörter“ in der deutschen Umgangssprache

Legt man Modell (1) als Hypothese für die Entwicklung der Begriffswörter des umgangssprachlichen Wortschatzes zugrunde, dann erhält man das folgende Ergebnis für die Anpassung des Modells an die beobachteten Daten:

Tabelle 1
Entwicklung der Begriffswörter der deutschen Umgangssprache

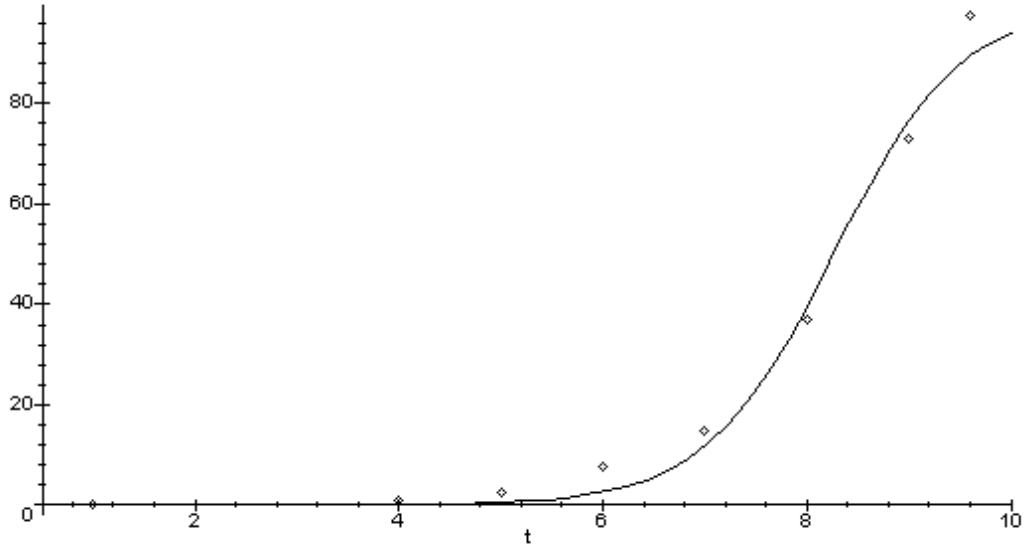
Jahrhundert	*t*	Wortschatzanteil beobachtet (in Prozent)	Wortschatzanteil kumuliert (in Prozent)	Wortschatzanteil berechnet
10. - 11.	1	0.1	0.1	0.00
12. – 14.	4	0.9	1.0	0.11
15.	5	1.4	2.4	0.54
16.	6	5.4	7.8	2.62
17.	7	7.1	14.9	11.76
18.	8	22.2	37.1	39.73
19.	9	36.0	73.1	76.52
20.	9.6	24.6	97.7	89.48
a = 542280.424 b = 1.5983 D = 0.99				

Legende zur Tabelle: Die erste Spalte benennt die Jahrhunderte, in denen die Begriffswörter erstmals belegt werden können; *t* sind die Zeitabschnitte, fortlaufend durchnummeriert; für das 20. Jahrhundert wurde statt $t = 10$ nur $t = 9.6$ angesetzt, da die Beobachtungen nur bis in

die 60er Jahre reichen. Mit „beobachtet“ wird angegeben, wie hoch der Anteil des umgangssprachlichen Wortschatzes ist, der dem betreffenden Zeitraum entstammt; „kumuliert“ summiert die Werte aus der Spalte „Wortschatzanteil beobachtet“ auf; „berechnet“ ist die Anzahl der Begriffswörter, die bei der Anpassung von Modell (1) an die kumulierten beobachteten Daten bestimmt wurden. *a* und *b* sind die Parameter. Die Anpassung von (1) an die beobachteten Daten ist mit $D = 0.99$ sehr gut. Das Piotrowski-Gesetz in seiner Version für den vollständigen Sprachwandel erweist sich damit ein weiteres Mal als erfolgreich, um auch diesen Sprachwandel zu modellieren.

In der Spalte „Wortschatzanteil kumuliert“ fällt auf, dass die Werte nicht, wie zu erwarten wäre, 100%, sondern nur 97.7% erreichen; wo die restlichen 2.3% bleiben, ist Meiers Ausführungen nicht zu entnehmen. Die Abweichung ist aber nicht als gravierend einzustufen und kann das Testergebnis nicht nennenswert beeinträchtigen.

Die folgende Graphik veranschaulicht den Verlauf des Sprachwandels und die gute Übereinstimmung zwischen berechneten und beobachteten Daten:



Graphik zu Tabelle 1: Entwicklung der Begriffswörter der deutschen Umgangssprache

5. Zusammenfassung

Die vorliegende Untersuchung auf der Grundlage der von Küpper & Meier erarbeiteten Daten zur Entwicklung des umgangssprachlichen Wortschatzes stützt erneut die Annahme, dass derartige Sprachwandel gesetzmäßig verlaufen und dabei dem Modell des vollständigen Sprachwandels folgen. Leider lässt sich nicht abschätzen, wie repräsentativ die Untersuchung tatsächlich ist, da Meier nur Prozentwerte, aber keine absoluten Zahlen zur Menge der beobachteten Daten bekanntgibt. Der Verlauf des Trends, besonders sein spätes Einsetzen, wird aber deutlich.

Literatur

Altmann, Gabriel (1983). Das Piotrowski-Gesetz und seine Verallgemeinerungen. In: Best, Karl-Heinz, & Kohlhase, Jörg (Hrsg.) (1983). *Exakte Sprachwandelforschung: 54-90.*

Göttingen: edition herodot.

Altmann, Gabriel, von Buttlar, H., Rott, W., & Strauß, U. (1983). A law of change in language. In: Brainerd, B. (ed.), *Historical linguistics: 104-115*. Bochum: Brockmeyer.

Best, Karl-Heinz (2003). Spracherwerb, Sprachwandel und Wortschatzwachstum in Texten. Zur Reichweite des Piotrowski-Gesetzes. *Glottometrics 6, 9-34.*

Best, Karl-Heinz (2007). Zur Ausbreitung einiger Konfixe und Suffixe im Französischen. *Göttinger Beiträge zur Sprachwissenschaft 14, 29-33.*

Körner, Helle (2004). Zur Entwicklung des deutschen (Lehn-)Wortschatzes. *Glottometrics 7, 25-49.*

Küpper, Heinz (1955). *Wörterbuch der deutschen Umgangssprache.* Hamburg: Classen.

Küpper, Heinz (1963). *Wörterbuch der deutschen Umgangssprache. Band 2: 10000 neue Ausdrücke von A – Z.* Hamburg: Classen.

Meier, Helmut ([2]1967). *Deutsche Sprachstatistik*. 2., erweiterte und verbesserte Auflage. Hildesheim: Olms.

Verwendete Software

MAPLE V Release 4 (1996). Berlin u.a.: Springer.

NLREG*. Nonlinear Regression Analysis Program.* Ph. H. Sherrod. Copyright (c) 1991 - 2001.

Adresse des „Göttinger Projekts“ im Internet (mit ausführlicher Bibliographie): http://wwwuser.gwdg.de/~kbest

Glottometrics 20, 2010, 38-58

How do Local Syntactic Structures Influence Global Properties in Language Networks?

Haitao Liu[1], *Yiyi Zhao*[1], *Wei Huang*[2]

Abstract. Language networks, which are built based on different principles, have small-worldness and scale-freeness. This fact has raised the questions of why the different local microscopic constituents produce similar behaviors of global macroscopic networks and how local syntactic structures influence global properties of a linguistic network. To answer the questions, we built up three language (dependency) networks with coordinating structures and investigated their global statistical properties. The results show that all three networks are small-world and scale-free. The experiment demonstrates that, if we want to make an explainable link between local syntactic structures and global behavior of a linguistic network, current global indicators are not sufficient for finding differences between global syntactic dependency trees constructed with different annotation schemes or for finding differences between linguistic and non linguistic networks.

Keywords: linguistic network; coordination; Chinese; small-world; scale-free

1. Introduction

A human language is a network or a system with interconnected elements, which essentially is not new to contemporary linguistics (Hudson 2007; Lamb 1998). Empirical studies show that language networks are small-world and scale-free (Čech and Mačutek 2009; Ferrer i Cancho 2005; Ferrer i Cancho et al. 2004; Li et al. 2005; Li and Zhou 2007; Liu 2008a, 2009; Solé et al. 2005; Zhou et. al. 2008). Although these language networks are built based on different principles, they still demonstrate very similar global statistical properties. From the view of complex network, this similarity has raised a number of questions: Why do the different local microscopic constituents produce similar behaviors of global macroscopic networks? How do local syntactic structures influence global properties of a linguistic network? If the syntactic annotation schemes in a treebank are changed intentionally, will it be reflected in global properties of the linguistic network? These questions are explored and answered in the present paper. Linguistically, these questions are very important, because "networks are means, not the goal" (Liu 2008a) for linguists.

[1] Address correspondence to: School of International Studies, Zhejiang University, 310058, Hangzhou, Zhejiang, China. Email address: lhtzju@gmail.com.
[2] Institute of Applied Linguistics, Communication University of China, CN-100024, Beijing, China. Email: zoyiyi@163.com.
[3] Chinese Proficiency Test Center (HSK) , Beijing Language and Culture University, CN-100083, Beijing China. Email: huangwei@blcu.edu.cn.

In a language network, a node often represents a word (type), and the edge refers to the relation between two words. The difference of the language networks is found in the means to build links between two words. The paper is focusing on syntactic networks, in which a vertex is often a word (type), and the edge is the relation between two words. The researchers often use the word (type) as vertex of a syntactic network, but various other means to build links between two words. While we construct a syntactic network, we should follow a syntactic theory. Otherwise, the built network will not be sufficiently convincing, at least if the study is linguistically oriented.

Constituency and dependency analyses are two main means to analyze the structure of a sentence. Constituency (or phrase structure) analysis tends to find a part-whole relation of a sentence, while dependency analysis consists of binary asymmetrical relations between words in a sentence. Therefore, dependency syntactic analysis is more network-friendly (Čech and Mačutek 2009; Ferrer i Cancho 2005; Ferrer i Cancho et al. 2004; Liu 2008a). In this way, dependency syntax is employed in this study to construct language networks.

For all theories of syntax, coordination is a difficult point. Hence, coordination is chosen as the object of local syntactic structure in this study. Compared with syntactic theories based on phrase structure, coordination greatly challenges dependency syntax, in which binary asymmetrical relation is working as basic element (Lobin 1993; Osborne 2003; Temperley 2005). This challenge makes it possible to easily select various methods for the same syntactic phenomenon during constructing a syntactic treebank.

In this research, three syntactic annotation schemes for coordination were extracted from available literature on dependency syntax. 1000 sentences with more than 2000 coordinating structures were annotated by dependency syntax. After the preprocessing, we built three language networks, which have the same size with different annotation schemes for coordinating structure.

Based on these three linguistic networks, this paper examines how local syntactic structures (coordination) influence global statistical properties of linguistic networks. This study also explores the relation between local and global properties in complex networks for finding the possibility of using complex networks in linguistic study.

Section 2 introduces the method to build syntactic dependency (linguistic) networks focusing on coordination. Section 3 presents the results of network analysis and the fourth section contains a sketch discussion.

2. Methods

Dependency syntactic theories are based on a directed relation between words (Tesnière 1959; Hudson 2007). Fig. 1 shows a dependency analysis of the sentence *The boy eats an apple*.

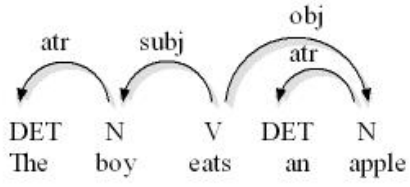


Fig. 1. Dependency structure of *The boy eats an apple*

In Fig. 1 all the words in the sentence are connected by asymmetrical syntactic relations. In each pair of connected words, one is called the *dependent* and the other is called the *head.* The labeled arc is directed from the *head* to the *dependent*. This "asymmetrical" feature makes it more difficult to process coordination in dependency syntax, because the words in a coordinating structure are equally related. For example, the sentence *The boy eats an apple and an orange* contains a coordination [*an apple and an orange*]. Here *and* is called conjunction, *an apple* and *an orange* conjuncts. Syntactically, *an apple* and *an orange* are equal or coordinated. Then, a question arises as how to process coordinated elements with unequal dependency relation in dependency syntax.

Following the available literature on dependency syntax, we propose three schemes to treat coordination in this section. In the following figures, *cc* is conjunction; *C1*, *C2* and *C3* are conjuncts in a coordinating structure; *X* is dependency label between coordinating structure and other constituents in the same sentence; *W1* is a word outside the coordination, but is syntactically related to it.

(1) Conjunction as head of coordinating structure

In this scheme, conjunction works as head of the whole coordinating structure (Schubert 1987; Liu and Huang 2006). Fig. 2 describes the relation between conjunctions and conjuncts.

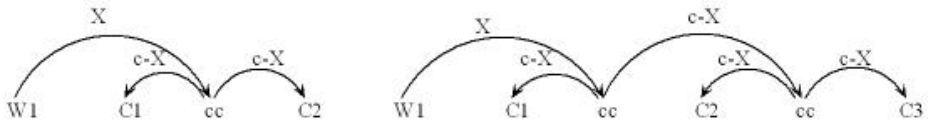


Fig. 2. Conjunction as head of coordinating structure
left panel: two conjuncts, right panel: three conjuncts

(2) The first conjunct as head of coordinating structure

This scheme is similar to that of Mel'čuk (1988), but with a little difference in the position of conjunction. In the scheme proposed in this study, conjunction depends on the second conjunct, which is probably easy for mapping syntactic structure to semantic analysis.

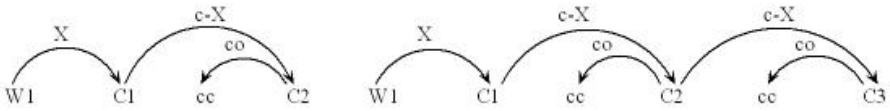


Fig. 3. The first conjunct as head of coordinating structure
left panel: two conjuncts, right panel: three conjuncts

(3) All conjuncts as heads of coordinating structure

This scheme gives all the conjuncts equal positions in a coordinating structure (the multi-head view). It is similar to Tesnière (1959) and Hudson (1998, 2007), but with different operation for conjunction.

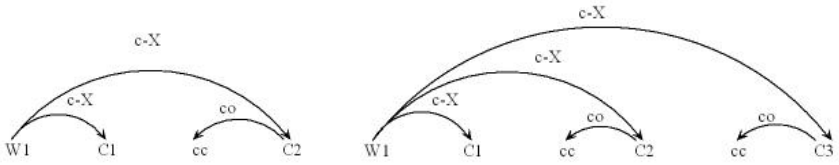


Fig. 4. All conjuncts as heads of coordinating structure
left panel: two conjuncts, right panel: three conjuncts

To better understand the three schemes, Fig. 5 provides three dependency analyses of the sentence *John bought an apple and a peach*. In this example, *W1* is *bought*, *cc* is *and*, *C1* is *apple* and *C2* is *peach*.

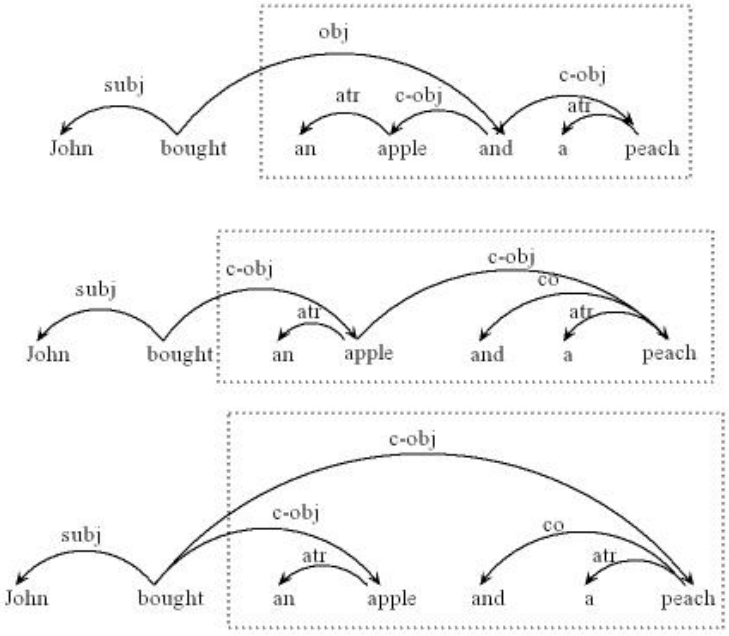


Fig. 5. Three dependency analyses of *John bought an apple and a peach*
Coordination is framed in dot line rectangle.

1000 Chinese sentences with more than 2000 coordinating structures were extracted from newspaper articles published in People's Daily in year 2000 to create a working corpus for the study of coordinating structures. We manually annotated the working corpus by dependency syntax (Liu and Huang 2006) and three schemes of annotating coordination. As a result, three treebanks (corpus with syntactic annotation) were built. Those three treebanks have the same size (32049 word tokens), and the mean sentence length is 33 words. There are 2197 conjunctions in the treebank. The most frequent conjunctions are 和 (and),、(the Chinese sign of coordination) and 与 (with). The proportion of dependency links involved in coordination is 31.4%[3], a statistic based on word types, because the nodes of the linguistic networks in this study are formed by word types. Zhao (2008) provides more detailed inform-

[3] This figure is very high, because the corpus was chosen with the aim of maximizing the number of complex coordinations.

ation about annotation schemes of coordination and how to build a treebank of coordinating structures.

Based on the method proposed in Liu (2008a), we have converted the three dependency treebanks into three undirected Chinese linguistic networks. *Coo1*, *Coo2* and *Coo3* are used to distinguish the three networks, which correspond with the annotation schemes in Fig.2, Fig.3, and Fig. 4 respectively.

Formally, these three annotation schemes produce different networks. Fig. 6 shows the difference in an example network, which consists of three sentences: *The woman and the child slept. John bought an apple and a peach. The child ate a pear and an apple*.

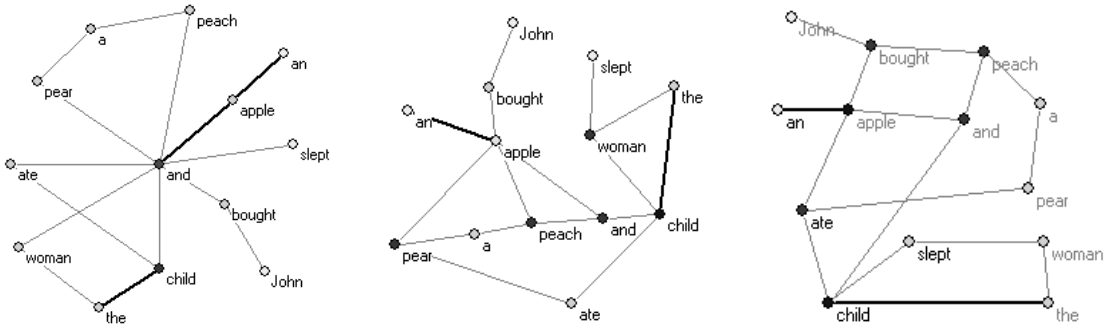
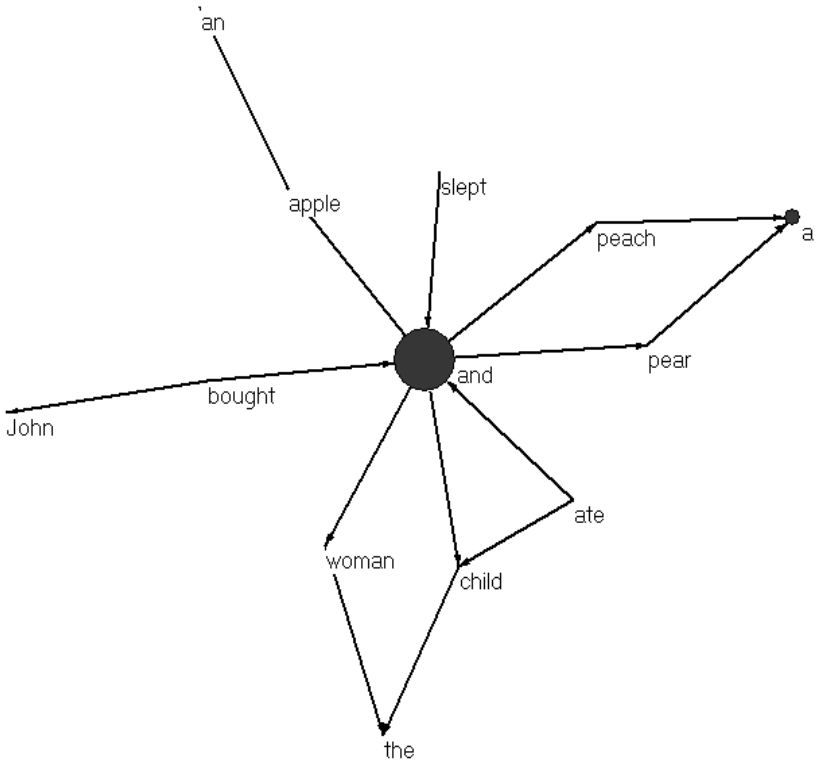


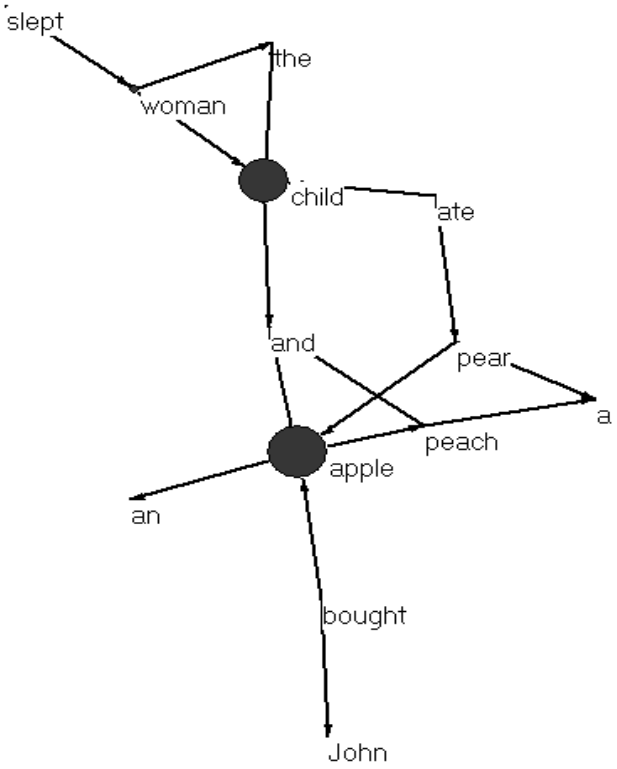
Fig. 6. Coordinating networks of three English sentences
left panel: Coo1, middle panel: Coo2, right panel: Coo3

In Fig. 6, the grey nodes mark the measurement of degree, and the thickness of the edges indicates the strength of the link. The first scheme gives prominence to conjunctions; in the other two schemes, conjuncts play more important roles than in the first. The network also displays the importance of the first conjunct in the second scheme. The third scheme increases the degree of governing node of coordinating structure.

The network analysis software Pajek was used to extract the Centers of the example networks (net -> vector -> centers)[4] as presented in Fig. 7.

[4] Pajek finds centers in a graph using 'robbery' algorithm, i.e. nodes that have higher degrees (are stronger) than their neighbors steal from them (Batagelj and Mrvar 2008).


an
apple
slept
peach
a
pear
and
bought
John
ate
woman
child
the


slept
the
woman
child
ate
and
pear
a
peach
apple
an
bought
John

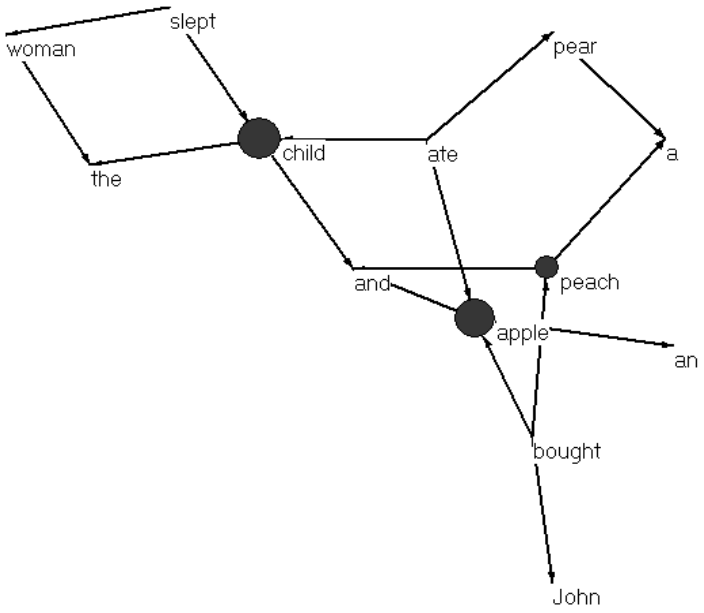


Fig. 7. Centers of coordinating networks with three English sentences
upper panel: Coo1, middle panel: Coo2, lower panel: Coo3

Fig. 6 and 7 clearly show that local syntactic changes may influence the properties of a language network. However, do these changes make evident differences from the view of a complex network? That will be discussed in the following section.

3. Results

The average path length, the clustering coefficients and the degree distribution of a network belong to the network indicators most frequently investigated for describing the basic properties of the complexity of a network (Börner et al. 2007; Brinkmeier and Schank 2005). Albert and Barabási (2002) call them the three robust measures of a network's topology.

In a (syntactic) language network, a node is a word (type), and the edge is the relation between two words.

The average shortest path length $<d>$ is the average shortest distance between any pair of nodes in a network. The three networks in the present study have the same number of nodes N (4259), but with different $<d>$'s. The path length $<d>$ of *Coo1* is 3.112, and those of *Coo2* and *Coo3* are 3.292 and 3.272 respectively. The differences among three networks are not significant (p-value = 0.997). The maximum shortest path in a network is defined as diameter D of the network. There are a few differences among these three diameters, whose value are 8, 9 and 9 respectively.

The distribution of the shortest path lengths (Fig. 8) is drawn based on the histogram of the length of the shortest paths between pairs of nodes of a network. The Kolmogorov-Smirnov two-sample tests show that three distributions of the shortest path length do not have significant differences (*Coo1* and *Coo2*, D = 0.1528, p-value = 0.9993; *Coo1* and *Coo3*, D =

0.1528, *p*-value = 0.9993; *Coo2* and *Coo3*, *D* = 0.1111, *p*-value = 1). This difference probably shows one of the differences between local syntactic analysis and global properties of a language network.

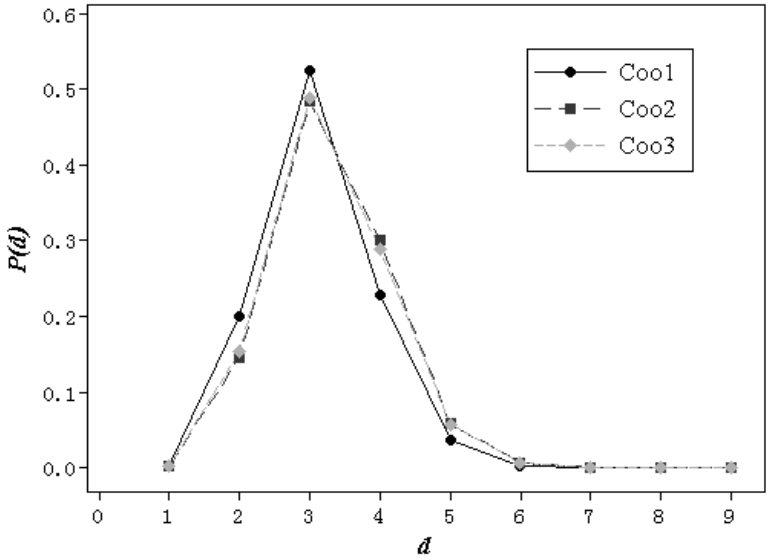
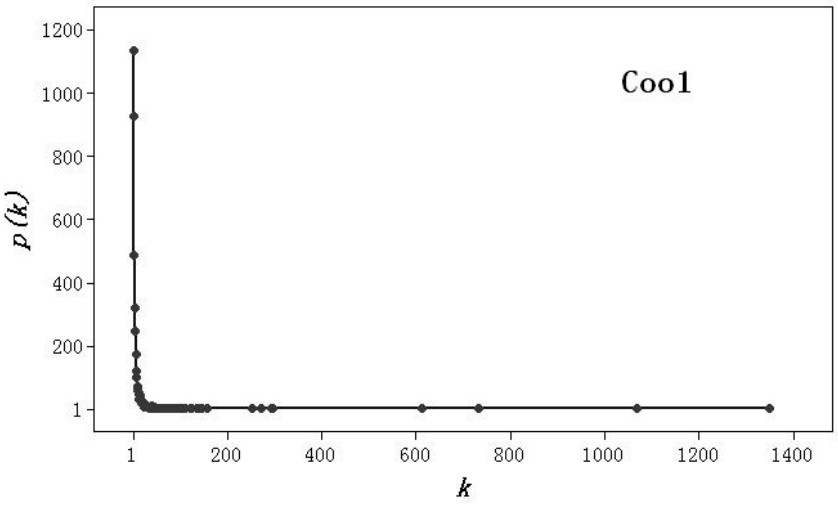


Fig. 8. Shortest Path Distribution of three networks

In a language network, the number of links of a given word type is called its degree *k*, which works in a very similar manner as valency in dependency grammar (Liu 2008a). <*k*> is the average degree of a network, which reflects the combining capacity of a word type with other words. From the three networks, their <*k*>'s are 7.565 (*Coo1*), 8.303 (*Coo2*) and 8.261 (*Coo3*) respectively, the difference being insignificant (*p*-value = 0.9789). Degree distributions are defined as the frequency *P(k)* of having a word type with *k* links. In other words, *P*(*k*) is the probability that a node chosen uniformly at random has degree *k*. The degree distributions (listed in Appendix) of the three networks are shown in Fig.9.



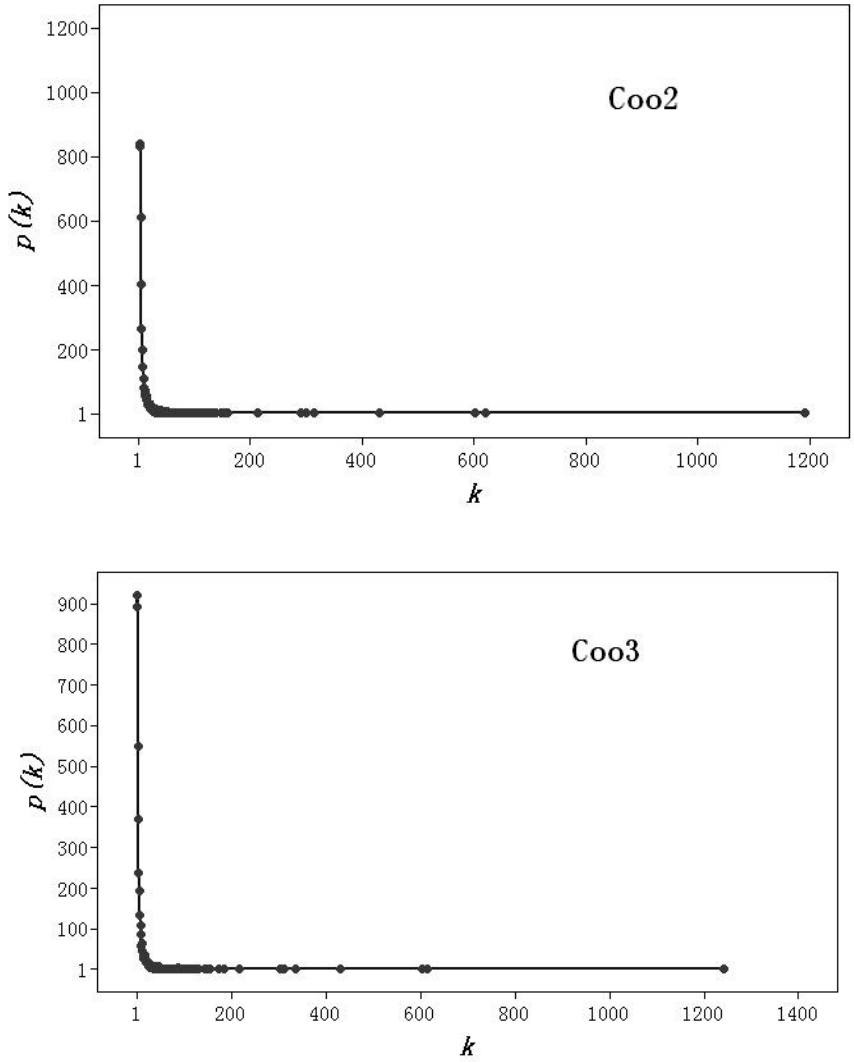


Fig. 9. Degree distributions of three networks. They were fitted by power functions with exponents -1.306 ($y = 674.58x^{-1.306}$, $R^2 = 0.789$, *Coo1*), -1.358 ($y = 950.8x^{-1.358}$, $R^2 = 0.828$, *Coo2*) and -1.364 ($y = 914.42x^{-1.364}$, $R^2 = 0.822$, *Coo3*)

The clustering coefficient *C* measures the average probability that two neighbors of the same node (word type) are also mutually connected. This can be better explained with Fig. 10, which is not simply the network structure of the example sentence in Fig. 5 but the tree of the 1st and 2nd neighbors of the word *and*.

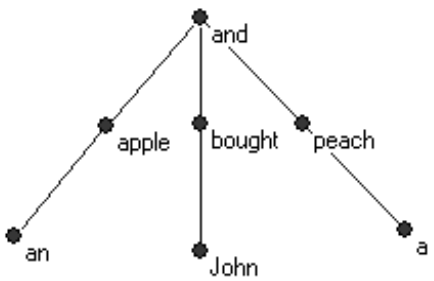


Fig. 10. Network structure of *John bought an apple and a peach* (*Coo1*)

For a syntactic network in Fig. 10, the clustering coefficient of the word *and* reflects the probability of two words that are adjacent and linked in turn, for example, the linking probability between *apple* and *bought*, *peach* and *apple*, *bought* and *peach*.

Let k_i denotes the degree of node i, E_i denotes the number of edges among the nodes in the nearest neighborhood of node i. Then, the clustering coefficient C_i of the node i is $2E_i/k_i(k_i\text{-}1)$. The clustering coefficient of the network is given by the average of C_i over all the nodes in the network. From the three networks, the clustering coefficient of *Coo1* is 0.199, that of *Coo2* is 0.154, and that of *Coo3* is 0.149. The differences among them are not signiciant (p-value = 0.9955).

These basic complexity properties of the three networks are summarized in Table 1.

Table 1

Main properties of three networks with coordinating structures.

N: the number of nodes, E: the number of edges, $<k>$: average degree, C: clustering coefficient, $<d>$: average path length, D: diameter, γ: exponent of power law, C_{rand} : clustering coefficient of random graph, $<d_{rand}>$: average path length of random graph.

network	***N***	***E***	***<k>***	***C***	***<d>***	***D***	***γ***	C_{rand}	$<d_{rand}>$
Coo1	4259	16714	7.565	0.199	3.112	8	1.306	0.0018	4.328
Coo2	4259	18227	8.303	0.154	3.292	9	1.358	0.0020	4.204
Coo3	4259	18007	8.261	0.149	3.272	9	1.364	0.0021	4.182

If a network has a high clustering coefficient C ($C >> C_{rand}$) and a very short path length $<d>$ ($<d> \sim <d_{rand}>$), it is a small world (SW) network (Watts and Strogatz 1998). Following this criterion, the three networks are small-worlds. If the degree distribution of a network follows a power law,

$$P(k) \sim k^{-\gamma} \tag{1}$$

the network is a scale-free network (Barabási and Albert 1999). Therefore, as shown in Fig. 9, the three networks are also scale-free.

Linguistically, it is understandable why the degrees of syntactic networks above must follow the power curve. In a text which aims at information communication, new information cannot be conveyed if the same words are always repeated. Therefore, new words must be used. However, new words can not be isolated from other words in the same sentence. Syntactically, they have to have a governing word, and probably also one or more dependency words. So, the degree of a new word will increase or change with its function in a sentence. Meanwhile, since the nodes in syntactic networks are word types, this also makes the degree of a node change based on all sentences including the word in the corpus. Hubs of a syntactic network are often the grammatical words, which play an important role in the growth of a linguistic network. All these factors contribute to a scale-free syntactic network.

Liu (2008a) hypothesizes that the small-world phenomenon of a linguistic syntactic network is closely related to dependency distance that tends to a minimum value (Liu 2007). This hypothesis connects small-world characteristics of a linguistic network with linguistics and cognitive science. Dependency distance is the linear distance between governor and the

dependent in a sentence. Based on the corpora of 20 languages, Liu (2008b) shows the average dependency distance of these languages ranges from 1.798 to 3.662. That is to say, the linear distance between two words in grammatical relationship is within three real text words. So, it is explainable to small-worldness of a syntactic network from the perspectives of human working memory capacity and grammar.

The networks in this study are based on the three treebanks including coordinating structures. Therefore, it is probably interesting and necessary to investigate hierarchical organization in networks. To reveal the hierarchy of a network, we need to study the correlation between the clustering coefficient and the degree of the nodes of a network. The correlation is expressed through the function *C(k)*, which represents the average clustering coefficient of all nodes with degree *k* (Ravasz and Barabási 2003). In a network, if *C*(*k*) follows a power-law distribution, then it signals the presence of hierarchical organization, in which most low degree nodes belong to well interconnected communities, and hubs connect many nodes that are not directly connected (Pastor-Satorras and Vespignani 2004).

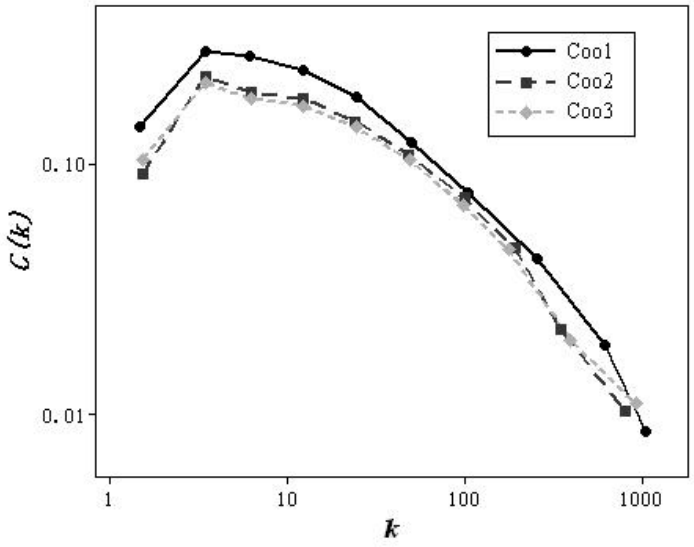


Fig. 11. Clustering coefficient C(k) vs degree *k* for the three networks[5].

The Kolmogorov-Smirnov two-sample tests show that there are no significant differences among distributions of *C(k)* in three networks (*Coo1* and *Coo2*, *D* = 0.3, *p*-value = 0.787; *Coo1* and *Coo3*, *D* = 0.3, *p*-value = 0.787; *Coo2* and *Coo3*, *D* = 0.1, *p*-value = 1).

Fig. 11 shows that the distributions of *C(k)* are skewed, i.e. they are not power laws as in other networks (Ravasz and Barabási 2003). The result is similar to that of Ferrer i Cancho et al. (2004) that makes the conclusion based on dependency treebanks of Czech, Romanian and German. The result of the present investigation is not compatible with Zhou et al. (2008) who

[5] *C(k)* is calculated by the modeling algorithms of the NWB (Network Workbench Tool, NWB Team 2006). This technique is suitable to study skewed correlation functions due to the fact that the size of the bins grows large for large values of the degree compensates for the fact that not many nodes have high degrees, so it suppresses the fluctuations that one would observe by using bins of equal size. On a double logarithmic scale, which is very useful to determine the possible power law behavior of the correlation function, the points of the latter will appear equally spaced on the *x*-axis.

obtained better results from two Chinese networks. It seems strange that the result of the present study is more similar to the results from other languages, and not to that of the same language -- Chinese. The common point between this study and the one by Ferrer i Cancho et al. (2004) is that both use language networks based on dependency syntax, while Zhou et al. (2008) built their networks non-syntactically. So, syntax may play some role here, but it needs to be proven in further investigation. Another factor, which probably brings about the skewed distribution, is that our corpora include many coordination structures, which makes the networks flatter, although we have treated them by the asymmetrical dependency link.

Another global indicator, which can characterize the real-world networks, is *K-Nearest-Neighbor* (Average neighbours degree) k_{NN} that measures the *correlation* between the degree of a node and that of its neighbors. If large (small) degree nodes tend to be linked with large (small) degree nodes in this network, then this network is *assortative mixing* or *assortativity*. If large (small) degree nodes tend to be linked with small (large) degree nodes, such network is *dissortative mixing* or *dissortativity*. Social networks are typical representatives of assortative mixing networks. Biological and technological networks are examples of dissortative mixing networks (Newman 2002). The reasons that social and biological networks have such difference are not completely understood (Caldarelli 2007). A probable direction to get an explanation is to explore whether this difference is the result of a conscious effort of human participants in these networks, because humans contribute consciously more during the forming of a social network than in that of a technological or biological network. However, more empirical data are needed for a solid explanation.

Pastor-Satorras and Vespignani (2004) and Caldarelli (2007) have argued that the correct mathematical way to quantify such a measure is the conditional probability $P(k'|k)$ of having a node with degree k' at one side of the edge given that at the other side of the edge the degree is k as shown in the formula (2):

$$k_{NN}(k) = \sum_{k'} k' P(k' \mid k) \qquad (2)$$

We used a less strict but more intuitive and simple approach proposed by Pastor-Satorras et al. (2001), who defined undirected *K-Nearest-Neighbor* as follows: a node is selected and the average degree of all its neighbors is calculated. By repeating the procedure for all nodes of the network one derives a pair (k_{NN}, k) for each node, where k is the degree of the node. By averaging over nodes with equal degree k one derives the function $k_{NN}(k)$, which allows to study the correlation. If $k_{NN}(k)$ grows with k, the network is assortative; if $k_{NN}(k)$ decreases with k, the network is disassortative. A flat curve would indicate the absence of correlation.

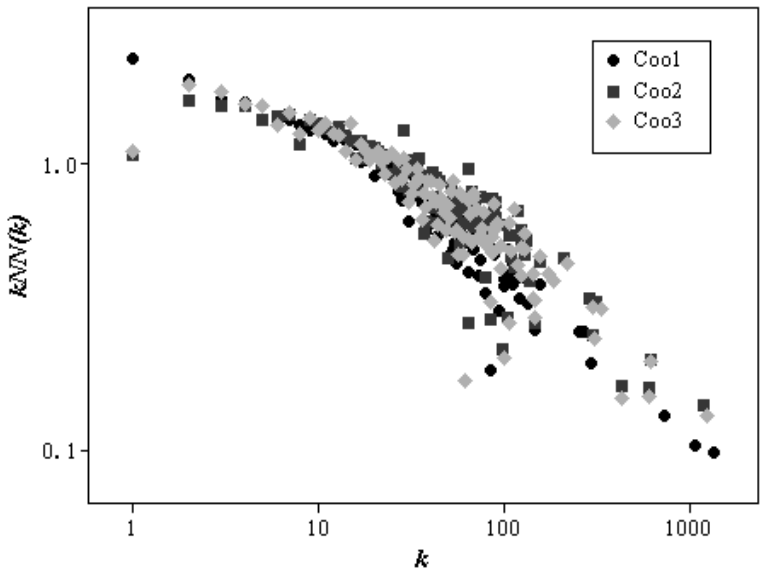


Fig. 12. The average nearest-neighbor degree as a function of the node degree.

The Kolmogorov-Smirnov two-sample tests show that there is no significant difference among distributions of $k_{NN}(k)$ in three networks (*Coo1* and *Coo2*, D = 0.182, p-value = 0.06932; *Coo1* and *Coo3*, D = 0.1192, p-value = 0.446; *Coo2* and *Coo3*, D = 0.0991, p-value = 0.6786), although the test between *Coo1* and *Coo2* almost makes a significant difference.

Fig. 12 indicates negative correlations between the degree of a node and that of its neighbors in the three networks. This dissortative property of a language network can be explained by the fact that functional (linking) words, which are often found in the list of the most frequent words, tend to avoid mutual connections in a syntactic network. Grammatically, the function of functional words is to link content words, which have very different degrees. Hence, the dissortative property of a language network adequately reflects the relation between content and functional words in a language. It is perhaps interesting to notice the similarity between language and biological networks, which demonstrates the biological foundations of language as claimed in biolinguistics (Boeckx and Grohmann 2007).

To understand the functions of local structures in a global network, we also investigated the closeness and betweenness centrality of seven words[6], which were chosen from the list of the most frequent words in the treebank. In those seven words,和 and 与 are conjunctions; 在 and 对 are prepositions; 的 is a structural particle; 是 is a verb; "、" is a sign serving as a conjunction.

Similarly to degree, closeness centrality also reflects the reachability of a word within a language network. Closeness can be regarded as a measure of how long it will take to spread information from a given node to the others in the network (Newman 2005). The closer a node is to all other nodes, the easier the information may reach it, the higher its centrality. The closeness centrality $C_c(i)$ of a node i can be formally defined according to Caldarelli (2007):

$$C_c(i) = \frac{1}{\sum_{t \in V \setminus i} d_G(i,t)} \qquad (3)$$

[6] They are 和, 与, 在(at, in, on), 对 (opposite, versus), 的 (a structural particle), 、and 是(be).

where *n* is the size of the network's connectivity component *V* reachable from the node *I*, $d_G(i, t)$ is the distance between the node *i* and *t*.

Table 2
Closeness Centrality of seven words

network	**和**	**、**	**与**	**的**	**是**	**在**	**对**	Mean	Max	Min
Coo1	0.559	0.498	0.450	0.528	0.457	0.456	0.434	0.329	0.559	0.166
Coo2	0.445	0.415	0.392	0.521	0.433	0.438	0.408	0.310	0.521	0.148
Coo3	0.441	0.412	0.397	0.531	0.442	0.443	0.416	0.312	0.531	0.163

Conjunctions have higher closeness centrality in *Coo1* than in *Coo2* and *Coo3*. Closeness centrality of the other four words (的, 是, 在, 对) seems more stable than conjunctions in the three networks. The centrality of all the seven words is higher than the mean score, while 和 has the highest in *Coo1* and 的 in *Coo2* and *Coo3*. So, the frequency and syntactic annotation scheme of a treebank may influence the reachability of a word in a linguistic network.

The betweenness centrality of a node is the proportion of all shortest paths between pairs of other nodes that include this node. The betweenness centrality of a word depends on the extent to which it is needed as a link that facilitates the spread of information within a language network. The more a word is an intermediator, the more central its position in the network. The betweenness centrality $B_c(i)$ of a node *i* can be calculated using the following formula (Caldarelli 2007):

$$B_c(i) = \sum_{j \neq k} \frac{b_{jk}(i)}{b_{jk}} \qquad (4)$$

where b_{jk} is the number of shortest paths between node *j* and *k*, while $b_{jk}(i)$ the number of shortest paths that include node *i*.

Table 3
Betweenness Centrality of seven words

network	**和**	**、**	**与**	**的**	**是**	**在**	**对**	Mean	Max	Min
Coo1	0.368	0.171	0.029	0.248	0.029	0.035	0.010	0.0005	0.368	0
Coo2	0.133	0.095	0.012	0.353	0.037	0.048	0.015	0.0005	0.353	0
Coo3	0.112	0.076	0.012	0.370	0.040	0.050	0.019	0.0005	0.370	0

The conjunction 和 has again the central position in *Coo1*. However, the particle 的 replaces 和 as the most central in *Coo2* and *Coo3*. It seems that the syntactic annotation

scheme has played a more important role to decide betweenness centrality of a word than its frequency in the corpus.

We extracted the centers of the three networks by the same method as in Fig. 7. The conjunction "he" is almost the unique center of *Coo1*, but the centers of *Coo2* and *Coo3* are 的,和 and 、.

Table 2 and 3 show that the changes of syntactic structure more significantly impact on closeness and betweenness centrality of the related words than global statistical patterns of a network do. Fig. 7 reflects the similar conclusion, although it is only exemplified in three English sentences. In Chinese language networks, the particle 的 is very important. It is not only the most frequent word in Chinese texts, but also has almost the highest centrality of closeness and betweenness. The phenomenon is worth further investigation.

4. Discussion

Based on the statistical patterns of the three networks and corresponding E-R random networks, we therefore conclude that all the three networks are small-world and scale-free. Kolmogorov-Smirnov two-sample tests also show that there is no significant difference among the three networks. However, this conclusion is not unproblematic, because the three networks not only contain the links related to coordination, but also other links that are built using exactly the same annotation scheme; these links were not changed in the three networks. So, we have to more precisely explore the influence of local coordinating structure upon the global properties of the language network.

According to the statistics in section 2, the different analyses of coordination in this study involve 31.4% of all links. So the question turns into whether the global differences are less that 31.4%. In order to test that, we set *Coo2* as the baseline of three networks and calculate the differences between *Coo2* and *Coo1*/*Coo3*. This result is shown in Table 4.

Table 4
The global differences of three networks (%)

Coo2 with	**<*k*>**	***C***	**<*d*>**	***D***	***γ***
Coo1	8.9	29.2	5.5	11.1	3.8
Coo3	0.5	3.2	0.6	0	0.4

Table 4 shows that, besides the difference between clustering coefficients of *Coo1* and *Coo2*, other differences are much smaller in proportion than the local ones. In this way, it seems reasonable to claim that global properties are not affected by local variations.

Table 1 and 4 show that, although *Coo2* and *Coo3* have closer parameters than that of *Coo1*, all the networks belong to the same basic network typology, hence classification as small-world and scale-free does not reflect the changes of syntactic annotation. However, we have also found that syntactic annotation has more significant influences on the closeness and betweenness centrality of the related words, which indicates that more attention should be paid to the role of a word (type) in language networks. It may suggest that these centrality

indices have great value in the analysis and understanding of the roles played by actors (words) as social networks (Wasserman and Faust 1994).

Fig. 6 and 7 show that local syntactic changes can be well reflected in network's form, but why do such differences disappear in global statistical patterns of the network? Are small-worldness and scale-freeness sufficient to indicate the syntactic properties of a language network? If the statistical properties of the microscopic (local) structures cannot explain the behavior of macroscopic (global) networks, does it mean that complex network is not an appropriate means to study language structure?

Our working corpora include 1000 sentences with more than 2000 coordinating structures. Three annotation schemes for coordination are different from the view of microscopic level. However, after linking 1000 sentences into the three networks, why do they have very similar global statistical properties? It seems reasonable to regard these global statistical properties as emergent properties of a network, but what are the factors which make the emergent property in our networks? Is it the number of the sentences in the treebank or syntactic annotation schemes? Or simply, we may just ignore these questions, "since networks are very large systems, whose collective behaviour cannot be understood from the elementary features" (Caldarelli 2007: xi). In other words, our experiment proves that language is a complex network (system), in which we can not explain the global behaviors of the system by local changes.

The small-worldness and scale-freeness perhaps can be understood as an emergent property of a linguistic network, but the empirical studies show that such property seems not to depend on the local structures in a language network; many other real-world networks are also small-world and scale-free (Albert and Barabási 2002). Therefore, we may have to search for other global statistical patterns to build a more solid link between the microscopic structures and the macroscopic networks.

This experiment also demonstrates that, if we want to make an explainable link between local syntactic properties and global behavior of a language network, current indicators are not sufficient for finding differences between global syntactic dependency trees constructed with different annotation criteria or for finding differences between linguistic and non linguistic networks. Further efforts need to be made to find other indicators for identifying this relationship. Except for finding new static global properties, further work is to investigate dynamical aspects of linguistic network based on general theory and method of dynamics in complex networks (Barrat et al. 2008), which will be helpful to find how local changes influence the global structure of a (linguistic) network.

In summary, our finding essentially shows that graph-theoretical results will not help theoretical linguists to choose between competing theories, which is perhaps disappointing for theoretical linguists. However, it is helpful to understand why (language) networks based on different building principles are often small-world and scale-free, because current statistical patterns cannot capture the changes of local syntactic structure in a complex (language) network. On the other hand, the study also shows the value of this kind of careful statistical analysis applied to a large corpus of sentences, although many questions are still open for further research. So, combining this empirical study with the findings in Liu and Hu (2008) and Ferrer i Cancho et al. (2004), further research should be carried out to explore other statistical patterns for finding linguistic and cognitive universals from the point of view of

complex networks.

Acknowledgements

We thank Richard Hudson and the referees for insightful comments. This work is partly supported by the National Social Science Foundation of China (Grant No. 09BYY024) and the Communication University of China (Project No. 21103010101).

References

Albert, R., Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics 74(1), 47-97.*

Barabási, A-L., Albert, R. (1999) Emergence of scaling in random networks. *Science 286, 509–12.*

Barrat, A., Barthelemy, M., Vespignani, A. (2008). *Dynamical processes in complex networks*. Cambridge University Press.

Batagelj, V., Mrvar, A. (2008). *Pajek: Program for Analysis and Visualization of Large Networks*. Reference Manual. version 1.23. http://vlado.fmf.uni-lj.si/pub/networks/pajek/doc/pajekman.pdf (2009-10-23)

Boeckx, C., Grohmann, K.K. (2007). The Biolinguistics Manifesto. *Biolinguistics 1, 001–008.* http://www.biolinguistics.eu (2009-7-21)

Börner, K., Sanyal, S., Vespignani, A. (2007). Network Science. In: Blaise Cronin (Ed.), *Annual Review of Information Science & Technology, Volume 41, 537-607.* Medford, NJ: Information Today, Inc./American Society for Information Science and Technology.

Brinkmeier, M., Schank, T. (2005). Network Statistics. In: U. Brandes and T. Erlebach (Eds.): *Network Analysis*: *293–317*. Berlin/Heidelberg: Springer-Verlag.

Caldarelli, G. (2007). *Scale-free networks: complex webs in nature and technology*. Oxford: Oxford University Press.

Čech, R., Mačutek, J. (2009). Word form and lemma syntactic dependency networks in Czech: a comparative study. *Glottometrics 19, 85-98*.

Ferrer i Cancho, R. (2005). The structure of syntactic dependency networks: insights from recent advances in network theory. In: Altmann, G.., Levickij, V., Perebyinis, V. (eds.). *The problems of quantitative linguistics: 60-75*. Chernivtsi: Ruta.

Ferrer i Cancho, R., Solé, R.V., Köhler, R. (2004). Patterns in syntactic dependency networks. *Physical Review E 69, 051915*.

Hudson, R. (1998). *English Grammar*. London: Routledge.

Hudson, R. (2007). *Language Networks: The New Word Grammar*. Oxford: Oxford University Press.

Lamb, S.M. (1998). *Pathways of the Brain: The Neurocognitive Basis of Language*. Amsterdam: John Benjamins.

Li, J., Zhou, J. (2007). Chinese Character Structure Analysis Based on Complex Networks. *Physica A. 380, 629-638.*

Li, Y., Wie, L.X., Li, W., et al. (2005). Small-world patterns in Chinese phrase networks. *Chinese Sci. Bulletin 50, 286-288.*

Liu, H. (2007). Probability distribution of dependency distance. *Glottometrics 15,1-12.*

Liu, H. (2008a). The complexity of Chinese dependency syntactic networks. *Physica A. 387, 3048-3058*.

Liu, H. (2008b). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*. *9(2), 159-191*.

Liu, H. (2009). Statistical properties of Chinese semantic networks. *Chinese Science Bulletin, 54, 2781-2785.*

Liu H., Huang, W. (2006). A Chinese Dependency Syntax for Treebanking. *Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation*: *126-133.* Beijing: Tsinghua University Press.

Liu, H., Hu, F. (2008). What role does syntax play in a language network? *Europhysics Letters 83, 18002*.

Lobin, H. (1993). *Koordinationssyntax als prozedurales Phänomen.* Tübingen: Narr.

Mel'čuk, I.A. (1988). *Dependency syntax: theory and practice*. Albany: State University Press of New York.

Newman, M.E.J. (2002). Assortative mixing in networks. *Physical Review Letters 89, 208701*.

Newman, M.E.J. (2005). A measure of betweenness centrality based on random walks. *Social Networks 27(1), 39-54.*

NWB Team (2006). *Network Workbench Tool*. Indiana University and Northeastern University, http://nwb.slis.indiana.edu . (2009-10-27)

Osborne, T. (2003). *The Third Dimension: A Dependency Grammar Theory of Coordination for English and German.* Ph.D Dissertation. The Pennsylvania State University.

Pastor-Satorras, R., Vespignani, A. (2004). *Evolution and structure of the Internet: A statistical physics approach*. Cambridge, UK: Cambridge University Press.

Pastor-Satorras, R., Vazquez, A., Vespignani, A. (2001). Dynamical and Correlation Properties of the Internet. *Physical Review* Letters 87, 258701.

Ravasz, E., Barabasi, A.-L. (2003). Hierarchical organization in complex networks. *Physical Review E 67, 026112*.

Schubert, K. (1987). *Metataxis: contrastive dependency syntax for machine translation*. Dordrecht: Foris.

Solé, R., Corominas, B., Valverde, S., Steels, L. (2005). *Language Networks: Their Structure, Function and Evolution.* Santa Fe Institute Working Paper (05-12-042).

Temperley, D. (2005). The Dependency Structure of Coordinate Phrases: A Corpus Approach. *Journal of Psycholinguistic Research 34, 577-601*.

Tesnière, L. (1959). *Eléments de la syntaxe structurale*. Paris: Klincksieck.

Watts, D.J., Strogatz,S.H. (1998). Collective dynamics of 'small-world' networks. *Nature 393, 440–442.*

Wasserman, S., Faust, K. (1994). *Social Network Analysis*. Cambridge: Cambridge University Press.

Zhao, Y. (2008). *Automatic parsing of Chinese coordination based on dependency grammar*. Master thesis, Communication University of China, Beijing.

Zhou, Sh., Hua, G., Zhang, Zh., Guan, J. (2008). An empirical study of Chinese language networks. *Physica A 387, 3039–3047.*

Appendix

Degree Distributions of three networks

Coo1		*Coo2*		*Coo3*	
degree	frequency	degree	frequency	degree	frequency
1	1133	1	830	1	892
2	928	2	838	2	923
3	487	3	610	3	550
4	318	4	402	4	371
5	245	5	265	5	240
6	175	6	200	6	195
7	120	7	147	7	134
8	102	8	108	8	108
9	63	9	81	9	87
10	70	10	73	10	60
11	58	11	55	11	64
12	47	12	59	12	65
13	50	13	54	18	47
14	31	14	44	14	34
15	42	15	37	15	29
16	26	16	26	16	26
17	27	17	33	17	36
18	26	18	23	18	26
19	17	19	23	19	23
20	19	20	25	20	19
21	18	21	16	21	20
22	13	22	16	22	19
23	18	23	13	23	18
24	6	24	16	24	12
25	10	25	7	25	10
26	13	26	17	26	14
27	11	27	11	27	14
28	8	28	13	28	12
29	7	29	4	29	7
30	9	30	13	30	6
31	5	31	10	31	5
32	6	32	10	32	9

33	6	33	7	33	6
34	6	34	3	34	7
35	4	35	7	35	6
36	7	36	3	36	2
37	6	37	6	37	6
38	3	38	4	38	5
39	9	39	9	39	6
40	2	40	7	40	6
41	6	41	6	41	8
42	5	42	6	42	4
43	4	43	4	43	5
44	1	44	5	44	3
45	6	45	3	45	9
46	2	46	4	46	4
47	2	47	5	47	3
48	4	48	5	48	8
49	1	49	5	49	4
50	1	50	1	50	1
51	4	51	6	51	7
52	2	52	4	52	1
53	2	54	2	53	3
54	2	56	3	54	1
55	2	57	4	55	1
56	2	58	1	56	1
57	1	60	1	57	2
58	4	61	1	58	2
59	3	62	4	59	3
60	2	63	3	60	2
61	1	64	1	61	1
62	4	65	1	62	1
63	1	66	4	64	3
64	3	67	3	65	3
66	1	68	1	66	3
69	1	69	1	67	2
70	1	70	3	68	1
71	1	74	1	69	2
72	2	76	1	71	2
73	1	77	3	74	1
75	1	78	1	75	3
78	1	79	3	76	1
79	1	80	1	77	1
80	1	84	2	78	2
81	1	86	1	79	1

84	2	88	2	84	2
85	1	89	2	86	1
89	1	90	1	87	4
94	2	91	1	88	1
100	1	96	2	89	2
101	2	98	1	90	1
105	1	105	1	93	1
106	1	107	2	96	1
107	1	109	1	97	2
108	1	113	1	101	1
111	2	119	1	106	1
112	1	124	1	108	1
121	1	125	1	109	1
125	2	128	1	113	1
135	1	132	2	119	1
137	1	138	3	124	1
147	1	148	1	125	1
158	1	155	1	128	2
252	1	158	1	130	1
273	1	212	1	144	1
294	1	289	1	145	1
297	1	299	1	146	1
612	1	314	1	148	1
732	1	430	1	155	1
1069	1	602	1	175	1
1350	1	620	1	185	1
		1192	1	218	1
				304	1
				311	1
				336	1
				430	1
				604	1
				614	1
				1243	1

Glottometrics 20, 2010, 59-69

The distribution of parts-of-speech in Russian texts

Andrei Beliankou, Trier
Reinhard Köhler, Trier

Abstract. This paper contributes to testing the hypothesis that many linguistic units display a stratified behaviour, which is a reason for data inhomogeneity and makes it difficult to model them using classical probability distributions such as the Zipf-Mandelbrot, the Zipf-Alekseev, or the hypergeometric distributions. Specifically, the distribution of parts-of-speech in a large Russian text corpus is studied. Instead of a probability distribution, the Popescu function is applied, which can be fitted to stratified rank distributions. The results support the hypothesis and the usefulness of the new model.

Keywords: Popescu function, parts-of-speech, Russian

Introduction

The distribution of word classes was studied in a relatively large number of publications. Most of them applied the classical method using probability distributions such as the Zipf-Aleekseev or the Hypergeometric distributions (Best 1994, 1997, 1998, 2000, 2001; Hammerl 1990, Schweers/Zhu 1991; Zhu/Best 1992; Ziegler 1998, 2001; Ziegler/Altmann 2001). Theoretically justified distributions are often derived with the help of Altmann/Köhler´s (1996) approach; its basic idea is that the probability of a frequency class x depends (proportionally) on the probability of the class x-1. Depending on the specific function which is used for $g(x)$ in the corresponding difference equation

$$P_{x+1} = g(x)P_x$$

one of the potentially appropriate distributions is obtained as the solution to the equation. This technique has proved to be quite successful, at least if the data are homogeneous. However, linguistic material is often inhomogeneous in various ways and for various reasons.

Recently, an alternative was therefore proposed by Popescu (cf. Popescu/Altmann/Köhler 2009) and applied in Tuzzi/Popescu/Altmann (2010: 116-126.). The model has the form of a function instead of a probability distribution and is based on the assumption that linguistic data are, in general, composed of several layers ('strata'). In the case of word class frequencies, these strata could reflect influences of e.g., grammatical, thematic, and stylistic factors. For each of the possible strata, a term with specific parameters is introduced in the formula. The relation between rank and frequency is assumed to be exponential, i.e. with a constant relative rate of (negative) growth (1). Originally, the model was designed for rank-frequency series of words where the smallest frequency is always 1. In other cases, instead of unity a parameter should be introduced which can be estimated from the smallest frequency in the sample, cf. formula (2).

$$f(r) = 1 + \sum_{i=1}^{k} A_i \exp(-r / r_i) \qquad (1)$$

$$y = k + a_1 e^{b_1 x} + a_2 e^{b_2 x} + \ldots \quad (2)$$

The model yields very good results also with non-stratified data, where classical probability distributions work well enough. Here, we will present an example where more than one exponential term is needed. The syntactically annotated corpus of Russian (cf. below) differentiates relatively few parts-of-speech; e.g. all kinds of non-inflecting word classes are tagged as "PART" (particle). Therefore, it seems likely that the distribution of these tags displays more strata than parts-of-speech do when more classes are differentiated. We obtain for a randomly chosen text from the corpus the rank-frequency distribution shown in Table 1. Fitting model (2) with one, two, three, and four exponential terms yields the better values of the determination coefficient the more terms we use (cf. Table 2).

Table 1
Rank-frequency distribution of the parts-of-speech in a Russian text

Rank	1	2	3	4	5	6	7	8	9
Frequency	182	63	54	50	19	14	11	10	4

Table 2
Adjusted coefficients of multiple determination (ACMD)
and estimated parameters of function (2) with one to four exponential terms

No. of exp. terms	**1**	**2**	**3**	**4**
ACMD	0.9322	0.9752	0.9702	0.9942
a1	326.7518	135.2902	9105.0244	-7487.1539
b1	-0.6619	-0.3662	-0.1266	-1.1530
a2		34849502.6	-8999.6489	1627.4647
b2		-12.9334	-0.1253	-0.8643
a3			431819.96	6072.0465
b3			-8.4131	-1.6684
a4				1686.3195
b4				-0.8637

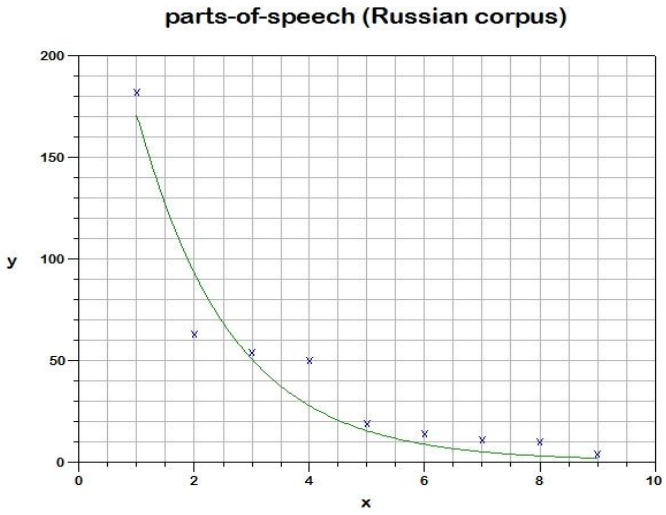


Fig. 1a: Plot of function (2) with one exponential term as fitted to the data from Table 1.

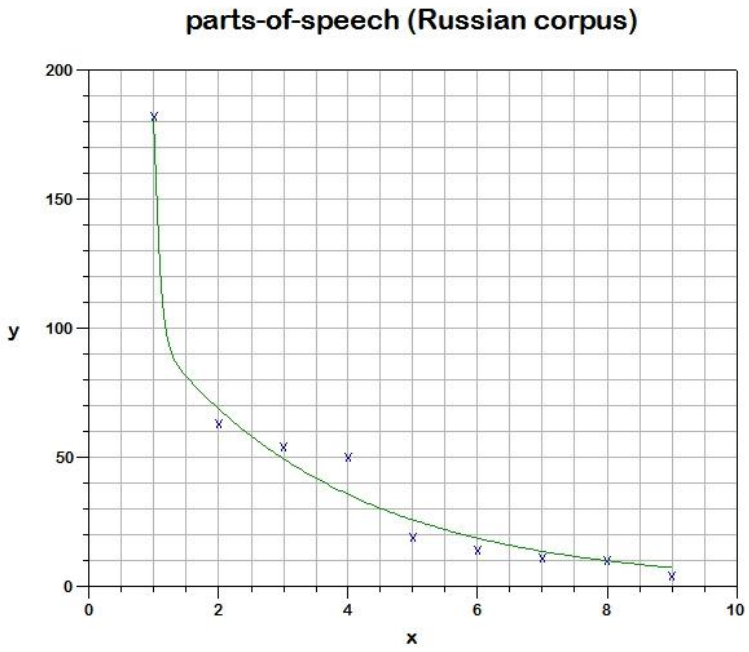


Fig. 1b: Plot of function (2) with two exponential terms as fitted to the data from Table 1.

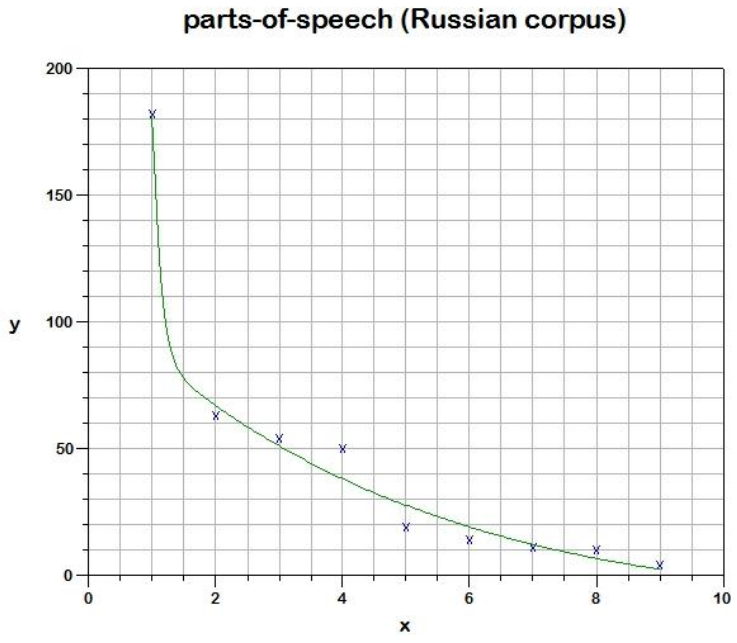


Fig. 1c: Plot of function (2) with three exponential terms as fitted to the data from Table 1.

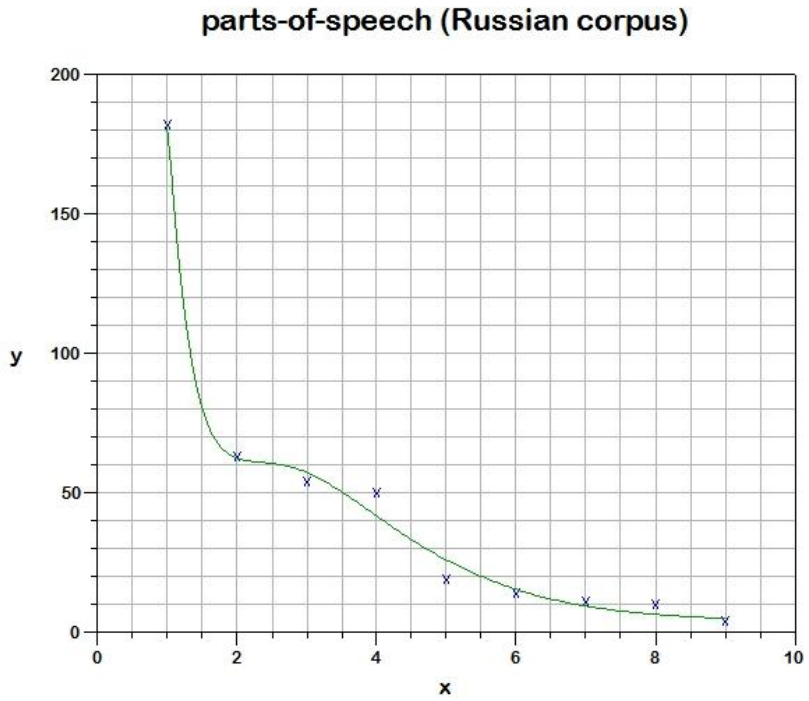


Fig. 1d: Plot of function (2) with four exponential terms as fitted to the data from Table 1.

Figures 1a–d show very clearly the stratification of the data and also the stepwise improvement of the function behaviour with respect to the configuration of the data elements. However, although the fourth exponential term brings another improvement of the determination coefficient and a smoother fit of the function to the data there is an important reason to reject the model with four terms: (1) It is determined by nine parameters while it describes just nine data elements whence the model has almost no descriptive advantage over the original data[1]. (2) We see that in the four terms variant a_2 and a_4 as well as b_2 and b_4 respectively are almost identical, a sign of redundancy. (3) We should be warned that an "abnormal" deviation from a model may be caused by modifications of the text after its completion either by the author himself or by editors, which could be considered as a case of manipulated data. In general, models with more than 3 parameters (even if you have hundreds or thousands of data elements) are seldom useful in linguistics because already for two or three empirically determined parameters it may be hard to find plausible linguistic interpretations. As a rule, a model with an additional parameter which gives a better result than another model with fewer parameters need not be the better one. A small improvement of the goodness-of-fit characteristic is of little value if you have no idea what the extra parameter stands for. In our case, the situation is different insofar as model (1) resp. (2) is a (well grounded) series of two-parameter functions where the pair-wise addition of parameters does not introduce any principally new aspects. Nevertheless, a trade-off between the two criteria – improvement of the goodness-of-fit on the one hand and number of parameters on the other – might lead to prefer the model version with only two components.

[1] We must not forget, however, that the model is theoretically founded. Its main purpose is not just to describe a single data set but infinitely many data and, even more important, it does not only describe but it explains the behaviour of the data.

In the following, a study on the rank-frequency data of parts-of-speech use in the first 193 texts from the “Deeply Annotated Corpus”, a part of the Russian National Corpus, using the described model will be presented.

Material and data source

The Russian National Corpus (RNC) is an electronic reference system of modern Russian. The corpus provides both spoken (transcribed) and written texts. The RNC consists primarily of original prose texts but includes also smaller amounts of translated works. We can find there some non-standard (colloquial and dialectal) forms.

The planned size should be approximately 200 Millions of tokens. As for the current state, the corpus contains about 150 Millions of annotated data tokens. The RNC consists of loosely coupled parts which in turn allow providing a linguistic annotation based on different theories. The corpus is realized by collaborating scholars from different schools; it is therefore crucial to have a common annotation platform for the different approaches.

The main and most interesting part of the RNC, the Deeply Annotated Corpus, provides morpho-syntactic and semantic annotations on the basis of the "Meaning ↔ Text" model by Igor A. Mel’čuk and Alexander K. Zholkovsky. The syntactic structures were labelled with a version of dependency relations with deep and shallow grammatical roles. The whole corpus has been manually disambiguated at the morphological and syntactical levels and forms a reliable source of linguistic knowledge.

The RNC has its roots in the machine translation project ETAP and served primarily as a collection of parallel linguistic data. A big part of the RNC, the parallel text corpus, contains a manually aligned English-Russian corpus, parsed with the same dependency based formalism.

At the morphological level, 16 parts-of-speech are distinguished, including such interesting categories as 'predicative' and 'parenthesis'. For the whole tagset cf. the project’s internet page (http://www.ruscorpora.ru).

Currently, the Corpus provides a search facility for lexical and semantic characteristics as the texts are semantically tagged. Most words in the corpus are tagged with a number of semantic and derivational parameters such as “person”, “substance”, “space”, “movement”, “diminutive”, “verbal noun”, etc. Every word can be assigned characteristics along several different parameters. Semantic homonymy is not disambiguated because of the manual character of this work; homonyms are assigned multiple semantic analyses.

Semantic tagging is based on a classification system developed for the database Lexicograph (http://lexicograf.ru/eng/main.html) beginning from 1992 under the supervision of E.V. Paducheva and E.V. Rakhilina at the Department of Linguistic Research at the All-Russian Institute of Scientific and Technical Information of the Russian Academy of Sciences. Since then, the dictionary was essentially expanded; several new semantic classes and the derivational parameters were added for the needs of the Corpus.

The semantic dictionary is based on the morphological dictionary of the DIALING system (120 thousand words) which is in turn an expansion of Zalizniak’s Grammatical Dictionary of Russian (2009).

Method and results

For a first study of parts-of-speech distribution in Russian texts based on model (2), the first 193 texts were selected from the Deeply Annotated Corpus (cf. Table 3); for each of the texts,

a rank-frequency count of the part-of-speech tags was performed. Formula (2) was fitted to each data set in four versions: with one to four exponential terms. As can be seen, most data sets were compatible with a single term function; all of them could be modelled in this way with an at least acceptable determination coefficient. However, all of them showed an improved fit when two terms were applied; some improved with a third and others even with a fourth term. Table 3 presents the adjusted coefficients of multiple determination[2] of all the goodness-of-fit tests for the 193 texts and for all cases where an additional term yielded an improved determination coefficient.

In some cases, Table 3 does not give a value. The "-" indicates that numerical problems occurred during the calculation.

Table 3

Determination coefficients of fitting experiments with formula (2). The function was fitted with two, three, and four terms to the data from 193 Russian texts. Column "Optimum" gives the number of terms needed for the best fit. *k* is the smallest frequency of the data set.

Text	**k**	**1 term**	**2 terms**	**3 terms**	**4 terms**	**Optimum**
Antiterror	1	0.9055	0.9846	-	0.9946	4
Assemb	5	0.9816	0.9797	0.9918	0.9884	3
Audit	1	0.9661	0.9536	-	0.9469	1
Auto	12	0.9476	0.9629	0.9506	0.9858	4
Avraam_Lincoln	2	0.9281	0.9809	0.9932	-	3
Beg_po_krugu	1	0.8898	0.9583	0.9683	-	3
Bezrab	1	0.8707	0.9676	0.9665	0.8907	2
Bezrybye	1	0.9426	0.9245	-	0.9666	4
Bitov1	1	0.9020	0.9140	0.8566	0.5698	2
blesnuli_masterstvom	3	0.9439	0.9789	-	-	2
Bob	2	0.9148	0.9789	0.9852	-	3
bolshe_kandidatov	6	0.9204	0.9716	-	-	2
Bronja	1	0.9145	0.9702	0.9503	-	2
Budget	29	0.9414	0.9734	-	-	2
Buran	3	0.9659	0.9848	-	0.9629	2
byt_modnym	1	0.9306	0.9138	0.8828	0.7871	1
Cats	17	0.9263	0.9798	0.9798	0.9798	2
Cennosti	1	0.8960	-	0.7987	0.5961	1
Chance	3	0.9534	0.9842	-	0.9652	2
Chelovek_na_tribune	1	0.9096	0.8804	0.8219	0.6423	1
Chuvstvo_spravedlivosti	1	0.9144	0.9795	0.9964	-	3
Cifr	28	0.8805	0.9750	-	-	2
firma_ne_garant	1	0.9357	0.9916	0.9959	0.9955	3
Fitoterapiya	1	0.9358	0.9917	0.9993	0.9962	3
Foto	12	0.9359	0.9918	0.9914	0.9967	4
Gadget	20	0.9360	0.9919	0.9915	0.9914	2
Gaidar	11	0.9361	0.9920	0.9759	0.9759	2
Gematologija	3	0.9362	0.9921	0.9824	0.9941	2
Golod	18	0.9363	0.9922	0.9838	0.9838	2

[2] This coefficient yields somewhat smaller values than the usually applied determination coefficient without adjustment.

Gorodoskoye_slonovodstvo	3	0.9364	0.9923	0.9935	0.9934	3
Grechko	2	0.9365	0.9924	0.9657	0.9657	2
Grekova1	3	0.9366	0.9925	0.9945	0.9970	4
Grekova2	1	0.9367	0.9926	0.9669	0.9828	4
Grekova3	1	0.9368	0.9927	0.9767	0.9956	2
Gubarev	2	0.9369	0.9928	0.9949	0.9820	3
Hleb	25	0.9370	0.9929	0.9912	0.9989	2
hol_voina	1	0.9371	0.9930	0.9912	0.9911	2
I k nim ne zarastet narodnaya tropa	6	0.9372	0.9931	0.9927	0.9927	2
I slepyye prozreiut	2	0.9373	0.9932	0.9558	0.9654	2
I_evro_takoi_molodoi	12	0.9374	0.9933	0.9827	0.9754	2
Igrushki_Bogov	1	0.9375	0.9934	0.9898	0.9732	2
inf_ob	2	0.9376	0.9935	0.9938	0.9949	4
Informtexnologii	1	0.9377	0.9936	0.9962	0.9916	3
Interaktiv	3	0.9378	0.9937	0.9857	0.9918	2
Internet-zavisimost´	2	0.9379	0.9938	0.9950	0.9859	3
Interv	1	0.9380	0.9939	0.9192	0.9188	2
istochniki_rosta	21	0.9381	0.9940	0.9797	0.9798	2
Istoriya_tualeta	9	0.9382	0.9941	0.9899	0.9898	2
iz_ognja	7	0.9383	0.9942	0.9937	0.9887	2
izobreteno_vo_sne	2	0.9384	0.9943	0.9932	0.9932	2
Journ	31	0.9385	0.9944	0.9897	0.9896	2
Kak_politikov	12	0.9386	0.9945	0.9913	0.9913	2
Kaprossija	6	0.9387	0.9946	0.9967	0.9967	3
Katastrofa	8	0.9388	0.9947	0.9936	0.9905	2
Khranilisha	4	0.9389	0.9948	0.9960	0.9967	4
Kobzon	2	0.9390	0.9949	0.9706	0.9706	2
Kodeks_molchanija	17	0.9391	0.9950	0.9760	0.9933	4
Kolybel	2	0.9392	0.9951	-	-	2
Kompartii	20	0.9393	0.9952	0.9859	0.9859	2
Komu dostanetsia Chingiskhan	7	0.9394	0.9953	0.9953	0.9953	2
Korp04_09	18	0.9395	0.9954	0.9957	0.9989	4
Korp06_01	1	0.9396	0.9955	0.9946	0.9985	4
Korp06_02	16	0.9397	0.9956	0.9897	0.9897	2
Korp06_03	6	0.9398	0.9957	0,9932	0.9961	4
Korp06_04	1	0.9399	0.9958	0.9938	0.9905	2
Korp06_05	1	0.9400	0.9959	0.9898	0.9896	2
Korp06_06	14	0.9401	0.9960	0.9849	0.9849	2
Korp06_07	17	0.9402	0.9961	0.9886	0.9886	2
Korp06_08	4	0.9403	0.9962	0.9950	0.9988	4
Korp06_09	20	0.9404	0.9963	0.9908	0.9908	2
Korp06_10	2	0.9405	0.9964	0.9961	0.9961	2
Korp06_11	1	0.9406	0.9965	0.9862	0.9862	2
Korp06_12	7	0.9407	0.9966	0.9883	0.9883	2
Korp06_13	17	0.9408	0.9967	0.9881	0.9953	2
Korp06_14	3	0.9409	0.9968	0.9849	0.9849	2

Korp06_15	2	0.9410	0.9969	0.9951	0.9989	4
Korp06_16	26	0.9411	0.9970	0.9939	0.9926	2
Korp06_17	19	0.9412	0.9971	0.9930	0.9956	2
Korp06_18	2	0.9413	0.9972	0.9947	0.9947	2
Korp06_19	62	0.9414	0.9973	0.9883	0.9883	2
Korp06_20	33	0.9415	0.9974	0.9917	0.9917	2
Korp06_21	1	0.9718	0.9719	0.9941	0.9981	4
Korp06_22	1	0.9287	0.9295	0.9936	0.9934	3
Korp06_23	1	0.9477	0.9913	0.9975	0.9913	3
Kopr06_24	6	0.9831	0.9941	0.9941	0.9941	2
Korp06_25	8	0.9431	0.9891	0.9891	0.9963	4
Korp06_26	1	0.9451	0.9864	0.9925	0.9972	4
Korp06_27	13	0.9602	0.9918	0.9917	0.9917	2
Korp06_28	4	0.9811	0.9886	0.9886	0.9936	2
Korp07_01	13	0.9758	0.9947	0.9947	0.9987	2
Korp07_02	5	0.9789	0.9947	0.9977	0.9977	3
Korp07_03	2	0.9687	0.9820	0.9915	0.9915	3
Korp07_04	3	0.9658	0.9677	0.9689	0.9910	4
Korp07_05	9	0.9684	0.9930	0.9953	0.9930	3
Korp07_06	1	0.9713	0.9957	0.9978	0.9978	3
Korp07_07	9	0.9362	0.9900	0.9900	0.9900	2
Korp07_08	4	0.9411	0.9455	0.9781	0.9789	4
Korp07_09	1	0.9300	0.9819	0.9923	0.9935	4
Korp07_10	16	0.9826	0.9956	0.9956	0.9956	2
Korp07_11	4	0.9696	0.9961	0.9961	0.9961	2
Korp07_12	1	0.9726	0.9883	0.9941	0.9946	4
Korp07_13	4	0.9620	0.9955	0.9985	0.9985	3
Korp07_14	2	0.9215	0.9230	0.9861	0.9916	4
Korp07_16	2	0.9470	0.9877	0.9877	0.9877	2
Korp07_17	10	0.9546	0.9923	0.9923	0.9923	2
Korp07_18	1	0.9576	0.9942	0.9942	0.9984	4
Korp07_19	8	0.9805	0.9812	0.9925	0.9959	4
Korp07_20	6	0.9738	0.9751	0.9914	0.9921	4
Korp07_21	3	0.9177	0.9185	0.9920	0.9920	3
Korp07_22	5	0.9369	0.9853	0.9852	0.9955	4
Korp07_23	11	0.9709	0.9965	0.9965	0.9965	2
Korp07_24	5	0.9819	0.9910	0.9910	0.9910	2
Korp07_25	6	0.9792	0.9918	0.9918	0.9918	2
Korp07_26	20	0.9775	0.9897	0.9897	0.9897	2
Korp2_1	43	0.9459	0.9921	0.9988	0.9920	3
Korp2_2	21	0.9367	0.9836	0.9912	0.9910	3
Korp2_3	1	0.9858	0.9916	0.9920	0.9940	4
Korp2_4	1	0.9327	0.9880	0.9938	0.9936	3
Korp2_5	3	0.9149	0.9863	0.9862	0.9930	4
korp2_6_1	2	0.9075	0.9084	0.9944	0.9943	3
korp2_6_2	7	0.9369	0.9864	0.9953	0.9964	4
korp2_6_3	1	0.9379	0.9849	0.9909	0.9909	3
korp2_6_4	10	0.9454	0.9932	0.9987	0.9986	3

korp2_6_5	14	0.9669	0.9955	0.9954	0.9979	4
korp2_6_6	6	0.9736	0.9969	0.9968	0.9968	2
kovka	1	0.9443	0.9838	0.9923	0.9921	3
krisis	17	0.9518	0.9850	0.9850	0.9850	2
kruglyj_stol	2	0.9062	0.9829	0.9829	0.9919	2
Krupnaya_kala	7	0.9036	0.9744	0.9744	0.9744	2
Kto_ubil	5	0.9434	0.9878	0.9878	0.9922	4
kumush	4	0.9043	0.9824	0.9944	0.9823	3
kvartira_na_depozit	59	0.9286	0.9813	0.9812	0.9812	2
led	2	0,9187	0.9695	0.9861	0.9859	3
lego	22	0.9432	0.9804	0.9964	0.9803	3
lesorub	45	0.9192	0.9760	0.9759	0.9758	2
levyj_povorot	50	0.9461	0.9919	0.9918	0.9982	2
Liubit_drakona	1	0.9364	0.9364	0.9895	0.9895	3
lunneye_kamni	4	0.9244	0.9871	0.9869	0.9869	2
lunokhod	1	0.9498	0.9498	0.9963	0.9962	3
lyzhi	6	0.9210	0.9725	0.9878	0.9876	3
mamleev_charlie	2	0.9611	0.9620	0.9628	0.9938	4
mamleev_svadba	1	0.9397	0.9599	0.9716	0.9807	4
Manekeny	7	0.9455	0,9458	0.9890	0.9887	3
Mars	7	0.9303	0.9728	0.9821	0.9726	3
Mars1	51	0.9475	0.9812	0.9810	0.9810	2
Mars2	1	0.9280	0.9774	0.9854	0.9859	4
Mech	22	0.9573	0.9887	0.9953	0.9951	3
Meinert-Ranks	1	0.9508	0.9830	0.9963	0.9964	3
metallovedenie	31	0.9764	0.9915	0.9914	0.9958	4
Metis	18	0.9598	0.9907	0.9907	0.9907	2
misha	1	0.9025	0.9675	0.9675	0.9849	4
molch	5	0,9419	0.9882	0.9882	0.9983	4
molod	6	0.9838	0.9924	0.9924	0.9924	2
Molotok	3	0.9374	0.9395	0.9867	0.9907	4
monopolizacija_kanalizacii	22	0.9161	0.9854	0.9852	0.9852	2
Moskva	3	0.9360	0.9902	0.9936	0.9936	3
msu	1	0.9178	0.9813	0.9864	0.9864	3
muzej	5	0.9833	0.9946	0.9968	0.9957	3
muzhej_ne_obeshchal	7	0.9089	0.9873	0.9873	0.9969	4
Na_levom_flange	6	0.9667	0.9905	0.9905	0.9947	2
nachal	10	0.9338	0.9893	0.9892	0.9946	4
Nagibin1	1	0.9733	0.9871	0.9947	0.9953	4
Nagibin2	15	0.9358	0.9828	0.9827	0.9900	4
Nagibin3	1	0.9395	0.9753	0.9963	0.9904	3
Nagibin4	2	0.9539	0.9539	0.9713	0.9713	3
nalog	2	0.9138	0.9857	0.9949	0.9948	3
Nalog_na_tainu	2	0.9051	0.9751	0.9870	0.9827	3
name	5	0.9328	0.9834	0.9934	0,9925	3
narkot	1	0.9056	0.9888	0.9944	0.9917	3
nds	1	0.9308	0.9881	0.9881	0.9525	3
Nebes	15	0.9154	0.9795	0,9834	0.9848	4

nedouch	2	0.9131	0.9874	0.9934	0.9931	3
neftjanye_kacheli	4	0.9953	0.9966	0.9968	0.9968	2
nehvat	27	0.9648	0.9922	0.9922	0.9964	4
nejron	1	0.9447	0.9903	0.9983	0.9982	3
Nelegalnaya_perepis´	2	0.9202	0.9861	0.9861	0.9960	4
nelishnyaya_formalnost	1	0.9414	0.9876	0.9875	0.9916	2
nelzia selbia delit_1	10	0.8973	0.9767	0.9767	0.9767	2
Nepotopliaemyye	7	0.9586	0.9894	0.9893	0.9893	2
nezhil	3	0.9374	0.9837	0.9900	0.9925	4
nichego_chelovecheskogo	1	0.9632	0.9885	0.9959	0.9959	3
nko	2	0.9335	0.9335	0.9954	0.9902	3
Nobelevskiye_premii	1	0.9755	0.9977	0.9994	0.9994	3
novyj_levyj	24	0.9548	0.9937	0.9937	0.9937	2
Obratnaya_reaktsiya	1	0.9719	0.9919	0.9919	0.9943	4
obrazov	5	0.9312	0.9864	0.9862	0.9940	4
obrush	24	0.9796	0.9935	0.9935	0.9935	2
Ochishenye_Olkhona	32	0.9205	0.9808	0.9809	0.9809	2
oka	4	0.9089	0.9793	0.9793	0.9963	4
Okhota	6	0.9578	0.9875	0.9875	0.9926	4
olenina	1	0.9160	0.9698	0.9882	0.9880	3
Opasnaya_blizost´	6	0.9140	0.9818	0.9817	0,9896	4
Optimizm	1	0.9279	0.9792	0.9957	0.9956	3

Our results support the assumptions made in (Popescu/Altmann/ Köhler 2009): As expected, the distributions of parts-of-speech in the Russian texts display a stratified behaviour, a fact which can be partly attributed to the small tagset used by the linguists who annotated the corpus. Texts annotated with a more differentiated tagset would probably show a reduced stratification but it would not disappear because some of the different parts-of-speech will still follow different distribution patterns.

Table 4
Frequency of the optimum number of exponential terms

Rank	**No. of terms**	**Frequency**
1	2	84
2	3	54
3	4	51
4	1	4

Table 4 shows the frequency distribution of the optimum number of terms in the fitted function. Only very few (four) texts suggest unstratified parts-of-speech whereas the overwhelming majority supports the above-cited hypotheses. Future studies will be needed to find out whether more terms than four would yield even better results (which the authors of the present paper do not assume), and – more interestingly – to determine the factors which lead to the observed differences in the degree of stratification.

References

Altmann, Gabriel, Köhler, Reinhard (1996). "Language Forces" and Synergetic Modelling of Language Phenomena. In: *Glottometrika 15*, 62-76. Trier: WVT.

Apresjan, J.D., Boguslavskij, I.M., Iomdin, L.L. et al. (1978). *Lingvističeskoe obespečenie v sisteme avtomatičeskogo perevoda tret´ego pokolenija*. Moskva: Sovet po kompleksnoj probleme „Kibernetika". AN SSSR.

Best, Karl-Heinz (1997). Zum Stand der Untersuchungen zu Wort- und Satzlängen. In: *Third International Conference on Quantitative Linguistics. Helsinki, 172-176.*

Best, Karl-Heinz (1994). Word class frequencies in contemporary German short prose texts. *Journal of Quantitative Linguistics 1, 144-147*.

Best, Karl-Heinz (1997). Zur Wortartenhäufigkeit in Texten deutscher Kurzprosa der Gegenwart. In: Best, Karl-Heinz (ed.), *Glottometrika 16, 276-285*. Trier: Wiss. Verlag Trier.

Best, Karl-Heinz (1998). Zur Interaktion der Wortarten in Texten. *Papiere zur Linguistik 58, 83-95.*

Best, Karl-Heinz (2000). Verteilungen der Wortarten in Anzeigen. *Göttinger Beiträge zur Sprachwissenschaft H. 4, 37-51.*

Best, Karl-Heinz (2001). Zur Gesetzmäßigkeit der Wortartenverteilungen in deutschen Pressetexten. *Glottometrics 1, 1-26.*

Hammerl, Rolf (1990). Untersuchungen zur Verteilung der Wortarten im Text. In: Hřebíček, Luděk (ed.), *Glottometrika 11: 142-156*. Bochum: Brockmeyer.

Kendall, Maurice G.; Babington Smith, B. (1939): The problem of m rankings. In: *The Annals of Mathematical Statistics 10(3), 275-287*.

Schweers, Anja; Zhu, Jinyang (1991). Wortartenklassifizierung im Lateinischen, Deutschen und Chinesischen. In: Rothe, Ursula (ed.), *Diversification processes in language: grammar: 157-165.* Hagen: Margit Rottmann Medienverlag.

Tuzzi, Arjuna; Popescu, Ioan-Iovitz; Altmann, Gabriel (2010). *Quantitative Analysis of Italian Texts*. Lüdenscheid: RAM-Verlag.

Zaliznjak, Andrej A. (2009). *Grammatičeskij slovar´ russkogo jazyka*. Moskva: AST-Press Kniga.

Ziegler, Arne (1998). Word Class Frequencies in Brazilian-Portuguese Press Texts. *Journal of Quantitative Linguistics 5, 269-280.*

Ziegler, Arne (2001). Word Class Frequencies in Portuguese Press Texts. In: Uhlířová, Ludmila, Wimmer, Gejza, Altmann, Gabriel & Reinhard Köhler (eds.), *Text as a Linguistic Paradigm: Levels, Constituents, Constructs. Festschrift in Honour of Luděk Hřebíček: 295-312.* Trier: Wissenschaftlicher Verlag Trier.

Ziegler, Arne; Best, Karl-Heinz; Altmann, Gabriel (2001). A contribution to text spectra. *Glottometrics 1, 97-108.*

"LexicoGraph," http://lexicograf.ru/eng/main.html, accessed on 2010-08-02 11:25:35

"Национальный корпус русского языка," http://www.ruscorpora.ru/, accessed on 2010-08-02 11:22:41

Glottometrics 20, 2010, 70-109

Dynamics of Word Length in Sentence

Fengxiang Fan, Dalian
Peter Grzybek, Graz
Gabriel Altmann, Lüdenscheid

Abstract. In the present article the length of words in individual positions in sentence is being scrutinized. The conspicuous increase of word length with increasing position in short sentences very regularly slows down with increasing sentence length.

Keywords: word length, sentence length, position in sentence, English, Hungarian, Indonesian, Latin, Russian, Slovak

1. Introduction

Linguistic laws may arise in two ways: either they are derived directly from an existing theory, or they represent an isolated hypothesis to be tested. If accepted, the hypothesis must be systematized, i.e. incorporated into a system of similar statements. This systematization can consist in showing that either the given hypothesis is a consequence of higher level laws, theories, or axioms, or that it can itself be used to draw consequences about other phenomena; in the latter case, the hypothesis serves as a kind of axiom or as the starting-point of a new theory. If none of these two procedures is possible, we are concerned with a local phenomenon, i.e. with testing an inductive hypothesis on restricted data. If we are lucky and the hypothesis can be maintained, this does not mean that it can simply be declared to be a law.

In searching for linguistic laws, there is a tendency to jump to generalizations, not sufficiently taking into account important boundary conditions. That is, we tend to tacitly accept the *ceteris paribus* condition because our argumentation seems to be linguistically well substantiated. Fortunately, occasional rejections of a hypothesis in some languages need not weaken the status of the law-candidate – on the contrary, they force us to refine it by includeing some necessary boundary conditions. Natural laws, too, hold only if some boundary conditions are fulfilled, and even mathematical theorems begin with famous sentences like "Let be given…, then …" or "Let hold that …, then…". With regard to linguistic studies on word length, Grzybek, Kelih and Stadlober (2008) pointed out a number of boundary conditions whose consideration may lead to different parameters in models, or even to different functions: such conditions are, among others, data scarcity, data homogeneity, language type, intra- vs. inter-textual approach, etc.

The present study focuses on word length distributions according to word positions within a sentence and according to the length of the sentence: We concentrate on the question if word length is stable in the course of a sentence, from its beginning to its end, or if there is a particular change or development of word length, depending on the length of the sentence. Since we proceed inductively, the first step consists of scrutinizing texts in several languages in order to achieve more corroboration so as to reduce language-specific boundary conditions possibly coming into play.

Here we want to study a hypothesis concerning the interrelation between word length and its position in sentence. In this context, word length is determined in terms of the number

of syllables per word, sentence length in terms of the number of words per sentence – though one should better measure it in terms of the number of clauses (or phrases), since these are the direct constituents of sentences. This means that we search for some specific form of self-regulation – if there is any – and for the boundary conditions under which it holds.

Most probably it was A. Niemikorpi (1991) who first observed that, in Finnish sentences, the longest words tend to occur at the end of a sentence, or a clause.[1] Niemikorpi's study was based on Finnish corpus material from the 1960s, including oral and written sources. From five basic text categories (non-fiction, press reportage, fiction, and informal standard language), which were further subdivided into 58 subcategories, 100 delimited, classified text samples consisting of five sentences or approximately 60 textual words were selected by random sampling. As a result, a corpus of ca. 430,000 running words was obtained and submitted to detailed analysis.

Niemikorpi's first calculation showed that, except for very short sentences, the average length of running words (measured in the number of graphemes per word) is greater at the end of a sentence than in the middle or at the beginning; it was also found that in a sentence the word in initial position is normally longer than the word that follows it.

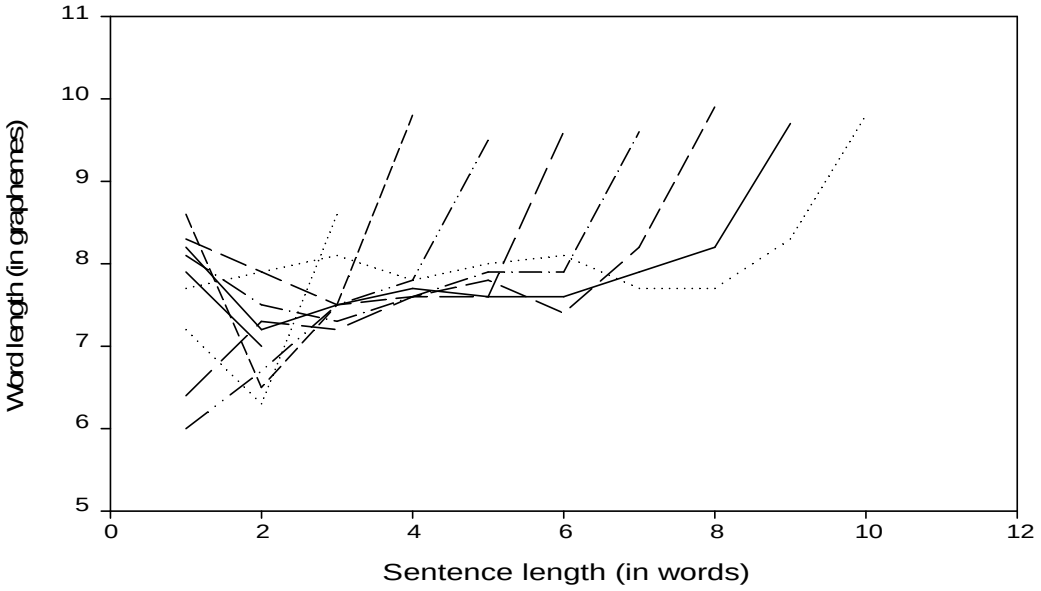


Figure 1. Average word length in different positions in sentences of different length (Niemikorpi 1991: 287)

Basing the same kind of calculation on clauses, not on sentences, Niemikorpi found that in this case, too, word length almost invariably increases towards the end. According to Niemikorpi, this regularity applies equally well to the length of running words in both main and subordinate clauses; Figure 3 shows the tendency for main clauses.

Given these observations, Niemikorpi suggested the English term "scoop law" to denote this phenomenon, considering this term as an equivalent for the Finnish term 'viskurilaki' an adaptation from the term "Wannmühlegesetz", previously used in Finnish folklor-

[1] Unfortunately, in Saukonen's (1994) summary of Niemikorpi's work, only the term 'clause' is used, ignoring Niermikorpi's detailed and comparative study of main and subordinate clauses as well as complete sentences.

istics, to describe the fact that in the verses of the Finnish folk epic Kaleva, the heaviest elements tend do occur at the end of stanzas.[2]

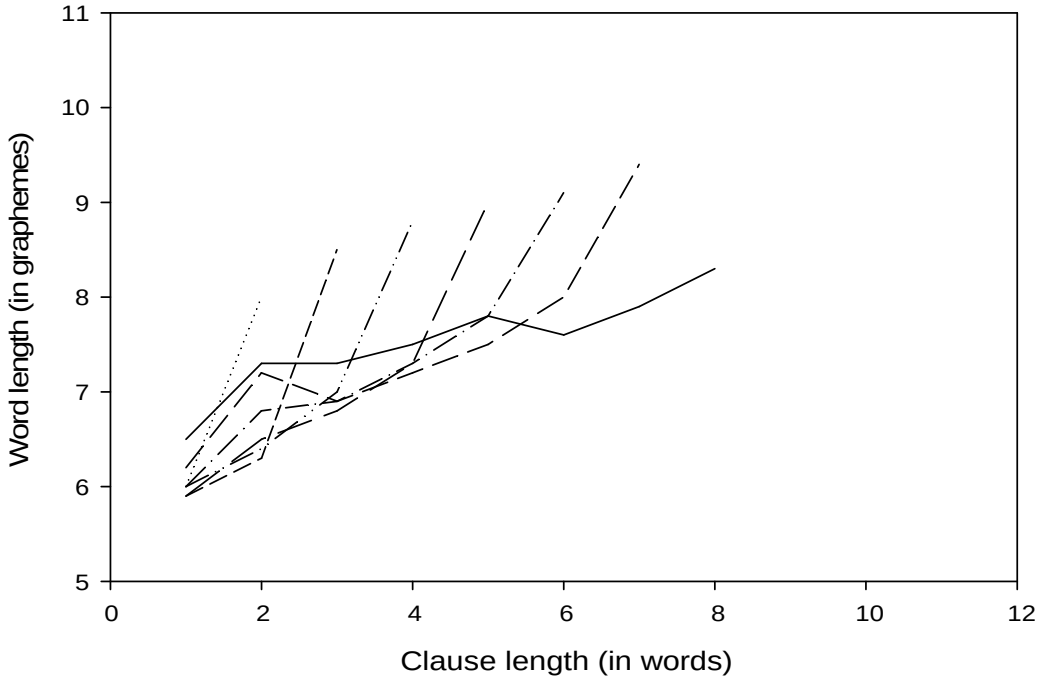


Figure 2. Average word length in different positions in clauses of different length (Niemikorpi 1991: 287)

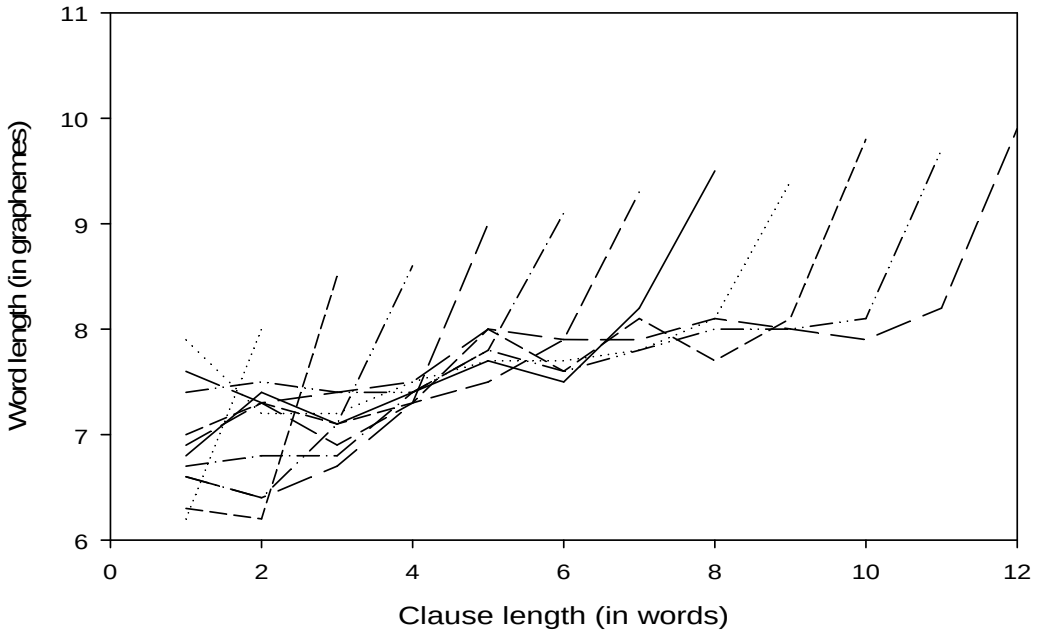


Figure 3. Average word length in different positions in main clauses of different length (Niemikorpi 1991: 292)

[2] By way of a general explanation, Niemikorpi referred to the fact that in case of Finnish, we are concerned with an SVO language, in which the subject precedes the predicate and the verb qualifiers are placed at the end. Also, the theme-rheme structure of clauses is likely to play a crucial role here.

There is, however, one major drawback with Niemikorpi's study: measuring word length in the number of graphemes per word may heavily skew the results obtained, or even lead to erroneous results. The reason for this is that graphemes (even less than phones or phonemes) are no direct constituents of the word; rather, units like syllables or morphemes should be taken as measuring units. Last not least, the necessity to proceed this way is related to the Menzerathian fact that long words tend to be composed of short syllables, and short words of long syllables. A "long" word measured in terms of graphemes, however, may consist either of many short syllables, or of a few long syllables – as a result, on the basis of Niemikorpi's study we cannot be sure about word length neither in general, nor with regard to sentence position, specifically.

In a subsequent study, Uhlířová (1997a,b) has analyzed the relation between word length and position in sentence with Czech material, in order to study a language typologically different from Finnish. For this purpose, she analysed ten short stories for schoolchildren by Czech writer Miloš Macourek (1926-2002), summing up to a corpus of ca. 10,000 running words. In contrast to Niemikorpi, Uhlířová did not analyze word length in clauses, but in sentences only[3], confining the presentation of results to the interval from 1 to 12 words per sentence, since longer sentences were considered to occur too rarely in the corpus to provide reliable results. Moreover, in her study word length was measured in the number of syllables per word.

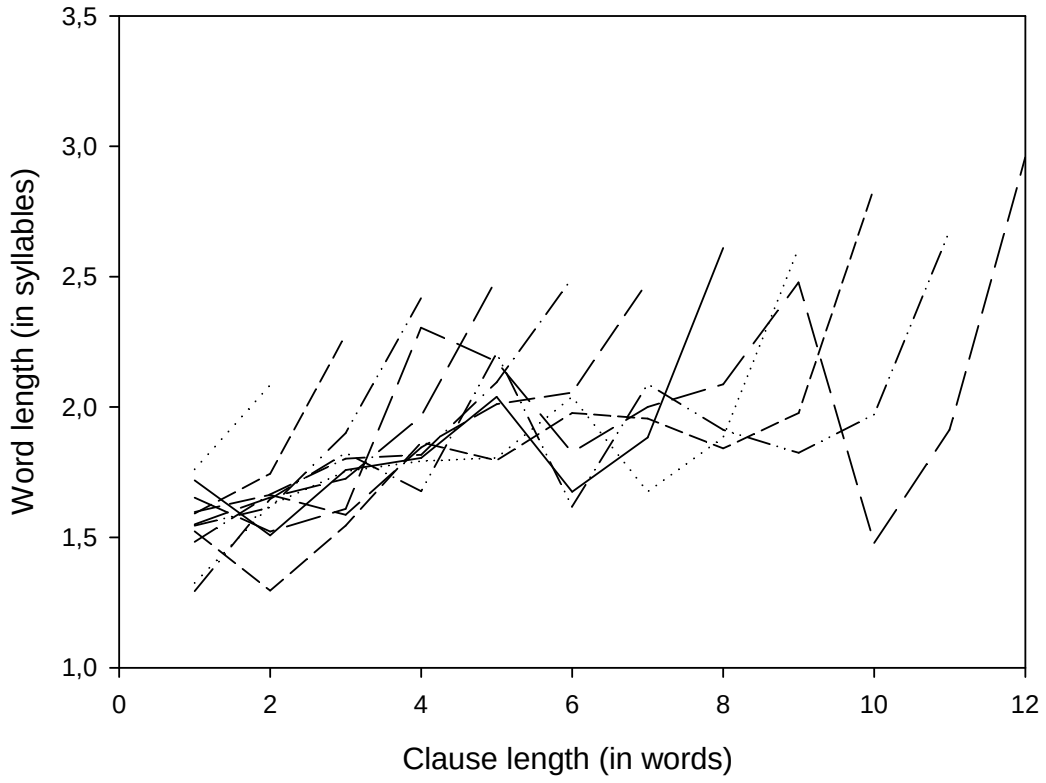


Figure 4. Average word length in sentences of different length in a corpus of Czech short stories (Uhlířová 1997a: 182, 1997b: 269)

[3] Unfortunately, the original term 'věta' (= sentence) used in Uhlířová's (1997a) original Czech text, has been translated as 'clause' in Uhlířová's (1997b) English version, thus given rise to serious misinterpretations.

According to Uhlířová, the two most salient points are those at the beginning of a sentence (i.e. the first position), and those at the end (i.e. the last position): the first position is the position of short words, whereas the end is the position of long words. With the exception of sentence length 1, the average word length is lowest in the first position with sentences of any length, word length then gradually increasing from position to position, at least up to sentence length 8, when occurrences become too rare to represent reliable results. Uhlířová (1997b: 269) suspected that the quantitative relationship between length and order of words in sentences need not be language specific, rather it may be of a more general nature, lending itself to be interpreted as a "quantitative manifestation of the structural principle of end weight".

In fact, a closer inspection of Figures 1 to 4 gives rise to the impression that not only the last, but also the first word in a sentence seem to play a specific role, as far as their length is concerned (at least on the average). We have no idea thus far, to what kind of linguistic factors this tendency might be related, in how far language specifics, discourse or sentence typological aspects, or other factors come into play, etc. Without a doubt, a more detailed study of this question is an obligatory task for future research, which cannot be tackled in the framework of this article.

Summarizing, one can say that, both for Finnish and Czech, similar tendencies seem to have been observed, in so far as words at the end of clauses and/or sentences tend to be longer. Such a tendency would seem to be reasonable, if one takes into account the fact that words occurring towards the end of an utterance carry more information than words at its beginning; hence these words do not occur frequently and, as a consequence of the well-known Zipfian rule, tend to be longer. As a result, we would be concerned with some specific kind of regulation between the position, information, length and frequency of words.

Given the situation outlined above, a number of open issues to be solved remain open. The first question is, if the above tendency holds true more generally for other languages, as well. The second question concerns the point, in how far the observed tendency characterizes clauses and sentences in the same way. On the one hand, it may be, theoretically speaking, possible that the above observations with regard to sentences are but a special case (i.e. a consequence) of tendencies ultimately characterizing clauses, since any sentence is, of course, composed of one or more clauses. On the other hand, this would mean that the course of a sentence composed of more than one clause would display some specific kind of oscillation, or steps, around clause boundaries, a phenomenon which has never before been reported about.

It goes without saying that, in the present study, we will not be able to give answers to all of these questions. We will therefore concentrate on sentences only, extending the empirical data base to more languages. In addition to this, in case the results will corroborate the findings reported above, we will attempt to model the overall trend common to all languages studied. Of course, whatever turns out to be the case, is likely to hold on the average only; further generalizations will automatically give rise to further hypotheses.

Linguistically, this hypothesis is reasonable: the "topic" introduced at the beginning of the sentence obtains "comments" or "predicates" whose length increases with increasing sentence. As has been said above, this phenomenon has been studied for Finnish (Niemikorpi 1991) and for Czech (Uhlířová 1997a,b) with regard to whole sentences. It should be mentioned that these two languages are both strongly synthetic. But what is the situation in languages tending towards analytism like English? How is the situation in languages like Japanese, having many long words, or in some American Indian languages, in which whole sentences can be formed of one word only? Do they conform to one of the alternatives of our hypothesis? And if not, how do we have to modify the hypothesis? Almost as a matter of fact, a second question emerges: If there is a trend, how can it be characterized: is it linear or not?

And, if it is linear, does the slope of the straight line increase with increasing sentence length? And if it is not linear, what form does it then have?

Taking into account that there are a number of uncertainty factors arising from previous research, the only way to get some security is to proceed inductively, by way of analyzing several languages and then elaborating on the results obtained. The languages chosen to test our hypothesis quite logically included languages with different typological profiles: Russian, Slovak, Latin are strongly inflecting, Hungarian is strongly agglutinative, English tends to analytism, Indonesian is moderately agglutinative.

For each of these languages, we shall use texts of different origin and different length, and we will study the progression of averages within sentences of a given length. In detail, the following hypotheses can be tested in isolation:

Hypothesis I:
Mean word length increases from the beginning to the end of a sentence

against the null hypothesis that no change of word length occurs. Given hypothesis (I) can be corroborated, we will then test the subsequent hypothesis (II):

Hypothesis II:
Word length increase from the beginning to the end of a sentence follows a common principle.

2. Testing

The hypothesis is very simple; it may, in fact, be scrutinized with longer individual texts, or a group of texts of the same sort (thus guaranteeing data homogeneity), maybe with non-homogeneous samples of texts from one and the same language. For each analysis performed in this study (see Appendix for the sources), groups of sentences with identical sentence length were formed, and for each individual length group, mean word length in the individual positions was calculated. Possible specifics of the length of first and last words in sentences (see above) are going to be ignored, a sentence rather being taken "as a whole", including first and last word.

In our first analyses it turned out quite soon that for short individual texts there seems to be no clearly overt tendency; this is most probably caused by the small number of sentences of equal length, displaying a vast variance and therefore rendering the count of averages non-reliable. In these cases – cf. the analyses of Russian and Slovak journalistic texts, and English below –, all individual texts were merged into a corpus of texts, and sentences of equal length from the whole corpus were pooled, in order to increase reliability; mean lengths in the given word position was then computed group-wise.

The results for all languages are presented in Tables 1 to 6: Column 1 contains the sentence length in terms of the number of words per sentence; Column 2 contains the mean word-lengths in a given position within a sentence; Column 3 shows the linear regression equation for word length depending on position; Column 4 (R^2) shows the determination coefficient of the regression; Column 5 contains the mean word length in the whole sentence; Column 6 (k) presents the number of sentences of length n in the given sample.

Still, even when merging individual texts to a larger corpus, the problem of data reliability remains, a problem which has not been solved even in statistics. What is a reliable sample size? In the present investigation we adhere to the following principles:

a. The number of sentences of a given length must be at least $k \geq 10$.

b. Sentences with less than 3 words are omitted from the analyses, and only sentences of length $n \geq 3$ are scrutinized, because lengths 1 and 2 do not make sense for regression analyses.
c. If sentences with length n = 3,4,… are represented by less than $k \geq 10$ sentences they, too, are excluded from analysis which begins with the "well represented" lengths.
d. At the upper end of length as well, sentences are scrutinized only up to the length n having minimal frequency of $k \geq 10$; more exactly, all those occurrences (possibly including such with less than 10) are taken account of up to that point, where there is no more occurrence of $k \geq 10$ in the whole sample.

2.1. Russian

As was mentioned above, no reliable results were obtained for relatively short individual journalistic texts; therefore data were pooled, and a corpus of 44 Russian journalistic texts, consisting of 860 sentences with 15001 words, was obtained and submitted to analysis. The results are presented in Table 1. The length n = 3, 4 are not well represented, with only 3 or 5 occurrences only, and are omitted from analysis. At length n = 31, there seems to be a break after which oscillation begins. As to an explanation of this break, it is possible that, in addition to statistical reasons – the data points above sentence length 30 not being represented by sufficient observations – linguistic factors come into play (cf. Grzybek et al. 2008); this problem will have to be studied in future with more voluminous data.

With regard to the regression analyses carried out in the present study, it should be mentioned that we are not so much interested in the goodness of the fitting results for the individual regression of each sentence length; rather, we are concerned with the tendency along sentences of different length. For the sake of simplicity we use a linear regression, even if in many cases different functions would be more appropriate to express the positional changes. Again, a thorough study as to which function(s) might be more appropriate, must be left to future research.

Table 1
Mean lengths (L) of words in individual positions (x) in sentences of length n in 44 Russian journalistic texts from the journal "Vremja" (M = mean of all positions, k = number of sentences)

n	**Mean syllabic lengths of words** ***L***	**Linear relation** $L = a + bx$	R^2	***M***	***k***
3	3.33, 1.87, 3.67	-	-	-	3
4	3.00, 2.80, 3.20, 2.40	-	-	-	5
5	2.75, 2.75, 2.92, 2.92, 3.00	2.6670 + 0.0670x	0.89	2.87	12
6	2.07, 2.72, 2.72, 2.69, 2.48, 3.07	2.2000 + 0.1214x	0.47	2.63	29
7	2.33, 2.40, 3.13, 3.13, 2.33, 3.10, 3.20	2.3443 + 0.1146x	0.34	2.80	30
8	2.26, 2.67, 2.67, 2.81, 2.56, 2.67, 2.63, 3.22	2.3504 + 0.0746x	0.47	2.69	27
9	2.65, 2.76, 2.81, 2.86, 3.14, 2.62, 2.62, 2.54, 3.11	2.7433 + 0.0093x	0.01	2.79	37
10	2.43, 2.64, 2.75, 2.77, 2.48, 2.77, 2.89, 2.77, 2.75, 3.36	2.4313 + 0.0599x	0.51	2.76	44
11	2.66, 2.90, 2.59, 3.17, 2.85, 2.71, 2.54,	2.7311 + 0.0116x	0.03	2.80	41

	2.66, 2.88, 2.61, 3.24				
12	2.40, 2.58, 2.37, 2.35, 2.40, 3.12, 2.86, 2.81, 2.56, 2.65, 2.98, 2.93	2.3627 + 0.0469x	0.40	2.67	43
13	2.22, 2.54, 3.02, 2.38, 2.82, 2.68, 2.74, 2.66, 2.66, 2.58, 2.76, 2.86, 3.24	2.4377 + 0.0381x	0.32	2.70	50
14	2.34, 2.83, 2.71, 3.12, 2.78, 3.15, 3.05, 2.61, 2.83, 2.71, 2.83, 2.63, 2.54, 3.02	2.7775 + 0.0025x	0.00	2.80	41
15	2.37, 2.61, 2.50, 2.76, 2.97, 2.63, 3.03, 3.11, 2.82, 2.45, 2.76, 3.08, 2.47, 2.89, 2.97	2.5953 + 0.0208x	0.14	2.76	38
16	2.31, 2.28, 2.90, 2.69, 2.90, 2.48, 3.00, 2.76, 3.03, 2.93, 3.00, 2.55, 2.90, 3.00, 3.38, 3.21	2.4450 + 0.0456x	0.51	2.83	29
17	2.57, 2.84, 2.68, 2.50, 2.98, 2.59, 2.86, 2.77, 2.57, 2.68, 2.36, 2.91, 2.75, 2.48, 2.80, 2.91, 3.18	2.6222 + 0.0121x	0.09	2.73	44
18	2.28, 2.69, 2.67, 2.82, 2.67, 3.05, 2.59, 2.77, 2.74, 3.08, 2.54, 2.44, 2.62, 2.69, 2.74, 2.79, 2.26, 2.95	2.6718 + 0.0017x	0.00	2.69	39
19	2.54, 2.80, 2.37, 2.93, 2.32, 2.37, 3.05, 2.71, 2.63, 2.63, 2.37, 2.93, 2.78, 3.10, 2.93, 2.71, 2.39, 2.78, 3.17	2.5414 + 0.0170x	0.13	2.71	41
20	2.47, 2.74, 2.91, 2.65, 3.32, 3.12, 3.12, 2.85, 2.62, 2.35, 2.56, 2.59, 2.85, 2.82, 2.79, 2.71, 2.82, 2.85, 2.65, 3.26	2.7778 + 0.0024x	0.00	2.80	34
21	2.30, 2.81, 2.85, 3.11, 2.48, 3.00, 2.74, 2.96, 2.63, 2.48, 3.11, 3.04, 2.33, 2.59, 2.59, 2.33, 2.63, 2.63, 2.15, 2.56, 3.07	2.7954-0.0100x	0.05	2.69	27
22	2.20, 2.80, 2.30, 3.17, 2.97, 2.83, 2.77, 2.87, 2.43, 2.77, 2.63, 2.23, 3.40, 2.80, 2.77, 2.70, 3.03, 2.40, 2.73, 2.50, 3.57, 2.90	2.6134 + 0.0129x	0.06	2.76	30
23	2.03, 2.47, 2.69, 2.56, 2.56, 2.38, 2.81, 2.88, 2.91, 2.88, 2.88, 2.84, 3.13, 2.47, 2.56, 2.84, 2.59, 2.94, 2.72, 2.59, 2.69, 2.91, 3.34,	2.4799 + 0.0204x	0.26	2.72	32
24	2.32, 2.64, 2.64, 2.77, 3.00, 2.86, 3.41, 2.36, 2.91, 2.77, 3.36, 3.05, 3.09, 2.50, 2.77, 2.77, 2.36, 2.95, 2.82, 2.50, 2.50, 2.95, 2.77, 3.41	2.7437 + 0.0054x	0.02	2.81	22
25	2.24, 2.52, 2.76, 2.36, 2.64, 2.36, 2.28, 2.88, 2.36, 2.80, 2.64, 2.68, 2.84, 2.40, 2.32, 2.60, 2.76, 2.32, 2.80, 2.68, 2.56, 2.64, 2.36, 3.04, 2.96	2.4432 + 0.0114x	0.13	2.59	25
26	2.67, 2.87, 2.53, 3.07, 2.73, 2.67, 2.47, 2.87, 3.27, 3.20, 2.20, 2.87, 3.07, 2.47, 3.13, 3.20, 3.33, 2.53, 3.27, 2.73, 3.07, 2.60, 2.93, 2.73, 2.80, 3.13	2.7584 + 0.0077x	0.04	2.86	15
27	2.94, 2.22, 2.72, 2.94, 2.94, 3.33, 3.44,	2.8913 - 0.0123x	0.09	2.72	18

	2.94, 2.72, 2.33, 2.22, 2.78, 3.00, 2.72, 2.72, 2.33, 2.89, 3.06, 2.72, 2.72, 2.33, 2.72, 2.17, 2.61, 2.56, 2.44, 2.89				
28	2.06, 2.65, 2.41, 2.59, 2.94, 2.65, 2.76, 2.82, 2.94, 2.71, 2.47, 2.29, 2.41, 2.76, 2.35, 2.53, 2.82, 3.12, 3.24, 2.12, 1.88, 2.18, 2.94, 2.47, 3.00, 2.88, 3.24, 3.18	2.4893 + 0.0116x	0.07	2.66	17
29	2.83, 2.11, 3.00, 2.83, 3.11, 3.00, 2.56, 3.28, 2.78, 2.89, 2.50, 2.50, 2.72, 2.44, 1.94, 2.67, 3.00, 2.89, 2.11, 2.17, 2.50, 3.11, 2.39, 3.39, 2.00, 3.17, 2.50, 2.33, 3.22	2.7538 - 0.0044x	0.01	2.69	18
30	2.22, 2.33, 3.11, 2.56, 3.11, 3.33, 2.44, 3.44, 3.56, 2.44, 2.44, 2.89, 2.67, 2.67, 2.78, 3.22, 3.56, 1.78, 2.89, 2.33, 3.00, 2.67, 2.89, 3.33, 3.11, 3.33, 3.56, 2.22, 1.67, 2.56	2.8300 - 0.0017x	0.00	2.80	9
31	1.80, 2.50, 2.70, 3.20, 2.50, 3.20, 3.00, 2.40, 2.80, 2.90, 1.80, 2.90, 3.10, 2.40, 2.90, 2.30, 2.80, 2.40, 2.30, 3.40, 2.30, 3.00, 2.00, 2.50, 2.90, 2.20, 1.60, 2.20, 2.00, 2.70, 2.50	2.7535 - 0.0124x	0.06	2.55	10

Considering the results presented in Table 1a, we may draw the following conclusions: Parameter *b* (cf. Column 3), indicating the positional change (increase or decrease) of word length, takes positive values with shorter sentences; this means that there is an increase of word length from the beginning of a sentence towards its end. To show the strength of the relation, we use the determination coefficient (R^2).[4] However, the increase gets slower as sentence length increases; this can be seen in Figure 5: at length $n = 22$, the values for *b* may become negative, and from this point on, *b* oscillates between positive and negative values, i.e. the observed tendency ceases to exist. This means that when a sentence attains a certain number of words, word length may even decrease towards the end of the sentence.

The trend is illustrated in Figure 5; with regard to the analyses of other texts and languages to follow below, the *x*- and *y*-axes are chosen in intervals identical for all Figures. In the given case, the trend can be captured by a straight line $b = 0.0859 - 0.0034n$ (in the interval $5 \geq n \leq 31$); a slightly better fitting can be attained by the function $b = 0.2349 \exp(-0.1616n)$, which will be used in the subsequent analysis, here yielding $R^2 = 0.69$.

Thus, as a result, hypothesis (I) is only partly corroborated: it holds for short sentences, but for long sentences (with the addition of subordinate clauses), the tendency disappears: towards the end of long sentences, average word length may cease to increase (i.e. increase may approximate zero).

It is possible of course that this conclusion will have to be modified by future research: after all, a long sentence consists of a number of clauses, and the addition of clause-specific rules may completely change the image.

[4] The values of the *t*- and *F*-tests were always satisfactory but mostly equal because of the software limitations

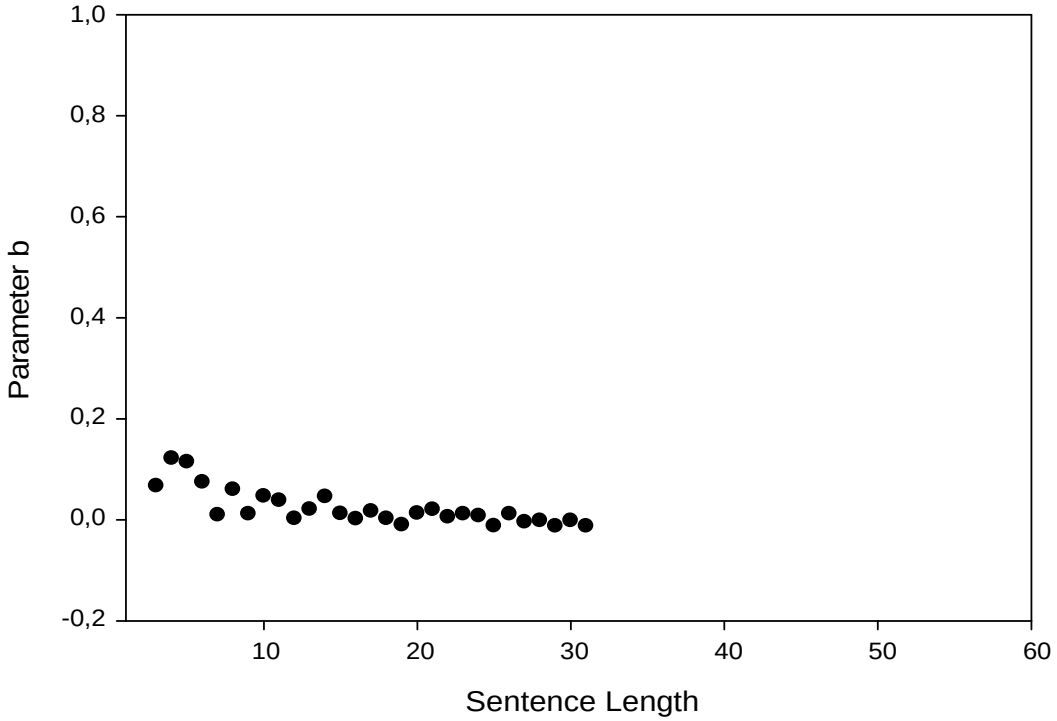


Figure 5. Decrease of coefficient *b* in a corpus of 44 Russian journalistic texts

In order to avoid the bias that we are concerned with a corpus-biased effect (due to data heterogeneity), it seems reasonable to test the hypothesis (and our above conclusion) on broader data material.

For this purpose, we now analyze a larger Russian text, namely L.N. Tolstoj's *Anna Karenina* in its complete length (with 19297 sentences and 258384 running words), in order to obtain a reliable number of sentences of all lengths from a single text (which still may be, of course, be composed of heterogeneous elements). This novel consists of 19297 sentences, so the data basis should be sufficient in case our hypothesis holds. The results are presented in Table 2. Since in *Anna Karenina*, more sentence lengths are (reliably) represented, we can take the interval $3 \leq n \leq 58$.

Table 2
Mean lengths (*L*) of words in individual positions (*x*) in sentences of length *n* in L.N. Tolstoj's *Anna Karenina* (*M* = mean of all positions, *k* = number of sentences)

n	**Mean syllabic lengths of words *L***	**Linear relation *L = a + bx***	R^2	***M***	***k***
3	1.64, 2.02, 2.48	1.2067 + 0.4200x	1.00	2.05	1093
4	1.63, 1.98, 2.06, 2.47	1.3850 + 0.2600x	0.95	2.04	1296
5	1.65, 2.06, 2.12, 2.17, 2.52	1.5490 + 0.1850x	0.89	2.10	1294
6	1.58, 2.01, 2.06, 2.14, 2.21, 2.60	1.5220 + 0.1651x	0.88	2.10	1145
7	1.66, 2.07, 2.05, 2.13, 2.18, 2.27, 2.59	1.6614 + 0.1186x	0.85	2.14	1115
8	1.68, 2.10, 2.10, 2.11, 2.18, 2.27, 2.22, 2.59	1.7518 + 0.0899x	0.77	2.16	992

9	1.73, 2.15, 2.10, 2.13, 2.20, 2.23, 2.24, 2.24, 2.65	1.8247 + 0.0722x	0.71	2.19	975
10	1.68, 2.10, 2.11, 2.17, 2.11, 2.15, 2.15, 2.22, 2.31, 2.67	1.8033 + 0.0661x	0.69	2.17	874
11	1.69, 2.08, 2.15, 2.21, 2.16, 2.19, 2.23, 2.13, 2.23 , 2.24, 2.75	1.8551 + 0.0554x	0.57	2.19	840
12	1.68, 2.14, 2.09, 2.22, 2.24, 2.31, 2.27, 2.15, 2.32, 2.22, 2.31, 2.67	1.9111 + 0.0473x	0.58	2.22	736
13	1.75, 2.15, 2.10, 2.14, 2.21, 2.25, 2.25, 2.29, 2.20, 2.25, 2.26, 2.37, 2.66	1.9312 + 0.0415x	0.66	2.22	648
14	1.76, 2.15, 2.17, 2.21, 2.21, 2.17, 2.26, 2.27, 2.25, 2.17, 2.29, 2.25, 2.34, 2.66	1.9765 + 0.0332x	0.57	2.23	643
15	1.76, 2.19, 2.20, 2.16, 2.18, 2.23, 2.24, 2.30, 2.35, 2.21, 2.23, 2.21, 2.28, 2.28, 2.66	2.0131 + 0.0274x	0.48	2.23	598
16	1.87, 2.24, 2.14, 2.27, 2.18, 2.18, 2.23, 2.22, 2.21, 2.17, 2.28, 2.26, 2.21, 2.25, 2.29, 2.71	2.0470 + 0.0218x	0.42	2.23	534
17	1.78, 2.10, 2.14, 2.18, 2.10, 2.25, 2.25, 2.37, 2.23, 2.29, 2.19, 2.26, 2.30, 2.13, 2.29, 2.31, 2.70	2.0046 + 0.0248x	0.48	2.23	489
18	1.77, 2.13, 2.16, 2.26, 2.25, 2.24, 2.22, 2.21, 2.28, 2.22, 2.23, 2.26, 2.23, 2.18, 2.26, 2.22, 2.27, 2.74	2.0444 + 0.0195x	0.37	2.23	489
19	1.86, 2.21, 2.32, 2.21, 2.09, 2.27, 2.31, 2.34, 2.25, 2.06, 2.28, 2.17, 2.28, 2.19, 2.24, 2.28, 2.37, 2.44, 2.75	2.0721 + 0.0187x	0.36	2.26	373
20	1.87, 2.23, 2.10, 2.15, 2.25, 2.17, 2.18, 2.36, 2.34, 2.20, 2.18, 2.18, 2.26, 2.18, 2.24, 2.27, 2.20, 2.17, 2.25, 2.81	2.0736 + 0.0148x	0.27	2.23	362
21	1.89, 2.27, 2.12, 2.27, 2.34, 2.23, 2.26, 2.18, 2.18, 2.20, 2.38, 2.33, 2.17, 2.31, 2.35, 2.27, 2.18, 2.28, 2.34, 2.22, 2.64	2.1293 + 0.0117x	0.28	2.26	320
22	1.82, 2.19, 2.19, 2.22, 2.20, 2.18, 2.19, 2.26, 2.33, 2.40, 2.25, 2.32, 2.18, 2.25, 2.36, 2.18, 2.27, 2.22, 2.31, 2.25, 2.52, 2.68	2.0832 + 0.0156x	0.41	2.26	330
23	1.75, 2.14, 2.18, 2.21, 2.27, 2.28, 2.35, 2.21, 2.32, 2.39, 2.30, 2.24, 2.35, 2.42, 2.37, 2.19, 2.21, 2.32, 2.30, 2.19, 2.24, 2.42, 2.75,	2.1109 + 0.0139x	0.31	2.28	271
24	1.90, 2.27, 2.20, 2.21, 2.32, 2.32, 2.22, 2.35, 2.33, 2.32, 2.39, 2.28, 2.32, 2.21, 2.09, 2.20, 2.11, 2.30,	2.1860 + 0.0060x	0.09	2.26	244

	2.20, 2.22, 2.32, 2.16, 2.34, 2.69				
25	1.94, 2.24, 2.17, 2.19, 2.26, 2.32, 2.37, 2.20, 2.30, 2.36, 2.27, 2.20, 2.21, 2.30, 2.24, 2.14, 2.04, 2.24, 2.20, 2.26, 2.23, 2.32, 2.36, 2.45, 2.72	2.1452 + 0.0089x	0.21	2.26	230
26	1.86, 2.06, 2.25, 2.23, 2.32, 2.16, 2.21, 2.29, 2.20, 2.31, 2.36, 2.36, 2.21, 2.16, 2.29, 2.23, 2.46, 2.17, 2.40, 2.24, 2.24, 2.30, 2.16, 2.25, 2.30, 2.71	2.1307 + 0.0095x	0.24	2.26	199
27	1.92, 2.22, 2.08, 2.19, 2.29, 2.14, 2.18, 2.12, 2.20, 2.29, 2.24, 2.17, 2.28, 2.26, 2.34, 2.32, 2.11, 2.25, 2.39, 2.19, 2.26, 2.31, 2.24, 2.48, 2.33, 2.27, 2.71	2.0873 + 0.0117x	0.43	2.25	194
28	1.88, 2.19, 2.06, 2.30, 2.13, 2.28, 2.22, 2.08, 2.12, 2.29, 2.26, 2.30, 2.06, 2.07, 2.30, 2.36, 2.18, 2.17, 2.24, 2.12, 2.13, 2.13, 2.19, 2.29, 2.08, 2.43, 2.26, 2.84	2.0866 + 0.0087x	0.18	2.21	189
29	1.94, 2.13, 2.43, 2.35, 2.21, 2.09, 2.26, 2.09, 2.26, 2.18, 2.14, 2.42, 2.58, 2.18, 2.22, 2.42, 2.35, 2.31, 2.21, 2.40, 2.27, 2.30, 2.25, 2.28, 2.12, 2.19, 2.27, 2.36, 2.67	2.1785 + 0.0062x	0.12	2.27	159
30	1.83, 2.29, 2.22, 2.27, 2.21, 2.23, 2.08, 2.17, 2.31, 2.29, 2.20, 2.12, 2.31, 2.37, 2.25, 2.23, 2.30, 2.37, 2.37, 2.32, 2.32, 2.23, 2.17, 2.12, 2.21, 2.38, 2.31, 2.20, 2.57, 2.61	2.1356 + 0.0082x	0.26	2.26	145
31	1.70, 2.12, 2.17, 2.16, 2.26, 2.21, 2.33, 2.23, 2.12, 2.21, 2.31, 2.40, 2.19, 2.28, 2.25, 2.48, 2.55, 2.23, 2.28, 2.20, 2.23, 2.18, 2.31, 2.19, 2.34, 2.26, 2.38, 2.32, 2.35, 2.44, 2.86	2.0939 + 0.0114x	0.33	2.28	130
32	2.03, 2.19, 2.12, 2.37, 2.09, 2.24, 2.12, 2.19, 2.24, 2.18, 2.08, 2.31, 2.43, 2.30, 2.30, 2.21, 2.45, 2.35, 2.26, 2.39, 2.15, 2.39, 2.27, 2.30, 2.04, 2.27, 2.07, 2.07, 2.33, 2.24, 2.19, 2.83	2.1676 + 0.0050x	0.09	2.25	115
33	1.88, 2.09, 2.07, 2.24, 2.41, 2.02, 2.25, 2.25, 2.18, 2.09, 2.21, 2.35, 2.04, 2.32, 2.12, 2.10, 2.25, 2.27, 2.36, 2.34, 2.51, 2.05, 2.39, 2.50, 2.31, 2.32, 2.12, 2.16, 2.35, 2.57, 2.42, 2.33, 2.62	2.0819 + 0.0103x	0.34	2.26	106

34	2.11, 2.13, 2.29, 2.25, 2.26, 2.38, 2.41, 2.30, 2.37, 2.35, 2.24, 2.42, 2.31, 2.31, 2.31, 2.41, 2.48, 2.31, 2.42, 2.29, 2.14, 2.13, 2.32, 2.47, 2.48, 2.45, 2.35, 2.26, 2.34, 2.12, 2.24, 2.23, 2.46, 2.81	2.2561 + 0.0041x	0.09	2.33	91
35	1.95, 2.15, 2.16, 2.28, 2.17, 2.48, 2.44, 2.40, 2.37, 2.33, 2.31, 2.28, 2.33, 2.32, 2.48, 2.32, 2.10, 2.43, 2.12, 2.17, 2.38, 2.15, 2.25, 2.25, 2.12, 2.07, 2.11, 2.16, 2.07, 2.41, 2.44, 2.10, 2.26, 2.58, 2.79	2.2339 + 0.0025x	0.02	2.28	81
36	1.99, 2.23, 2.26, 2.27, 2.21, 2.45, 2.47, 2.35, 2.13, 2.34, 2.30, 2.23, 1.99, 2.16, 2.15, 2.20, 2.37, 2.10, 2.41, 2.15, 2.26, 2.38, 2.12, 2.08, 2.12, 2.30, 2.38, 2.23, 2.37, 2.19, 2.29, 2.35, 2.17, 2.08, 2.29, 2.53	2.2246 + 0.0012x	0.01	2.25	86
37	1.92, 2.28, 2.10, 2.21, 2.12, 2.32, 2.06, 2.37, 2.22, 2.21, 2.32, 2.06, 2.13, 2.17, 2.19, 2.14, 2.51, 2.17, 2.29, 2.18, 2.29, 2.42, 2.15, 1.90, 2.37, 2.01, 2.35, 2.14, 2.03, 2.22, 2.31, 2.37, 2.37, 2.32, 2.21, 2.55, 2.77	2.1171 + 0.0063x	0.16	2.24	78
38	1.78, 2.35, 2.40, 2.10, 2.47, 2.19, 2.36, 2.29, 2.18, 2.47, 2.14, 2.25, 2.19, 2.13, 2.29, 2.29, 2.11, 2.01, 2.32, 2.38, 2.25, 1.92, 2.17, 2.14, 2.24, 2.04, 2.32, 2.29, 2.22, 2.00, 2.38, 2.58, 2.11, 2.28, 2.40, 2.24, 2.13, 2.63	2.1960 + 0.0021x	0.02	2.24	72
39	1.89, 2.07, 1.85, 1.93, 2.25, 2.02, 2.11, 2.42, 2.22, 1.96, 2.15, 2.33, 2.27, 2.07, 2.36, 2.36, 2.40, 2.38, 2.31, 2.24, 1.98, 2.02, 2.40, 1.98, 2.07, 2.29, 2.33, 2.04, 2.16, 2.00, 2.11, 2.09, 2.31, 2.05, 2.45, 2.24, 2.47, 2.29, 2.67	2.0677 + 0.0063x	0.15	2.19	55
40	1.72, 2.17, 2.13, 2.15, 2.06, 2.34, 1.96, 2.34, 2.15, 2.06, 2.38, 2.08, 2.19, 2.40, 2.15, 2.30, 2.42, 2.58, 2.38, 2.38, 2.23, 2.17, 2.38, 2.02, 2.15, 2.38, 2.36, 2.42, 2.21, 2.53, 2.62, 1.92, 2.21, 2.28, 2.19, 2.19, 1.91, 2.42, 2.47, 2.85	2.1148 + 0.0069x	0.14	2.26	53
41	2.06, 2.37, 2.00, 2.31, 2.10, 2.33, 1.96, 2.24, 1.96, 2.45, 2.02, 2.16, 2.41, 2.53, 2.18, 2.06, 2.47, 2.29,	2.1536 + 0.0038x	0.06	2.23	49

	2.04, 2.08, 2.37, 2.20, 2.04, 2.39, 2.45, 1.98, 2.31, 2.16, 2.08, 2.39, 2.33, 2.10, 2.24, 2.00, 2.20, 2.43, 2.39, 2.18, 2.31, 2.27, 2.76				
42	1.80, 2.30, 1.94, 2.18, 2.78, 2.40, 2.64, 2.38, 2.50, 2.06, 2.14, 2.44, 2.20, 2.36, 2.44, 1.88, 2.32, 2.26, 2.10, 2.08, 2.28, 2.52, 2.36, 2.34, 2.34, 2.34, 2.10, 2.12, 2.32, 2.38, 2.74, 2.36, 2.48, 2.60, 2.08, 2.48, 2.14, 2.18, 2.16, 2.64, 2.52, 2.96	2.2081 + 0.0054x	0.08	2.32	50
43	1.85, 2.46, 2.03, 2.49, 2.15, 1.95, 2.03, 2.28, 2.41, 2.38, 2.44, 2.05, 2.00, 2.10, 2.56, 2.05, 2.33, 2.03, 2.05, 2.28, 2.23, 2.36, 2.54, 2.36, 2.08, 1.95, 2.21, 2.10, 2.54, 2.33, 2.28, 1.79, 2.00, 1.92, 2.36, 2.18, 2.56, 2.23, 2.08, 2.31, 2.28, 2.64, 2.64	2.1566 + 0.0033x	0.04	2.23	39
44	2.10, 2.17, 2.34, 2.07, 2.49, 2.46, 2.00, 1.85, 2.68, 2.29, 2.68, 2.41, 2.78, 1.80, 2.22, 2.34, 2.34, 2.61, 2.37, 2.32, 2.49, 2.44, 1.98, 2.24, 2.20, 2.12, 2.63, 1.90, 2.05, 2.10, 2.41, 2.24, 2.27, 2.24, 2.22, 2.10, 2.17, 2.29, 2.24, 2.46, 2.34, 2.27, 2.27, 2.98	2.2710 + 0.0011x	0.00	2.29	41
45	2.07, 2.30, 2.13, 2.40, 2.00, 2.63, 2.10, 2.00, 2.87, 2.63, 2.37, 2.40, 2.27, 2.37, 2.17, 2.40, 2.33, 2.60, 2.33, 2.27, 2.10, 2.43, 2.40, 2.03, 2.30, 2.50, 2.30, 2.10, 2.33, 2.27, 2.23, 2.07, 2.33, 2.47, 2.17, 2.07, 2.67, 2.03, 2.17, 2.23, 2.57, 2.33, 2.20, 2.60, 3.03	2.2742 + 0.0022x	0.02	2.32	30
46	2.00, 2.18, 2.23, 1.92, 2.46, 2.05, 2.44, 2.21, 2.38, 2.23, 2.41, 2.21, 2.21, 2.28, 2.38, 2.46, 2.23, 2.26, 2.18, 2.46, 2.05, 2.13, 2.31, 2.13, 2.10, 2.56, 2.46, 2.13, 2.26, 2.21, 2.26, 2.10, 2.56, 2.18, 2.08, 2.18, 2.62, 2.44, 2.64, 2.21, 1.92, 2.44, 2.28, 2.36, 2.05, 2.77	2.2003 + 0.0031x	0.05	2.27	39
47	1.96, 2.11, 2.59, 1.85, 2.26, 2.30, 2.37, 2.11, 2.30, 2.00, 2.26, 2.48, 2.44, 2.26, 2.48, 1.93, 2.04, 2.04, 2.26, 1.96, 2.04, 2.07, 2.07, 2.22, 1.85, 2.44, 1.96, 2.15, 2.30, 2.19, 2.41, 1.96, 2.33, 1.78, 2.59, 2.11,	2.1734 + 0.0006x	0.00	2.19	27

	2.00, 2.22, 2.00, 2.00, 2.59, 2.33, 1.74, 2.26, 1.96, 2.48, 2.74				
48	1.69, 1.73, 2.42, 2.15, 2.19, 1.88, 2.19, 1.88, 2.31, 2.31, 2.12, 1.73, 2.04, 2.00, 2.58, 2.42, 1.96, 2.58, 2.23, 2.15, 2.50, 2.42, 2.23, 2.23, 2.08, 2.23, 1.92, 2.35, 2.31, 2.19, 2.27, 2.54, 2.50, 1.81, 2.12, 2.50, 2.19, 1.85, 2.35, 2.38, 1.96, 2.19, 2.19, 2.12, 2.35, 2.42, 2.73, 2.73	2.0496 + 0.0067x	0.13	2.21	26
49	1.88, 2.67, 2.13, 2.21, 1.92, 1.96, 1.92, 2.54, 2.04, 3.04, 2.17, 2.17, 2.71, 2.25, 2.13, 2.21, 2.25, 2.42, 2.42, 2.33, 2.25, 2.13, 2.17, 2.04, 2.21, 2.13, 2.25, 2.38, 2.33, 2.38, 2.13, 2.00, 2.04, 1.71, 2.50, 1.71, 2.54, 2.13, 2.50, 2.29, 2.08, 2.54, 2.13, 2.38, 2.33, 2.25, 2.79, 2.50, 2.83	2.1786 + 0.0035x	0.03	2.27	24
50	1.62, 2.50, 1.85, 1.88, 2.54, 2.08, 2.08, 1.73, 2.27, 2.27, 2.08, 2.46, 2.38, 2.31, 2.27, 1.85, 2.08, 1.96, 1.85, 2.12, 2.23, 2.27, 2.00, 2.23, 2.15, 2.62, 2.00, 1.88, 2.38, 2.19, 2.23, 2.27, 2.19, 2.35, 2.12, 2.19, 2.58, 2.31, 1.96, 2.58, 1.92, 1.73, 2.00, 2.00, 2.08, 2.27, 2.12, 2.12, 2.31, 2.96	2.0837 + 0.0033x	0.04	2.17	26
51	2.48, 2.05, 1.86, 2.14, 2.52, 2.43, 2.86, 2.29, 2.71, 2.33, 2.48, 2.19, 2.29, 2.29, 2.67, 2.33, 2.10, 2.48, 2.33, 2.62, 2.05, 2.43, 2.33, 2.05, 2.48, 2.57, 2.62, 2.38, 1.90, 1.95, 2.29, 1.81, 2.81, 1.81, 2.57, 2.48, 2.38, 2.62, 2.19, 1.57, 2.14, 1.90, 2.38, 1.95, 2.00, 2.52, 2.14, 2.43, 2.38, 2.52, 2.62	2.3663 - 0.0022x	0.01	2.31	21
52	1.53, 2.12, 2.18, 2.12, 2.29, 1.82, 2.35, 2.41, 2.53, 1.59, 2.59, 2.59, 3.00, 2.00, 3.00, 2.18, 2.12, 2.18, 2.59, 2.88, 2.29, 2.35, 2.12, 2.65, 1.88, 2.24, 2.76, 2.35, 2.06, 2.24, 2.12, 2.29, 2.06, 2.12, 2.82, 2.12, 1.94, 2.65, 1.88, 2.82, 2.29, 2.00, 2.00, 2.29, 2.18, 2.53, 2.29, 2.94, 2.47, 2.47, 1.88, 2.59	2.2290 + 0.0028x	0.02	2.30	17
53	1.76, 2.44, 2.52, 2.28, 2.36, 2.48, 1.60, 2.28, 1.92, 2.44, 2.08, 2.84,	2.2562 + 0.0014x	0.01	2.29	25

	2.48, 1.92, 2.20, 2.60, 2.68, 2.12, 2.56, 2.28, 2.16, 2.48, 2.24, 2.48, 2.00, 1.84, 2.36, 2.28, 2.40, 2.68, 2.20, 2.36, 2.20, 2.24, 2.32, 2.40, 2.16, 1.92, 2.40, 2.16, 2.68, 2.32, 2.48, 2.32, 2.16, 2.16, 2.48, 2.12, 1.96, 2.40, 2.28, 2.28, 2.80				
54	2.33, 1.87, 1.60, 2.27, 2.40, 2.27, 2.07, 2.27, 2.53, 1.80, 2.07, 1.93, 1.87, 2.00, 2.20, 1.93, 2.13, 2.07, 1.60, 2.00, 2.00, 2.00, 1.60, 2.27, 1.93, 2.13, 1.87, 2.73, 1.93, 2.67, 2.07, 2.33, 2.27, 1.87, 2.67, 1.87, 1.87, 2.20, 2.07, 2.27, 2.20, 1.93, 2.33, 2.47, 2.33, 2.07, 2.67, 2.00, 1.67, 2.07, 2.33, 2.33, 2.93, 2.60	1.9959 + 0.0054x	0.08	2.14	15
55	1.73, 2.00, 2.91, 3.09, 2.18, 1.82, 1.73, 1.82, 2.73, 2.27, 2.91, 2.73, 2.82, 3.18, 2.45, 2.18, 1.82, 2.55, 1.55, 2.09, 2.27, 2.45, 2.00, 1.55, 3.64, 1.91, 2.73, 1.91, 2.00, 2.45, 2.36, 2.27, 2.18, 2.00, 2.91, 2.45, 2.55, 2.09, 2.91, 2.73, 2.09, 2.45, 2.09, 2.00, 1.91, 2.00, 1.64, 3.00, 2.09, 2.27, 3.18, 2.45, 2.09, 2.00, 3.00	2.3097 + 0.0007x	0.00	2.33	11
56	1.58, 2.58, 2.25, 2.25, 2.50, 2.75, 1.92, 2.17, 2.58, 2.08, 1.67, 2.17, 2.08, 1.92, 1.83, 1.92, 1.75, 2.25, 2.58, 2.08, 1.92, 2.25, 2.50, 2.00, 2.17, 2.17, 2.17, 2.33, 2.08, 1.92, 2.42, 3.17, 2.17, 2.92, 1.75, 2.00, 2.25, 2.00, 2.42, 2.42, 1.58, 1.83, 2.75, 2.42, 2.17, 1.92, 2.42, 2.33, 2.67, 2.50, 2.25, 2.42, 1.83, 2.42, 2.33, 2.50	2.1207 + 0.0035x	0.03	2.22	12
57	1.64, 1.82, 2.09, 1.64, 2.55, 2.64, 2.64, 2.91, 1.27, 3.00, 1.73, 1.64, 3.09, 2.18, 2.27, 3.09, 1.64, 3.27, 1.91, 2.09, 1.64, 2.73, 1.91, 2.27, 2.18, 2.55, 1.64, 1.82, 1.82, 2.09, 2.18, 1.82, 2.36, 3.18, 1.91, 2.27, 1.55, 1.91, 2.27, 2.09, 2.09, 2.09, 2.18, 2.09, 2.09, 2.00, 1.91, 2.18, 2.27, 1.73, 1.91, 2.27, 2.27, 2.45, 2.18, 2.18, 2.64	2.2076 - 0.0012x	0.00	2.17	11
58	1.67, 2.08, 2.50, 2.08, 2.17, 1.92, 2.42, 2.25, 2.42, 2.42, 2.92, 2.42, 2.67, 2.00, 2.67, 2.67, 2.33, 1.83, 2.00, 2.50, 2.33, 2.33, 2.00, 2.42,	2.2970 - 0.0009x	0.00	2.27	12

	1.92, 1.92, 2.42, 2.17, 2.17, 2.75, 2.25, 2.08, 3.08, 2.33, 2.17, 2.58, 2.17, 2.92, 2.42, 2.33, 2.33, 1.67, 2.00, 2.50, 2.33, 1.92, 1.75, 2.17, 2.33, 1.83, 2.08, 2.33, 2.08, 2.33, 2.58, 1.92, 2.08, 2.83				

In Figure 6 one can see that the course of the parameter *b* is very smooth. This is most likely caused by the great number of sentences of each length. As can be seen from Figure 6, the decrease of parameter *b* is very regular. It can be expressed by $b = 0.8737 \exp(-0.2763n)$, with $R^2 = 0.97$. The results show that long texts display the expected tendency much more clearly. Interestingly enough, we again see that regression coefficient *b* may become negative for longer sentences, although in this case, for sentences with $n > 51$ only. The question remains open if this effect has to be considered as some usual dispersion, including the possibility that ultimately, there is no decrease of word length towards the end of a sentence, but rather some kind of "swinging in" around zero.

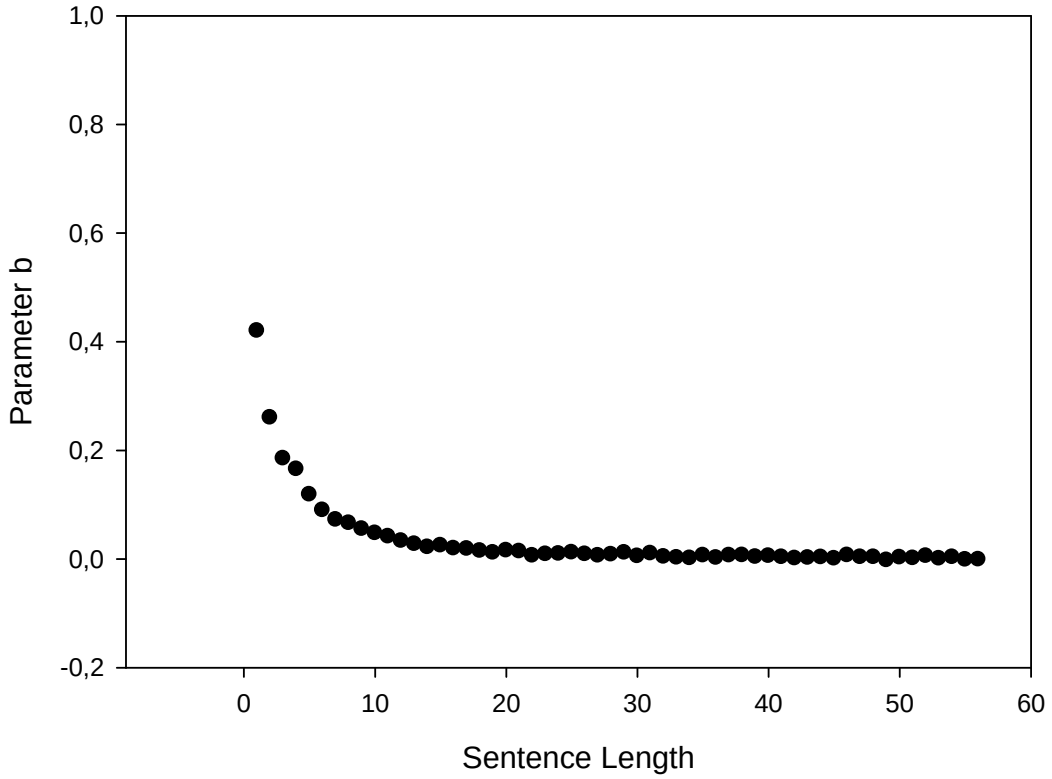


Figure 6. The decrease of parameter *b* in Tolstoj´s *Anna Karenina*

In any case, it seems obvious that the *regularity* of the decrease of parameter *b* with increasing sentence length very much depends on sample size.

Given these findings, the next step must include the analysis of data from further languages – in our case in Slovak, Hungarian, English, Latin and Indonesian texts – in order to arrive at more general conclusions. We therefore continue our analyses with Slovak texts

because this language, despite its overall differences, still displays great kinship and similarity to Russian.

2.2. Slovak

Like in case of our analyses of Russian texts, no reliable data were obtained for short individual texts; we therefore merged 30 Slovak journalistic texts to one overall corpus of 433 sentences (with 6072 words). The results for this corpus are presented in Table 3. As can be seen, reliable data can be obtained for $4 \leq n \leq 23$ (with $k \geq 10$).

Table 3
Mean lengths (L) of words in individual positions (x) in sentences of length n in Slovak press texts (M = mean of all positions, k = number of sentences)

n	**Mean syllabic lengths of words** ***L***	**Linear relation** ***L = a + bx***	R^2	***M***	***k***
3	1.67, 2.67, 3.33	-	-	-	3
4	2.18, 1.82, 2.82, 3.27	1.4550 + 0.4270x	0.72	2.52	11
5	2.47, 2.13, 2.53, 2.33, 2.73	2.2220 + 0.0720x	0.26	2.44	15
6	2.52, 2.10, 2.43, 2.33, 2.38, 3.05	2.1293 + 0.0969x	0.33	2.47	21
7	2.69, 2.15, 2.38, 1.96, 2.73, 2.77, 3.00	2.1657 + 0.0900x	0.27	2.53	26
8	2.48, 2.62, 1.90, 2.71, 2.38, 2.38, 2.71, 3.05	2.2314 + 0.0661x	0.23	2.53	21
9	2.80, 1.95, 2.50, 2.20, 2.50, 2.30, 1.60, 3.05, 3.10	2.2111 + 0.0467x	0.07	2.44	20
10	2.35, 2.29, 2.77, 2.84, 2.19, 2.77, 2.55, 2.61, 2.71, 2.68	2.4153 + 0.0292x	0.15	2.58	31
11	2.59, 2.32, 2.45, 2.45, 2.55, 2.05, 2.32, 2.23, 2.68, 2.18, 2.68	2.4140 – 0.0008x	0.00	2.41	22
12	2.54, 2.54, 2.21, 2.25, 2.88, 2.04, 2.17, 2.75, 2.13, 2.71, 2.54, 2.75	2.3467 + 0.0173x	0.05	2.46	24
13	2.81, 2.48, 2.38, 2.29, 2.52, 2.19, 2.43, 2.29, 2.19, 2.52, 2.76, 2.48, 2.62	2.4388 + 0.0028x	0.00	2.46	21
14	2.91, 1.91, 2.14, 2.82, 2.91, 2.95, 2.95, 2.68, 2.18, 2.41, 2.23, 2.64, 2.50, 2.91	2.5521 + 0.0039x	0.00	2.58	22
15	2.52, 2.81, 2.33, 2.81, 2.48, 2.71, 2.33, 2.86, 2.48, 2.43, 2.86, 2.62, 2.52, 2.43, 2.81	2.5809 + 0.0024x	0.00	2.60	21
16	2.76, 2.33, 1.94, 2.85, 2.55, 2.27, 1.91, 2.73, 2.39, 2.97, 2.58, 2.39, 2.33, 2.61, 2.94, 2.88	2.3308 + 0.0231x	0.11	2.53	33
17	2.57, 2.67, 2.27, 2.80, 2.40, 2.27, 2.60, 2.23, 2.30, 2.53, 2.20, 2.40, 2.13, 2.03, 2.30, 2.27, 2.53	2.5584 – 0.0196x	0.23	2.38	30
18	3.43, 1.86, 2.00, 2.57, 2.57, 2.43, 2.57, 2.57, 2.43, 2.71, 3.00, 2.57, 3.00, 1.71, 1.71, 2.14, 2.71, 2.43	2.6052 – 0.0145x	0.03	2.47	7
19	2.06, 2.28, 2.67, 2.28, 3.06, 2.50, 2.00,	2.3428 + 0.0197x	0.09	2.54	18

	2.56, 2.78, 2.56, 3.06, 2.56, 2.00, 2.44, 2.67, 2.22, 2.39, 3.22, 2.94				
20	2.72, 2.83, 2.33, 2.56, 3.00, 3.06, 2.39, 1.83, 2.17, 2.72, 2.78, 2.61, 2.67, 2.61, 2.44, 3.00, 2.50, 2.56, 2.72, 2.67	2.5975 + 0.0010x	0.00	2.61	18
21	2.67, 2.47, 2.40, 1.93, 3.73, 3.13, 3.07, 2.33, 2.60, 2.67, 2.47, 1.80, 3.40, 1.67, 2.20, 2.27, 2.13, 2.27, 2.27, 2.47, 2.67	2.7527 – 0.0225x	0.08	2.50	15
22	3.50, 2.10, 2.90, 2.60, 2.50, 2.20, 2.70, 3.40, 2.00, 3.50, 2.10, 2.50, 2.50, 3.00, 2.20, 2.80, 2.20, 1.90, 2.20, 2.80, 3.00, 3.00	2.7091 – 0.0079x	0.01	2.62	10
23	2.45, 3.09, 2.55, 2.00, 3.27, 2.18, 2.64, 2.18, 1.73, 3.18, 2.91, 2.27, 2.00, 2.64, 2.73, 2.09, 2.00, 2.73, 2.00, 2.00, 2.36, 1.91, 3.00	2.6158 – 0.0154x	0.05	2.43	11

Although Slovak and Russian are typologically quite similar to each other, the Slovak results strongly differ from those for our Russian data: In the Slovak texts, the regression coefficient quite soon takes values of $b < 0$, as can be seen in Figure 7.

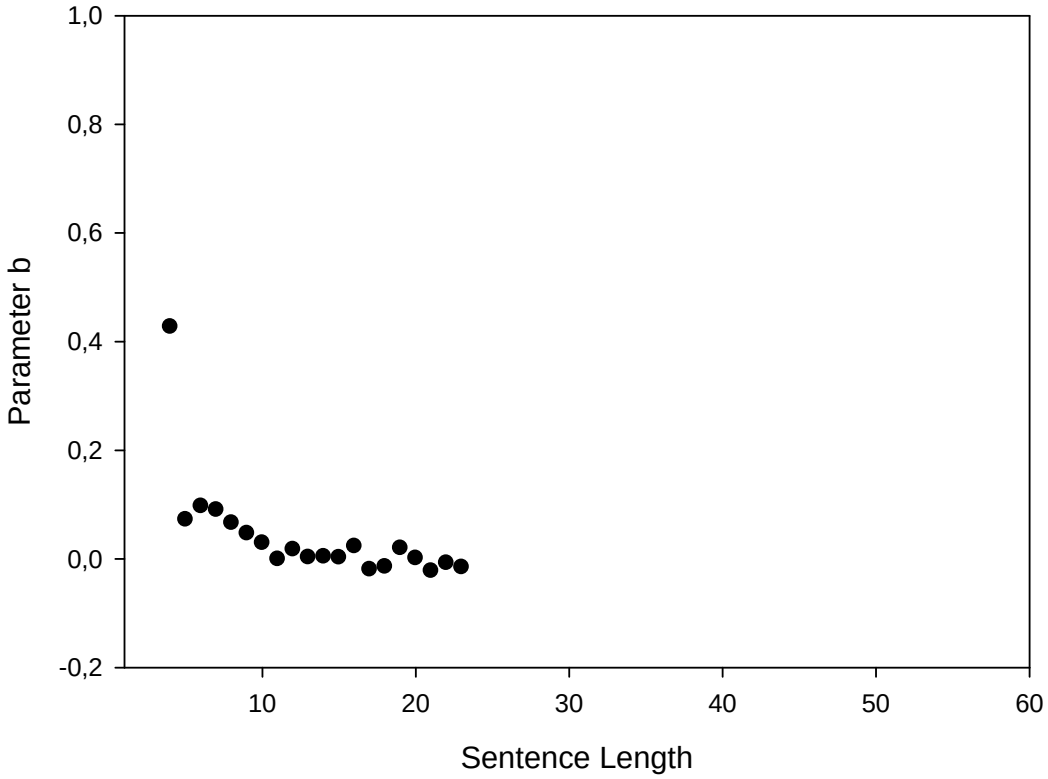


Figure 7. Dependence of *b* on sentence length in Slovak

Here we obtain $b = 25.2594 \exp(-1.0298n)$ with $R^2 = 0.88$. Although the tendency is, in principle, the same as observed before, the rapid decrease of parameter *b* is conspicuous. Since there is no easy explanation for this phenomenon, which may be caused by the text

material, a mixture of press texts, let us additionally analyze a longer Slovak text, which is more likely to be homogeneous, the novel *Veterná ružica* (1995), written by Vincent Šikula. For this text, which comprises 13961 words (in 881 sentences), reliable results are obtained for $3 \leq n \leq 25$, which are presented in Table 4.

Table 4
Mean lengths (L) of words in individual positions (x) in sentences of length n in Slovak text *Veterná ružica* by Šikula
(M = mean of all positions, k = number of sentences)

n	**Mean syllabic lengths of words *L***	**Linear relation *L = a + bx***	R^2	***M***	***k***
3	1.95, 1.59, 2.40	1.5300 + 0.2250x	0.31	1.98	58
4	1.91, 1.40, 1.74, 2.50	1.3600 + 0.2110x	0.35	1.89	78
5	1.83, 1.43, 1.59, 1.91, 2.62	1.2580 + 0.2060x	0.51	1.88	90
6	1.78, 1.51, 1.71, 1.75, 1.79, 2.63	1.3487 + 0.1466x	0.49	1.86	68
7	1.71, 1.40, 1.42, 1.54, 1.83, 1.77, 2.69	1.1814 + 0.1461x	0.51	1.77	48
8	1.71, 1.42, 1.97, 2.06, 1.81, 1.81, 1.90, 3.00	1.3868 + 0.1274x	0.46	1.96	31
9	1.87, 1.70, 1.78, 1.78, 1.83, 1.70, 1.26, 1.96, 2.70	1.5939 + 0.0497x	0.13	1.84	23
10	2.08, 1.33, 1.61, 1.67, 1.75, 2.00, 1.75, 1.83, 2.22, 2.44	1.4993 + 0.0670x	0.40	1.87	36
11	1.44, 1.56, 1.44, 2.00, 1.93, 1.85, 1.81, 1.67, 1.67, 1.59, 2.63	1.4547 + 0.0544x	0.29	1.78	27
12	2.07, 1.75, 1.82, 1.79, 1.89, 1.96, 1.79, 1.82, 2.00, 1.79, 2.04, 2.82	1.7044 + 0.0396x	0.24	1.96	28
13	2.03, 1.41, 1.93, 1.83, 2.00, 1.86, 2.03, 1.76, 1.90, 1.62, 1.97, 1.83, 2.59	1.7242 + 0.0258x	0.14	1.90	29
14	1.90, 1.70, 1.60, 1.75, 2.15, 2.00, 2.20, 1.90, 1.90, 2.40, 1.50, 1.75, 2.10, 2.75	1.7126 + 0.0345x	0.19	1.97	20
15	1.62, 1.71, 1.67, 2.05, 2.14, 1.76, 2.05, 2.05, 1.48, 1.76, 1.62, 1.86, 2.14, 1.52, 2.95	1.6740 + 0.0273x	0.11	1.89	21
16	2.00, 1.14, 1.90, 2.19, 2.19, 2.00, 2.14, 2.14, 2.19, 1.43, 1.90, 1.95, 1.71, 1.76, 1.86, 2.90	1.8033 + 0.0187x	0.05	1.96	21
17	2.06, 1.71, 1.35, 2.12, 1.76, 1.94, 1.76, 2.71, 1.53, 1.71, 1.88, 2.12, 1.82, 1.59, 1.94, 1.76, 2.88	1.7475 + 0.0192x	0.06	1.92	17
18	1.75, 1.83, 1.75, 2.42, 1.75, 1.83, 2.25, 1.83, 1.75, 2.33, 1.58, 2.25, 2.00, 1.42, 2.08, 2.00, 1.83, 2.75	1.8239 + 0.0150x	0.06	1.97	12
19	2.07, 1.93, 1.47, 2.47, 2.40, 1.73, 1.93, 1.93, 2.07, 1.93, 1.87, 1.87, 1.27, 2.20, 2.20, 2.00, 1.80, 1.73, 2.60	1.9504 + 0.0022x	0.00	1.97	15
20	2.00, 1.55, 1.95, 1.95, 2.20, 2.00, 1.85, 1.95, 1.75, 1.90, 1.75, 2.10, 1.65, 1.95, 1.70, 1.35, 1.65, 1.65, 2.20, 2.20	1.9084 - 0.0041x	0.01	1.87	20

21	1.86, 1.43, 2.14, 2.07, 1.86, 1.86, 2.43, 1.79, 2.50, 2.36, 2.57, 2.07, 1.86, 2.00, 1.64, 1.64, 1.79, 1.57, 1.36, 1.86, 3.07	1.9784 + 0.0008x	0.00	1.99	14
22	2.38, 1.75, 1.50, 1.88, 2.13, 2.38, 2.00, 1.88, 1.88, 1.38, 2.25, 2.75, 1.63, 2.00, 1.38, 1.75, 2.25, 1.50, 2.50, 2.25, 2.75, 2.88	1.8008 + 0.0215x	0.10	2.05	8
23	2.44, 1.33, 1.33, 2.11, 2.78, 2.11, 2.00, 1.22, 2.56, 2.11, 2.33, 1.56, 1.78, 2.00, 1.56, 1.44, 1.89, 2.11, 2.44, 2.22, 1.89, 2.00, 3.11	1.8459 + 0.0140x	0.04	2.01	9
24	1.83, 1.58, 1.67, 1.58, 2.00, 1.83, 1.67, 2.17, 2.33, 1.58, 2.42, 2.25, 1.67, 2.25, 2.00, 1.92, 1.75, 2.25, 2.33, 1.50, 1.75, 1.75, 2.50, 2.75	1.7295 + 0.0194x	0.16	1.97	12
25	2.00, 1.75, 1.67, 1.92, 1.92, 2.00, 2.25, 1.83, 1.75, 1.83, 2.00, 1.83, 1.83, 2.08, 1.67, 1.83, 2.42, 1.75, 1.58, 2.00, 1.92, 1.75, 2.00, 2.42, 2.50	1.8101 + 0.0100x	0.09	1.94	12

The course of parameter *b* can be modelled by the regression equation $b = 0.4325 \exp(-0.1812n)$ with $R^2 = 0.95$. Here the decrease is not as rapid as in the press texts, and in the analyzed domain ($3 \leq n \leq 25$), $b < 0$ in only one case (at sentence length 20). Figure 8 illustrates the results of the regression analyses.

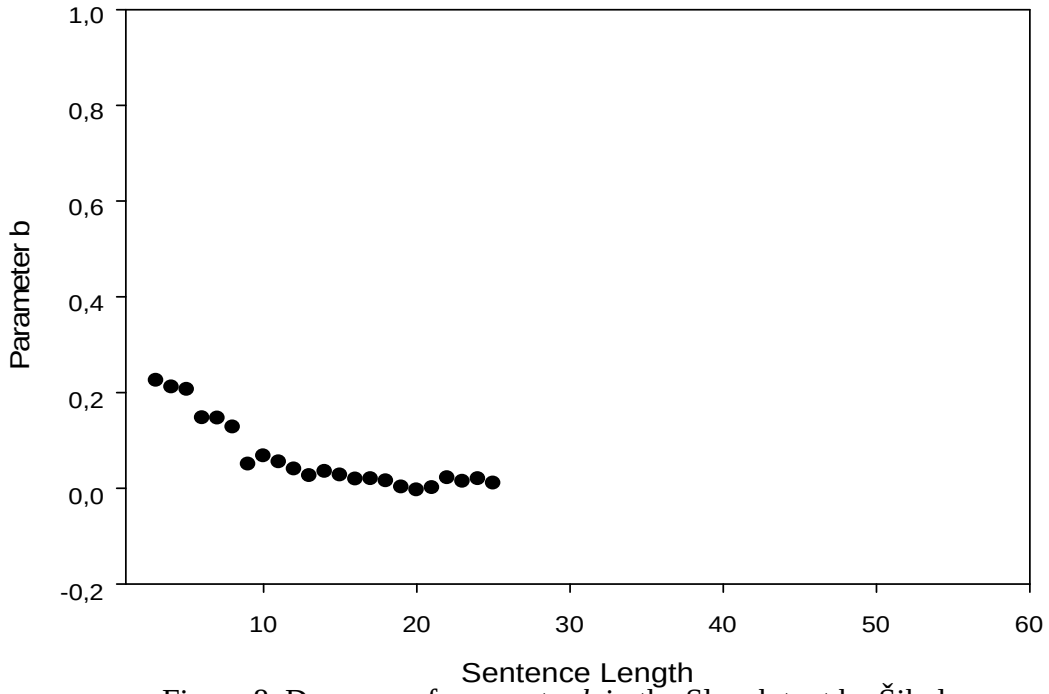


Figure 8. Decrease of parameter *b* in the Slovak text by Šikula

We may therefore conclude that the general tendency predicted by hypothesis (I) – i.e., decrease of *b* with increasing text length – holds also for Slovak.

2.3. Hungarian

For Hungarian, we analyzed two longer texts, both of approximately equal length, to compare the results with other languages, on the one hand, and to get additional information as to the degree of the variation within a given language.

The first data were taken from the novel *Az ezüst kecske* (1897) by Sándor Bródy (1863-1924), with 3356 sentences (and 46049 words), the second Hungarian text was a novel by Frigyes Karinthy (1887-1938), *Utazás a koponyám körül* (1937), with 44125 words (from 3440 sentences).

The results for the first text are presented in Table 5; all length classes up to the point when *k* was 10 for the last time are taken into consideration (in this case $n = 40$).

Table 5
Mean lengths (L) of words in individual positions (x) in sentences of length n in the Hungarian text *Az ezüst kecske* by Sándor Bródy
(M = mean of all positions, k = number of sentences)

n	**Mean syllabic lengths of words**	**Linear relation $L = a + bx$**	**R^2**	***M***	***k***
3	1.71, 1.87, 2.41	1.2967 + 0.3500x	0.91	2.2115	118
4	1.72, 1.97, 1.97, 2.50	1.4550 + 0.2340x	0.85	2.086	159
5	1.75, 2.04, 1.99, 1.93, 2.63	1.5730 + 0.1650x	0.61	1.9972	191
6	1.82, 1.86, 2.03, 2.00, 1.99, 2.66	1.6040 + 0.1303x	0.64	2.044	213
7	1.69, 2.03, 2.12, 1.97, 2.00, 2.04, 2.56	1.7000 + 0.0896x	0.56	2.0681	188
8	1.65, 2.00, 1.92, 2.15, 1.92, 1.94, 2.03, 2.81	1.6186 + 0.0964x	0.49	2.0595	198
9	1.80, 2.18, 1.93, 2.12, 2.06, 2.08, 2.09, 2.29, 2.61	1.8081 + 0.0642x	0.59	2.0593	179
10	1.76, 2.09, 2.17, 2.02, 2.04, 2.06, 1.96, 2.17, 2.04, 2.78	1.8200 + 0.0525x	0.37	2.0524	187
11	1.74, 2.11, 2.07, 2.23, 2.03, 2.20, 2.10, 2.12, 1.92, 2.34, 2.72	1.8589 + 0.0475x	0.40	2.1291	179
12	1.71, 2.19, 2.11, 2.13, 2.14, 2.07, 1.91, 2.04, 2.24, 2.26, 2.07, 2.64	1.8920 + 0.0360x	0.35	2.108	140
13	1.76, 2.29, 2.18, 2.21, 2.07, 1.97, 2.13, 2.16, 2.01, 2.08, 2.13, 2.18, 2.71	1.9665 + 0.0254x	0.21	2.1422	143
14	1.94, 2.39, 1.92, 2.30, 2.15, 2.07, 2.03, 2.14, 2.10, 2.31, 2.42, 2.04, 2.17, 2.72	2.0175 + 0.0234x	0.20	2.1262	105
15	1.75, 2.20, 2.22, 2.13, 2.31, 2.22, 2.10, 2.13, 1.87, 2.20, 2.13, 2.02, 2.04, 2.15, 2.65	2.0313 + 0.0138x	0.09	2.1452	134
16	1.71, 2.12, 2.17, 2.12, 2.09, 2.04, 2.09, 2.12, 2.04, 2.07, 2.03, 2.10, 2.17, 2.05, 2.25, 2.78	1.9125 + 0.0246x	0.31	2.1939	139
17	1.80, 2.09, 2.35, 2.00, 1.97, 2.10, 1.99, 2.23, 2.16, 1.95, 2.03, 1.97,	1.9236 + 0.0223x	0.20	2.1408	115

	2.07, 2.16, 2.04, 2.23, 2.97				
18	1.75, 2.10, 2.07, 2.23, 2.20, 2.25, 2.14, 2.25, 2.06, 2.04, 2.26, 2.24, 1.96, 2.41, 2.20, 1.98, 2.07, 2.66	2.0233 + 0.0143x	0.16	2.1223	102
19	1.84, 2.09, 2.21, 2.16, 2.06, 2.08, 2.37, 2.17, 2.08, 2.15, 2.26, 2.10, 2.07, 2.21, 2.13, 2.18, 2.11, 2.11, 2.79	2.0240 + 0.0143x	0.19	2.1228	89
20	1.80, 2.00, 2.28, 2.05, 2.14, 2.01, 2.17, 2.19, 2.08, 2.05, 1.93, 2.23, 2.00, 2.06, 2.29, 1.89, 2.08, 2.13, 2.12, 2.71	1.9854 + 0.0119x	0.14	2.158	83
21	1.61, 2.18, 2.11, 2.05, 2.09, 2.02, 1.95, 2.26, 2.36, 2.32, 1.94, 2.21, 2.03, 2.02, 2.36, 2.14, 2.03, 2.15, 2.42, 2.26, 3.08	1.8997 + 0.0247x	0.31	2.1668	66
22	1.77, 1.97, 2.23, 2.12, 1.94, 2.17, 2.30, 2.52, 2.22, 1.93, 2.30, 2.25, 2.03, 2.06, 2.16, 2.09, 2.22, 2.33, 2.39, 2.12, 2.32, 2.90	1.9816 + 0.0188x	0.27	2.1108	69
23	1.91, 2.17, 1.92, 2.17, 2.28, 2.09, 2.09, 2.03, 2.16, 2.05, 2.41, 2.22, 2.13, 2.11, 2.03, 2.31, 1.89, 2.02, 2.19, 2.45, 2.25, 2.31, 2.97	1.9818 + 0.0166x	0.25	2.1703	64
24	1.54, 2.10, 2.08, 2.17, 1.81, 2.20, 1.93, 2.00, 2.07, 2.41, 2.12, 1.95, 1.95, 2.19, 2.07, 2.29, 2.02, 2.08, 2.31, 2.44, 2.12, 2.44, 2.29, 2.83	1.8572 + 0.0228x	0.41	2.197	59
25	1.83, 2.29, 2.02, 2.14, 1.74, 2.43, 2.17, 1.93, 2.17, 1.95, 2.17, 2.19, 1.76, 2.50, 2.14, 2.26, 2.05, 2.00, 2.10, 2.52, 2.21, 2.38, 2.19, 2.21, 3.14	1.9424 + 0.0182x	0.22	2.1807	42
26	1.70, 2.16, 2.14, 2.05, 2.14, 1.88, 2.44, 2.23, 1.93, 2.09, 2.09, 2.44, 2.26, 2.16, 2.14, 1.93, 2.21, 2.02, 2.12, 2.16, 2.02, 2.26, 2.05, 2.19, 2.07, 2.56	2.0418 + 0.0067x	0.08	2.1419	43
27	1.96, 2.33, 2.38, 2.09, 1.93, 1.78, 2.40, 2.20, 2.07, 2.11, 2.00, 2.09, 2.07, 2.31, 2.22, 2.00, 2.00, 2.18, 2.27, 1.89, 1.98, 2.49, 2.24, 2.09, 2.24, 1.93, 2.84	2.0658 + 0.0061x	0.05	2.18	45
28	1.74, 2.20, 2.26, 2.09, 2.40, 2.09, 2.29, 2.09, 1.97, 2.03, 2.34, 1.94, 2.20, 1.89, 2.31, 2.06, 2.31, 2.14, 2.11, 2.17, 2.37, 2.31, 2.00, 1.94, 1.71, 2.63, 1.77, 3.23	2.0566 + 0.0074x	0.04	2.1324	35
29	1.71, 1.92, 2.21, 2.17, 2.42, 2.08,	2.0890 + 0.0080x	0.06	2.1514	24

	2.42, 2.67, 2.08, 2.13, 2.17, 2.13, 2.38, 2.13, 1.83, 2.08, 2.25, 2.42, 2.25, 1.88, 2.75, 2.00, 1.96, 2.50, 1.92, 2.00, 2.17, 2.50, 2.92				
30	1.88, 2.19, 1.96, 2.19, 2.27, 2.23, 2.23, 2.00, 2.23, 2.42, 1.96, 2.00, 2.19, 1.81, 2.00, 2.15, 2.62, 2.12, 2.00, 1.92, 2.27, 2.15, 2.73, 2.23, 2.00, 2.46, 2.23, 1.92, 2.50, 2.38	2.0625 + 0.0072x	0.08	2.1643	26
31	2.05, 2.11, 2.16, 2.37, 2.16, 1.84, 1.89, 2.47, 1.89, 2.37, 2.32, 2.53, 2.05, 1.63, 2.47, 2.37, 1.63, 2.84, 2.26, 1.79, 2.42, 1.68, 2.26, 2.47, 2.32, 2.11, 1.74, 1.79, 1.95, 2.47, 2.89	2.1187 + 0.0033x	0.01	2.2069	19
32	1.77, 2.77, 2.00, 1.69, 2.08, 2.23, 2.38, 2.38, 2.15, 2.54, 2.46, 2.00, 2.23, 2.15, 2.54, 2.38, 2.00, 2.62, 2.38, 2.00, 2.31, 1.77, 2.08, 1.77, 2.62, 2.15, 2.46, 2.62, 2.08, 2.54, 2.00, 3.00	2.1337 + 0.0073x	0.05	2.1756	13
33	1.64, 2.09, 2.27, 2.00, 1.82, 2.18, 2.27, 2.55, 2.18, 2.36, 2.00, 1.82, 2.18, 2.09, 1.91, 2.09, 2.27, 2.00, 2.00, 1.73, 2.09, 2.45, 1.82, 2.45, 2.36, 2.09, 2.45, 1.73, 2.18, 2.09, 2.00, 2.00, 2.73	2.0423 + 0.0044x	0.03	2.1715	11
34	1.69, 1.85, 2.23, 1.85, 2.69, 2.23, 2.23, 2.23, 1.69, 2.08, 2.38, 2.00, 2.23, 1.85, 2.15, 2.00, 2.54, 1.77, 2.62, 2.69, 2.38, 3.08, 2.46, 2.31, 1.92, 1.85, 1.92, 1.85, 2.54, 2.23, 2.46, 1.92, 2.23, 3.15	2.0307 + 0.0105x	0.08	2.2548	13
35	1.73, 2.53, 1.93, 2.20, 1.93, 2.67, 2.33, 2.47, 2.13, 2.07, 2.47, 1.93, 2.47, 2.40, 2.20, 1.73, 2.47, 2.47, 1.53, 1.73, 2.33, 2.60, 1.87, 2.00, 1.93, 1.80, 2.47, 1.73, 2.47, 2.27, 2.20, 3.07, 1.87, 2.47, 3.1	2.1109 + 0.0058x	0.03	2.1185	15
36	1.78, 2.33, 1.67, 1.67, 2.44, 1.78, 2.00, 2.22, 2.33, 1.78, 2.00, 2.44, 2.22, 1.78, 2.78, 2.22, 1.67, 2.44, 1.78, 2.33, 2.67, 2.00, 2.56, 1.89, 2.00, 1.56, 2.56, 1.89, 1.67, 2.44, 2.56, 1.67, 1.44, 2.00, 2.10, 3.10	2.0238 + 0.0044x	0.01	2.2149	9
37	1.71, 1.71, 1.71, 1.71, 2.14, 2.00, 1.29, 2.14, 2.14, 2.29, 2.43, 1.86, 1.86, 2.29, 2.86, 2.57, 2.14, 2.29, 2.43, 2.71, 1.71, 2.00, 1.71, 2.86,	1.9008 + 0.0129x	0.12	2.2152	7

	1.86, 2.43, 1.86, 2.57, 2.29, 2.29, 2.14, 1.86, 2.14, 2.00, 1.90, 2.10, 3.40				
38	1.60, 1.40, 2.20, 2.20, 3.10, 1.90, 1.80, 3.00, 2.20, 2.60, 1.80, 1.80, 2.20, 2.00, 2.10, 1.60, 2.60, 2.40, 2.50, 2.00, 2.30, 2.20, 1.60, 2.40, 2.20, 1.10, 1.60, 2.50, 2.00, 1.80, 2.20, 2.50, 3.30, 2.20, 2.10, 1.70, 2.40, 2.80	2.0539 + 0.0052x	0.02	2.1049	10
39	2.67, 3.00, 3.00, 2.33, 2.33, 3.00, 2.67, 2.67, 2.67, 1.00, 2.33, 1.33, 2.33, 3.67, 1.67, 1.33, 2.33, 3.67, 1.33, 2.00, 2.00, 2.33, 3.00, 1.67, 2.67, 1.33, 1.33, 2.00, 1.00, 1.33, 1.67, 3.33, 2.00, 2.33, 2.00, 2.30, 2.70, 2.00, 3.00	2.4878 - 0.0124x	0.04	2.1467	3
40	1.50, 2.00, 2.10, 2.00, 2.20, 2.00, 3.00, 2.10, 2.50, 1.50, 2.70, 2.00, 1.80, 1.80, 2.30, 2.40, 2.10, 1.80, 1.70, 2.10, 1.50, 1.50, 1.90, 2.20, 2.10, 1.80, 1.60, 1.80, 1.50, 2.40, 1.70, 1.70, 2.00, 1.90, 2.20, 2.00, 2.50, 2.30, 2.20, 2.50	2.0262 - 0.0002x	0.00	2.1553	10

Again one sees that in the linear regression the slope (*b*) decreases from highly significant positive values to small negative ones as shown in Figure 9. The *b*-sequence is very regular; a straight line seems to be no good model, but again, with $b = 0.7432 \exp(-0.2756n)$, the determination coefficient is very satisfying ($R^2 = 0.97$).

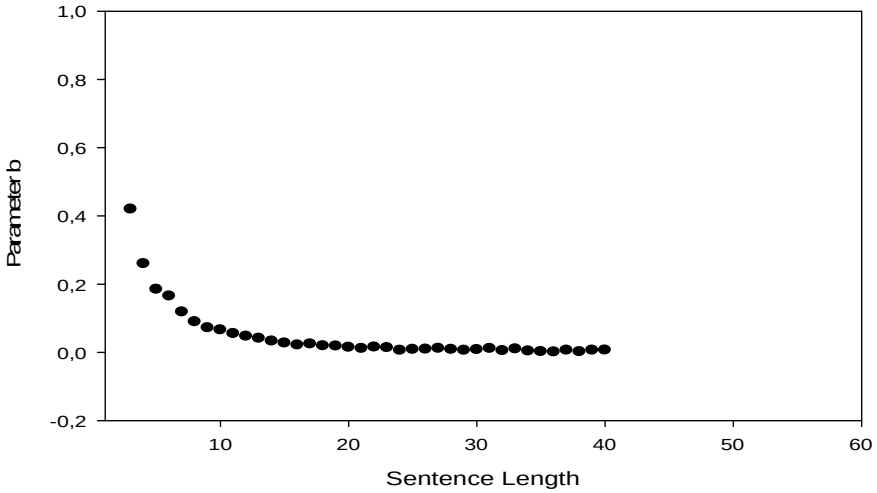


Figure 9. Decrease of *b* in Bródy's Hungarian text

The results for second Hungarian text are represented in Table 6. The function expressing the decrease of *b* in Figure 10 is $b = 0.3061\exp(-0.1778n)$, with $R^2 = 0.92$.

Table 6
Mean lengths (*L*) of words in individual positions (*x*) in sentences of length *n* in the Hungarian text *Utazás a koponyám körül* by Frigyes Karinthy
(*M* = mean of all positions, *k* = number of sentences)

n	**Mean syllabic lengths of words**	**Linear relation** $L = a + bx$	R^2	***M***	***k***
3	2.11, 2.08, 2.50	1.8400 + 0.1950x	0.69	2.23	210
4	2.04, 1.89, 2.02, 2.58	1.6950 + 0.1750x	0.55	2.13	223
5	1.95, 2.05, 2.16, 1.94, 2.45	1.8430 + 0.0890x	0.45	2.11	195
6	1.98, 2.13, 2.09, 2.09, 2.10, 2.61	1.8607 + 0.0874x	0.54	2.17	215
7	2.13, 2.06, 2.10, 1.95, 2.20, 2.16, 2.67	1.9071 + 0.0686x	0.42	2.18	188
8	1.90, 2.03, 2.19, 2.01, 2.16, 2.09, 2.24, 2.80	1.7918 + 0.0857x	0.59	2.18	182
9	1.87, 2.25, 2.07, 2.17, 2.14, 2.24, 2.25, 2.22, 3.01	1.8383 + 0.0817x	0.52	2.25	153
10	2.18, 2.43, 2.05, 2.18, 2.08, 2.19, 2.07, 2.27, 2.13, 2.90	2.0727 + 0.0319x	0.14	2.25	153
11	2.09, 2.02, 2.18, 2.01, 2.26, 2.30, 2.20, 2.32, 1.98, 2.13, 2.84	1.9855 + 0.0377x	0.27	2.21	128
12	1.88, 2.06, 2.27, 2.07, 2.32, 2.13, 2.24, 2.44, 2.13, 2.19, 2.15, 2.72	1.9835 + 0.0359x	0.37	2.22	125
13	1.96, 2.33, 2.19, 1.98, 2.39, 2.12, 2.24, 2.28, 2.27, 2.34, 2.18, 2.13, 2.96	2.0300 + 0.0327x	0.27	2.26	113
14	1.91, 2.10, 2.25, 2.20, 1.98, 2.27, 2.20, 2.19, 2.11, 2.13, 2.10, 2.18, 2.17, 2.92	1.9821 + 0.0282x	0.26	2.19	119
15	1.76, 2.08, 2.16, 2.15, 2.27, 2.35, 2.17, 2.26, 2.17, 2.02, 2.28, 2.23, 2.03, 2.17, 2.78	2.0000 + 0.0240x	0.25	2.19	92
16	2.00, 2.31, 2.21, 2.23, 2.07, 2.19, 2.48, 2.22, 2.26, 2.35, 2.16, 2.30, 2.23, 2.19, 2.46, 2.91	2.0795 + 0.0243x	0.31	2.29	96
17	2.03, 2.22, 2.32, 2.03, 2.36, 2.06, 2.31, 2.41, 2.03, 2.24, 2.20, 2.57, 2.22, 2.24, 2.16, 2.28, 2.87	2.0954 + 0.0191x	0.21	2.27	87
18	1.89, 2.16, 2.04, 2.40, 2.41, 2.15, 2.15, 2.36, 2.19, 2.24, 2.35, 2.44, 2.03, 2.32, 2.23, 2.00, 2.31, 2.88	2.0909 + 0.0170x	0.17	2.25	75
19	2.09, 2.11, 2.32, 2.33, 1.97, 2.08, 2.05, 2.15, 2.24, 2.30, 2.12, 2.05, 2.03,2.29,2.27,2.12,2.18,2.36,3.02	2.0377 + 0.0177x	0.19	2.21	66
20	2.04, 1.96, 2.21, 2.02, 2.16, 2.00, 2.13, 2.32, 2.39, 2.52, 2.45, 2.27, 2.57, 2.13, 2.43, 2.30, 1.95, 2.34, 2.45, 3.18	1.9968 + 0.0280x	0.34	2.29	56
21	2.10, 1.84, 2.20, 2.31, 2.17, 2.17, 2.29, 2.14, 2.17, 2.01, 2.29, 2.31, 2.00, 2.23, 2.11, 2.34, 2.04, 2.04, 2.36, 1.97, 2.91	2.0728 + 0.0107x	0.10	2.19	70
22	1.95, 2.49, 2.11, 2.49, 2.10, 2.13, 2.48, 2.30, 2.15, 2.43, 2.34, 2.15, 2.26, 2.18, 2.25, 2.15, 2.15, 2.07, 2.30, 2.36, 2.84, 2.85	2.1481 + 0.0129x	0.13	2.30	61
23	2.12, 2.21, 2.40, 2.27, 2.50, 2.31, 2.40, 2.02, 2.23, 2.21, 2.06, 2.02, 2.04, 2.15, 1.92, 2.31,	2.1686 + 0.0070x	0.05	2.25	52

	2.27, 2.13, 2.33, 2.23, 2.46, 2.23, 2.98				
24	1.89, 2.27, 1.95, 2.14, 1.93, 2.36, 2.09, 2.23, 1.98, 2.09, 2.52, 2.07, 2.07, 2.48, 2.57, 2.11, 2.07, 2.55, 2.48, 2.23, 2.39, 2.32, 2.18, 3.00	1.9890 + 0.0208x	0.32	2.25	44
25	1.66, 2.28, 2.14, 1.79, 2.10, 2.24, 2.14, 2.24, 2.55, 2.17, 2.38, 2.21, 2.21, 2.24, 2.07, 2.31, 2.21, 2.38, 2.14,2.03,1.90,2.24,2.14,2.10,2.90	2.0603 + 0.0100x	0.10	2.19	29
26	1.84, 2.30, 1.95, 2.16, 2.27, 1.97, 2.35, 2.38, 2.51, 2.05, 1.97, 2.59, 2.19, 2.22, 2.14, 2.30, 2.22, 2.08, 2.08, 2.08, 2.03, 2.22, 2.19, 2.14, 2.11, 3.11	2.1042 + 0.0078x	0.06	2.21	37
27	2.05, 2.11, 2.22, 2.00, 2.24, 2.43, 1.95, 2.38, 2.35, 2.51, 2.27, 2.00, 2.00, 2.22, 2.22, 2.62, 2.24, 2.46, 2.38, 2.68, 2.03, 2.41, 2.22, 1.97, 2.16, 2.49, 2.86	2.1274 + 0.0107x	0.13	2.28	37
28	2.03, 2.13, 2.17, 2.17, 2.27, 2.40, 2.13, 2.10, 2.33, 2.20, 2.10, 2.13, 2.10, 2.20, 1.93, 2.20, 2.50, 2.00, 2.00, 2.03, 2.33, 1.87, 2.27, 2.30, 2.13, 2.37, 2.13, 3.10	2.0968 + 0.0072x	0.07	2.20	30
29	2.24, 2.08, 2.00, 1.80, 2.28, 2.32, 1.96, 2.48, 1.92, 2.12, 2.20, 1.96, 1.76, 2.52, 2.16, 2.04, 2.12, 1.88, 2.40, 2.44, 2.32, 2.32, 2.28, 2.60, 2.40, 2.20, 2.36, 2.92, 2.92	1.9523 + 0.0193x	0.33	2.24	25
30	1.85, 2.47, 2.09, 2.06, 2.09, 2.47, 2.00, 2.18, 2.06, 2.21, 2.18, 2.12, 2.24, 2.24, 2.29, 2.47, 2.00, 1.94, 2.15, 2.24, 2.44, 2.12, 2.21, 2.03, 2.38, 2.15, 2.44, 2.56, 2.29, 3.00	2.0547 + 0.0115x	0.20	2.23	34
31	2.16, 2.24, 1.92, 2.12, 2.20, 2.84, 2.16, 2.36, 2.72, 2.20, 2.08, 2.24, 2.56, 2.08, 1.92, 2.48, 2.00, 2.28, 2.56, 2.16, 2.04, 2.00, 2.24, 2.08, 2.76, 2.24, 2.36, 2.20, 1.88,1.92,2.92	2.2446 + 0.0007x	0.00	2.26	25
32	1.78, 2.33, 2.00, 2.22, 2.15, 2.37, 2.04, 2.04, 2.63, 2.22, 2.26, 2.48, 2.19, 1.89, 1.96, 2.52, 2.19, 2.07, 2.07, 2.22, 2.70, 2.63, 2.22, 1.59, 2.11, 2.33, 2.59, 2.00, 2.37, 2.11, 2.56, 2.48	2.1244 + 0.0063x	0.05	2.23	27
33	2.22, 2.00, 2.39, 1.94, 2.06, 1.89, 2.06, 2.50, 2.28, 2.28, 2.67, 2.00, 2.22, 2.28, 2.33, 2.22, 1.67, 3.17, 2.39, 2.50, 2.33, 2.06, 2.56, 2.56, 2.00, 1.94, 2.28, 2.22, 2.33, 2.11, 2.17, 2.28, 2.83	2.1589 + 0.0062x	0.04	2.26	18
34	2.18, 2.23, 2.45, 2.05, 1.82, 2.45, 2.41, 1.91, 1.86, 2.18, 1.68, 2.23, 2.68, 2.14, 2.32, 2.23, 2.23, 1.91, 2.05, 2.68, 2.27, 2.23, 2.36, 2.09, 1.95, 2.50, 2.32, 2.50, 2.05, 2.27, 2.14, 2.05, 2.36, 2.82	2.1174 + 0.0061x	0.06	2.22	22
35	2.27, 1.91, 2.64, 1.55, 2.09, 1.91, 1.73, 1.64, 2.73, 1.73, 3.18, 1.82, 1.82, 1.91, 2.18, 2.18, 2.09, 1.45, 2.64, 2.55, 2.55, 2.18, 2.00, 2.18,	2.0832 + 0.0012x	0.00	2.11	11

	1.64, 2.55, 1.73, 2.27, 2.00, 2.55, 1.91, 1.73, 1.91, 1.82, 2.64				
36	1.64, 2.14, 2.50, 2.50, 1.93, 2.79, 2.64, 2.43, 1.79, 2.50, 2.07, 2.50, 2.36, 2.29, 2.14, 2.00, 2.00, 2.07, 2.36, 2.43, 1.93, 1.79, 1.86, 2.36, 2.50, 2.43, 2.07, 2.21, 1.93, 2.86, 2.14, 1.86, 2.36, 1.79, 2.50, 2.64	2.2351 - 0.0002x	0.00	2.23	14
37	3.00, 2.50, 2.50, 2.75, 1.63, 2.50, 2.50, 1.63, 2.88, 2.13, 2.38, 2.75, 3.00, 2.63, 2.88, 2.75, 2.63, 2.00, 3.13, 2.38, 1.75, 2.50, 2.88, 2.13, 2.50, 2.25, 1.75, 1.63, 1.25, 2.38, 2.13, 2.25, 1.75, 2.25, 1.75, 2.13, 2.63	2.6394 - 0.0159x	0.14	2.34	8
38	1.80, 2.20, 2.20, 1.80, 2.70, 2.20, 2.00, 1.70, 1.90, 2.00, 2.00, 2.30, 2.20, 2.00, 2.10, 1.70, 3.10, 2.70, 2.60, 2.30, 1.60, 3.00, 2.10, 2.80, 2.10, 2.20, 2.70, 2.40, 2.00, 2.40, 2.30, 2.20, 2.50, 2.40, 3.20, 3.10, 2.00, 3.00	1.9653 + 0.0173x	0.21	2.30	10
39	2.14, 2.86, 2.00, 2.29, 2.00, 2.29, 1.57, 1.86, 1.86, 2.71, 1.71, 1.71, 3.14, 2.14, 2.71, 2.29, 1.86, 1.71, 2.29, 2.00, 2.71, 2.00, 1.86, 2.43, 1.29, 1.71, 2.14, 2.00, 2.57, 1.43, 2.57, 2.00, 1.57, 2.43, 1.57, 1.14, 2.43, 1.86, 3.00	2.2024 - 0.0052x	0.02	2.10	7
40	1.50, 2.67, 2.83, 2.00, 1.83, 1.33, 2.50, 2.17, 2.83, 2.00, 2.00, 2.67, 3.33, 2.50, 3.00, 2.00, 2.83, 1.17, 2.00, 2.00, 2.50, 1.67, 2.50, 2.00, 2.00, 2.50, 1.50, 2.50, 3.17, 1.83, 3.50, 2.67, 2.83, 2.33, 2.33, 3.00, 2.67, 2.17, 1.67, 3.00	2.1448 + 0.0094x	0.04	2.34	6
41	2.38, 2.38, 2.25, 2.00, 1.50, 2.13, 2.25, 2.13, 2.63, 2.25, 2.88, 2.50, 1.75, 1.50, 2.13, 2.63, 2.38, 2.00, 2.00, 2.75, 2.50, 2.13, 2.25, 1.63, 2.25, 2.13, 2.25, 2.00, 2.63, 2.75, 1.88, 2.50, 2.50, 2.00, 2.38, 2.13, 2.13, 1.88, 2.13, 2.63, 2.25	2.1965 + 0.0015x	0.00	2.23	8
42	1.50, 1.80, 2.10, 2.00, 2.90, 2.50, 2.10, 1.80, 2.10, 2.90, 2.50, 1.90, 1.60, 2.30, 3.30, 2.20, 2.70, 2.30, 3.00, 2.00, 2.40, 2.00, 2.40, 2.00, 2.70, 1.90, 2.10, 2.30, 2.40, 3.00, 2.20, 2.20, 1.80, 2.50, 1.90, 2.40, 2.00, 2.20, 2.60, 1.60, 2.00, 3.10	2.2050 + 0.0029x	0.01	2.27	10
43	2.67, 1.67, 1.83, 2.67, 2.33, 1.75, 1.83, 3.00, 1.83, 2.50, 2.08, 2.67, 2.25, 3.08, 1.83, 2.42, 1.92, 2.08, 1.67, 2.08, 1.83, 2.00, 2.17, 2.25, 2.92, 2.58, 2.08, 1.83, 2.42, 2.42, 2.17, 2.17, 1.67, 2.67, 2.42, 1.92, 2.75, 2.25, 1.92, 2.42, 1.92, 2.17, 2.75	2.1940 + 0.0016x	0.00	2.23	12

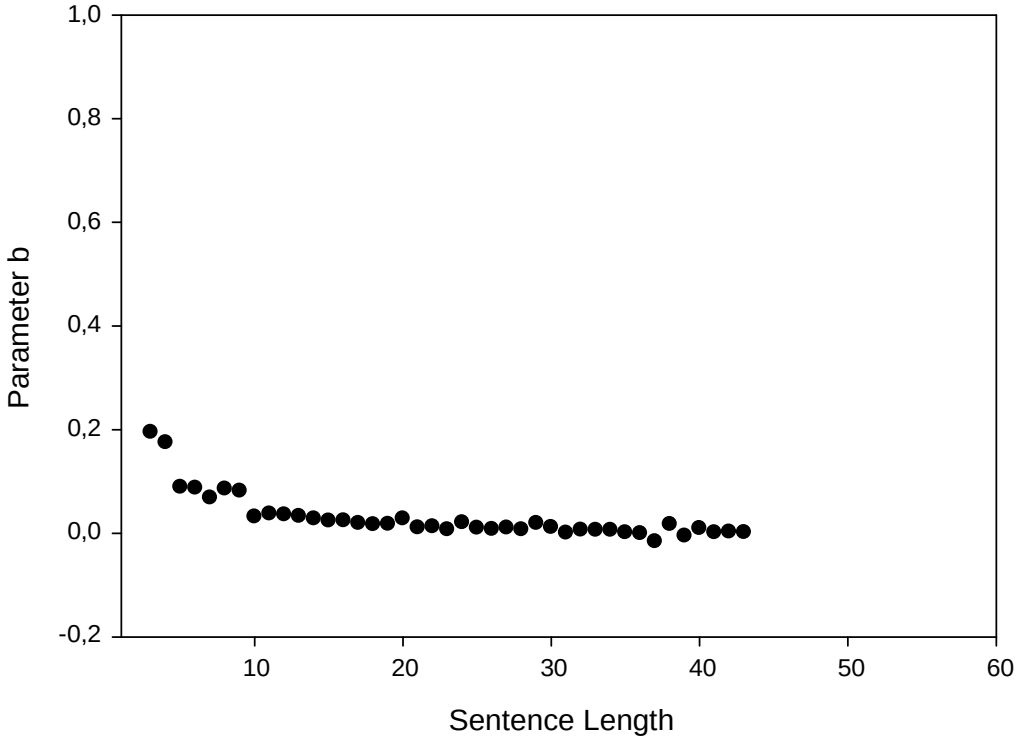


Figure 10. The decrease of *b* in Karinthy's Hungarian text

2.4. Latin

In order to check the hypothesis in a more extreme linguistic situation, we next analyze a Latin text by Seneca (ca. 1-65), *Epistularum moralium ad Lucilium liber primus*, summing up to 6226 words (in 365 sentences). Latin is a strongly inflecting language; therefore one might expect that the results strongly differ from the ones observed above for languages which are synthetic to a lesser degree. The results are presented in Table 7.

Table 7
Mean lengths (*L*) of words in individual positions (*x*) in sentences of length *n* in the Latin text by Seneca

n	**Mean syllabic lengths of words**	**Linear relation** $L = a + bx$	R^2	***M***	***k***
3	2.5, 2.5, 2.0	-	-	-	-
4	1.4, 2.3, 2.6, 2.9	1.1600 + 0.4580x	0.91	2.305	9
5	1.9, 1.9, 1.7, 2.3, 2.6	1.5340 + 0.1780x	0.60	2.068	9
6	2.2, 2.2, 2.5, 2.5, 2.9, 2.9	1.9693 + 0.1554x	0.96	2.513	13
7	2.2, 1.9, 2.5, 2.4, 2.5, 2.7, 2.8	1.9557 + 0.1157x	0.75	2.419	12
8	2.2, 2.1, 2.3, 2.7, 2.1, 1.9, 2.7, 2.5	2.1232 + 0.0418x	0.12	2.311	14
9	2.2, 2.3, 2.0, 2.1, 2.0, 2.5, 3.0, 2.6, 2.3	2.0036 + 0.0622x	0.27	2.314	12
10	1.8, 1.9, 1.9, 2.1, 2.1, 2.3, 2.1, 2.7, 2.6, 2.8	1.6340 + 0.1078x	0.89	2.227	16

11	2.3, 2.2, 2.1, 1.9, 2.3, 1.9, 2.5, 1.8, 2.3, 2.4, 2.8	2.0373 + 0.0300x	0.13	2.217	17
12	2.0, 1.9, 1.7, 2.1, 2.7, 2.0, 3.1, 2.9, 2.5, 2.3, 2.6, 2.7	1.8786 + 0.0752x	0.37	2.368	15
13	2.2, 2.3, 2.8, 2.4, 2.2, 2.1, 2.4, 2.3, 2.5, 2.1, 2.2, 3.1, 2.7	2.2100 + 0.0259x	0.12	2.392	12
14	2.3, 2.4, 2.6, 2.3, 2.1, 2.2, 2.5, 2.5, 2.3, 2.1, 2.6, 2.3, 2.4, 2.6	2.3186 + 0.0077x	0.04	2.376	18
15	2.2, 2.4, 2.2, 2.0, 2.2, 2.3, 2.8, 2.4, 2.5, 2.4, 2.2, 2.2, 2.6, 2.4, 2.85	2.1176 + 0.0282x	0.29	2.343	20
16	1.6, 2.0, 2.0, 2.2, 2.8, 2.4, 2.2, 1.9, 2.0, 2.4, 1.8, 2.3, 2.7, 1.7, 2.89, 2.6,	1.9493 + 0.0313x	0.14	2.216	9
17	1.9, 1.7, 2.6, 2.4, 2.6, 2.2, 2.2, 2.4, 2.5, 1.9, 2.4, 2.4, 2.5, 2.0, 2.72, 2.5, 2.7	2.0782 + 0.0284x	0.22	2.334	25
18	2.0, 1.8, 2.2, 2.4, 2.3, 2.3, 2.5, 2.7, 1.8, 2.0, 3.1, 2.4, 2.2, 2.1, 2.42, 1.8, 2.3, 2.3	2.1841 + 0.0069x	0.01	2.25	12
19	2.0, 2.3, 2.0, 2.5, 2.4, 2.4, 2.0, 2.1, 2.2, 2.4, 2.3, 2.8, 2.3, 2.3, 2.42, 2.4, 2.6, 2.5, 2.8	2.0960 + 0.0251x	0.35	2.347	12
20	1.5, 1.8, 2.1, 2.6, 2.5, 2.8, 2.2, 2.1, 2.9, 2.3, 1.8, 2.4, 2.9, 2.0, 2.58, 2.4, 2.4, 2.4, 2.4, 2.8	2.0602 + 0.0272x	0.17	2.346	12
21	2.5, 2.1, 1.8, 2.3, 2.2, 2.7, 1.9, 2.1, 2.1, 2.0, 2.9, 3.2, 2.3, 2.6, 2.08, 2.5, 1.7, 2.5, 2.3, 2.9, 2.9	2.1118 + 0.0220x	0.11	2.353	12
22	1.9, 2.0, 2.2, 2.6, 2.0, 2.3, 2.7, 2.5, 2.3, 2.1, 2.3, 2.3, 2.9, 2.1, 2.92, 2.0, 2.9, 2.3, 1.8, 2.0, 2.2, 2.3	2.2673 + 0.0025x	0.00	2.295	12
23	1.9, 2.6, 2.7, 1.9, 1.9, 2.5, 2.6, 3.1, 2.6, 2.1, 2.3, 2.6, 2.6, 2.2, 2.6, 1.9, 1.7, 2.3, 2.6, 2.4, 2.6, 2.7, 3.4	2.2516 + 0.0147x	0.06	2.428	11
24	2.8, 2.3, 2.8, 1.8, 2.7, 1.8, 1.8, 1.8, 2.8, 1.8, 2.7, 2.7, 2.0, 1.3, 2.5, 1.8, 3.3, 2.3, 2.2, 1.8, 1.8, 2.3, 2.5, 2.3,	2.3433 - 0.0065x	0.01	2.262	6
25	2.0, 2.4, 2.1, 2.1, 1.5, 1.8, 2.1, 1.9, 2.6, 2.2, 2.4, 1.9, 2.3, 1.9, 2.1, 1.8, 2.1, 2.4, 2.0, 2.2, 1.7, 2.1, 2.2, 2.1, 2.0	2.0740 + 0.0002x	0.00	2.076	10

The course of b is shown in Figure 11. As can be seen, b decreases again, $b = 0.8698 \exp(-0.1786n)$ with $R^2 = 0.75$. The reliable interval of Seneca's text (i.e., $k \geq 10$) is relatively short ($6 \leq n \leq 25$) for this text.

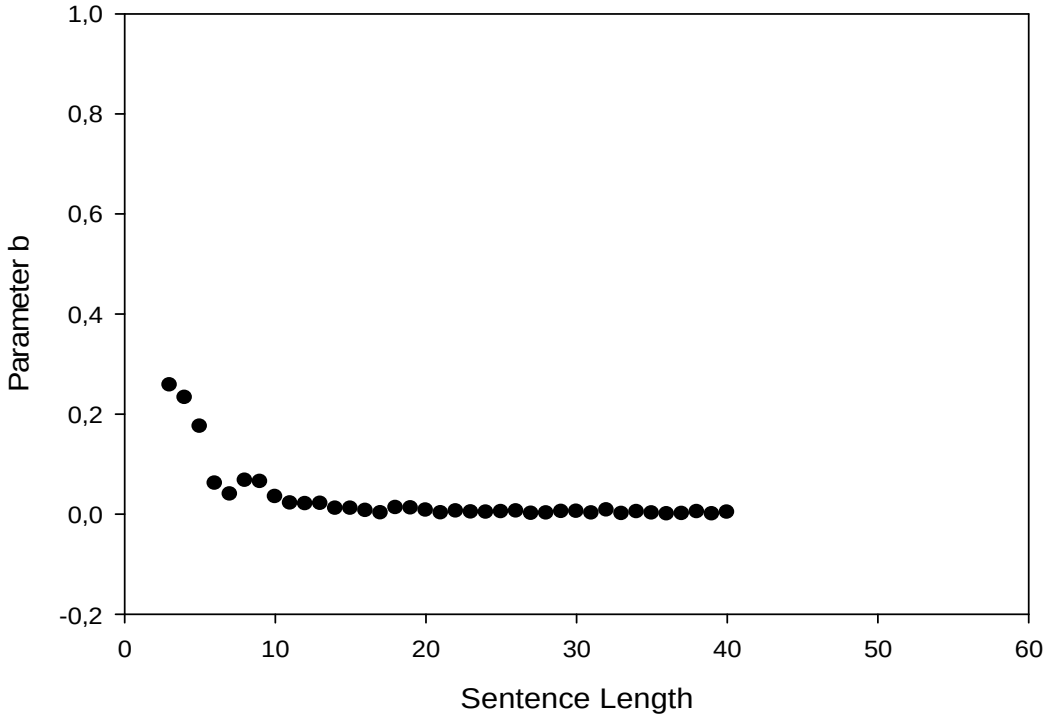


Figure 11. Decrease of *b* in Seneca's Latin text

2.5. English

Let us now consider English, a language which, from an evolutionary point of view, can be regarded as to be increasingly analytic. For our analysis, journalistic texts were combined to a corpus of 2268 sentences and 43320 words, and submitted to analysis. The results obtained are presented in Table 8. We considered the data for $3 \leq n \leq 40$.

Table 8
Mean lengths (L) of words in individual positions (x) in sentences of length n in English press texts (M = mean of all positions, k = number of sentences)

n	**Mean syllabic lengths of words**	**Linear relation** $L = a + bx$	R^2	***M***	***k***
3	1.18, 1.48, 1.70	0.9393 + 0.2576x	0.99	1.45	33
4	1.17, 1.25, 1.28, 1.94	0.8282 + 0.2328x	0.72	1.41	64
5	1.19, 1.48, 1.50, 1.73, 1.94	1.0417 + 0.1750x	0.96	1.57	48
6	1.34, 1.46, 1.34, 1.50, 1.58, 1.66	1.2648 + 0.0614x	0.78	1.48	74
7	1.62, 1.39, 1.42, 1.62, 1.37, 1.50, 1.93	1.3928 + 0.0395x	0.19	1.55	76
8	1.23, 1.52, 1.44, 1.49, 1.59, 1.54, 1.46, 2.01	1.2334 + 0.0671x	0.55	1.54	71
9	1.19, 1.44, 1.50, 1.33, 1.66, 1.52, 1.61, 1.53, 1.98	1.2048 + 0.0646x	0.63	1.53	64
10	1.49, 1.48, 1.39, 1.47, 1.52, 1.60, 1.36, 1.57, 1.71, 1.87	1.3567 + 0.0346x	0.47	1.55	77
11	1.34, 1.36, 1.37, 1.45, 1.56, 1.46, 1.53,	1.3329 + 0.0218x	0.30	1.46	107

	1.50, 1.32, 1.43, 1.79				
12	1.40, 1.46, 1.51, 1.52, 1.56, 1.50, 1.39, 1.60, 1.51, 1.47, 1.57, 1.88	1.3980 + 0.0204x	0.35	1.53	104
13	1.44, 1.54, 1.63, 1.63, 1.57, 1.61, 1.61, 1.58, 1.69, 1.69, 1.63, 1.64, 1.93	1.4829 + 0.0211x	0.55	1.63	72
14	1.60, 1.76, 1.50, 1.58, 1.56, 1.39, 1.53, 1.69, 1.52, 1.56, 1.53, 1.60, 1.68, 1.97	1.5222 + 0.0110x	0.11	1.60	62
15	1.24, 1.50, 1.47, 1.58, 1.50, 1.59, 1.46, 1.46, 1.50, 1.59, 1.35, 1.38, 1.39, 1.49, 1.93	1.4072 + 0.0111x	0.11	1.50	74
16	1.37, 1.54, 1.73, 1.56, 1.64, 1.49, 1.47, 1.56, 1.54, 1.40, 1.29, 1.63, 1.39, 1.49, 1.70, 1.91	1.4857 + 0.0068x	0.04	1.54	70
17	1.45, 1.62, 1.64, 1.66, 1.52, 1.70, 1.51, 1.74, 1.55, 1.68, 1.51, 1.61, 1.47, 1.55, 1.47, 1.44, 1.98	1.5737 + 0.0021x	0.01	1.59	87
18	1.38, 1.64, 1.55, 1.58, 1.58, 1.55, 1.64, 1.55, 1.55, 1.58, 1.53, 1.54, 1.53, 1.49, 1.66, 1.57, 1.64, 2.13	1.4747 + 0.0126x	0.20	1.59	76
19	1.33, 1.50, 1.45, 1.63, 1.43, 1.41, 1.53, 1.46, 1.41, 1.58, 1.61, 1.68, 1.55, 1.66, 1.47, 1.67, 1.41, 1.46, 1.91	1.4150 + 0.0119x	0.24	1.53	76
20	1.39, 1.61, 1.62, 1.69, 1.71, 1.55, 1.50, 1.43, 1.68, 1.64, 1.50, 1.42, 1.57, 1.70, 1.40, 1.50, 1.48, 1.88, 1.62, 1.98	1.5146 + 0.0075x	0.08	1.59	84
21	1.33, 1.60, 1.65, 1.54, 1.63, 1.39, 1.55, 1.44, 1.49, 1.54, 1.43, 1.56, 1.49, 1.69, 1.63, 1.46, 1.56, 1.51, 1.29, 1.45, 1.83	1.5014 + 0.0021x	0.01	1.53	80
22	1.73, 1.29, 1.52, 1.68, 1.48, 1.64, 1.68, 1.41, 1.59, 1.38, 1.65, 1.77, 1.53, 1.45, 1.70, 1.45, 1.85, 1.55, 1.80, 1.42, 1.35, 2.00	1.5217 + 0.0057x	0.04	1.59	66
23	1.47, 1.41, 1.33, 1.47, 1.57, 1.47, 1.51, 1.35, 1.92, 1.61, 1.90, 1.35, 1.55, 1.71, 1.55, 1.61, 1.35, 1.35, 1.73, 1.37, 1.37, 1.59, 1.76	1.4875 + 0.0039x	0.02	1.53	49
24	1.36, 1.52, 1.43, 1.71, 1.57, 1.45, 1.62, 1.64, 1.69, 1.67, 1.53, 1.41, 1.71, 1.47, 1.59, 1.66, 1.52, 1.26, 1.57, 1.50, 1.55, 1.57, 1.50, 1.95	1.5162 + 0.0035x	0.03	1.56	58
25	1.47, 1.44, 1.65, 1.38, 1.62, 1.58, 1.58, 1.67, 1.55, 1.75, 1.25, 1.56, 1.67, 1.69, 1.71, 1.73, 1.78, 1.35, 1.47, 1.42, 1.64, 1.45, 1.65, 1.84, 1.65	1.5238 + 0.0045x	0.05	1.58	55
26	1.28, 1.48, 1.60, 1.60, 1.51, 1.52, 1.61, 1.52, 1.63, 1.54, 1.93, 1.52, 1.58, 1.58, 1.64, 1.51, 1.42, 1.82, 1.42, 1.67, 1.51, 1.70, 1.60, 1.48, 1.52, 1.94	1.5039 + 0.0058x	0.09	1.58	67
27	1.55, 1.61, 1.78, 1.61, 1.43, 1.57, 1.84,	1.5907 + 0.0011x	0.00	1.61	51

	1.71, 1.67, 1.39, 1.69, 1.47, 1.51, 1.82, 1.49, 1.59, 1.51, 1.53, 1.45, 1.61, 1.49, 1.67, 1.61, 1.69, 1.39, 1.69, 2.00				
28	1.51, 1.35, 1.60, 1.60, 1.42, 1.49, 1.60, 1.81, 1.84, 1.47, 1.67, 1.77, 1.60, 1.58, 1.53, 1.65, 1.23, 1.60, 1.47, 1.51, 1.56, 1.65, 1.51, 1.53, 1.53, 1.56, 1.60, 1.84	1.5519 + 0.0016x	0.01	1.58	43
29	1.32, 1.50, 1.80, 1.52, 1.64, 1.43, 1.61, 1.70, 1.39, 1.77, 1.64, 1.61, 1.50, 1.41, 1.68, 1.34, 1.64, 1.66, 1.59, 1.57, 1.66, 1.57, 1.66, 1.55, 1.48, 1.75, 1.48, 1.86, 1.77	1.5200 + 0.0046x	0.08	1.59	44
30	1.36, 1.57, 1.55, 1.50, 1.60, 1.50, 1.50, 1.50, 1.33, 1.52, 1.86, 1.48, 1.60, 1.52, 1.45, 1.62, 1.79, 1.74, 1.57, 1.43, 1.71, 1.45, 1.79, 1.74, 1.57, 1.36, 1.71, 1.45, 1.45, 1.95	1.4915 + 0.0052x	0.09	1.57	42
31	1.43, 1.60, 1.40, 1.45, 1.50, 1.85, 1.63, 1.55, 1.48, 1.75, 1.53, 1.40, 1.68, 1.65, 1.55, 1.58, 1.73, 1.50, 1.70, 1.63, 1.55, 1.53, 1.70, 1.63, 1.53, 1.48, 1.65, 1.43, 1.38, 1.63, 1.83	1.5466 + 0.0018x	0.02	1.58	40
32	1.17, 1.61, 1.33, 1.42, 1.42, 1.64, 1.47, 1.47, 2.00, 1.58, 1.42, 1.28, 1.47, 1.44, 1.44, 1.53, 1.56, 1.69, 1.47, 1.72, 1.56, 1.69, 1.61, 1.58, 1.86, 1.61, 1.53, 1.39, 1.83, 1.33, 1.50, 2.03	1.4236 + 0.0078x	0.14	1.55	36
33	1.19, 1.65, 1.73, 1.46, 1.62, 1.54, 1.73, 1.62, 1.62, 1.65, 1.77, 1.31, 1.81, 1.46, 1.50, 1.96, 1.62, 1.69, 1.42, 1.42, 2.04, 1.31, 1.54, 1.31, 1.50, 1.69, 1.42, 1.65, 1.54, 1.85, 1.15, 1.50, 2.08	1.5727 + 0.0008x	0.00	1.59	26
34	1.40, 1.40, 1.65, 1.60, 1.30, 1.40, 1.30, 1.50, 1.55, 1.50, 1.45, 1.25, 1.65, 1.55, 1.55, 1.35, 1.55, 1.90, 1.85, 1.50, 1.85, 1.55, 1.40, 1.50, 1.65, 1.35, 1.75, 1.55, 1.35, 1.85, 1.45, 1.40, 1.60, 1.65	1.4529 + 0.0045x	0.07	1.53	20
35	1.73, 1.57, 1.63, 1.47, 1.37, 1.70, 1.70, 1.50, 1.47, 1.47, 1.63, 1.63, 1.50, 1.47, 1.93, 1.77, 1.53, 1.43, 1.67, 1.70, 1.30, 1.87, 1.53, 1.50, 1.53, 1.50, 1.27, 1.80, 1.57, 1.50, 1.80, 1.33, 1.53, 1.67, 2.10	1.5599 + 0.0017x	0.01	1.59	30
36	1.54, 1.85, 1.54, 1.58, 1.46, 1.69, 1.73, 1.42, 1.69, 1.85, 1.46, 1.35, 1.96, 1.46, 1.46, 1.85, 1.38, 1.77, 1.46, 1.69, 1.54, 1.35, 1.38, 1.46, 1.65, 2.15, 1.62, 1.50, 1.46, 1.85, 1.23, 1.54, 1.54, 1.46, 1.73, 1.92	1.6017 - 0.0001x	0.00	1.60	26
37	1.53, 1.63, 1.32, 1.21, 1.68, 1.79, 1.63,	1.4969 + 0.0006x	0.00	1.51	19

	1.32, 1.42, 1.89, 1.37, 1.74, 1.68, 1.37, 1.53, 1.68, 1.26, 1.16, 1.32, 1.63, 1.47, 1.21, 1.37, 1.58, 1.37, 1.53, 2.00, 1.21, 1.26, 1.58, 1.26, 1.58, 1.95, 1.79, 1.26, 1.26, 1.95				
38	1.41, 1.06, 1.35, 1.41, 1.47, 1.59, 1.47, 1.76, 1.59, 1.35, 1.94, 1.24, 1.82, 1.65, 1.53, 1.59, 1.76, 1.76, 1.35, 1.35, 1.35, 1.53, 1.71, 1.76, 1.59, 1.47, 1.71, 1.47, 1.65, 1.41, 1.47, 1.24, 1.53, 1.71, 1.88, 1.06, 1.82, 1.94	1.4610 + 0.0044x	0.05	1.55	17
39	1.33, 2.17, 1.58, 1.58, 1.92, 1.08, 1.42, 2.17, 1.67, 1.58, 1.67, 1.42, 1.67, 2.08, 1.08, 1.75, 1.75, 1.50, 1.67, 1.75, 1.58, 1.25, 2.00, 1.58, 2.08, 1.67, 1.58, 1.08, 1.50, 1.83, 1.33, 1.75, 1.50, 1.42, 1.92, 1.58, 1.75, 1.75, 1.75	1.6343 + 0.0000x	0.00	1.63	12
40	1.53, 1.60, 1.33, 1.33, 1.80, 1.60, 1.87, 1.80, 1.20, 1.40, 1.67, 1.93, 1.60, 1.40, 1.80, 1.47, 1.73, 1.53, 1.20, 1.33, 1.40, 1.67, 1.47, 1.80, 1.67, 1.53, 1.53, 1.67, 1.33, 1.27, 2.07, 1.67, 2.00, 1.20, 1.47, 1.40, 1.60, 1.93, 1.40, 2.40	1.5187 + 0.0035x	0.02	1.59	15
41	1.78, 1.44, 1.11, 2.22, 1.89, 1.00, 1.56, 1.56, 1.67, 1.56, 1.33, 1.67, 1.78, 1.22, 1.89, 1.33, 1.78, 1.67, 1.33, 1.67, 1.56, 1.78, 1.33, 1.33, 1.11, 1.44, 1.22, 1.44, 1.22, 1.22, 1.78, 1.11, 2.22, 1.56, 1.11, 1.67, 1.44, 1.67, 1.22, 1.89, 1.67	1.5678 - 0.0021x	0.01	1.52	9
42	1.42, 1.75, 1.67, 1.33, 1.67, 1.33, 2.00, 1.83, 1.58, 1.83, 1.33, 1.67, 1.58, 1.50, 1.42, 1.50, 1.33, 1.42, 1.33, 1.58, 1.83, 1.42, 1.42, 1.67, 1.83, 1.67, 1.25, 1.50, 1.58, 1.33, 1.33, 1.33, 1.67, 1.83, 1.33, 1.58, 1.25, 1.67, 1.25, 1.75, 1.75, 2.00	1.5660 - 0.0005x	0.00	1.56	12
43	1.27, 1.73, 1.36, 1.18, 1.82, 1.64, 1.82, 1.64, 1.18, 1.73, 1.91, 2.09, 1.82, 1.64, 1.82, 1.55, 1.73, 1.64, 1.45, 1.64, 1.09, 1.82, 2.00, 1.09, 1.45, 1.27, 1.45, 1.27, 1.36, 1.91, 1.18, 1.82, 1.55, 1.64, 1.55, 2.09, 1.45, 2.09, 1.82, 1.45, 1.00, 1.55, 2.55	1.5575 + 0.0022x	0.01	1.61	11
44	1.70, 1.50, 1.30, 1.50, 1.60, 2.00, 1.70, 1.80, 1.60, 1.50, 1.30, 1.80, 2.00, 1.90, 1.30, 1.70, 1.50, 1.90, 1.40, 1.20, 1.20, 1.50, 1.50, 1.30, 1.90, 1.20, 1.60, 1.60, 1.20, 1.90, 1.90, 1.90, 1.40, 1.50, 1.80, 1.07, 1.30, 1.60, 1.40, 1.30, 1.20, 1.50, 1.40, 2.30	1.6094 - 0.0015x	0.01	1.58	10

The course of *b* is illustrated in Figure 12: In English, the decrease of parameter *b* can be expressed by the function $b = 0.6980 \exp(-0.3118n)$, with $R^2 = 0.95$.

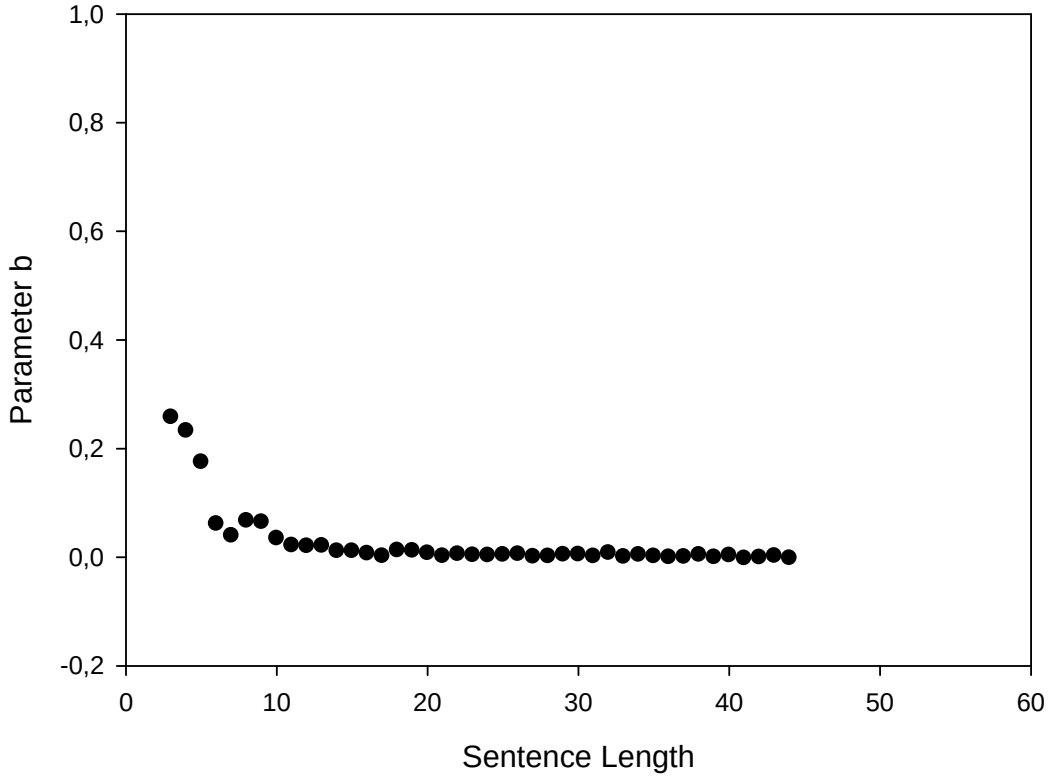


Figure 12. Decrease of *b* in a corpus of English journalistic texts

2.6. Indonesian

The last language analyzed here is a longer Indonesian text containing the well-known fairy tale *Burung api* [= *The fire bird*] by Pak Ojik (1971); this text consists of 409 sentences (4260 words). Indonesian is a language with a low degree of inflection and a higher degree of composition and derivation. Since the text is a fairy tale, the sentences are not too long; yet, we obtained reliable results for the sentence length interval $3 \leq n \leq 17$. The results are presented in Table 9 and the decrease of *b* in Figure 13.

Table 9
Mean lengths (*L*) of words in individual positions (*x*) in sentences of length *n*
in the Indonesian text *Burung api* - (*M* = mean of all positions, *k* = number of sentences)

n	**Mean syllabic lengths of words** ***L***	**Linear relation** $L = a + bx$	R^2	***M***	***k***
3	2.76, 3.12, 3.35	2.4867 + 0.2950x	0.98	3.08	17
4	2.56, 2.28, 2.67, 2.78	2.3100 + 0.1050x	0.40	2.57	18
5	2.57, 2.62, 2.71, 2.57, 3.00	2.4510 + 0.0810x	0.50	2.69	21
6	2.53, 2.53, 2.22, 2.22, 2.53, 2.72	2.3633 + 0.0271x	0.07	2.46	32
7	2.32, 2.68, 2.53, 2.32, 2.62, 2.38, 2.88	2.3657 + 0.0418x	0.19	2.53	34

8	2.44, 2.59, 2.30, 2.41, 2.22, 2.52, 2.52, 2.85	2.3211 + 0.0356x	0.21	2.48	27
9	2.50,2.33,2.30,2.43,2.57,2.47, 2.47, 2.37, 2.53	2.3894 + 0.0103x	0.10	2.44	30
10	2.67, 2.67, 2.27, 2.36, 2.58, 2.21, 2.18, 2.27, 2.52, 3.06	2.4273 + 0.0094x	0.01	2.48	33
11	2.45, 2.64, 2.21, 2.42, 2.64, 2.70, 2.42, 2.36, 2.42, 2.48, 2.48	2.4855 - 0.0018x	0.00	2.47	33
12	2.26, 2.57, 2.61, 2.26, 2.52, 2.39, 2.26, 2.65, 2.48, 2.52, 2.17, 2.96	2.3611 + 0.0169x	0.08	2.47	23
13	2.85, 2.50, 2.74, 2.47, 2.85, 2.65, 2.59, 2.50, 2.12, 2.62, 2.29, 2.71, 2.94	2.6550 - 0.0075x	0.02	2.60	34
14	2.43, 2.64, 2.29, 2.36, 2.29, 3.00, 2.50, 2.71, 2.29, 2.36, 2.07, 2.21, 2.29, 2.64	2.5240 - 0.0120x	0.04	2.43	14
15	2.08, 2.92, 3.25, 2.17, 2.58, 2.92, 2.42, 2.25, 2.58, 2.50, 2.58, 2.42, 2.25, 3.00, 2.83	2.5533 + 0.0038x	0.00	2.58	12
16	2.67, 2.56, 2.22, 2.11, 2.56, 2.78, 3.22, 3.22, 2.67, 2.11, 2.78, 2.33, 2.89, 1.89, 2.67, 3.00	2.5515 + 0.0063x	0.01	2.61	9
17	2.47, 2.60, 2.47, 2.40, 2.47, 3.00, 2.47, 2.33, 2.93,2.67,2.00,2.33,2.73,2.53, 2.07, 2.53, 2.47	2.5823 - 0.0093x	0.03	2.50	15

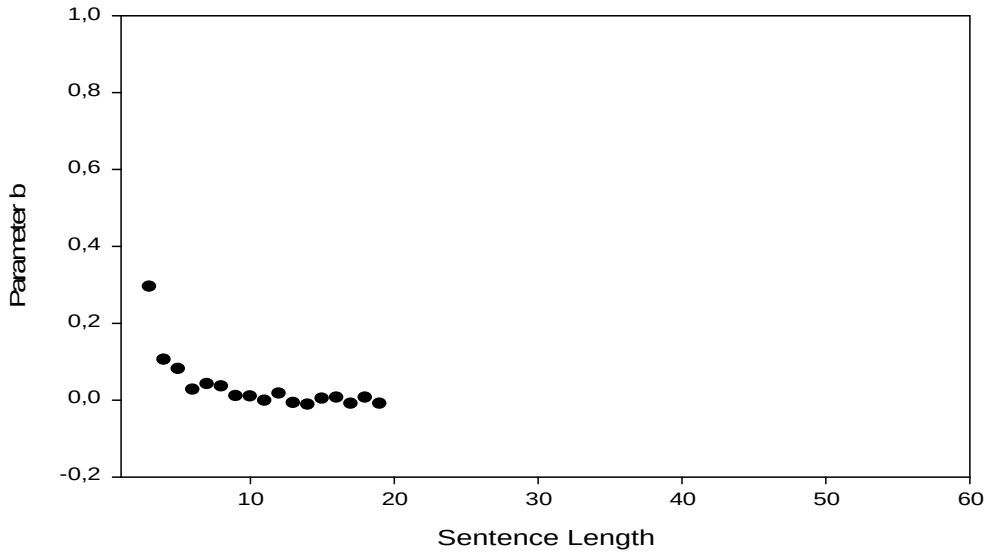


Figure 13. The decrease of parameter *b* in the Indonesian text *Burung api*

Though the number of well represented lengths is rather small, parameter *b* clearly decreases exponentially with increasing sentence length, just as in all other cases above. Here we obtain $b = 2.5394 \exp(-0.7295n)$ with $R^2 = 0.97$.

3. Conclusions

Summarizing the results from all texts and languages, we can state that a common tendency can be observed across (typologically different) languages. According to this tendency, word length is not stable in the course of a sentence, but changes; specifically, there seems to be an increase from the beginning of a sentence to its end, the degree of increase following a straight line (if one ignores possible specifics of first and last words, as was pointed out

above). As a result, this increase is slowed down for longer sentences (signalized by decreasing parameter *b*), resulting in some kind of "zero-increase" towards the end of longer sentences (signalized by parameter $b \to 0$). The overall exponentially decreasing tendency of *b* across the languages studied in this text can be presented lucidly in Table 10. The data are ordered according to the increasing exponent, which is clearly associated with the decreasing multiplicative constant.

Table 10
The decrease of parameter *b* in individual languages

Language	***b***
Slovak (press texts)	25.2594 exp(-1.0298n)
Indonesian (Ojik)	2.5394 exp(-0.7295n)
Latin (Seneca)	2.5420 exp(-0.4556n)
English (press texts)	0.6980 exp(-0.3116n)
Russian (Tolstoj)	0.8737 exp(-0.2763n)
Hungarian (Bródy)	0.7432 exp(-0.2756n)
Slovak (Šikula)	0.4325 exp(-0.1812n)
Hungarian (Karinthy)	0.3061 exp(-0.1778n)
Russian (press texts)	0.2349 exp(-0.1616n)

As can be seen from Table 10, the parameter values of *b* do not seem to represent some constant for a given language; rather, *b* may significantly vary with a given language. For the time being, we do not have any reliable interpretation as to factors influencing *b*. It seems most reasonable to assume that the coefficients in the exponential function are related not only to characteristics of the given language, but also to certain text sort specifics, and by way of this to average word and sentence length, possibly influenced by the number (and type) of clauses constituting the sentences.

Though we do not consider the number of analyzed texts to be sufficient to arrive at general conclusions, hypothesis (I) has been corroborated in all cases: Word length increases towards the end of a sentence; this conclusion must be specified, however, in one important aspect: the longer the sentence, the more this increase is being braked (i.e., the slower is the increase); in very long sentences it can even turn negative, although it is as well possible that values of $b > 0$ do not significantly differ from zero. The slight oscillation in very long sentences may as well indicate the loss of control of word length. In any case, we can conjecture that the change of *b* according to the change of *n*, i.e. *db/dn* is slowed down to the extent almost achieved by *b*. Re-writing this assumption in a formal way, we obtain

$$\frac{db}{dn} = -cb. \tag{1}$$

Solving this elementary differential equation we obtain

$$b = ae^{-cn} \tag{2}$$

where *a* is the integration constant. In fact, function (2) has turned out to be adequate in the above analyses; the numerical values of fitting it to individual samples are presented in Table 10.

Since formula (2) can be also be written as $b = aq^n$ (writing $e^{-c} = q$), which represents a geometric sequence, we may also conclude that the increasing tendency of word length diminishes geometrically as sentence length increases. However, the fact that frequently the first word is longer than the second, and the last word longer than the last but one suggests that the mechanism controlling word length in sentence is not that simple as presented above. There are surely other factors influencing the smooth course of the regression. We consider them as boundary conditions which must be studied for each case separately.

Our data suggest that the phenomenon observed is neither language specific nor restricted to particular genres; most probably, we are concerned with a latent cross-linguistic mechanism. Its genesis can be directly derived from the unified theory of linguistic laws (cf. Wimmer, Altmann 2005) as shown above. Thus we can consider it as a candidate for a text law. However, more texts in more languages must be scrutinized in order to find all the boundary conditions determining the values of the parameters *a* and *c* in (2).

References

Ahrens, H. (1965). *Verborgene Ordnung: Die Beziehung zwischen Satzlänge und Wortlänge in deutscher Erzählprosa von Barock bis heute*. Düsseldorf: Schwann.

Altmann, G. (1983). H. Arens´ "Verborgene Ordnung" und das Menzerathsche Gesetz. In: Faust, M., et al. (eds.), *Allgemeine Sprachwissenschaft, Sprachtypologie und Textlinguistik: 31-39*. Tübingen: Narr.

Grzybek, P., Kelih, E., Stadlober, E. (2007). Long sentences, long words – short sentences, long words? *Presentation at the 31. Jahrestagung der Gesellschaft für Klassifikation: "Data Analysis, Machine Learning, and Application"*. (Freiburg, Germany 2007.)

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

Grzybek, P., **Stadlober, E.** (2007). Do we have problems with Arens´ law? A new look at the sentence-word relation. In: Grzybek, P., Köhler, R. (eds.), *Exact Methods in the Study of Language and Text: 205-218*. Berlin: de Gruyter.

Grzybek, P., Stadlober, E., Kelih, E. (2007). The relationship of word length and sentence length: the inter-textual perspective. In: R. Decker and H.-J. Lenz (eds.), *Advances in Data Analysis: 611-618*. Berlin: Springer.

Grzybek, P., Stadlober, E., Kelih, E., Antić, G. (2005). Quantitative text typology: the impact of word length. In: C. Weihs, and W. Gaul (eds.), *Classification – The Ubiquitous Challenge: 53-64*. Berlin: Springer.

Kelih, E., Grzybek, P., Antić, G., Stadlober, E. (2006). Quantitative text typology: the impact of sentence length. In: M. Spiliopoulou et al. (eds.), *From Data and Information Analysis to Knowledge Engineering: 382-389*. Berlin: Springer.

Niemikorpi, A. (1991). Suomen kielen sanaston dynamiikkan. *Acta Wasaensia 26, Kielitiede 2*. Vaasa: Vaasan yliopisto.

Niemikorpi, A. (1997). Equilibrium of Words in the Finnish Frequency Dictionary. *JQL 4(1-3), 190-196*.

Saukkonen, P. (1994). Main trends and results of quantitative linguistics in Finland. *JQL 1(1), 2-15*.

Strauss, U., Fan, F., Altmann, G. (2008). *Problems in Quantitative Linguistics 1*. Lüdenscheid: RAM:

Uhlířová, L. (1997a). O vztahu mezi délkou slova a jeho polohou ve větě. *Slovo a slovesnost 58, 174-184*.

Uhlířová, L. (1997b). Length vs. Order: Word Length and Clause Length from the Perspective of Word Order. *Journal of Quantitative Linguistics 4(1-3), 266-275.*

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

Texts and their sources

English

1. *A Few Good Shirts*, Jonathan Tepperman, Newsweek, March 27, 2009, www.newsweek.com/2009/03/27/a-**few-good-shirts**.html
2. *All Things Greek: To Hellenic and Back*, Jeremy McCarter, Newsweek, March 18, 2010, http://www.newsweek.com/2010/03/18/all-things-greek-to-hellenic-and- back. html
3. *Calming the Bear*, Barton Biggs, Newsweek, March 6, 2009, www.newsweek.com/2009/03/06/**calming-the-bear**.html
4. *Don't talk to aliens, warns Stephen Hawking*, Jonathan Leake, the Sunday Times, April 25, 2010, http://www.timesonline. co.uk/tol/news/science /space/article 7107207. ece
5. *Dr. No*, Owen Matthews, Newsweek, March 11, 2010, www.newsweek.com/2010/03/11/**dr-no**.html
6. *Geoducks: Happy as Clams*, Craig Welch, Smithsonian Magazine, March 2009, http://www.smithsonianmag.com/science-nature/Happy-As-Clams.html
7. *If the Stones Could Speak, Searching for the Meaning of Stonehenge*, Caroline Alexander, National Geographic, June 2008, http://ngm.nationalgeographic.com/2008/06/stonehenge/geiger-photography
8. *J. D. Salinger Outlived His Legend*, Malcolm Jones, Newsweek Web Exclusive, January, 27, 2010, http://www.newsweek.com/2010/01/27/j-d-salinger-outlived- his-legend.html
9. *Letter from China, Crazy English*, Evan Osnos, the New Yorker, April 28, 2008, http://www.newyorker.com/reporting/2008/04/28/080428fa_fact_osnos
10. *Mourning the Death of Handwriting*, Claire Suddath, August 03, 2009, Time Magazine, http://www. time.com/time/magazine/article/0,9171,1912419,00.html
11. *One Small Step*, George Alexander, July 17, 2009, Newsweek, www.newsweek.com/2009/07/16/**one-small-step**.html
12. *Palin With a Pedigree*, Michael Isikoff and Michael Hirsh, Newsweek, March 12, 2010, www.newsweek.com/2010/03/11/**palin-with-a-pedigree**.html
13. *Pregnant? Eat Chocolate!* Rob Stein, Washington Post, May 8, 2008, http://voices.washingtonpost.com/checkup/2008/05/bad_news_for_pregnant_women_ea.html
14. *Scanning a Stradivarius*, Erica R. Hendry, May 2010, Smithsonian Magazine, http://www.smithsonianmag.com/arts-culture/Scanning-a-Stradivarius.html
15. *The Case Against Goldman Sachs*, Stephen Gandel, Time Magazine, April 22, 2010, http://www.time.com/time/business/article/0,8599,1983747,00.html
16. *The Greatest Shakespeare Hoax*, Doug Stewart, Smithsonian Magazine, June 2010, http://www.smithsonianmag.com/history-archaeology/The-Greatest-Shakespeare-Hoax.html
17. *The iPad Launch: Can Steve Jobs Do It Again?* Stephen Fry, Time Magazine, April 01, 2010. http://www.time.com/time/business/article/0,8599, 1976935,00.html
18. *The Pill at 50: Sex, Freedom and Paradox*, Nancy Gibbs Thursday, Time Magazine, April 22, 2010, http://www.time.com/time/health/article/0,8599,1983712,00.html

19. *Trading Unit Propels Morgan Stanley to Profit*, Michael J. de la Merced, New York Times, April 21, 2010, http://www.nytimes.com/2010/04/22/business/22 morgan.html
20. *Unlocking Mysteries of the Parthenon*, Evan Hadingham, Smithsonian Magazine, February 2008, http://www.smithsonianmag.com/ history-archaeology/ Unlocking-Mysteries-of-the-Parthenon.html

Latin

Seneca, L.A., *Epistularum moralium ad Lucilium liber primus*
http://www.thelatinlibrary.com/sen/seneca.ep1.shtml (28.7.2010)

Hungarian

Bródy, S., *Az ezüst kecske*. 1898.
Karinthy, F., *Utazás a koponyám körül.* 1937.

Slovak

Journalistic texts from a Slovak corpus
Šikula, V. *Veterná ružica*. 1995.

Russian

44 journalistic texts from the journal Vremja (http://quanta-textdata.uni-graz.at/).
L.N. Tolstoj, *Anna Karenina*. 1877.

Indonesian

Pak Ojik, *Burung Api*. Djakarta: Pustaka Jaya. 1971

Glottometrics 20, 2010, 110-114

History of Quantitative Linguistics

Since a historiography of quantitative linguistics does not exist as yet, we shall present in this column short statements on researchers, ideas and findings of the past – usually forgotten – in order to establish a tradition and to complete our knowledge of history. Contributions are welcome and should be sent to Peter Grzybek, peter.grzybek@uni-graz.at.

Laut- und Buchstabenzählungen im frühen 19. Jahrhundert

Karl-Heinz Best

1. Vorbemerkung

Bei der Suche nach der Vorgeschichte der Quantitativen Linguistik stößt man auf sprachstatistische Erhebungen, die im Zusammenhang mit der Entwicklung der Stenographie entstanden. Für das 19. Jahrhundert ist Kaedings kommentierte Bibliographie (1897/98: 37-40) eine wichtige Quelle dafür. Geht man diesen Hinweisen nach, macht man unterschiedliche Erfahrungen, abgesehen davon, dass sie auch nicht vollständig sind: In Einzelfällen handelt es sich um bekannte Autoren, deren Werke leicht verfügbar sind; in anderen Fällen um solche, die nur bei Stenographen mehr oder weniger bekannt sind und deren Werke in den wissenschaftlichen Bibliotheken meistens fehlen und nur in Spezialsammlungen wie vor allem der Forschungs- und Ausbildungsstätte für Kurzschrift und Textverarbeitung in Bayreuth E.V., der Universitätsbibliothek Dresden und der Bibliothek des niedersächsischen Landtags in Hannover an Ort und Stelle einzusehen sind.

2. Zwei frühe Zählungen

In diesem Beitrag sollen zwei solcher Werke vorgestellt werden; dies geschieht in aller Kürze, da über die Autoren bisher fast nichts in Erfahrung zu bringen war. Immerhin können die bibliographischen Angaben ergänzt und die Statistiken, die die beiden Autoren vorgelegt haben, vorgestellt und daraufhin getestet werden, ob ihre Daten Altmanns Modell für beliebige Rangordnungen (1993: 62)

$$y_x = \frac{\binom{b+x}{x-1}}{\binom{a+x}{x-1}} c, \qquad x = 1, 2, 3, \ldots$$

entsprechen, das keine Verteilung, sondern eine Folge darstellt. Hier sind *a,b* und *c* Parameter. Das Modell hat sich schon mehrmals für ähnliche Rangordnungen bewährt (Best 2004/5; 2005; 2006: 57-60), u.a. bei der 100000-Laute-Zählung“, die Meier (1967: 249ff.) präsentierte.

2.1. XLII. François Dujardin aîné (1834)

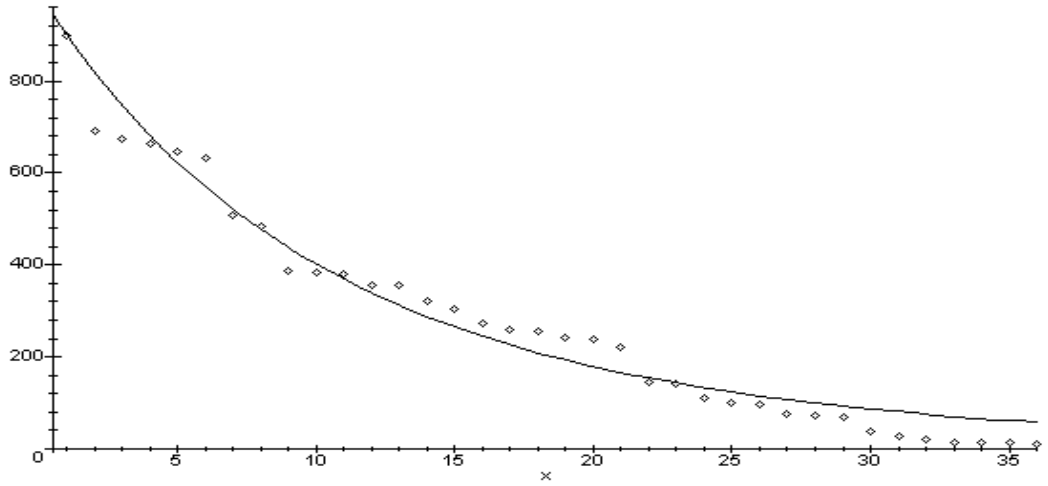
Im ersten Fall handelt es sich um eine Lautzählung zum Französischen durch Auswertung von 10117 Lauten, die in Dujardin aîné (1834) enthalten ist und von Faulmann (1887: 327) mitgeteilt wird. Das Werk von Dujardin aîné (1834) ist ein „Extrait“ aus einer Zeitschrift; es konnte noch nicht geklärt werden, ob der Originalbeitrag im gleichen Jahr oder früher erschienen ist. Die Tabelle 1 führt die Zählung nach den Angaben von Faulmann (1887: 327) an[1].

Tabelle 1
Anpassung des Modells für beliebige Rangordnungen
an die von F. Dujardin ainé mitgeteilten französischen Lauthäufigkeiten

Rang	Lautbe-zeichnung	n_x	NP_x	Rang	Lautbe-zeichnung	n_x	NP_x
1	e	900	900.00	19	ou	241	193.57
2	i	690	819.20	20	s[2]	237	179.24
3	a	674	746.51	21	v	222	166.11
4	è	664	681.04	22	f	145	154.05
5	s	646	622.00	23	b	140	142.97
6	t	634	568.68	24	ein	110	132.79
7	é	507	520.48	25	j	99	123.42
8	k	485	476.85	26	eu	97	114.80
9	an	387	437.32	27	g	76	106.84
10	o	382	401.46	28	ll	74	99.51
11	l	380	368.89	29	oi	68	92.74
12	p	357	339.28	30	ch	39	86.49
13	d	356	312.33	31	un	27	80.71
14	n	322	287.78	32	h	19	75.36
15	on	303	265.39	33	gn	15	70.41
16	u	274	244.95	34	gs	13	65.83
17	m	258	226.28	35	oin	12	61.58
18	r	254	209.20	36	x	10	57.64
a = 82.4138			b = 74.8351		D = 0.96		

Legende zu den Tabellen:
n_x: beobachtete Häufigkeit der betreffenden Einheit;
NP_x: aufgrund des Modells berechnete Häufigkeit der betreffenden Einheit;
a, b, c: Parameter des Modells von Altmann; bei den Berechnungen wird $c = y_1$ gesetzt.
D: Determinationskoeffizient. Der Determinationskoeffizient ist akzeptabel, wenn $D \geq 0.80$, und sehr gut mit $D \geq 0.90$. In diesem Fall wurde also mit $D = 0.96$ ein sehr gutes Testergebnis erzielt.

[1] Die Lautbezeichnungen wurden von Faulmann (1887: 320) übernommen; seine Angaben konnten bisher nicht anhand des Textes von Dujardin überprüft werden, da beide Ausgaben noch nicht erreichbar waren.
[2] Das Zeichen sieht bei Faulmann wie ein <s> aus.



Graphik zu Tabelle 1: Anpassung des Modells für beliebige Rangordnungen an die von F. Dujardin ainé mitgeteilten Lauthäufigkeiten

2.2. XLIII. Josef Nowak (1848)

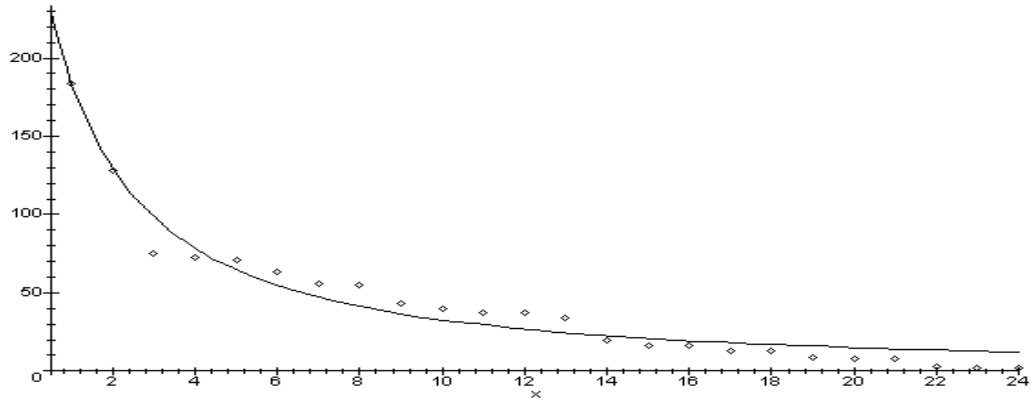
Eine frühe[3], wenig umfangreiche Zählung mit nur 1006 deutschen Buchstaben findet man bei Nowak[4] (1848: 20f.). Auch in diesem Fall lässt sich Altmanns Modell mit einem sehr guten Ergebnis anpassen.

Tabelle 2
Anpassung des Modells für beliebige Rangordnungen
an die von J. Nowak mitgeteilten deutschen Buchstabenhäufigkeiten

Rang	Lautbe-zeichnung	n_x	NP_x	Rang	Lautbe-zeichnung	n_x	NP_x
1	E	184	184.00	13	H	34	24.59
2	N	128	129.62	14	C	20	22.64
3	I	75	98.41	15	W	16	20.95
4	R	73	78.42	16	O	16	19.47
5	U	71	64.65	17	B	13	18.17
6	S	63	54.65	18	F	13	17.01
7	T	56	47.09	19	K	9	15.98
8	D	55	41.21	20	V	8	15.05
9	A	43	36.51	21	Z	8	14.22
10	L	40	32.69	22	P	3	13.46
11	G	37	29.52	23	X	2	12.77
12	M	37	26.85	24	Y	2	12.14
a = 2.3937			b = 1.0951		D = 0.96		

[3] Vor Nowak muss es mindestens eine Zählung zum Deutschen von Kerndörffer gegeben haben, deren Daten aber anscheinend nicht veröffentlicht wurden (Best 2008).
[4] Über Josef Nowak konnte bisher nur in Erfahrung gebracht werden, dass er österreichischer Militärarzt war (Schneider & Blauert 1936: 276).

Die Lautbezeichnungen sind im Original groß geschrieben; dies wurde übernommen. J und I hat Nowak zusammengefasst; er bemerkt zu J: „Es kann eigentlich nicht als Mitlaut betrachtet werden, da das i von einem andern Selbstlaut, mit einer Oeffnung des Mundes ausgesprochen, ihm seinen eigenthümlichen Laut geben muss.“ (Nowak 1948: 48)



Graphik zu Tabelle 2: Anpassung des Modells für beliebige Rangordnungen an die von J. Nowak mitgeteilten Buchstabenhäufigkeiten

3. Schlussbemerkung

Man sieht, dass auch diese frühen, nicht sehr umfangreichen Zählungen die Hypothese bestärken, dass Rangordnungen sich gesetzmäßig verhalten.

Abschließend sei angemerkt, dass es sich sicher lohnt, weiter im Bereich der Kryptologen und Stenographen zu suchen; es gibt Hinweise auf weitere Zählungen, die aber noch nicht verifiziert werden konnten.

Literatur

Altmann, Gabriel (1993). Phoneme Counts. In: Altmann, Gabriel (ed.), *Glottometrika 14* (S. 54-68). Trier: Wissenschaftlicher Verlag Trier.

Best, Karl-Heinz (2004/5). Laut- und Phonemhäufigkeiten im Deutschen. *Göttinger Beiträge zur Sprachwissenschaft 10/ 11, 21-32.*

Best, Karl-Heinz (2005). Buchstabenhäufigkeiten im Deutschen und Englischen. *Naukovyj Visnyk Černivec'koho Universytetu: Hermans'ka filolohija. Vypusk 231, 119-127.*

Best, Karl-Heinz (2006). Quantitative Untersuchungen zum Niederdeutschen und Niederländischen. *Göttinger Beiträge zur Sprachwissenschaft 13, 51-71.*

Best, Karl-Heinz (2008). Heinrich August Kerndörffer (1843-1928). *Glottometrics 18, 94-96.*

Dujardin aîné, François (1834). *Essai sur la sténographie et sur l'écriture en général.* Paris: Imprimerie de Everat. (Extrait du *Journal des connaissances usuelles et pratique, recueil des notions immédiatement utiles aux besoins et aux jouissances de toutes les classes de la société, et mises à la portée de toutes les intelligences.* Es konnte bisher nicht festgestellt werden, in welchem Band der Beitrag zuerst erschien.)

Faulmann, Karl (1887). *Historische Grammatik der Stenographie. Übersichtliche Darstellung der Systeme der Stenographie von der ältesten Zeit bis auf die Gegenwart auf Grundlage von Originalstudien.* Wien: Verlag von Bermann & Altmann.

Kaeding, Friedrich Wilhelm [Hrsg.] (1897/98). *Häufigkeitswörterbuch der deutschen Sprache. Festgestellt durch einen Arbeitsausschuß der deutschen Stenographie-Systeme. Erster Teil: Wort- und Silbenzählungen. Zweiter Teil: Buchstabenzählungen.* Steglitz bei Berlin: Selbstverlag des Herausgebers. Teilabdruck: *Beiheft zu Grundlagenstudien aus Kybernetik und Geisteswissenschaften. Bd. 4/ 1963.* (Anmerkung: Die Titel von Band 1 und Band 2 enthalten nur die Jahreszahl 1897; der Gesamttitel das Jahr 1898.)

Meier, Helmut (1967). *Deutsche Sprachstatistik.* Zweite erweiterte und verbesserte Aufl. Hildesheim: Olms.

Nowak, Josef (1848). *Leicht lesbare Geschwindschrift (Tachygraphie, Stenographie), oder: Ausführliche Anleitung zum Selbstunterrichte in der Kunst, so schnell zu schreiben, als ein öffentlicher Redner spricht. Für alle Stände.* Dritte, umgearbeitete Auflage. Wien: Sallmayer und Comp.

Schneider, L., Blauert, G. [Hrsg.] (1936). *Geschichte der deutschen Kurzschrift.* Wolfenbüttel: Heckners Verlag.