Glottometrics 28
2014

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.**

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. 28 (2014), Lüdenscheid: RAM-Verlag, 2014. 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 28 (2014) **ISSN 2625-8226**

Contents Glottometrics 28, 2014

Bibliography

Book Review

Glottometrics 28, 2014, 1-12

Some macro quantitative features of low-frequency word classes

Fan Fengxiang[1]*, Wang Yaqin, Gao Zhao*

Abstract. This contribution examines the macro quantitative features of 15 low- frequency word classes. The relationship between word frequency classes and the sizes of the frequency classes obeys Altmann's power law, and the sizes of low- frequency word classes increase along with the increase of text length. The relationship between text length and the sizes of low-frequency word classes also obeys Altmann's law. For text of the same length, the relationship between vocabulary size and the sizes of hapax legomena and dis legomena is linear, but this sort of relationship does not hold for other low-frequency word classes. The relationship between vocabulary/low-frequency word class ratio and text length can be captured with reparametrized Tuldava's model.

Key words: *low-frequency word class, Altmann's law, reparametrized Tuldava's model, corpus*

1. Introduction

Low-frequency words are of interest both to language researchers and language teachers. Perhaps the most well-known low-frequency word class is the so called hapax legomenon (hereafter referred to as hapax or hapaxes) because of its importance in language study, such as vocabulary richness and author identification (Holmes 1991), language typology (Popescu & Altmann 2008; Fan 2009), the degree of analytism (Popescu, Mačutek, & Altmann 2009), and so on. Hapaxes and other low-frequency words are of concern to the EFL (English as a Foreign Language) teacher because of their inter-textual distributional characteristics, i.e., they have zero or near-zero inter-textual repetition, thus causing great trouble both to the teacher and the learner, since the acquisition of a word needs 5—15 inter-textual exposures to the word (Nation & Waring 1997). However, there seems to be a lack of systematic study on the macro quantitative features of low-frequency words, which is the collective quantitative behavior of the words belonging to certain low-frequency classes on the frequency spectrum. Therefore, instead of studying the quantitative behavior of the individual low-frequency words, this paper attempts to investigate the macro quantitative behavior of low-frequency word classes. Specifically, the present contribution aims to (i) examine the relationship between low-frequency words and text length; (ii) examine the relationship between vocabulary size and low-frequency words; (iii) formulate mathematical models capturing these relationships.

The data source of this study is the British National Corpus (hereafter referred to as the

[1] School of Foreign Languages, Dalian Maritime University; Dalian, 116026, China. fanfengxiang@yahoo.com

BNC). It has 100 million words consisting of spoken English and written English. The latter has nine domains: *applied science*, *arts*, *belief*, *commerce*, *imaginative*, *leisure*, *natural science*, *social science* and *world affairs*. The vocabulary size of the BNC is 346,578 lemmas. Considering the size of the BNC and the number of the low frequencies in it, we study only the frequencies between 1 and 15.

2. The frequency spectrum of the BNC

If word frequencies are grouped into classes and arranged in the following form:

1	18
2	12
3	6
4	3
5	2
6	1
7	1
8	1
9	1

it is called a word frequency spectrum. The left column contains frequency classes, and the right column the number of different lemmas in the corresponding frequency class, which is also called frequency of frequencies or the size of the corresponding frequency classes. For example, *1 18* means there are 18 different lemmas that have a frequency of 1, or that the size of the frequency class is 18; while 9 *1* means there is one word with a frequency of *9*, i.e., the size of frequency class 9 is 1. The frequency class that consists of only one lemma is referred to as frequency hapaxes (Fan 2012), and like the lexical hapaxes, more than half of the frequency classes in a word frequency spectrum are frequency hapaxes.

All the frequencies of the BNC lemmas were turned into a frequency spectrum like the example just given, arranged in ascending order of frequency class, i.e., from low to high. There are 5,072 different frequency classes in the spectrum, averaging 68.33 different lemmas per frequency class, i.e., the mean size of the frequency classes is 68.33; of these frequency classes, 2,991 are frequency hapaxes. The size of a frequency class decreases as the frequency class rank increases. For example, the size of the first frequency class in the spectrum, i.e., the lowest frequency class composed of hapaxes, is 154,403 different lemmas, accounting for 44.55% of the entire vocabulary of the BNC. While the size of the 5072th frequency class is 1, containing only one word, *the*, whose frequency is 604,2931. The 1st frequency class to the 15th frequency class accounts for only 0.3% of the 5,072 different frequencies, but they contain 287,635 different lemmas, accounting for 83% of the entire vocabulary, averaging 19175.67 different lemmas per frequency class; while the remaining 5,057 frequency classes have only 58,943 different lemmas, averaging 11.66 different lemmas per frequency class. From the 2000th frequency class (whose frequency is 2164) to the 5072th frequency class (whose frequency is 6,042,931) in the frequency spectrum, the total number of different lemmas is 3,587, averaging only 1.17 lemmas per frequency class. The 287,635 different lemmas included in the lowest 15 frequencies represent 785,145 word tokens, account for only 0.79% of all the word tokens of the BNC, but the 58,943 lemmas account for 99.21% of the total. Table 1 displays the first 40 frequency classes and the last 40 frequency classes of the BNC's word frequency spectrum.

Table 1
The first 40 frequency classes and the last 40 frequency classes
of the BNC's word frequency spectrum
(*F Class*: frequency class, *Size*: number of different lemmas the corresponding frequency class has. *F Class* 242331—*F class* 6042931 are the 5033th—5072th frequency class in the BNC's frequency spectrum)

F Class	**Size**	**F Class**	**Size**		**F Class**	**Size**	**F Class**	**Size**
1	154403	21	1094		242331	1	536076	1
2	46125	22	1039		253426	1	613515	1
3	23149	23	1043		254039	1	639729	1
4	14656	24	929		261121	1	654115	1
5	10353	25	953		281340	1	658477	1
6	7747	26	866		303683	1	667770	1
7	6402	27	825		316491	1	734206	1
8	5037	28	780		319245	1	767650	1
9	4108	29	741		319637	1	870429	1
10	3636	30	693		350221	1	879640	1
11	2963	31	667		352724	1	1054846	1
12	2628	32	615		365374	1	1119375	1
13	2332	33	577		369325	1	1255602	1
14	2153	34	614		409334	1	1950796	1
15	1943	35	561		419661	1	2510050	1
16	1802	36	573		424871	1	2600184	1
17	1589	37	508		446062	1	2619912	1
18	1490	38	471		453540	1	3045314	1
19	1360	39	492		514918	1	3818487	1
20	1323	40	454		522071	1	6042931	1

Graphically, the relationship between the frequency classes and the sizes of the corresponding classes is L-shaped, with a steep drop from the start to the 939th frequency class, after which the curve remains practically level until the end. This is caused by the vast number of frequency hapaxes and other low-frequency word classes. This is very similar to the high percentage of low frequency words in the wordlist of a mega-corpus.

Altmann's (1980) power model $y = Ax^{-b}$ can describe the relationship between the frequency classes and the sizes of the frequency classes. The model fit is shown in Figure 1. The fit is excellent, with $R^2 = 1$, A =154066.635 and b = 1.699. The x axis represents the frequency classes and the y axis the sizes of the corresponding frequency class (in number of different lemmas). The solid line is the observed value and the dotted line the model fit.

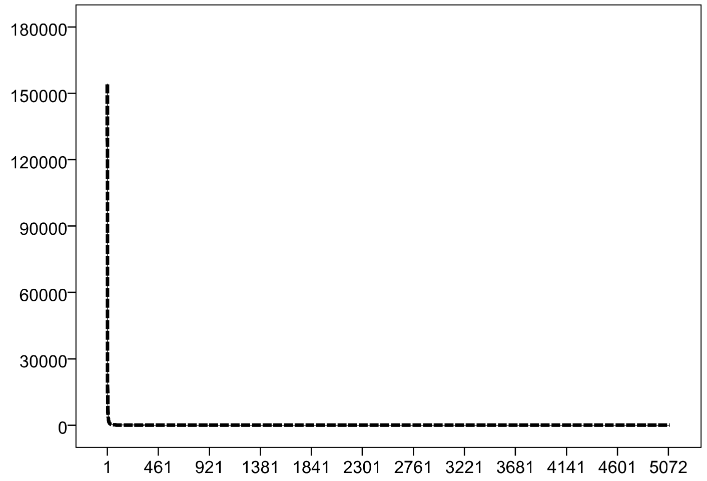


Figure 1. The model fit to the relationship between frequency classes and sizes of the corresponding frequencies.

3. The relationship between text length and low-frequency word class size

To study the relationship between text length and the number of words in each of the low frequency classes, the vocabulary growth and the increase of the number of words in the low frequency classes were computed along with the increase of text length at an interval of about 100,000 words. The result is shown in Figure 2.

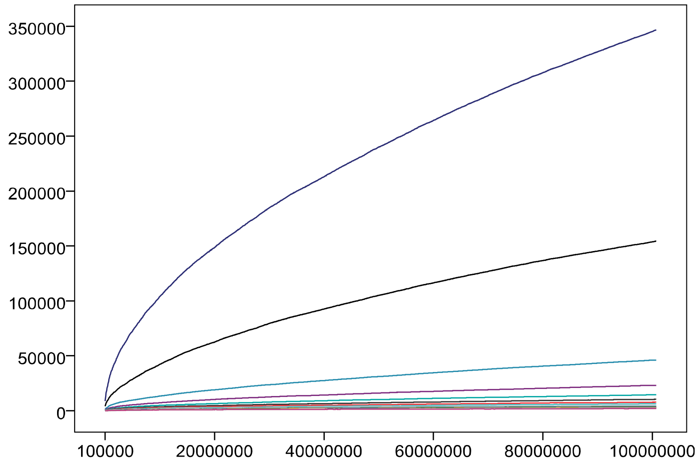
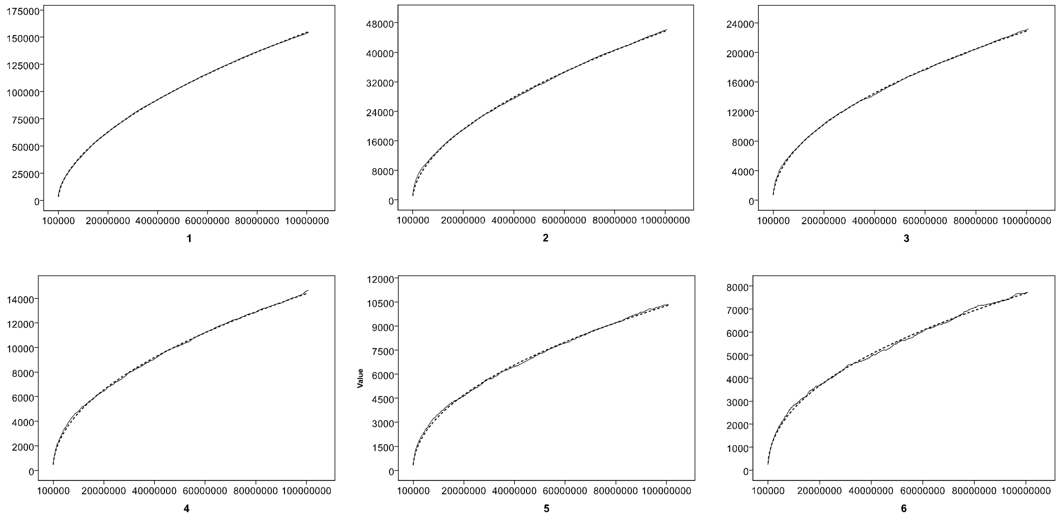


Figure 2. The relationship between text length, vocabulary size and low-frequency word class sizes. The curves from the top to bottom are vocabulary growth curve and the curves of sizes of frequency class 1—15

As shown in Figure 2, as text length increases, so does vocabulary size and the size of each of the 15 low-frequency word classes. In the low frequency word classes, the size of the preceding frequency class is larger than that of the following frequency class. For example, the number of different lemmas in frequency class 1 is larger than that in frequency class 2, as shown in Table 1. Again Altmann’s power model $y = Ax^{-b}$ can best describe the relationship between text length and the size of each of the 15 low frequency classes. The determination coefficients and the parameters are listed in Table 2. Figure 3 displays the curves of the fits and the observed values.

Table 2
The determination coefficients and the parameters of the power model fit
for low frequency word class 1—15

Frequency class	***R*^{2}**	***α***	***β***
1	1.000	5.322	−0.559
2	0.999	2.087	−0.543
3	1.000	2.315	−0.5
4	0.999	1.783	−0.489
5	0.999	1.357	−0.486
6	0.998	1.629	−0.46
7	0.997	0.901	−0.481
8	0.998	1.062	−0.459
9	0.999	1.047	−449
10	0.998	0.877	−0.451
11	0.998	1.118	−0.429
12	0.995	0.776	−0.442
13	0.995	1.063	−0.419
14	0.992	0.712	−0.435
15	0.995	0.676	−0.432



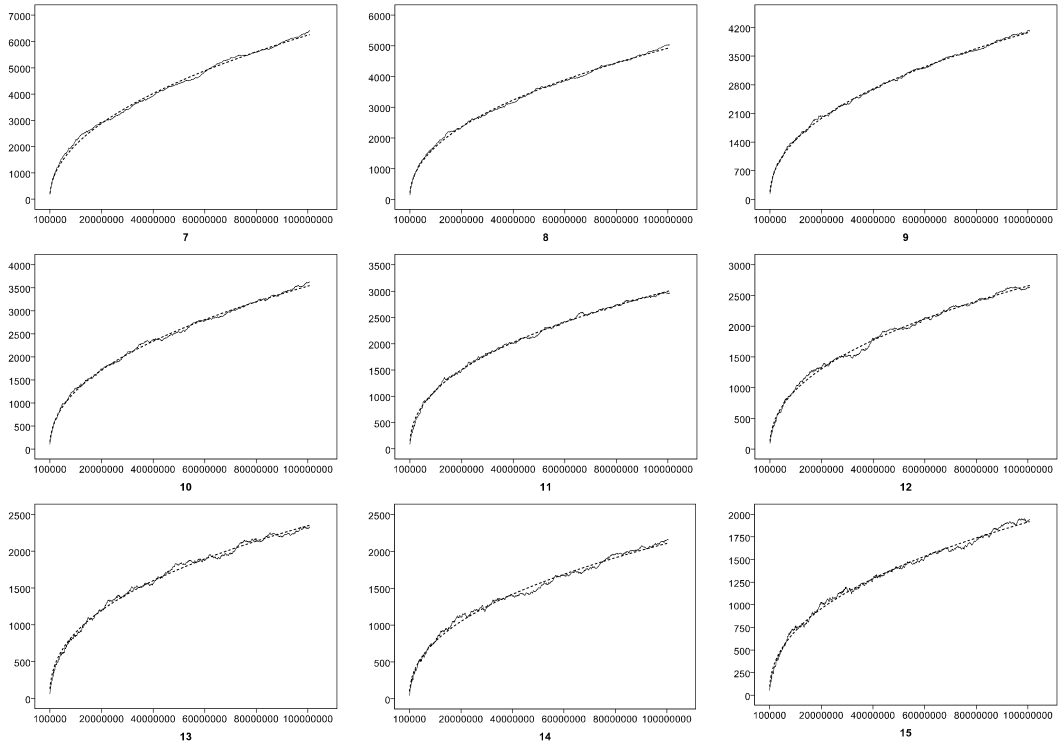


Figure 3. The power model fit for low frequency word class 1—15 (panel 1—15) and the observed values. The dotted lines are the model fit and the solid lines the observed values

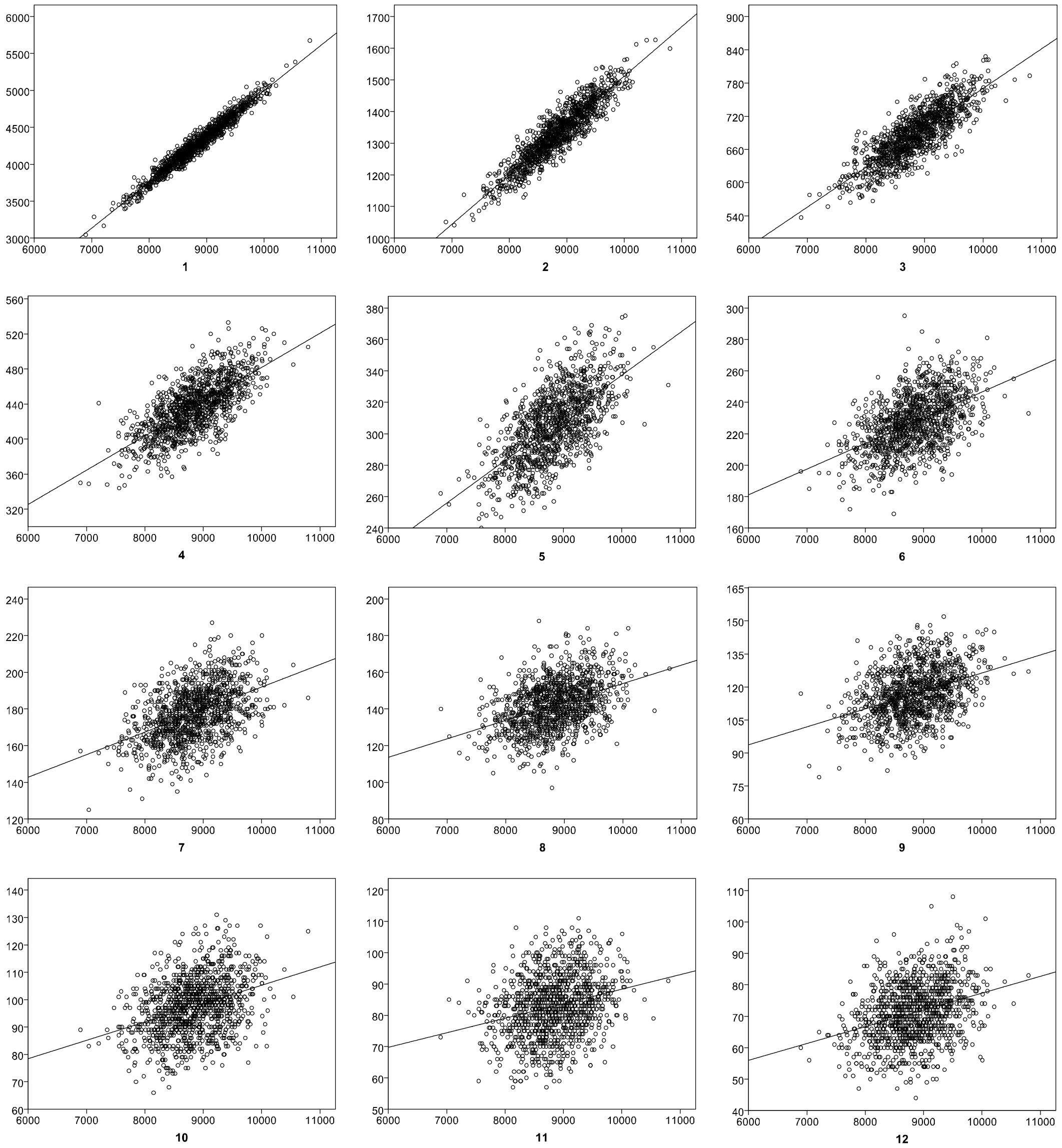
4. The relationship between vocabulary richness and the size of low-frequency word class

To examine the relationship between vocabulary richness and the size of low-frequency word class, the entire BNC was divided into 1,000 text chunks, each about 100,000 words in length. The vocabulary size and the sizes of the low-frequency classes of each of the text chunks were computed. The relationship between the vocabulary size and the sizes of frequency class 1—3 is basically linear and can be captured with a linear regression equation $y = \alpha + \beta x$; larger sizes of these low-frequency word classes generally indicate larger vocabulary size. However, this linear regression model does not fit well to the rest of the low-frequency word classes. Table 3 displays the determination coefficients and the parameters for the linear equation and Figure 4 shows the model fits and the observed values.

Table 3
The determination coefficients and the parameter values
for the linear regression model fit

Frequency class	R^2	α	β
1	0.955	-1206.418	0.62
2	0.843	-48.226	0.156
3	0.674	55.544	0.071

4	0.481	90.882	0.039
5	0.413	66.089	0.027
6	0.245	83.071	0.016
7	0.208	69.061	0.012
8	1.64	53.403	0.01
9	0.141	44.723	0.008
10	0.119	38.238	0.007
11	0.065	41.911	0.005
12	0.098	23.941	0.005
13	0.079	25.108	0.004
14	0.037	30.355	0.003
15	0.035	25.772	0.003



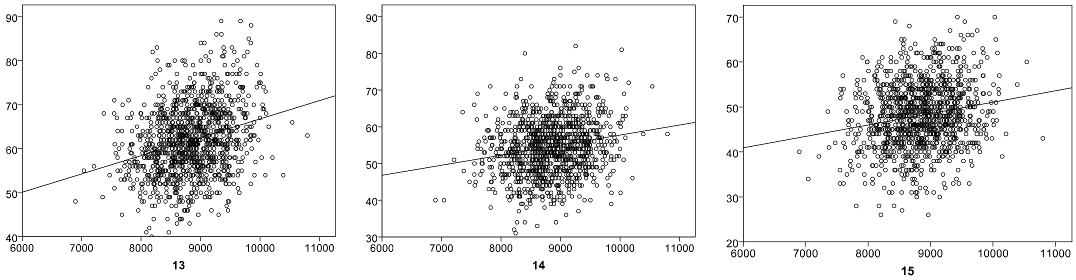


Figure 4. The relationship between vocabulary size and the sizes of low frequency classes. Panel 1—15 displays the linear regression equation fit to the observed values. The *x* axis is the size of a low frequency class and the *y* axis the size of vocabulary. The straight lines are the model fit and the dots the observed values

5. The ratio between vocabulary size and the sizes of the low-frequency classes

The ratio between vocabulary size and number of hapaxes has been studied by several linguists (Baayen 1996; Tweedie & Baayen 1998; Baayen 2001; Kornai 2002; Fan 2010). The general conclusion is that this ratio is about 0.4—0.5 in a text. Fan also examined the shape of the vocabulary/hapax curve and showed that it is U-shaped in the BNC; that is, as text length increases the ratio decreases until the text length reaches a certain value, whereupon the ratio starts to rise until the end. As to the ratio between vocabulary size and the sizes of other low-frequency word classes, Honoré (1979) regards the ratio between vocabulary size and the number of dis legomena (words occurring twice) as a constant. Apart from the foregoing, literature on the ratio between vocabulary size and the sizes of low-frequency word classes is far and few between. We examined such ratios and their relationship with text length and find that the vocabulary/hapax ratio curve is indeed U-shaped and the vocabulary/dis legomena ratio curve is similar in shape; but the rest of the ratio curves are generally decreasing as text length increases. The reparametrized Tuldava’s (1995) model can describe the relationship between text length and the ratios between vocabulary size and the sizes of low-frequency word classes. The original Tuldava’s model was for the description of the relationship between text length and vocabulary size. It is shown below:

$$V = Ne^{-\alpha\,(\ln N)^{\beta}} \qquad (1)$$

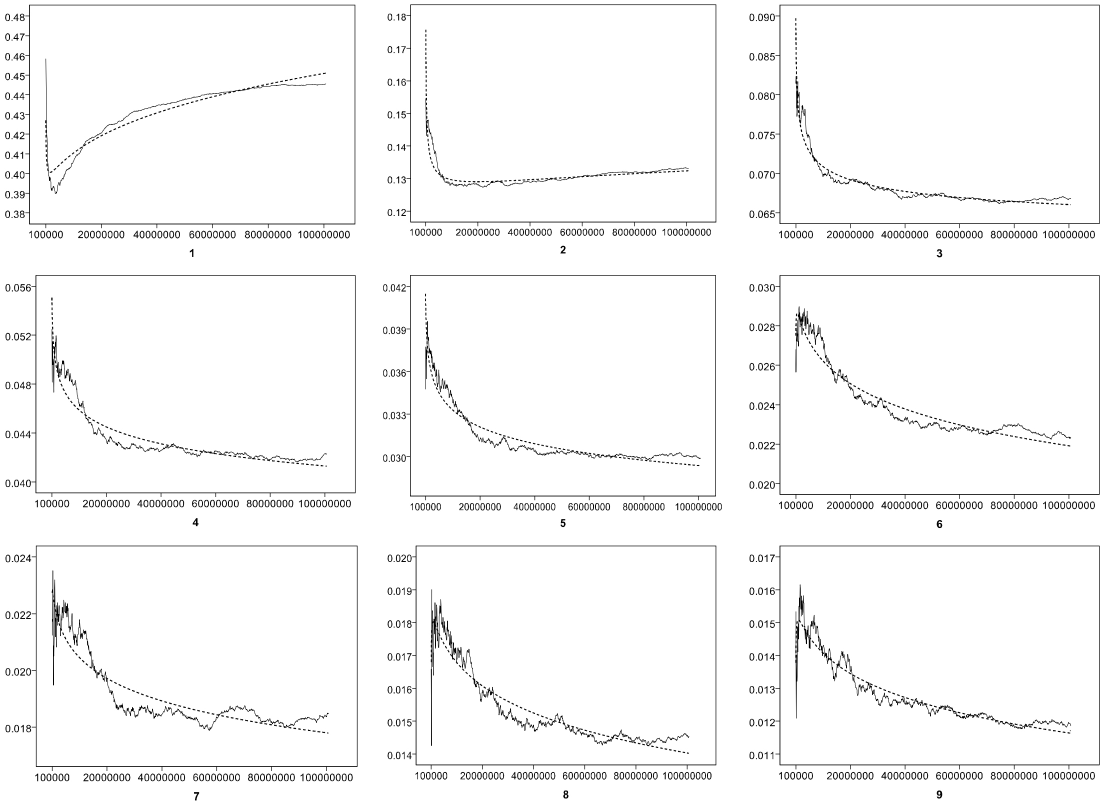
(1) was adjusted by multiplying it with a parameter *γ* to fit the ratio data:

$$V = \gamma Ne^{-\alpha\,(\ln N)^{\beta}} \qquad (2)$$

(2) can capture the relationships between text length and the vocabulary/low-frequency word class ratios. Table 4 is the determination coefficients and the values of the parameters, and Figure 5 displays the model fit to the observed values.

Table 4
The determination coefficients and the values of the parameters
of the reparametrized Tuldava’s model

Frequency class	R^2	α	β	γ
1	0.927	2.352	0.775	25.723
2	0.7067	3.416	0.694	210.382
3	0.916	1.405	0.920	0.540
4	0.876	0.9397	1.027	0.058
5	0.886	0.941	1.030	0.047
6	0.913	0.396	1.268	0.002
7	0.804	0.600	1.151	0.005
8	0.892	0.342	1.309	0.001
9	0.914	0.299	1.347	0.0004
10	0.895	0.286	1.359	0.0003
11	0.927	0.144	1.558	0.00006
12	0.822	0.204	1.455	0.0001
13	0.928	0.135	1.575	0.00004
14	0.845	0.105	1.649	0.00002
15	0.860	0.129	1.611	0.00003



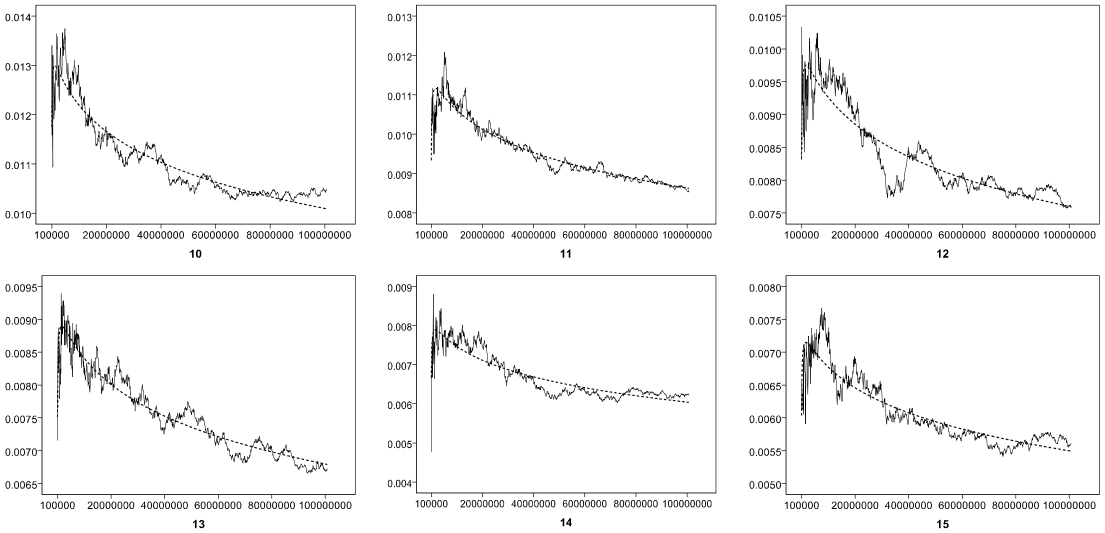


Figure 5. The reparametrized Tuldava's model fit to the ratio between the vocabulary size and the sizes of the 15 low-frequency word classes. The *x* axis is text length and the *y* axis the ratio. The solid lines are the observed values and the dotted lines the model fit

Generally the model fit is good, with 14 of the determination coefficients being higher than 0.8 and only one being 0.7067. According to Fan (2010), the U-shaped vocabulary/hapax ratio curve is caused by the reduction of early hapaxes and the accumulation of late hapaxes which have extremely low probability of occurrence. The pattern of the vocabulary/dis legomena ratio curve is fall-rise like that of the vocabulary/hapax ratio curve, only its rising section is not so steep. The cause of this sort of pattern is the same as that of the vocabulary/hapax ratio curve, i.e., the reduction of the early dis legomena and the accumulation of the late dis legomena. The rest of the ratio curves have a rise-fall pattern, as shown in Panel 3—15 of Figure 5, and generally the higher the frequency rank, the longer the rising section. For example, the rising section of the vocabulary/low-frequency word class 3 ratio curve is barely noticeable compared with other higher-ranking ratio curves. The mechanism behind this phenomenon needs further examination.

6. Summary and conclusion

The present study reveals some major macro quantitative lexical features of low-frequency word classes. The relationship between the frequency classes and the sizes of the frequency classes obeys Altmann's power law, and the sizes of low-frequency word classes increase along with the increase of text length, and this relationship can also be captured with Altmann's power model; in addition, the size of the preceding frequency class is larger than that of the following frequency class. There is a linear relationship between vocabulary richness and the sizes of frequency class 1—3. The ratios between vocabulary and the sizes of the low-frequency word classes are regular and can be captured with the reparametrized Tuldava’s model. The results prove that the lexicon is a complex and self-regulating system (Köhler 1990; Köhler and Altmann1993). If the sizes of high-frequency word classes are as large as those of the low-frequency word classes, i.e., instead of having only one word with a frequency of 604,2931, there are 154,403 different words with this frequency, 46,125 different words with a frequency of 3,818,487, 23,149 different words with a frequency of 3,045,314 and so on, human brains would be out of memory for such a vast stock of commonly used words, and a language with such a lexical system would be chaotic. The macro quantitative lexical features revealed in this study possibly hold in other human languages.

References

Altmann, G. (1980). Prolegomena to Menzerath’s law. In: Grotjahn, R (ed.), *Glottometrika 2: 1-10*. Bochum: Brockmeyer.

Baayen, H. (1996). The effects of lexical specialization on the growth curve of the vocabulary. *Computational Linguistics 22(4), 455–480.*

Baayen, H. (2001). *Word Frequency Distributions*. Dordrecht: Kluwer Academic Publishers.

Fan, F. (2009). Hapax Legomena and language typology, a case study In: Emmerich Kelih, Viktor Levickij, Gabriel Altmann (eds.), *Methods of Text Analysis: Omnibus volume: 63-84.* Chernivtsi: ČNU.

Fan, F. (2010). An asymptotic model for the English hapax/vocabulary ratio. *Computational Linguistics 36(4), 631-637.*

Fan F. (2012). A study on word frequency spectra. In: N. Naumann, P. Grzybek, R. Vulanović, G. Altmann (eds.). *Synergetic Linguistics, Text and Language as Dynamic Systems: 39-46.* Wien: Praesens.

Holmes, D. (1991). Vocabulary richness and the Prophetic Voice. *Literary & Linguistic Computing 6, 259-68.*

Honoré, A. (1979). Some simple measures of richness of vocabulary. *Association of Literary and Linguistic Computing Bullitin 172-179.*

Kennedy, G. (1998). *An Introduction to Corpus Linguistics.* London: Addison Wesley.

Kornai, A. (2002). How many words are there? *Glottometrics 4, 61–86.*

Köhler, R. (1990). Elemente der synergetischen linguistik In: Hermel, R. (ed.). *Glottometrika* 12: 179-187. Bochum: Brockermeyer.

Köhler, R., Altmann, G. (1993). Begriffsdynamik und Lexikonstruktur. In: Beckmann, F., Heyer, G. (eds.). *Theory und Praxis des Lexikon: 173-190.* Berlin, New York: de Gruyter.

Nation, P., Waring, R. (1997). Vocabulary size, text coverage and word lists. In: N. Schmitt, McCarthy, M. (eds.), *Vocabulary: Description, Acquisition and Pedagogy: 6—19.* Cambridge: Cambridge University Press.

Popescu, I.-I. Altmann, G. (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.

Tuldava, J. (1995). *Methods in quantitative linguistics*. Trier: WVT.

Tweedie, J., Baayen, H. (1998). How variable may a constant be? Measure of lexical richness in perspective. *Computer and the Humanities 32, 323-52*.

Glottometrics 28, 2014, 13-36

Clause centrality

Ioan-Iovitz Popescu
Gabriel Altmann

Abstract. The aim of this article is to analyze clause centrality and perform some elementary comparisons both within the same text sort and between two different text sorts. Various indicators and their relations are scrutinized.

Keywords: *clause, sentence, centrality, indicators, Slovak*

Clause centrality can be defined in various ways. The most common way is considering the finite verb as the center of the clause and characterize centrality taking into account the number of words preceding and following this verb. Sometimes, the verb may be absent: frequently it is the case in languages having copula and omitting it, e.g. in Hungarian, Indonesian, Russian, etc. but there are also sentences with elliptic verbs. In that case, one may, but need not, take this fact into account. Nevertheless, ellipses may be restored. In some other cases, there is merely an infinitive in the clause, a fact met frequently in the poetry; in such cases it can be considered a bearer of a finite form. Frequently, an auxiliary verb has a finite form and the main verb is given in infinitive. In that case one can decide according to the aspect of investigation and the role of auxiliaries in the given language, or mechanically, considering the finite form as the relevant one. In some languages, there are reflexive verbs whose reflexive part is another verb. In that case one can count both of them or merely the main one – here the latter case is accepted. Occasionally, different other circumstances may appear: they must be solved by setting up some criteria which have the power of boundary conditions. It is recommended to consider names consisting of several words as one word; similarly, dates. A good analysis is given by Köhler, Naumann (2008)

A quite different problem is the measurement of *sentence centrality*. Here the same formula (see below) may be used and the center of the sentence may be the main noun or another entity considered as topic which is commented by the rest of the sentence. A part of the comment precedes the topic, another part follows it. If the topic consists of several words, they are considered as one unit. The topic-comment problem exists since many years, it is very complex, and had been analyzed rather qualitatively.

The perspectives of this procedure are manifold: (1) One can characterize a text using the expression of centrality; (2) one can compare texts of the same text-sort in order to state whether text-sort exerts influence on clause/sentence formation; (3) one can compare the same text or the same text-sort in different languages in order to state whether there are typological differences; (4) one can study the development of a writer using his texts written in different years; the same can be done for the study of the given text-sort or language. All these questions lead to inductive procedures which must be performed before one begins to set up hypotheses.

The results of the investigation may be used for text-typological purposes, for the study of text-sorts, for the comparison of languages, i.e. for language typology, and for the study of the development of child language or of the stylistic changes in the texts of a unique author.

Since centrality is not sufficiently scrutinized, one can make conjectures or ask questions which can be later on formulated as hypotheses. Here we shall simply mention some directions of possible investigation. First, *clause* centrality is stated using the sequence

$a_f, a_{f\text{-}1}, \ldots, a_1 V a_1, a_2, \ldots, a_b$

where V is the (finite) verb and a_i are the individual words. If the words in front of and behind the verb are numerated, then f is the greatest index of words in front of the verb and b is the number of words behind it. The centrality is defined as (cf. Altmann, Lehfeldt 1973; Wimmer et al. 2003: 178)

$$C = 1 - \frac{|b - f| - \delta}{b + f} \qquad (1)$$

where b and f are the greatest indices,

$$\delta = \begin{cases} 0 \;\; if \;\; (b+f) \; is \; an \; even \; number \\ 1 \;\; if \;\; (b+f) \; is \; an \; odd \;\; number \end{cases}$$

and

C = 1 if $b + f = 0$.

This indicator varies in the interval [0; 1], 1 meaning maximal centrality. Computing it for all clauses of the text, we obtain a sequence of numbers which has special properties. We may ask (1) what is the average of these values? (2) What is the distribution of these values? Can the distribution, presented either in discrete or in continuous form, be substantiated linguistically? (3) What are the properties of the sequence? Is it regular? What are the properties of the oscillation? (4) What is the distribution of distances between equal centralities? (5) What are the factors – linguistic or stylistic – linked with at least one of the above properties? (6) Can one set up a control cycle including some of the above properties of centrality? (7) Can this cycle or at least one of the above properties be incorporated in Köhler’s (2005) synergetic control cycle?

Let us consider at least some of these problems. To this end we first analyze 20 texts by S. Svoraková (see Appendix) written in the same text-sort and style in Slovak. They concern cultural affairs like expositions, reviews of books about painting, articles about the problems of modern painting, etc. They have the advantage of being short and not controlled by a special form (like e.g. poetry).

The results of computing formula (1) are presented in Table 1. Here, the mean is the average of the computed values, SSQ is the sum of squared deviations from the mean.

Table 1
Clause centrality values in 20 texts by S. Svoráková

Text	Values	C	Mean	SSQ
T1	[0.75, 0.08, 0.33, 1, 0.69, 0.54, 0.26, 0.47, 0.67, 0.50, 0.39, 1, 0.67, 0.33, 0.56, 0.54, 0.94, 0.29, 0.60, 0, 0.73, 0.60, 0.60, 0.71, 0.50, 1, 1, 1, 1, 0.75, 0.5, 0.38, 0.11, 1, 1, 0.78, 0.69, 1, 0.5, 0.23,	67	0.5795	8.0875

	1, 0, 1, 0.75, 0, 0.87, 1, 0.71, 0.5, 0, 0.58, 0, 0.6, 1, 0.2, 1, 0, 0, 0.88, 1, 1, 0.43, 1, 0, 0.60, 0, 0]			
T2	[0.73, 0.67, 1, 0.77, 0, 0.87, 0.14, 0.63, 0.28, 0, 0.33, 0, 0.50, 1, 0, 0.25, 0.33, 0.5, 1, 1, 0.6, 0.5, 1, 0.55, 1, 0.2, 0.54, 0.3, 0.33, 0.56, 0, 0.43, 0.25, 0.82, 1, 1, 1, 0.67, 0.33, 0.67, 0.14, 0.45, 0.67, 0, 0, 0.33, 0.63, 0.71, 0.24, 0.64, 0.76, 0, 0.67, 0, 1, 0.86, 0.82, 0.5, 1, 0.67, 1, 1, 0.27, 1, 0.60, 0, 1, 0.33, 1, 0.5, 0.09, 0.25, 0.43, 0.70, 0.5, 1, 0.27, 0.20, 1, 0.27, 0.20, 0, 1, 0, 1, 0, 0.38, 0.33, 0.5, 0, 0.5, 0.14, 0.25, 0.20, 0.25, 0, 0.33, 1, 0.25, 0.5, 0.38, 0.75, 0.21]	103	0.4992	12.0287
T3	[0.54, 0.91, 0.56, 1, 0.33, 0.90, 0, 0.25, 0.75, 0.69, 0.68, 1, 0, 0.71, 0.75, 0.5, 0.56, 0.25, 1, 0.25, 0.33, 0.07, 0.56, 0.33, 1, 0.54, 1, 0.67, 1, 0.67, 1, 0, 0.82, 0.40, 0.72, 0.40, 0.72, 0.40, 0.15, 1, 0.37, 0.11, 0.5, 0.80, 1, 0, 0.60, 0.60, 0, 1, 0.38, 0.89, 0, 0.33, 0.89, 0, 0, 1, 0.38, 0.75, 0.75, 0.80, 0.33, 0.43, 0.80, 0.33, 0.43, 0.80, 0.86, 0.45, 1, 0.43, 1, 0.88, 0.22, 0, 1, 1, 1, 1, 0]	81	0.5743	9.2142
T4	[0.69, 0.14, 0.20, 0.75, 0.56, 0.69, 0.29, 0.41, 1.0, 0.80, 0.47, 0.73, 0.71, 0.85, 0.60, 0.38, 0.43, 0.5, 0.0, 0.82, 1.0, 0.33, 0.33, 1.0, 0.5, 1.0, 0.79, 0.57, 1.0, 0.5, 0.60, 0.18, 0.43, 0.38, 0.45, 0.26, 0.42, 0.33, 0.60, 0.33, 0.71, 0.78, 0.91, 0.56, 0.86, 0.88, 0.17]	47	0.5721	3.1976
T5	[0.45, 0.30, 0.76, 0.50, 0, 0.56, 0.25, 0.86, 0.54, 1, 0, 1, 0.45, 0, 0, 0.92, 0.25, 1, 1, 0.71, 0.29, 0, 0, 0.69, 0, 0.79, 0.50, 0.67, 0.87, 0.67, 0, 0.33, 0.75, 0.71, 0.82, 0.54, 0.29, 0.67, 0, 1, 0, 0.20, 0.50, 1, 0, 1, 1, 0.71, 0.80, 1, 0]	51	0.5167	6.7243
T6	[0.71, 0.60, 0.78, 0.50, 0.29, 0.71, 0.57, 1, 0.30, 0.14, 0.78, 1, 0.60, 0.75, 0.76, 0.52, 0, 0.89, 0.35, 0.43, 0.58, 1, 0.14, 0.79, 0, 1, 0.50, 0.68, 0.40, 0.39, 1, 0.54, 0.07, 1, 0.60, 0.50, 0.20, 0.22, 0.33, 0, 0.60, 0.22, 1, 0.25, 0.74, 0.20, 1, 0.63, 0.09]	49	0.5378	4.6266
T7	[0.88, 0, 0.73, 0.17, 0, 0.52, 0.71, 1, 0.78, 0, 0.50, 0, 0.33, 1, 0.83, 1, 0.67, 1, 0.33, 0.83, 0.67, 0.60, 0.50, 0.65, 0.33, 1, 0.45, 0.47, 1, 0.50, 0.66, 0.56, 0.37, 0.64, 0.78, 0.33, 0.50, 0.64, 1, 0.25, 0.27, 0.67, 0.87, 0.50, 0.67, 0.20, 0.75, 0.67, 1]	49	0.5873	4.1375
T8	[1, 0.60, 0.25, 0.85, 0.75, 1, 0.45, 0.60, 0.75, 0, 1, 1, 0.91, 0, 0, 1, 0.60, 0.20, 0.50, 1, 0.23, 0.50, 0.25, 0.36, 0, 0.25, 0, 0.71, 0.69, 1, 0.86, 0.50, 0.25, 0.38, 1, 0.17, 1, 1, 0.14, 0, 0.91, 0.71, 1, 1, 1, 0.73, 0.25, 0.50, 0, 0, 0.23, 1, 1, 0.78, 1, 1, 1, 0.75, 0.71, 1, 0.50, 0, 0.82, 0.50, 0.80, 0.57, 0.48, 0.56, 0.80, 1, 1, 0.73, 1, 0.71, 0.43, 0.71, 1, 0.67, 1, 0.75, 0.71, 1, 0.50, 1, 1, 1, 0.33, 0.60, 1, 0.43, 0, 1, 1, 0.33, 0.60, 0, 0.78, 1, 1, 1, 0.56, 0.45, 0.33, 0, 1]	105	0.6378	12.4536

T9	[1, 0.65, 0.67, 0.50, 0.58, 0.80, 1, 0.67, 0.88, 0.33, 0.88, 1, 1, 1, 0.45, 1, 0.40, 1, 0.60, 0.33, 1, 1, 0.20, 0.20, 0.40, 0.13, 1, 1, 0.17, 1, 0.56, 0.56, 0.38, 0.67, 0.67, 0, 1, 1, 0, 1, 0.50, 1, 0.33, 1, 1, 0.20, 0.20, 0.71, 0.56, 0.68, 0.75, 0.25, 1, 1, 0, 0.50, 0.33, 0.50, 0.43, 0.60, 0.33, 1, 0.18, 0.67, 0.33, 0.33, 0.56, 0.71, 0.11, 0.50, 0.67, 0.14, 1, 0, 1, 1, 1, 1, 0, 0.63]	80	0.6173	8.7622
T10	[0.53, 0, 0.73, 0.80, 0.22, 0.78, 0.78, 1, 0.80, 0, 0.29, 0.50, 1, 0.54, 1, 0.27, 0.67, 0.56, 1, 0.75, 0, 0.29, 1, 0.20, 0.27, 1, 0.78, 1, 0.50, 0.50, 1, 1, 0.75, 1, 0.11, 0.45, 0.83, 0.69, 0.60, 0.80, 1, 0.71, 1, 1, 0, 0.20, 0.73, 0.67, 1, 0.50, 0.43, 1, 1, 0.75, 1, 0.38, 0.24, 0.80, 1]	59	0.6508	6.0028
T11	[0.5, 0.33, 1, 0.64, 1, 0.67, 0.80, 0.20, 1, 0.33, 0, 1, 0.33, 0.71, 1, 1, 0, 0.33, 0.33, 0.69, 1, 1, 0.67, 0.67, 0.33, 0.78, 0.33, 0.33, 1, 0.14, 1, 0.45, 0.80, 0.56, 0.5, 0.73, 0. 0.5, 1, 0.60, 0.14, 0, 0.20, 0.33, 0.20, 0.60, 1, 1, 0.60, 0.83, 0.5, 1]	52	0.5894	5.5163
T12	[0.67, 1, 0.05, 0.17, 0.27, 0.81, 1, 0.57, 0.43, 0, 1, 1, 1, 0.43, 1, 0.33, 1, 0.6, 0.25, 1, 0.33, 1, 1, 0.33, 0.33,0.67, 1, 0.29, 0.78, 0.43, 0.07, 1, 1, 0.33, 0.69, 0.40]	36	0.6175	4.1113
T13	[0.80, 0.47, 0.67, 0.78, 0.80, 0.67, 0.67, 0.60, 0.33, 1, 0.78, 0.71, 0.20, 0.71, 0.71, 0.75, 0.75, 0.43, 0.5, 1, 0.33, 0.8, 0.43, 1, 0.09, 0.45, 0.54, 1, 0.26, 0.17, 0.5, 1, 0, 1, 1, 0.67, 0.64, 0.67, 0.33, 0.60, 0.6, 0.71, 1, 0.25, 0.6, 0, 0.67, 0.75, 0.75, 0.6, 0.5, 0.33, 0.33, 1, 0.75, 0.25, 1, 1, 0.75, 0.47, 0.71, 1, 0.75, 1, 0.46, 0.56, 1, 1, 1, 1, 0.33, 1, 0.64, 0.67, 0.78, 0.60, 0.33, 0.71, 1, 0.60]	80	0.6533	5.6628
T14	[0.89, 0.16, 1, 0.80, 1, 0.83, 0.88, 0.50, 0.33, 0.67, 1, 0.17, 0.67, 0.75, 1, 0.33, 0.63, 1, 1, 1, 0.43]	21	0.7162	1.6883
T15	[0.78, 0.71, 0.33, 1, 0.6, 0.25, 0.17, 0.69, 0.67, 0.33, 0.71, 0.67, 0.33, 0.71, 0.67, 1, 1, 0.85, 0, 0.48, 1, 0.40, 0.23, 0.73, 0, 0.45, 0.83, 0, 0.38, 0.89, 0.83, 0.33, 0.78, 0.24, 1]	35	0.5726	3.2377
T16	[0.45, 0.33, 1, 1, 1, 0.71, 1, 0.67, 0.67, 0.71, 0.86, 0.88, 0.11, 1, 0, 0.64, 1, 0.33]	18	0.6867	1.7748
T17	[0.08, 1, 1, 1, 1, 1, 1, 0.79, 0.25, 0.67, 1, 0.64, 0.2, 0.75, 0.69, 0.29, 0.22, 0.2, 0.67, 0.67, 0.60, 0.87, 0.33, 1, 1, 0.67, 0.38, 0.50, 0.75, 0.71, 0.43, 0.14, 1, 0.25, 0.43]	35	0.6337	3.2322
T18	[0.5, 1, 0.5, 0.25, 1, 1, 0.3, 0, 0.43, 1, 0.78, 0.69, 1, 0.4, 0.33, 1, 0.83, 1, 0.82, 0.5, 0.23, 0.20, 0.6, 1, 0.05, 0.33, 0.25, 0.5, 0.5, 0.25, 0, 1, 0.5, 1, 0.43, 1, 0.14, 0.2, 1, 1, 0.4, 0, 0, 0.33, 0, 1, 1]	47	0.5583	6.1051
T19	[0.69, 0.69, 0.78, 0.31, 0.10, 0.11, 0.5, 0, 0.5, 0, 0.5, 0.67, 1, 1, 0.82, 0.22, 0.40, 0.50, 0.43, 0.78, 0,	166	0.6067	13.5650

	0.78, 0.80, 1, 0.5, 0.60, 0.80, 0.33, 0.89, 1, 0.86, 0.50, 0.33, 0.56, 0.36, 0.38, 1, 0.65, 0.33, 0.89, 0.5, 0.56, 0.45, 1, 0.5, 1, 0.44, 1, 1, 0.29, 0.73, 0.83, 0.71, 0.6, 0.78, 1, 1, 0.2, 1, 1, 0, 0.67, 1, 0.33, 0.2, 0.75, 0, 0.52, 0.89, 0.80, 0.27, 0.67, 0.83, 0.78, 0.50, 0.43, 0.82, 1, 0.50, 1, 0.5, 0.91, 0.75, 0.73, 0.54, 0.43, 0.5, 0.87, 0.69, 0.5, 0.82, 0.71, 1, 0.56, 0.29, 0.14, 0.33, 0.50, 0.40, 1, 0.5, 0.75, 0.18, 0.20, 0.53, 0.43, 0.20, 0.67, 0.50, 0.65, 0.57, 0.07, 0.5, 0.71, 0.40, 0.43, 0.33, 0.33, 0.53, 1, 0.53, 0.60, 0.71, 0.6, 0.86, 0.79, 0.6, 0.25, 1, 0.78, 0.86, 1, 0.52, 1, 0.5, 0.78, 1, 0.08, 1, 0.67, 0, 0.67, .71, 0.07, 1, 0.86, 0.43, 1, 0.67, 1, 0.75, 0.08, 0.47, 0, 1, 0.5, 0.6, 0.6, 0.57, 0.82, 1, 0.45, 1, 0.27, 0.56, 0.75]			
T20	[0.6, 0.82, 0.78, 0.6, 0.25, 0, 0.30, 0.23, 1,0.85, 0.33, 0.75, 1, 1, 0.40, 0.64, 1, 0.56, 0.80, 0.38, 0.45, 0.6, 0.56, 1, 1, 0.43, 0.43, 0.43, 0.75, 0.67, 0, 0.33, 0.33, 1, 0.40, 0, 0.67, 0.25, 0, 0.33, 0.64, 0.67, 0, 0.67, 1, 1, 1, 0.67, 0.25, 0.75, 1, 0.75, 0.78, 1, 1, 0.06, 1, 0.75, 0.71, 0.08, 0, 0.33, 0, 0.76, 1, 0.75, 0.56, 0, 0.60, 0, 0.50, 0.80, 0.56, 0.88, 1, 0.43, 0.25, 0.25, 0.43, 0.14, 0.5, 1, 0.75, 0.40, 0.67, 0.43, 0.50, 0.38, 0.20, 0.33, 0, 0.33, 1, 0.71, 1, 0.71, 1, 0.5, 0.8, 0.14, 0.40, 0.45, 0.33, 0.43, 0.5, 1, 0.5, 0, 0.82, 0.47]	110	0.5563	10.9394

Of course, one can also test the averages, here we restrict ourselves to the test of two extreme averages, namely with Text 2 and Text 14. Using the asymptotic normal criterion (obtained from the t-distribution because our sizes are large)

$$t = \frac{|\, m'_{1,1} - m'_{1,2} \,|}{\sqrt{\dfrac{SSQ_1 + SSQ_2}{n_1 + n_2 - 2}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}\,, \tag{2}$$

we obtain inserting the values from Table 1

$$t = \frac{|\, 0.4992 - 0.7162 \,|}{\sqrt{\dfrac{12.0287 + 1.6883}{103 + 21 - 2}\left(\dfrac{1}{103} + \dfrac{1}{21}\right)}} = 0.08,$$

which is not significant. However, in spite of very small differences the centralities may significantly differ from one another. The mean clause centralities in texts by Svoráková vary in a very small interval which may characterize either her personal style, the property of the given text sort or of the language.

Normalizing the mean. All mean centralities can be transformed into standardized normal variables. This can be done by subtracting the expectation (i.e. 0.5) from the computed mean and divided by the standard deviation of the mean which can be obtained as the square root of *SSQ/n*^{2}. For example, for Text 1 we obtain

$$u = \frac{0.5795 - 0.5}{\sqrt{\frac{8.0875}{67^2}}} = 1.87$$

The standard normal values of centralities are presented in Table 2. The results show the significance of a trend, e.g. in T3 the value u = 1.98 displays the significant tendency to centrality while T5 with *u* = 0.33 does not display any tendency. The majority of texts prefer centrality of the clause.

Table 2
Standard normal N(0;1) values of mean centralities with S. Svoráková

T1	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**	**T8**	**T9**	**T10**
1.87	-0.02	1.98	1.90	0.33	0.86	2.10	4.10	3.17	3.63
T11	**T12**	**T13**	**T14**	**T15**	**T16**	**T17**	**T18**	**T19**	**T20**
1.98	2.09	5.15	3.49	1.41	2.52	2.60	1.11	2.60	1.87

The normalized values vary in a relatively large interval <-0.02, 5.15> but except for one case (T2) all u-values are positive. Those greater than 1.64 symbolize significant centrality.

Before comparing the results with other texts, we look at some properties of centrality.

Distribution of centrality

Texts with small sizes do not display a monotonous distributional regularity. Even if the values are pooled in classes, e.g. in 0-1.0, 0.11-2.0,…,0.91-1, they display a waveform distribution. But all of them, if the frequencies in individual classes are ranked, display a very regular Zipf-Mandelbrot distribution. That means, there are specific preferences for constructing clauses. If the text size is sufficiently large, one can compare the texts using a simple chi-square test for homogeneity.

In the sequel, we shall analyze the 20 texts presented in Table 1. The values, pooled in classes 0 to 0.1 as 1; 0.11 to 0.2 as 2; …0.91 to 1 as 10 are presented in Table 3. As can be seen, no text displays a clear tendency, all distributions oscillate in different ways. The only stochastic regularity is perhaps the tendency of the clause to display an extreme value, i.e. strong centrality or strong non-centrality, and somewhat stronger mean value. Looking at the average values in Table 1 we see that all texts tend to the mid of the interval <0, 1>. The question whether all texts are homogeneous in this sense can be tested by means of a chi-square test. We have a 10x20 contingency table and obtain a value of X^2 = 250.4433 which is with 9(19) = 171 degrees of freedom highly significant, since P = 7.16337E-05.

Table 3
Frequencies of centralities pooled in classes

Class	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**	**T8**	**T9**	**T10**
1	11	16	11	1	12	5	4	12	5	4
2	2	7	2	4	2	5	2	3	9	3
3	3	13	4	2	4	5	2	7	1	6
4	4	11	12	9	4	6	5	5	10	1
5	7	11	6	9	1	5	7	12	7	6
6	9	5	6	3	3	4	3	8	7	4
7	4	10	4	3	4	2	9	2	9	3
8	7	5	11	6	6	8	5	16	4	13
9	2	4	6	4	2	1	4	3	2	1
10	18	21	18	6	10	8	8	37	26	18

Class	**T11**	**T12**	**T13**	**T14**	**T15**	**T16**	**T17**	**T18**	**T19**	**T20**
1	4	3	3	0	3	1	1	6	12	13
2	5	1	2	2	1	1	3	3	7	3
3	0	3	3	0	3	0	4	5	6	7
4	9	6	7	2	6	2	2	6	13	14
5	5	3	9	2	2	1	3	7	29	16
6	4	2	10	0	1	0	1	1	19	8
7	5	2	9	3	4	3	6	1	12	8
8	5	1	18	2	6	2	4	1	23	16
9	1	0	0	3	4	2	1	2	14	4
10	4	11	19	7	5	6	10	15	31	21

That means, either the author is quite free in sentence structuring or it is the language permitting this variation.

Taking the means of 20 texts in each class from 1 to 10 we obtain the sequence

[6.35, 3.35, 3.90, 6.70, 7.40, 4.90, 5.15, 7.95, 3.00, 14.95]

as displayed in Figure 1.

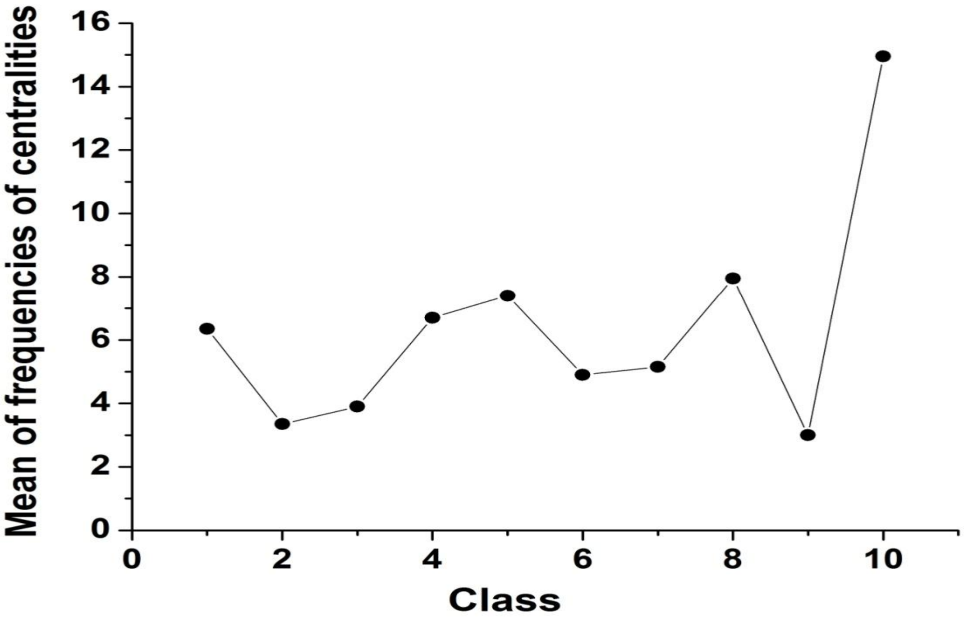


Figure 1. Oscillation of the centrality classes

Since this is the general trend (= means of 20 values of selected texts), one can see an oscillation which can be captured e.g. by a Fourier series. The fitting of Fourier series or other polynomials can be found in a software package or on the Internet. A simple description of the procedure can be found in Altmann (1988). But even if we obtain a preliminary inductive result, we cannot substantiate linguistically the parameters of the given function. Further one can ask whether the construction of classes is "correct": other subdivision of the scale (0 to 100) may have yielded smoother results.

At the present state of affairs, we cannot know whether the given trend is a property of language, of text sort or of author's style. The oscillation shows that there are preferences. Languages with preferred SOV (subject-object-verb) sequence would have a quite different "clause centrality". But in Slovak where all six sequences are allowed, the result may depend either on style or text sort. Further investigations are necessary.

Rank distribution

Though there is no clear tendency for the use of sentences displaying a special centrality (adding all frequencies of a class we obtain a curve with 4 maxima), there may be a common tendency concerning the ranking of frequencies. If we reorder the values in each column of Table 1 according to decreasing tendency, we obtain in all cases a very strongly expressed right truncated Zipf-Mandelbrot distribution as can be seen in Table 3. It must be remarked that there are also other distributions fitting well to the data but the Zipf-Mandelbrot's distribution is well known in linguistics and can easily be substantiated by the unified theory having a linguistic background (cf. Wimmer, Altmann 2005). We apply here the right-truncated distribution (x = 1,2,…,10) in which the normalizing constant is simply 1/(sum of all frequencies). One could apply also the Zipf-Mandelbrot function without normalization. Needless to say, one could apply several other distributions because the data are very regular and display a feature of personal style.

Table 3a
Rank-distribution of centralities in individual texts

Rank	Text 1		Text 2		Text 3		Text 4		Text 5	
	f_x	ZM	f_x	ZM	f_x	ZM	f_x	ZM	f_x	ZM
1	18	17.03	21	19.28	18	17.14	9	9.84	12	12.00
2	11	12.41	16	16.28	12	13.77	9	8.00	10	8.87
3	9	9.30	13	13.80	11	11.12	6	6.53	6	6.71
4	7	7.14	11	11.72	11	9.04	6	5.35	4	5.17
5	7	5.59	11	9.98	6	7.38	4	4.39	4	4.06
6	4	4.46	10	8.51	6	6.06	4	3.62	4	3.23
7	4	3.61	7	7.27	6	5.00	3	2.99	3	2.61
8	3	2.96	5	6.23	4	4.15	3	2.48	2	2.13
9	2	2.45	5	5.25	4	3.45	2	2.06	2	1.76
10	2	2.05	4	4.60	2	2.88	1	1.72	1	1.47
	a = 3.2168 b = 8.6767 n = 10 X^2_6 = 0.76 P = 0.9932		a = 11.9975 b = 69.9687 n = 10 X^2_6 = 0.97 P = 0.9866		a = 7.8246 b = 34.1845 n = 10 X^2_6 = 1.52 P = 0.9581		a = 11.9984 b = 56.5338 n = 10 X^2_6 = 0.81 P = 0.9919		a = 3.6712 b = 10.6358 n = 10 X^2_6 = 0.92 P = 0.9886	

Table 3b

Rank	Text 6		Text 7		Text 8		Text 9		Text 10	
	f_x	ZM	f_x	ZM	f_x	ZM	f_x	ZM	f_x	ZM
1	8	7.52	9	9.44	37	33.86	26	21.83	18	17.57
2	8	6.76	8	7.90	16	20.50	10	14.97	13	11.68
3	6	6.07	7	6.62	12	13.72	9	10.79	6	8.15
4	5	5.46	5	5.57	12	9.81	9	8.07	6	5.91
5	5	4.92	5	4.70	8	7.36	7	6.22	4	4.42
6	5	4.43	4	3.98	7	5.72	6	4.92	4	3.39
7	5	4.00	4	3.38	5	4.57	5	3.97	3	2.65
8	4	3.61	3	2.87	3	3.73	4	3.25	3	2.11
9	2	3.26	2	2.45	3	3.11	2	2.71	1	1.71
10	1	2.95	2	2.09	2	2.63	1	2.28	1	1.41
	a = 11.9998 b = 110.0209 n = 10 X^2_6 = 2.45 P = 0.87		a = 10.4109 b = 56.7731 n = 10 X^2_6 = 0.33 P = 0.9993		a = 2.0315 b = 2.5710 n = 10 X^2_6 = 2.67 P = 0.85		a = 2.5097 b = 5.1655 n = 10 X^2_6 = 4.53 P = 0.61		a = 3.0561 b = 6.0043 n = 10 X^2_6 = 1.71 P = 0.94	

Table 3c

Rank	Text 11		Text 12		Text 13		Text 14		Text 15	
	f_x	ZM	f_x	ZM	f_x	ZM	f_x	ZM	f_x	ZM
1	14	12.96	11	10.61	19	21.98	7	5.77	6	6.40
2	9	8.72	6	6.05	18	15.59	3	4.20	6	5.45
3	5	6.52	3	4.00	10	11.34	3	3.08	5	4.65
4	5	5.18	3	2.88	9	8.43	2	2.28	4	3.98
5	5	4.28	3	2.20	9	6.39	2	1.70	4	3.41
6	5	3.64	2	1.75	7	4.92	2	1.28	3	2.93
7	4	3.16	2	1.43	3	3.85	2	0.97	3	2.52
8	4	2.79	1	1.19	3	3.05	0	0.74	2	2.17
9	1	2.49	1	1.02	2	2.45	0	0.56	1	1.87
10	0	2.25	0	0.88	0	1.98	0	0.44	1	1.62
	a = 1.1059 b = 1.3257 n = 10 $X^2_6 = 4.97$ P = 0.5475		a = 1.5773 b = 1.3363 n = 10 $X^2_5 = 1.28$ P = 0.9367		a = 4.3297 b = 11.1118 n = 10 $X^2_6 = 5.17$ P = 0.5224		a = 11.0000 b = 33.0001 n = 10 $X^2_4 = 2.16$ P = 0.7073		a = 11.9914 b = 73.0689 n = 10 $X^2_6 = 0.96$ P = 0.9871	

Table 3d

Rank	Text 16		Text 17		Text 18		Text 19		Text 20	
	f_x	ZM	f_x	ZM	f_x	ZM	f_x	ZM	f_x	ZM
1	6	5.65	10	9.75	15	13.96	31	33.53	21	21.74
2	3	3.55	6	6.42	7	9.36	29	27.12	16	18.04
3	2	2.41	4	4.59	6	6.55	23	22.22	16	15.02
4	2	1.73	4	3.47	6	4.74	19	18.43	14	12.53
5	2	1.29	3	2.72	5	3.53	14	15.44	13	10.49
6	1	0.99	3	2.21	3	2.69	13	13.06	8	8.80
7	1	0.79	2	1.83	2	2.09	12	11.14	8	7.40
8	1	0.64	1	1.54	1	1.66	12	9.57	7	6.24
9	0	0.52	1	1.32	1	1.33	7	8.28	4	5.27
10	0	0.44	1	1.15	1	1.09	6	7.21	3	4.47
	a = 2.3683 b = 3.6177 n = 10 $X^2_3 = 0.86$ P = 0.8349		a = 1.7184 b = 2.6369 n = 10 $X^2_6 = 0.81$ P = 0.9918		a = 3.3642 b = 6.9209 n = 10 $X^2_6 = 2.07$ P = 0.9144		a = 3.1838 b = 13.5080 n = 10 $X^2_6 = 1.59$ P = 0.9538		a = 11.9994 b = 62.8538 n = 10 $X^2_6 = 2.10$ P = 0.9104	

As can be seen, the Zipf-Mandelbrot distribution is an adequate model, even if in some cases other distributions display even a “better” result. In order to make decisions about the rise of this structure many texts in many languages and text sorts must be analyzed.

What more, the parameters *a* and *b* display a correlation of 0.92, that means, one has to interpret only one of the parameters. Nevertheless, this must be left to future research. The relationship is visualized in form of a power function in Figure 2.

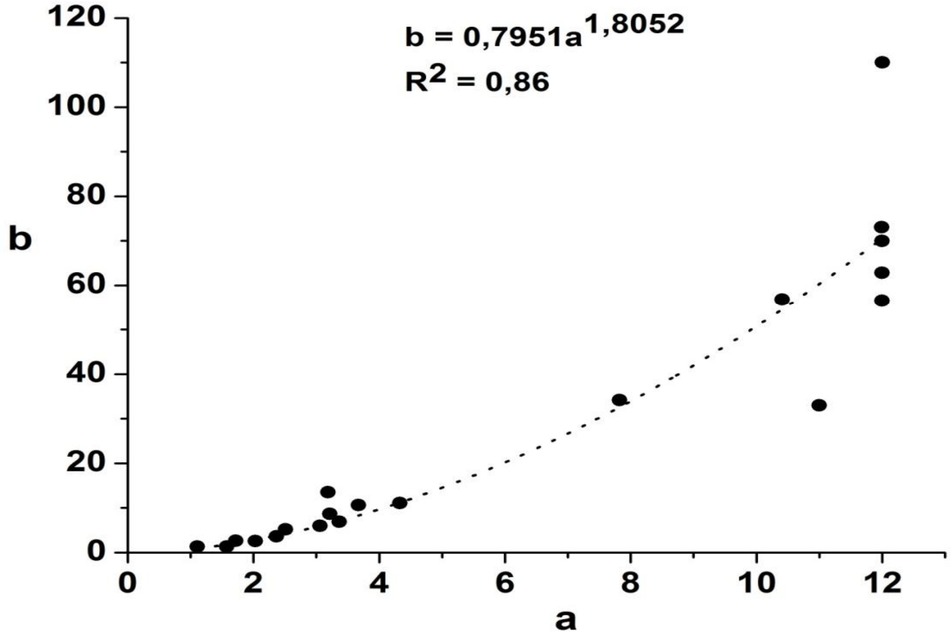


Figure 2. The relation between parameters *a* and *b*

Since ranking seems to be an adequate ordering of the centrality, one may characterize it using some indicators that can be computed from the distribution, in order to see some tendency. We shall use some of the possibilities.

Pearson’s excess. The excess of a distribution can be computed according to Pearson as

$$\beta_2 = \frac{m_4}{m_2^2}$$

Without comparing it with the normal distribution. Here, m_4 is the fourth central moment and m_2 is the variance of the distribution. For the examined texts we obtain the results presented in Table 4.

Ord’s indicators. Ord (1972) introduced the placing of a distribution into a Cartesian system <I, S>, where

$$I = \frac{m_2}{m_1^{'}} \quad \text{and} \quad S = \frac{m_3}{m_2},$$

where m_3 is the third central moment and m_1’ the mean of the distribution.

Popescu’s lambda indicator expresses the Euclidean distance between the rank-frequencies of classes. Since it depends on the sample size, the indicator has been relativized in

$$\Lambda = \frac{L(\log_{10} N)}{N},$$

where N is the sample size and L is the arc length joining the individual frequencies from the first to the last one. The arc is defined as

$$L = \sum_{x=1}^{k-1} \sqrt{(f_x - f_{x+1})^2 + 1}$$

where k is the greatest x value (here 10). For text T1 we compute

$L = [(18 – 11)^2 + 1]^{1/2} + [(11 – 9)^2 + 1\}^{1/2} + … + [(3 – 2)^2 + 1^{1/2}] + [(2 - 2)^2 + 1]^{1/2} =$
$= 20.5339$

Finally, we obtain $\Lambda = 20.5339(\log_{10}67)/67 = 0.5596$.

The values of all these indicators are displayed in Table 4. The relations are visualized in Figures 3 and 4.

Table 4
Some indicators of clause centrality

	T1	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**
N	67	103	80	47	35	49	49
m₁’	3.6418	4.1262	3.9375	3.9574	4.0857	4.3265	4.1837
m_2	6.4988	7.0617	6.7586	6.4663	6.2498	6.6281	7.0479
m_3	13.5668	10.7319	11.4050	10.4851	8.7637	6.0438	10.4980
m_4	113.6433	113.2951	107.2553	97.5442	91.4340	86.6708	110.5941
I	1.78	1.71	1.72	1.63	1.53	1.53	1.68
S	2.09	1.52	1.69	1.62	1.40	0.91	1.49
β_2	2.69	2.27	2.35	2.33	2.34	1.97	2.23
Λ	0.5596	0.4050	0.5012	0.4853	0.5547	0.4386	0.4245
	T8	**T9**	**T10**	**T11**	**T12**	**T13**	**T14**
N	105	79	59	52	32	80	21
m₁’	3.2476	3.4810	3.2881	3.6731	3.1875	3.3875	3.1429
m_2	5.9577	6.1737	5.8661	5.9893	5.4648	4.8123	4.3129
m_3	14.4353	11.2836	13.8037	7.4283	11.1694	8.0850	4.8630
m_4	108.8400	93.7922	100.3990	70.1117	78.8439	61.5726	35.8831
I	1.83	1.77	1.78	1.63	1.71	1.42	1.37
S	2.42	1.83	2.35	1.24	2.04	1.68	1.13
β_2	3.06	2.46	2.92	1.95	2.64	2.66	1.93
Λ	0.7425	0.6864	0.6620	0.6340	0.7763	0.5448	0.8672

	T15	T16	T17	T18	T19	T20	
N	35	18	35	47	166	110	
m_1'	4.0857	3.1667	3.5429	3.2979	4.1265	4.0727	
m_2	6.2498	4.8056	6.3053	5.4857	7.0623	6.4129	
m_3	8.7637	7.7593	13.4285	12.3480	11.1069	9.3095	
m_4	91.4340	54.8588	110.8910	97.4656	110.6591	96.5610	
I	1.53	1.52	1.78	1.66	1.71	1.57	
S	1.40	1.61	2.13	2.25	1.57	1.45	
β_2	2.34	2.38	2.79	3.24	2.21	2.35	
Λ	0.4884	0.8651	0.6442	0.6744	0.3729	0.4053	

The values of Ord's indicators are presented in Figure 3. As can be seen, all values are positioned below the line $S = 2I - 1$ which represents the upper boundary of the negative hypergeometric distribution. The exact boundaries can be found only after having analysed many texts. There will surely be differences between languages according to the SVO permutation.

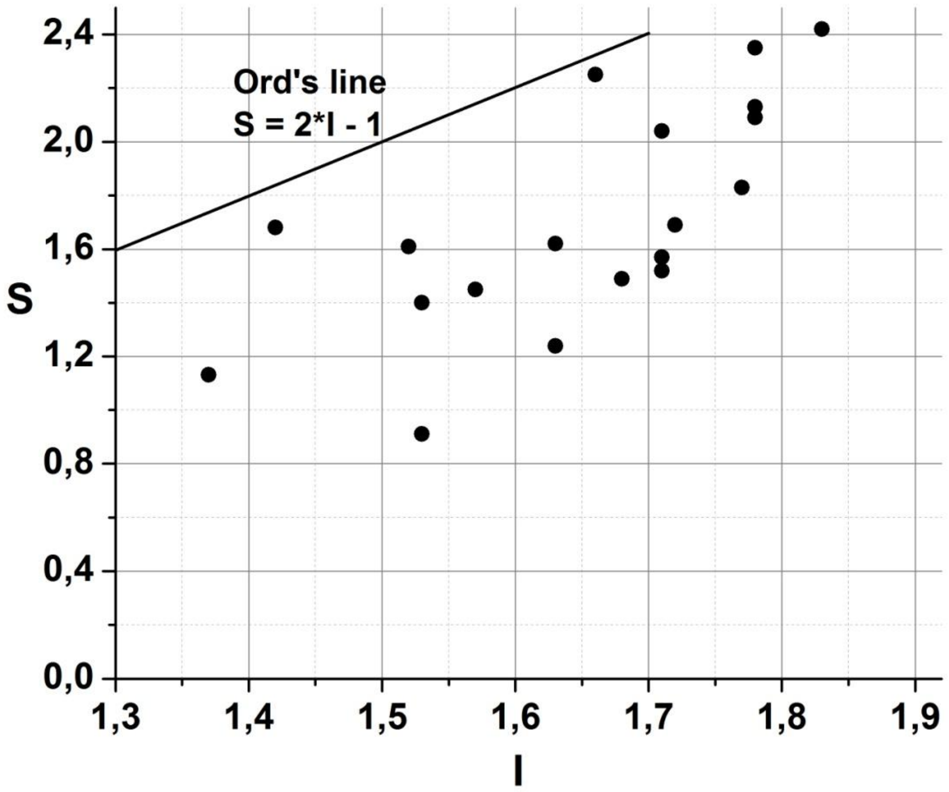


Figure 3. The relation between I and S

As can be seen in Figure 4, there is no relation between β_2 and Λ. Though the variation of these indicators is restricted, they seem to be independent of one another, however, further texts and languages must be analyzed.

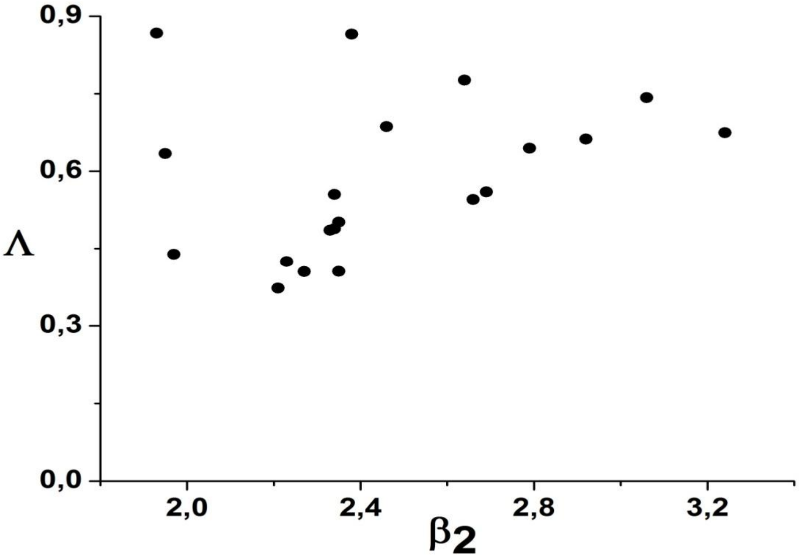


Figure 4. The relation between β_2 and Λ (Svoráková)

Comparison

In order to obtain at least a look at other texts comparison we analyzed 10 prosaic texts by Eva Bachletová (2012) (see Appendix). The texts are written in a poetic language, many sentences do not have a verb and we avoided subjective insertion of an elliptic verb. The values of centrality are presented in Table 5, the discrete transformation in Table 6, the fitting of the Zipf-Mandelbrot distribution in Table 8 and the individual indicators in Table 9.

Table 5
Centrality in texts by E. Bachletová

Text	Centrality	N	Mean	SSQ
T1	[0.56, 0.33, 0.33, 0, 1, 0.78, 0.33, 0.71, 0.50, 1, 0, 0.67, 0, 0.5, 0.67, 1, 0, 0.25, 0.75, 0.5, 0.5, 0.27, 1, 0, 0, 1, 1, 1, 1, 0.33, 0.43, 0.2, 0, 0, 0.85, 1, 1, 1, 0.43, 0.5, 0, 0.6, 0.3, 0.2, 0, 0.71, 0.33, 1, 0.82, 0.6, 0, 1, 0, 1, 0.5]	55	0.5172	7.9026
T2	[1,0, 0.5, 0.67, 1, 1, 1, 0.67, 1, 1, 0, 0, 0, 0, 0, 0.5, 1,1,0.33, 0.33, 0.33, 0.43, 1, 0, 0.2, 0.33, 1, 0, 0.33, 0.67, 1, 0, 1, 0.33, 1, 1, 0.67, 1, 0.33, 0.20, 0, 0, 0.75, 0.5, 0.17, 1, 0.71, 1, 0.5, 0.5, 1, 1, 1, 1, 1, 0, 0.25, 0.33, 0.2, 0, 1, 1, 1, 0, 0, 1, 1, 1,0,0.18, 0.33, 1, 0.67, 1, 1, 1, 1, 0.33, 1, 1, 0.33, 0, 1, 0.6, 0.6, 1, 1, 1, 0.5, 1, 0.5]	91	0.6019	14.4434
T3	[0.6, 0.14, 0, 1, 0.2, 0, 1, 1, 1, 0, 0, 1, 1, 1, 0.40, 0.33, 0.6, 0.14, 0.6, 1, 1, 0.71, 1, 1, 1, 0.67, 0.25, 1. 0.78, 0.75, 1, 0.5, 1, 1, 1, 1, 1, 0,1, 0,1, 1, 0.6, 1, 1, 1, 0, 1, 0, 0.5, 0, 0.83, 0.33, 0.5, 1, 0.2, 1, 0.2, 0, 1, 0.5, 0, 0.14]	63	0.6276	10.1447
T4	[1, 1, 1, 0, 0.67, 0.67, 1, 1, 1, 0.33, 0, 0.6, 0, 1, 0.33,	108	0.5125	18.7920

	1, 0.33, 0.67, 0.2, 1, 0.5, 1, 0.5, 0,0, 0.33, 0.11, 1, 1, 1, 1, 0, 0.33, 1, 0.33, 0.33, 0, 1,0.33, 0, 0.33, 0,0.14, 0, 0.67, 0.2, 1, 0, 0.25, 0, 0, 0, 0.71, 0, 0.71, 1, 1, 1, 0.33, 0.56, 1, 1, 0, 0, 1, 0, 0.33, 0, 1, 0.2, 1, 0, 1, 0.33, 0, 1, 1, 1, 1, 0, 1, 1, 1, 0, 0.33, 0.6, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0.67, 1, 0.5, 0.5, 0.67, 0.33, 0, 1, 0.43, 0.5,0.5]			
T5	[0.5, 0, 0.75, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0.6, 0, 0, 0, 0, 1, 0.33, 0, 0.2, 1, 0, 1, 0.33, 0.14, 0, 0, 0.33, 0, 1, 0, 0, 1, 0.6, 0, 1, 0.33, 1, 1, 0.25, 0.67, 1, 0.6, 0, 0, 0.5, 1, 1, 1, 0.6, 1, 0.67, 1, 1, 1, 0, 0, 0.67, 0.33, 0.33, 0.33, 1, 0, 0, 0]	67	0.5084	12.4189
T6	[1, 1, 0.80, 1, 0.25, 1, 0.33, 1, 0, 0.5, 1, 0.33, 0.33, 0.85, 0.6, 0.4, 0.25, 1, 0.45, 0.25, 1, 0.45, 0.5, 0.9, 0.83, 0.25, 0.82, 0.17, 0.67, 0.78, 0, 0.08, 0.43, 0.5, 0.33, 1, 0, 0, 1, 0.6, 1, 0.5, 0.33, 0.5, 0.6, 0.5, 0.69, 0.2, 1, 1, 0.71, 0.4, 0.75, 0.33, 0.53, 0.71, 0.67, 1, 0.75, 1, 0.5, 0.80, 1, 1]	64	0.6113	6.3767
T7	[0, 0.56, 1, 1, 0, 0, 0.5, 0.8, 0, 0, 0, 1, 0,0,0, 1, 0.73, 0, 1, 0.33, 0.67, 1, 1, 0.6, 0.20, 0.78, 0.5, 0, 1, 0, 0, 0.71, 0.6, 0, 0, 0, 0, 1, 0.23, 1, 1, 0.2, 0, 1, 0.75, 0.8, 0.5, 0, 0.4, 0.4, 0.33, 0, 0, 0, 0, 1, 0.8, 1, 1]	59	0.4473	10.7272
T8	[0, 1, 0.33, 0.43, 0.14, 0.5, 0, 1, 0.2, 0.5, 0, 0.5, 0.09, 0.33, 1, 0.33, 0.5, 1, 0.23, 1, 1, 1, 0, 0, 1, 0.78, 1, 0.33, 0.5, 1, 1, 1, 1, 1, 0, 0.6, 1, 0.11, 0.33, 1, 0.3, 0.2, 1, 0, 0.33, 0.14, 0.5, 0.2, 1, 0.5, 0.56, 0.14, 1, 1, 0.5, 1,1,1,1, 0.2, 0.67, 0.83, 0, 0.38, 1, 1]	66	0.5785	9.6978
T9	[1, 0.33, 0.2, 0.2, 1, 0.2, 0, 1, 0, 0.67, 0, 0, 0.67, 1, 1, 0, 1, 0, 0.5, 0, 1, 0, 0.5, 0, 0, 0, 0.33, 1, 1, 0.17, 1, 0.33, 0.29, 0.71, 0.71, 1, 1, 1, 1, 1, 1, 0.33, 0.67, 0.33, 0, 0.82, 1, 0.5, 0.67, 0.33, 0, 0.5, 0.5, 1, 0.33, 1, 0.14, 1, 0, 1, 0.5, 0.43, 0.71, 0, 0.5, 0.33, 1, 0.33, 0.5, 1, 0, 0, 0.43, 1, 0.33, 0.71, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0.71, 1, 0.33, 0.2, 1, 0, 0, 1, 1, 0.2, 0.6, 0.33, 0.5, 1, 0.5, 1, 1, 0, 1, 1, 1, 0.71, 1, 0]	110	0.5707	17.5315
T10	[0, 1, 1, 1, 0.33, 0.33, 0.09, 1, 1, 1, 0.2, 0, 1, 0.43, 1, 0.67, 1, 0, 1, 0, 0.5, 0, 1, 0, 1, 0.5, 1, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0.5, 0.71, 0.50, 0.38, 0.33, 1, 0.67, 1 0, 0.71, 0.43, 1, 0, 0]	52	0.6015	8.9789

For Bachletová the interval of centralities is <0.4473; 0.6276>. Testing for the difference of the extreme texts T7 and T3 yields u = (0.6276 – 0.4473)/((10.1447 + 10.7272)/(63 + 59))$^{1/2}$ = 0.44, that means, a not significant difference. Though with Bachletová *u* is slightly greater than with Svoráková, the centrality seems to be relatively stable within the work of the individual author.

The normalized values of centrality in texts by Bachletova are presented in Table 6

Table 6
Standard normal N(0;1) values of mean centralities with E. Bachletová

T1	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**	**T8**	**T9**	**T10**
0.33	2.43	2.52	0.31	0.16	2.82	-0.95	1.66	1.86	1.76

In Svoráková, there were 13 significant deviations from the theoretical mean (out of 20), in Bachletová, there are merely 3 out of 10. That means, the work of Bachletová (poetic prose) has a stronger trend to equilibrate the text in this sense than Svorakova (technical language) whose texts are factual.

In order to compare the two authors directly we consider the mean centralities of individual texts as simple observations. That means, we have 20 texts (values) for Svoráková and 10 texts (values) for Bachletová. With Svoráková we obtain the average of her 20 texts as c_S = 0.5982 and SSQ_S = 0.0571, and for Bachletová c_B = 0.5577 and SSQ_B = 0.0305. Using formula (2) we obtain t = 1.87 which is, with 28 degrees of freedom, not significant. It is to be remarked that in long run, texts must tend to the value prescribed by the grammar but to find this value would mean the analysis of an enormous number of texts in a single language.

Another possibility of comparison is the use of the chi-square method which enables us to state whether the texts of an individual author are homogeneous and whether two authors are homogeneous taking into account the frequency in individual classes (1 to 10). The values of Svoráková are presented in Table 3, those of Bachletová in Table 7.

Table 7
Centrality classes in texts by E. Bachletová

x	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**	**T8**	**T9**	**T10**
1	11	17	11	31	22	5	23	9	23	13
2	2	5	6	5	2	2	2	8	7	1
3	3	1	1	1	1	4	1	2	1	0
4	3	11	3	14	7	8	4	7	12	4
5	7	6	4	7	2	10	3	9	12	6
6	2	2	3	3	4	4	3	2	1	0
7	2	4	2	6	3	3	1	1	4	2
8	3	2	3	2	1	7	7	1	6	2
9	2	0	1	0	0	4	0	1	1	0
10	11	35	29	39	25	17	15	26	43	24

The chi-square for the homogeneity of classes among the texts of one author is given as

$$X^2 = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{(N_{ij} - \frac{N_{i.}N_{.j}}{N})^2}{\frac{N_{i.}N_{.j}}{N}}$$

where N is the total sum, N_{ij} are the values in individual cells (i,j), $N_{i.}$ is the marginal sum of rows, $N_{.j}$ the marginal sums of columns, and the chi-square is distributed with (R-1)(C-1) degrees of freedom..

Computing the values using the two tables we obtain for Svorakova, X^2 = 250.4433 with 171 degrees of freedom and P = 7.16337E-05, and for Bachletová, X^2 = 122.1404 with 81 DF and P = 0.0021. That means, Bachletová is more uniform than Svoráková.

The above test could be performed also using a non-parametric variant, simply by ranking the values in each column and compare the ranks.

Comparing the two authors we take into account only the marginal sums of rows in the two tables and obtain the results in Table 8.

Table 8
Row sums

Author	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Sum**
Svoráková	127	67	78	134	148	98	103	159	60	299	**1273**
Bachletová	165	40	15	73	66	24	28	34	9	264	**718**
Sum	**292**	**107**	**93**	**207**	**214**	**122**	**131**	**193**	**69**	**563**	**1991**

The chi-square with 9 DF is X^2 = 171.07, yielding P = 2.90546E-32, testifying to a drastic difference between the authors as to the verb placing in the clause.

Since the individual authors have internal differences and compared with one another differ significantly, we may conclude that Slovak is a language with few restrictions as to verb placing in sentence. This special aspect of syntax may serve the style for evoking special effects. The classes displaying the greatest differences can be found as the greatest components of the chi-square. But since chi-square increases with increasing sample size, it is simpler to have a look at the proportion of each class of the given author and if necessary to test the difference between the given proportions. For example, class 1 yields for Svorákova 127/1273 = 0.09976 and for Bachletová 165/718 = 0.2298. The results for Table 8 are presented in Table 9 – multiplied by 100.

Table 9
Proportions of individual classes of the two authors

Author	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Svoráková	9.97	5.26	6.13	10.53	11.63	7.70	8.09	12.49	4.71	23.49
Bachletová	23.00	6.67	2.09	10.17	9.19	3.34	3.90	4.74	1.25	36.77

The differences in individual classes can be tested using the exact binomial test or the asymptotic normal test, but one can decide here empirically by choosing the first five greatest differences in classes (1, 10, 8, 6, 7). Classes 1 and 10 are those of extreme asymmetry/non-centrality and extreme centrality respectively and are conditioned by the type of clause, e.g. *povedal som mu* (told-am-I to him) vs. *Keď' som mu povedal* (when am-I to him said). It must be emphasized that we do not analyze a language but its products.

As can be seen in Table 10, the Zipf-Mandelbrot distribution is quite adequate also for texts by Bachletová. The relationship between the parameters *a* and *b* are presented in Figure 5, and together with Svorakova in Figure 6. As can be seen, there are two outliers in the data of Bachletová: T3 and T10. The causes of these deviations can be sought in other domains concerning rather the theme, the subject of the text. For literary scientists it means a search for boundary conditions; for historical literary scientists it means the study of the life of the author. However, it is just the detection of outliers which may be interesting both for researchers and for the authors.

Table 10a
Fitting the Zipf-Mandelbrot distribution to the texts by E. Bachletová

Rank	**T1**		**T2**		**T3**		**T4**		**T5**	
	f_x	**ZM**	**f_x**	**ZM**	**f_x**	**ZM**	**f_x**	**ZM**	**f_x**	**ZM**
1	11	12.72	35	34.35	29	26,59	39	45,75	25	29,87
2	11	8,39	17	18.07	11	12,44	31	24,12	22	15,01
3	7	6,01	11	10.52	6	7,26	14	13,84	7	8,31
4	3	4,56	6	6.61	4	4,78	7	8,47	4	4,95
5	3	3,60	5	4.39	3	3,40	6	5,46	3	3,12
6	3	2,92	4	3.05	3	2,55	5	3,66	2	2,06
7	2	2,43	2	2.19	3	1,99	3	2,55	2	1,41
8	2	2,06	2	1.62	2	1,59	2	1,82	1	1,00
9	2	1,77	1	1.23	1	1,31	1	1,34	1	0,73
10	2	1,54	0	0.96	1	1,10	0	1,00	0	0,54
	a = 1,6571 b = 2,4975 n = 10 X^2_6 = 2.09 P = 0.9116		a = 3,3822 b = 3,7788 n = 10 X^2_5 = 1,28 P = 0,9366		a = 1,8662 b = 0,9891 n = 10 X^2_6 = 1.56 P = 0.9555		a = 4,2231 b = 5,1113 n = 10 X^2_5 = 4,62 P = 0,4644		a = 4,1885 b = 4,5992 n = 10 X^2_4 = 4.65 P = 0.3300	

Table 10b

Rank	**T6**		**T7**		**T8**		**T9**		**T10**	
	f_x	**ZM**	**f_x**	**ZM**	**f_x**	**ZM**	**f_x**	**ZM**	**f_x**	**ZM**
1	17	15,94	23	24.16	26	23.58	43	42,33	24	24,42
2	10	11,09	15	12.87	9	13.86	23	23,66	13	11,33
3	8	8,30	7	7.55	9	8.81	12	14,42	6	6,06
4	7	6,52	4	4.76	8	5.94	12	9,37	4	3,57
5	5	5,30	3	3.16	7	4.18	7	6,40	2	2,27
6	4	4,43	3	2.19	2	3.06	6	4,54	2	1,52
7	4	3,77	2	1.58	2	2.30	4	3,33	1	1,06
8	4	3,26	1	1.16	1	1.77	1	2,51	0	0,76
9	3	2,86	1	0.88	1	1.39	1	1,93	0	0,57
10	2	2,54	0	0.68	1	1.11	1	1,51	0	0,43
	a = 1,4458 b = 2,5068 n = 10 X^2_6 = 0.58 P = 0.9967		a = 3,4914 b = 4,0593 n = 10 X^2_5 = 1.21 P = 0.9436		a = 3,0718 b = 4,2913 n = 10 X^2_6 = 5.43 P = 0.4898		a = 3,3393 b = 4,2562 n = 10 X^2_6 = 3.36 P = 0.7627		a = 3,3906 b = 2,9360 n = 10 X^2_4 = 1,86 P = 0,7616	

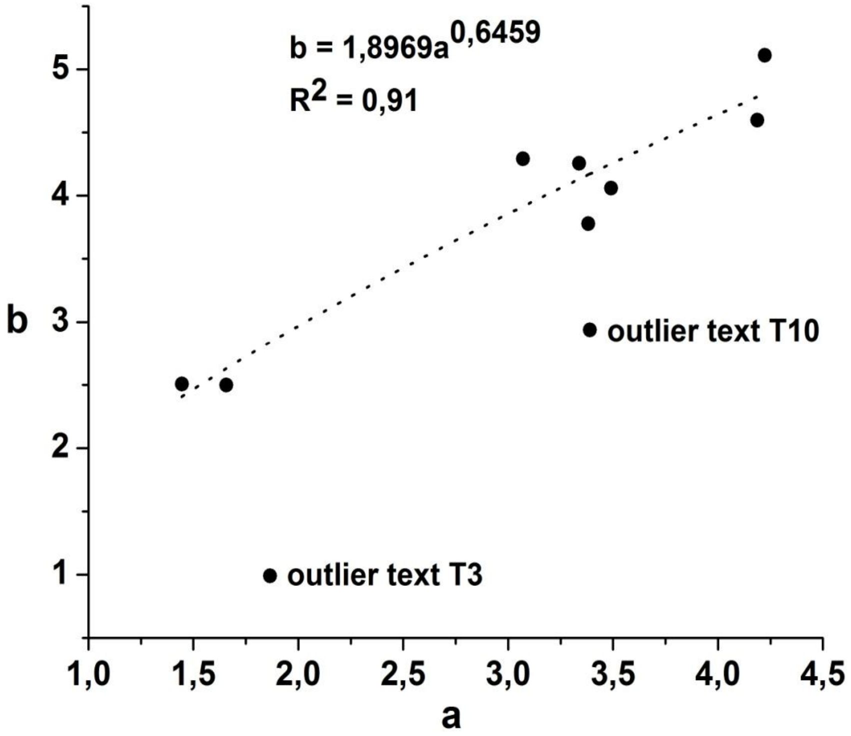


Figure 5. The parameters <a,b> of the Zipf-Madelbrot distribution for Bachletová based on data in Table 10

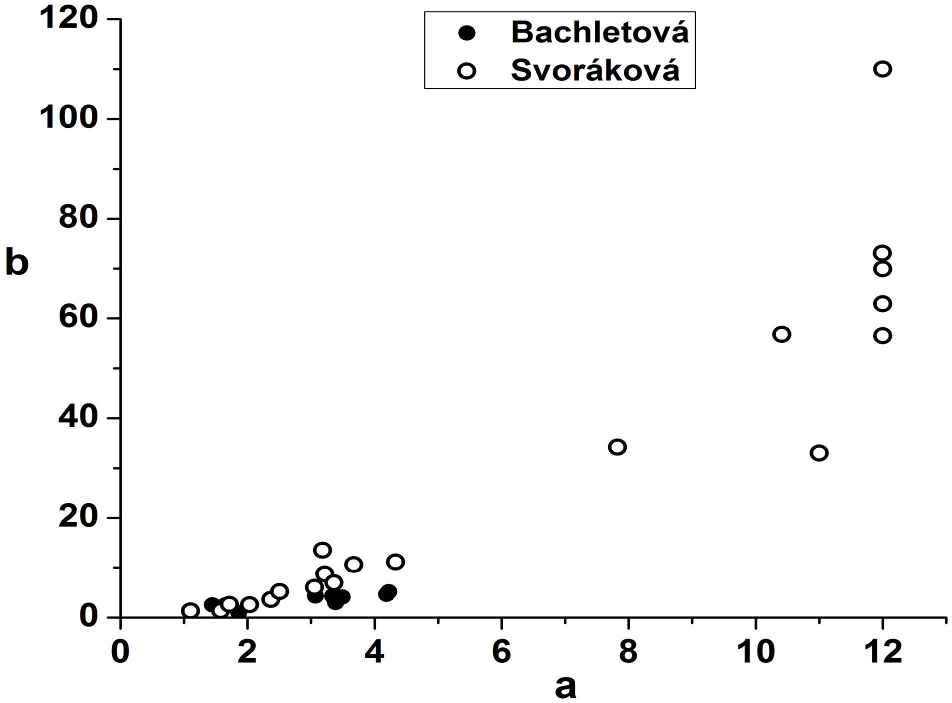


Figure 6. Parameters <a,b> for both authors

Pooling the data obtained from both authors, as visualized in Figure 6, we detect a well expressed power trend which can be captured by the relation $b = 0.4696a^{2.0198}$ yielding $R^2 = 0.8850$ and a highly significant F-value. The data and the fitting are presented in Table 11.

Table 11
Fitting the power function to the relation $b = f(a)$ of the Zipf-Mandelbrot distribution

a	b	b = f(a)		a	b	b = f(a)
1.1059	1.3257	0.5755		3.3822	3.7788	5.5028
1.4458	2.5068	0.9888		3.3906	2.9360	5.5304
1.5773	1.3363	1.1789		3.4914	4.0593	5.8675
1.6571	2.4975	1.3024		3.6712	10.6358	6.4939
1.7184	2.6369	1.4016		4.1885	4.5992	8.4749
1.8662	0.9891	1.6557		4.2231	5.1113	8.6169
2.0315	2.5710	1.9654		4.3297	11.1118	9.0619
2.3683	3.6177	2.6791		7.8246	34.1845	29.9440
2.5097	5.1655	3.0121		10.4109	56.7731	53.3107
3.0561	6.0043	4.4838		11.0000	33.0001	59.5793
3.0718	4.2913	4.5305		11.9914	73.0689	70.9237
3.1838	13.5080	4.8703		11.9975	69.9687	70.9965
3.2168	8.6767	4.9728		11.9984	56.5338	71.0073
3.3393	4.2562	5.3627		11.9994	62.8538	71.0193
3.3642	6.9209	5.4438		11.9998	110.0203	71.0240

The results show that there is an attractor for the formation of centralities. It had been possible to choose a different well fitting distribution for the frequencies but before doing it, many texts and languages must be analyzed. The values of outliers can easily be seen both in Table 11 and in Figure 6. One can see a very striking difference between the two text sorts.

The other indicators for Bachletová are presented in Table 12.

Table 12
Rank-frequency indicators for the texts by Bachletová

	T1	T2	T3	T4	T5
N	46	83	63	108	67
m_1’	3,6304	2,5783	2,7619	2,5648	2,4478
m_2	7,1895	3,8824	5,5465	3,4865	3,4115
m_3	18,8693	10,4202	17,7372	9,3102	10,7307
m_4	144,6139	62,1374	117,9393	53,6627	63,1742
I	1,9803	1,5058	2,0082	1,3594	1,3937
S	2,6246	2,6839	3,1979	2,6703	3,1455
β_2	2,7978	4,1224	3,8337	4,4145	5,4283
Λ	0,5661	0,8810	0,9312	0,7808	0,7908

	T6	**T7**	**T8**	**T9**	**T10**
N	64	59	66	110	52
m_1'	3,8281	2,6102	2,8939	2,7182	2,1731
m_2	7,2986	3,9867	4,6708	4,1297	2,2970
m_3	14,0965	11,1220	12,2903	10,8138	4,9752
m_4	124,4750	66,4316	88,8533	70,8250	22,9303
I	1,9066	1,5197	1,6139	1,5193	1,0570
S	1,9314	2,8038	2,6314	2,6186	2,1660
β_2	2,3367	4,2220	4,0732	4,1529	4,3461
Λ	0,5419	0,8087	0,8373	0,8533	0,9378

The relation <I, S> for Bachletová is visualized in Figure 7. In contrast to Svoráková, the points are dispersed both in the beta-binomial (= negative hypergeometric, below the line) and in the beta-Pascal domain (above the line), hence the mechanism evoking the given values may be more complex and the Zipf-Mandelbrot distribution will – after checking it on many texts – prove inadequate. But for the time being it is sufficient.

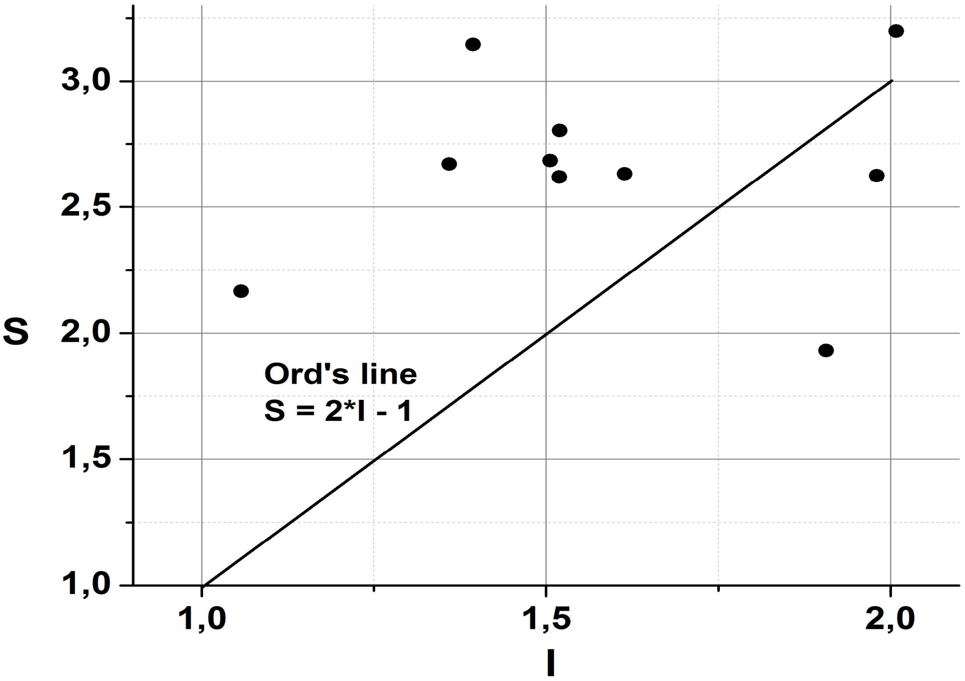


Figure 7. Ord’s relation <I, S> for Bachletová

The relationship between β_2 and Λ as presented in Figure 8 seems to have an unknown background. Either we reject the existence of this relation or we continue analyzing texts in other languages.

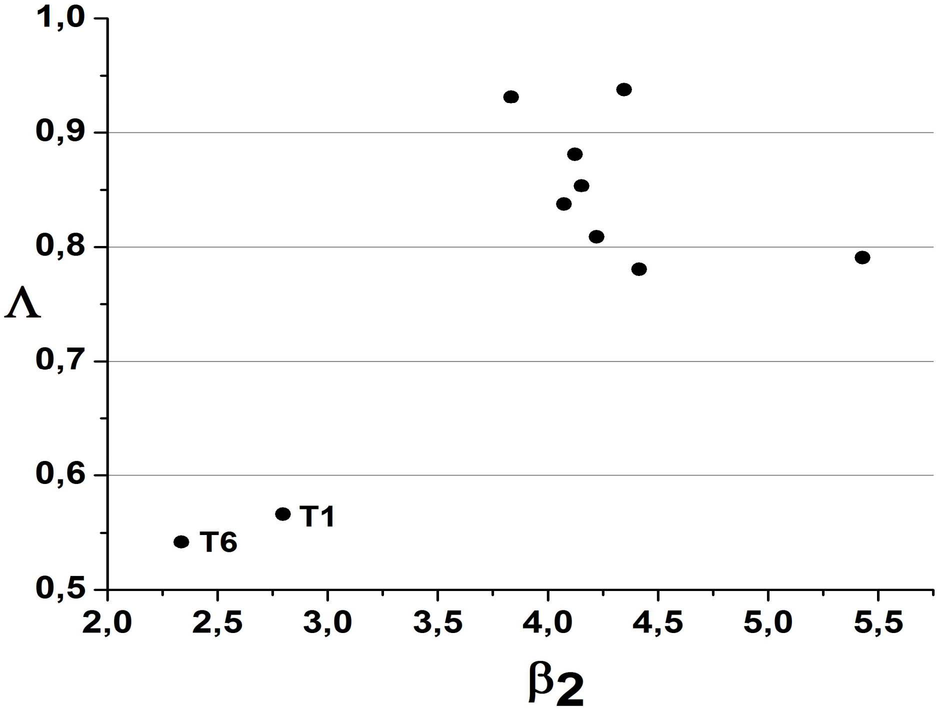


Figure 8. The relation between β_2 and Λ for Bachletová

Summary

Just as any other entity of language, clause has a set of properties which increases with the advancement of science. Here merely the centrality has been defined and evaluated. Since each property has at least one link to some other property, further research may lead to a more complex control cycle and, automatically, to a support of a very general theory. However, before the links will be found, many texts in many languages must be analyzed and the respective boundary conditions must be sought.

References

Altmann, G. (1988). *Wiederholungen in Texten*. Bochum: Brockmeyer.

Altmann, G., Lehfeldt, W. (1973). *Allgemeine Sprachtypologie*. München: Fink.

Bortz, J., **Lienert, G.A., Boehnke, K.** (1990). *Verteilungsfreie Methoden in der Biostatistik.* Berlin: Springer.

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

Köhler, R., Naumann, S. (2008). A contribution the quantitative studies on the sentence level. In: Köhler, R. (ed.), *Issues in Quantitative Linguistics: 34-45.* Lüdenscheid: RAM-Verlag.

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

Popescu, I.-I., Čech, R., Altmann, G. (2011). *The Lambda-structure of texts.* Lüdenscheid: RAM.

Svoráková, S. (1990a). Majstrovstvo bulharských ikon. *Výtvarný život*. 3, p. 60-65. **(Text 19)**

Svoráková, S. (1990b) Nahlas o jednom areáli – Pamätník SNP po novej úprave. In: *Priekopník* 10(4), p. 3. **(Text 20)**

Svoráková, S. (1991). Kabaret života. In: *Kabaret života – Kamila Štanclová: Obrazy, grafika*. Zvolen: Vlastivedné múzeum. **(Text 14)**

Svoráková, S. (1997). Národná múza Ladislava Dunajského. *Priekopník* 10(4), p. 3. **(Text 18).**

Svoráková, S. (1998). Alternatívy slovenskej grafiky. Review of the exposition. *Literárny týždenník 5, p. 14.* **(Text 9)**

Svoráková, S. (1999). Veľké ambície malej grafiky. Review of the exposition: XIV. Ročník Medzinárodného trienále drevorezu a drevorytu. *Literárny týždenník 6, p.14.* **(Text 10)**

Svoráková, S. (1998a). Štefan Prukner Bartůšek v zahraničí a doma. *Slovenská republika 8(9), p. 10* **(Text 17)**

Svoráková, S. (1998b). Štefan Prukner Bartušek: Mágia obrazu. In: *Originál* 3/1998, *p. 8.* **(Text 15)**

Svoráková, S. (1999). Plenér Liptov 1999. *Plenér Liptov 1999*. Úvodný text v katalógu Medzi národného sympózia. Banská Bystrica: Akadémia umení (**Text 12).**

Svoráková, S. (2000). Margita. In: *Margita*. Úvodný text katalógu Jaroslava Uhela.2000 (**Text 16)**

Svoráková, S. (2001). Plenér Liptov 2001. In: *Plenér Liptov 2001*. Úvodný text v katalógu Medzinárodného sympózia Vyd.: Norami pre Galériu P&P, Mesto Liptovský Mikuláš, Rotary Club Liptovský Mikuláš a FVU Bratislava (**Text 13)**

Svoráková, S. (2003a). Čakanie na Štraussa. Review of: Tomáš Štrauss, Metamorfózy umenia XX. storočia. Bratislava: Kalligram, 2001. *Dart - Revue súčasného výtvarného umenia 10, p. 37* (**Text 1)**

Svoráková, S. (2003b). Dvojhlasné dejiny a univerzálna kultúra ? Review of: Mária Orišková: Dvojhlasné dejiny umenia. Bratislava: Petrus, 2002. *Dart – Noviny o súčasnom výtvarnom umení. 2, p. 3* (**Text 2)**

Svoráková, S. (2004a). 200 plechoviek Campbellovej polievky. *Literárny dvoj)týžden ník) 26-27, p. 14* (**Text 11)**

Svoráková, S. (2004b). Stratená moderna. Review of: Tomáš Štrauss: Zo seba vystupujúce umenia. Príspevok k stratifikácii stredoeurópskych avantgárd. Bratislava: Kalligram, 2003. D*art - Noviny o súčasnom výtvarnom umení 1, p. 3.* (**Text 3)**

Svoráková, S. (2007). Znovuobjavené klenoty. Review of: Ján Hollý – Emil Makovický: Selanky. (Úvodný text K. Szmudová). Banská Bystrica: Štátna vedecká knižnica, 2007. *Slovenské pohľady 7- 8, p. 276 -277* (**Text 4)**

Svoráková, S. (2009a). Voľným okom – List zo Slovenska. Review of: Voľným okom. (Úvodný text Ľ. Hološka). Martin: Vydavateľstvo Matice slovenskej, s.r.o., 2006. *Mecenat i mir - Literarno-chudožestvennyj i kulturnyj magazin 41-44. Moskva, 2009. p. 356- 358* (**Text 8)**

Svoráková, S. (2009b). Smrť jej nepristane. Review of: Nová krv. (Úvodný text I. Jančár). Bratislava: Galéria Mesta Bratislavy, 2008. *Literárny (dvoj)týždenník 5-6, p. 13.* (**Text** 5)

Svoráková, S. (2011a). Ruská interpretácia slovenského naturizmu. Review of: Alla Mašková: Slovenský naturizmus v časopriestore. (Prel. Hedviga Kubišová) Bratislava, 2009. *Literárny (dvoj)týždenník* 13-14, p.12. (**Text 6)**

Svoráková, S. (2011b) ...a poslední nie sú prví – Na margo výstavy. Review of the exposition: Bienále v čase normalizácie v Stredoslovenskej galérii v B. Bystrici. *Literárny (dvoj)týždenník 37-38, p. 13.* (**Text 7)**

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

Wimmer, G., Altmann, G., Hřebíček, L., Ondrejovič, S., Wimmerová, S. (2003). *Úvod do analýzy textov*. Bratislava: VEDA.

Appendix

Texts by Eva Bachletová: *Riadky bytia*. Kežmarok: Vivit 2013.

T1. Sila ľudského ducha
T2. Moje (nájdené) bytie
T3. Zranená a milujúca zem
T4. Súkromná rana
T5. Tváre
T6. Túžba po múdrosti
T7. Ľudské Vianoce
T8. Zrýchlený čas
T9. Môj august
T10. Božia príroda

Glottometrics 28, 2014, 37-55

Hreb-like analysis of Eminescu's poems

Doina Tatar, Mihaiela Lupea, Gabriel Altmann

Abstract. The aim of the article is to show the hreb-like construction of poems of the Romanian poet M. Eminescu. Hrebs are constructed, ranked, and their stratification. denotational properties (topicality, concentration, diffuseness, compactness) as well as text concentration and hreb chains are scrutinized.

1. Introduction

Denotation analysis is a complex discipline concerned with the mutual relationships of sentences or verses. A relatively well developed aspect is that of references elaborated mostly from the qualitative point of view, i.e. as to their identification, description and classification in disciplines called text theory or discourse analysis or pragmatics or cohesion (cf. Agricola 1969; Halliday, Hasan 1976; Viehweger 1978; Brown, Yule 1983; Levinson 1983; Stubbs 1983; Schiffrin 1987; Palek 1988; Numan 1993; Vater 1994; Hoffmannová 1996). It was L. Hřebíček (1985, 1992, 1993, 1995, 1996, 1997) who introduced measurement in this domain and laid the basis of a theory. The development of the quantitative aspect has been advanced by Ziegler and Altmann (2002) (cf. also Schwarz 1995; Christmann 2004; Ziegler 2005; Köhler, Naumann 2007).

The basic concept is the so-called *hreb* baptized in this way in honor of L.Hřebíček who used originally the more specific term *sentence aggregate*. In our conception, hreb is a discontinuous text unit that can be presented in a set form. The set contains all entities denoting the same real entity or referring to one another in text, i.e. concerning the same textual entity. One can distinguish morpheme-hrebs, word-hrebs, phrase-hrebs and sentence-hrebs.

A *morpheme-hreb* is the set of morphemes referring semantically (not grammatically) to the same entity. In the sentence *I work* there are two morphemes belonging to two different herbs: {*I*}, {*work*}. But in the identical Russian sentence *ja rabotaju*, there are three morphemes and the morpheme *–ju* refers semantically to the named entity, hence {*ja, -ju*}, {*rabotat'*}. Some affixes refer, other ones do not. In the Hungarian objective conjugation there may be several referring morphemes, e.g. *látlak* ("I see you"). The expression *házban* ("in the house") need not be partitioned, but *bent a házban* ("in the house-in"), the morpheme *–ban* refers to *bent*. Nevertheless, such an analysis is mostly too redundant and the depth of the analysis is sometimes a personal decision.

A *word-hreb* contains all words which are synonyms or refer to one of the synonyms. In Goethe's poem *Der Erlkönig* one finds a hreb containing {*Kind, Sohn, Knabe, du, mein, er, es,…*}. However, the same word may be element of another set, too, e.g. *mein* ("my") refers always to the speaking person. Complete word-hreb analyses of several texts can be found in Ziegler, Altmann (2002).

A *phrase-hreb* ignores so to say auxiliaries and synsemantics which are part of the given phrase. In this way one obtains a smaller number of sets and can perform all operations used with the above hrebs. Hrebs of this sort were introduced by Köhler and Naumann (2007) and analyzed using press texts by Christmann (2004).

Sentence-hrebs are the greatest hrebs. A sentence hreb is based on an autosemantic contained in a sentence; all sentences containing the same autosemantic or a reference to it belong to the given hreb. As above, every sentence can belong to several hrebs. Sets containing hrebs of this sort have many intersections, one can easily follow the thematic concentration of the text, the cohesion of the text, one can set up the graph of the text, etc.

There is no problem with defining *verse-hrebs*. A verse-hreb contains all verses having a common entity which can be defined as morpheme, word, sign, synonym, or reference.

Hrebs have nothing common with syntactic dependence, they are semantic entities though in some languages they must be supported by morphology. The higher the unit we take into account, the smaller number of hrebs will be obtained, and consequently, the indicators of text properties will take on smaller values. One may conjecture that the number of hrebs increases linearly with lowering the hierarchic level of the analysis. This conjecture must, of course, still be tested but it explains the discrepancy between the phrase-based and word-based denotative analyses (Köhler, Naumann 2007 vs. Ziegler, Altmann 2002).

According to the information and ordering of entities, Ziegler and Altmann (2002: 31) defined five kinds of hrebs:

(1) Data-hreb containing the raw data, e.g. words, and the position of each unit in text, symbolized with ().

(2) List-hreb containing the data but without the positions of the units, symbolized with [].

(3) Set-hreb containing only the lemmas, morphemes, phrase heads etc., symbolized with {}.

(4) Ordered set-hreb is identical with (3) but the units are ordered according to a certain principle, e.g. alphabetically, or according to length, frequency, etc.

(5) Ordered position-hreb containing only the positions of units in the given text, symbolized with <>.

In the sequel we shall analyze some poems by the Romanian poet Mihai Eminescu in order to see some aspects of the semantic structure of his texts. We choose here the word-hreb, with the possibility of morphological reference, and the Data-hreb and Set-hreb type description of hrebs.

The hrebs will be constructed using the following rules.

1. Words can belong to one or more hrebs. For example verbs with personal ending belong both to the given verb and to the person (subject) they overtly refer to.
2. References belong to the hreb of the word they refer to, e.g. pronouns and proper nouns belong to the basic word.
3. Synonyms constitute a common hreb.
4. Articles and prepositions are parts of the noun phrase and are not considered as separate hrebs.
5. Adverbs may coincide with adjectives, e.g. in German *schön*, and may belong to the same hreb.

For some Eminescu’s poems we will describe the Data-hrebs and/or Set-hrebs below.

2. Rules of hreb formation for the Romanian language

The Rules for hrebs are of the form: “a ∈ B”. Here “a” is an expression containing a special element called *POS* indicator which is written in italic (*POS* is for “part of speech”). More exactly, the Rule "a ∈ B" means: "a" (or *POS* indicator of "a") is an element of hreb "B". The connection between "a" and "B" will result from the word used for the *POS* indicator (written in italic).

Since a word-form could be contained in more than one hreb, in the application of rules it is possible to obtain a result as: "a ∈ B,C,…" meaning: "a" is an element of all the hrebs in that enumeration: "B" and "C" and …

The rules are valid only for nouns, verbs, adjectives, adverbs, pronouns. Thus the *POS* indicator in "a" could be only one of the following types of words: *noun, verb, adjective, adverb, pronoun*.

RULES:

R1. *verb* ∈ VERB
R2. personal ending of the *verb* (noun or pronoun) ∈ NOUN or PRONOUN
R3. synonym of a verb ∈ VERB
R4. *pronoun* referring to a noun ∈ NOUN
R5. non-referring *pronoun* ∈ PRONOUN
R6. *noun* ∈ NOUN
R7. synonym of a *noun* ∈ NOUN
R8. article + *noun* ∈ NOUN
R9. preposition + *noun* ∈ NOUN
R10. *adjective* ∈ ADJECTIVE
R11. synonym of an *adjective* ∈ ADJECTIVE
R12. *adverb* ∈ ADVERB
R13. synonym of an *adverb* ∈ ADVERB
R14. compound word: w1w2...wn ∈ W1; W2;....Wn

The rules R1-R14 could be summarized as follows: a noun, its synonyms, referring pronouns and personal endings in a verb belong all to the hreb of the given noun; a verb in all its forms, its synonyms belong to the hreb of the given verb, however, the personal endings belong also to the hreb of the respective noun (or pronoun); an adjective (adverb) and its synonyms belong all to the hreb of the given adjective (adverb).

We illustrate the rules as applied to the poem *Lacul* presented below.

Lacul (tokens)

Lacul (1) codrilor (2) albastru (3)
Nuferi (4) galbeni (5) îl (6) încarcă (7);
Tresărind (8) în cercuri (9) albe (10)
El (11) cutremură (12) o barcă (13).

Şi eu (14) trec (15) de-a lung (16) de maluri (17),
Parc-ascult (18) şi parc-aştept (19)
Ea (20) din trestii (21) să răsară (22)
Şi să-mi (23) cadă (24) lin (25) pe piept (26);

Să sărim (27) în luntrea (28) mică (29) ,

Îngânați (30) de glas (31) de ape (32),
Şi să scap (33) din mână (34) cârma (35),
Şi lopețile (36) să-mi (37) scape (38);

Să plutim (39) cuprinşi (40) de farmec (41)
Sub lumina (42) blândei (43) lune (44
Vântu-n (45) trestii (46) lin (47) foşnească (48),
Unduioasa (49) apă (50) sune (51)!

Dar nu vine (52)... Singuratic (53)
În zadar (54) suspin (55) şi sufăr (56)
Lângă lacul (57) cel albastru (58)
Încărcat (59) cu flori (60) de nufăr (61).

Applying the above mentioned rules the following relations are obtained:
lacul $\in$ LAC (R6)
codrilor $\in$ CODRU (R6)
albastru $\in$ ALBASTRU (R10)
nuferi $\in$ NUFĂR (R6)
galbeni $\in$ GALBEN (R10)
îl $\in$ LAC (R4)
încarcă $\in$ A ÎNCĂRCA; NUFĂR (R1,R2)
tresărind $\in$ A TRESĂRI; LAC (R1,R2)
în cercuri $\in$ CERC (R9)
albe $\in$ ALB (R10)
el $\in$ LAC (R4)
cutremură $\in$ A CUTREMURA; LAC (R1,R2)
o barcă $\in$ BARCĂ (R8)
eu $\in$ EU (R5)
trec $\in$ A TRECE; EU (R1,R2)
de-a lung $\in$ LUNG (R9)
de maluri $\in$ MAL (R9)
parc-ascult $\in$ A PĂREA;
A ASCULTA; EU (R1,R1, R2, R14)
parc-aştept $\in$ A PĂREA;
A AŞTEPTA; EU (R1,R1, R2, R14)
ea $\in$ EA (R5)
din trestii $\in$ TRESTIE (R9)
să răsară $\in$ A RĂSĂRI; EA (R1,R2)
să-mi cadă $\in$ A CĂDEA; EA; EU (R1, R2, R5,R14)
lin $\in$ LIN (R12)
pe piept $\in$ PIEPT (R9)
să sărim $\in$ A SĂRI; NOI (R1, R2)
în luntrea $\in$ BARCĂ (R9, R7)
mică $\in$ MIC (R10)

îngânați ∊ ÎNGÂNAT (R10)
de glas ∊ GLAS (R9)
de ape ∊ APĂ (R9)
să scap ∊ A SCĂPA;EU (R1,R2)
din mână ∊ MÂNĂ (R9)
cârma ∊ CÂRMĂ (R6)
lopeţile ∊ LOPATĂ (6)
să-mi scape ∊ A SCĂPA; EU (R1, R2, R5, R14)
să plutim ∊ A PLUTI; NOI (R1, R2)
cuprinşi ∊ CUPRINS (R10)
de farmec ∊ FARMEC (R9)
sub lumina ∊ LUMINĂ (R9)
blândei ∊ BLÂNDĂ (R10)
lune ∊ LUNĂ (R6)
vântu-n ∊ VÂNT (R6)
trestii ∊ TRESTIE (R6)
lin ∊ LIN (R12)
foşnească ∊ A FOŞNI; VÂNT (R1,R2)
unduioasa ∊ UNDUIOASĂ (R10)
apa ∊ APĂ (R6)
sune ∊ A SUNA; APĂ (R1, R2)
nu vine ∊ A VENI; EA (R1, R2)
singuratic ∊ SINGURATIC (R10)
în zadar ∊ ZADAR (R9)
suspin ∊ A SUSPINA; EU (R1, R2)
sufăr ∊ A SUFERI; EU (R1, R2)
lânga lacul ∊ LAC (R9)
cel albastru ∊ ALBASTRU (R10)
încărcat ∊ ÎNCĂRCAT (R10)
cu flori ∊ FLOARE (R9)
de nufăr ∊ NUFĂR (R9)

In the following we will denote by *n* the number of hrebs in the studied poem. The data-hrebs (where all the positions in text are displayed) are presented in Table 1. The symbol ∅ is used as the sign of the zero-morpheme.

Table 1
Data-hrebs of the poem *Lacul* (*n* = 51)

Hreb	**Elements**	**Size of data-hreb**	**Size of set-hreb**
EU	(eu 14, trec∅ 15, parc-ascult∅ 18, parc-aştept∅ 19, să scap∅ 33, să-**mi** 23, să-**mi** 37, suspin∅ 55, sufăr∅ 56)	9	8

LAC	(lacul 1, îl 6, tresărind∅ 8, el 11, cutremur-**ă** 12, lacul 57)	6	5
EA	(ea 20, să rasar-**ă** 22, să-mi cad-**ă** 24, nu vin-**e** 52)	4	4
NUFĂR	(nuferi 4, încarc-**ă** 7, de nufăr 61)	3	2
APĂ	(de ape 32, apă 50, sun-**e** 51)	3	2
NOI	(să săr-**im** 27, să plut-**im** 39)	2	2
BARCĂ	(o barcă 13, în luntrea 28)	2	2
TRESTIE	(din trestii 21, trestii 46)	2	1
ALBASTRU	(albastru 3, cel albastru 58)	2	1
A PĂREA	(parc-ascult 18, parc-aştept 19)	2	1
LIN	(lin 25, lin 47)	2	1
A SCĂPA	(să scap 33, să-mi scape 38)	2	1
A ASCULTA, A AŞTEPTA, A ÎNCĂRCA, A CĂDEA, A CUTREMURA, A FOŞNI, A PLUTI, A RĂSĂRI, A SĂRI, A SUFERI, A SUNA, A SUSPINA, A TRECE, A TRESĂRI, A VENI, ALB, BLÂNDĂ, CÂRMĂ, CERC, CODRU, CUPRINS, FARMEC, FLOARE, GALBEN, GLAS, ÎNCĂRCAT, ÎNGÂNAT, LOPATĂ, LUMINĂ, LUNĂ, LUNG, MAL, MÂNĂ, MIC, PIEPT, SINGURATIC, UNDUIOASĂ, VÂNT, ZADAR (all occurring once and thus having only one element)			

For the poem *Peste vârfuri* a result similar to that in Table 1 will be presented in Table 2.

Table 2
Data-hrebs for the poem *Peste varfuri* (*n* = 26)

Hreb	**Elements**	**Size of data-hreb**	**Size of set-hreb**
CORN	(cornul 11, sun-**ă** 12, Îndulcind**Ø** 19, suna-**vei** 27, corn 28)	5	3
EU	(-**mi** 17, -**mi** 24, -ntorn**Ø** 26, **mine** 30)	4	2
LUNĂ	(trec-**e** 2, lună 3)	2	2
CODRU	(codru-şi 4, bat-**e** 5)	2	2
TU	(tac-**i** 22, **tine** 25)	2	2
SUFLET	(sufletu-mi 17, inima-mi 24)	2	2
A SUNA	(sună 12, suna-vei 27)	2	1
A BATE, A ÎNDULCI, A ÎNTOARCE, A TĂCEA, A TRECE, ARIN, DEPARTE, DOR, DULCE, FERMECAT, FRUNZĂ, ÎNCET, LIN, MELANCOLIC, MOARTE, NEMÂNGÂIAT, RAMURĂ, VÂRF, VREODATĂ (all occurring once and thus having only one element)			

Table 3 contains the data-hrebs for the poem *Dintre sute de catarge,* where the positions are indicated only for the first and the last element.

Table 3
Data–hrebs for the poem *Dintre sute de catarge* (n = 26)

Hreb	**Elements**	**Size of data-hreb**	**Size of set-hreb.**
VÂNT	(vor sparg-**e**(5), vânturile, o să le-nec-**e**, vânturile, urmeaz-**ă**, vânturile, zboar-**ă**, îngânând**Ø**, vânturile(33))	9	6
VAL	(vor sparg-**e**(5), valurile, o să le-nec-**e**, valurile, urmeaz-**ă**, valurile, zboar-**ă**, îngânând**Ø**, valurile(32))	9	6
GÂND	(rămân-**e**(24), străbat-**e**, gândul, îngânându-**l**(31))	4	4
CATARG	(catarge(2), las-**ă**, **le**(4))	3	3
TU	(de-**i** goni(15), te, ce-**ți**(26))	3	2
PASĂRE	(păsări(8), străbat**Ø**, o să **le**-nece(12))	3	3
A STRĂBATE	(străbat (10), străbate (27))	2	1
A GONI, A ÎNGÂNA, A ÎNNECA, A LĂSA, A RĂMÂNE, A SPARGE, A STRĂBATE, A URMA, A ZBURA, CĂLĂTOR, CÂNT, DEAL, LOC, MAL, NEÎNȚELES, NOROC, PĂMÂNT, SUTĂ, VECINIC (all occurring once and thus having only one element)			

Some other poems have been analysed too, but they will not be presented in details. For some results, see Table 4.

Table 4
Sizes of data-hrebs in some poems

Poem	Sizes of data-hrebs
Lacul	9,6,4,3,3,2,2,2,2,2,2,2,(39)1
Dintre sute de catarge	9,9,4,3,3,3,2, (19)1
La mijloc de codru	6,3,2,(22)1
Pe langă plopii fără soț	19,16,5,4,3,3,3,3,3,3,2,2,2,2,2,2,2,2,2,2,2,2,2,2,(58)1
Peste vârfuri	5,4,2,2,2,2,2,(19)1
Somnoroase păsărele	5,4,3,3,3,2,2,2,2,(26)1
Atât de fragedă	23,15,4,4,3,3,3,3,2,2,2,2,2,2,(64)1
La steaua	7,4,4,4,4,3,3,2,2,2,(20)1
Trecut-au anii…	10,7,4,3,3,3,3,3,2,2,2,(24)1
Ce te legeni?	11,10,4,3,3,2,2,2,2,2,2,2,(21)1
Mai am un singur dor	17,6,5,3,3,3,3,2,2,2,2,2,2,2,2,2,2,2,2,2,35(1)

3. Ranking of hrebs

As it is well known, word-forms or lemmas, if ranked according to their frequency, follow some ranking law. The law, called Zipf's law, has a great number of forms (models). Here, we shall adhere to the proposal of Popescu et al. (2010) based on the stratification of the text. It has been shown that the text is stratified according to all entities occurring in it. This is caused by the fact that some entities are "basic", other ones are concomitant, dependent, complementary, etc. Every text is stratified in different points of view.

This is why ranking can be captured using a formula consisting of several components differing only by parameters. According to Popescu et al. (2010) the frequency of an entity having rank r is given by

(1) $f = 1 + a\text{*exp}(-r/b) + c\text{*exp}(-r/d) + \ldots$

where the number of exponential components shows the extent of stratification. Here f is the data-hreb size, r is the rank, and $a,b,c,d,\ldots$ are parameters which must be obtained from the data. Since here we are concerned with hrebs, this is a kind of semantic ranking of the entities of text.

We illustrate the procedure using the size of data-hrebs of *Lacul* (Table 1). The ranked data are

1	2	3	4	5	6	7	8	9	10	11	12	13-51
9	6	4	3	3	2	2	2	2	2	2	2	1

Computing the parameters iteratively we obtain

$$f = 1 + 11.2417\text{*exp}(-r/1.1849) + 3.8644\text{*exp}(-r/5.5174)$$

yielding a determination coefficient $R^2 = 0.9843$. The poem has two semantic strata whose identification must be left to literary scientists.

In Table 5 the results for some other poems are presented.

Table 5
Semantic stratification of some poems

Poem	***a***	***b***	***c***	***d***	**R^2**
Lacul	11.2417	1.1849	3.8644	5.5174	0.98
La mijloc de codru	12.6744	1.0789	-	-	0.99
Pe lângă plopii fără soț	32.5540	1.5404	2.3390	14.7647	0.94
Peste vârfuri	3.6282	1.0781	3.5985	3.2739	0.94
Somnoroase păsărele	5.0416	4.0443	-	-	0.96
Atât de fragedă	45.2550	1.2411	2.7289	8.2684	0.97
La steaua	6.6639	4.4656	-	-	0.93
Trecut-au anii…	13.1228	0.9819	5.3249	5.1912	0.97
Ce te legeni?	15.8981	1.8863	1.5679	7.5631	0.94
Mai am un singur dor	70.5681	0.5804	3.7767	8.9517	0.99

The advantage of this kind of capturing the data is the fact that one can detect the number of strata mechanically: if two exponents are approximately equal, then one of the components is redundant and can be omitted (e.g. in *La mijloc de codru* etc.).

We see that in spite of the brevity of some poems there are mostly two semantic strata. That means that there are two (or more) semantic focuses.

4. Denotational properties of texts

Creating hrebs means a reduction of the text to its fundamental semantic components. Having defined them one can make statements both about the text and the hrebs themselves and obtain new indicators.

4.1. Topicality

If a set-hreb contains at least two elements, it belongs to the *Kernel* of the text. That is, if $|\{hreb_i\}| \geq 2$, where $\{hreb_i\}$ is the set representation of $hreb_i$, then $hreb_i \in Kernel$. The hrebs of a kernal will be called *kernel-* hrebs. The size of the *Kernel*, denoted by $|Kernel|$, represents the number of elements of hrebs in the *Kernel* (the sum of the sizes of kernel-hrebs). In the poem *Peste vârfuri* there are *kernel*-hrebs with sizes 3,2,2,2,2,2, hence the size of the *Kernel* is

|*Kernel*(*Peste varfuri*)| = 3+2+2+2+2+2 = 13.

The contribution of a hreb to the kernel can be called *topicality* of a hreb. It can be computed as a simple proportion

$$(2) \quad T(hreb_i) = \frac{|\{hreb_i\}|}{|Kernel|}$$

where $\{hreb_i\}$ is the set-hreb representation of $hreb_i$, e.g. *T*(CORN) = 3/13 = 0.23. All other hrebs in *Peste vârfuri* have the topicality 2/13 = 0.15.

The *Kernel* itself takes a special place in the text. It may contain more or fewer herbs which can have different topicality. The weight of the *Kernel* in the text is called *Kernel concentration* (*KC*), defined as the size of the *Kernel* divided by the number of hrebs in the texts (*n*), i.e.

$$(3) \qquad KC = \frac{|Kernel|}{n}$$

In the poem *Peste vârfuri* we obtained |*Kernel*| = 13, and there are *n* = 26 hrebs, hence

KC(*Peste vârfuri*) = 13/26 = 0.5.

It is to be mentioned that this is not a simple proportion because the |*Kernel*| can be greater than *n*. The greater *KC*, the more is the text concentrated to a small number of thematic words.

For some analyzed texts we obtain *KC* as shown in Table 6.

Table 6
Kernel concentration of some poems

Poem	|*Kernel*|	*n*	*KC*
Atât de fragedă	38	78	0.48
Ce te legeni?	33	33	1.00
Dintre sute de catarge	24	26	0.92
La mijloc de codru	6	25	0.24
La steaua	29	30	0.97
Lacul	25	51	0.49
Mai am un singur dor	46	55	0.84
Pe lângă plopii fără soț	46	82	0.56
Peste vârfuri	13	26	0.50
Somnoroase păsărele	17	35	0.48
Trecut-au anii…	31	35	0.89

As can be seen, several poems are outliers in different directions: while *La steaua* is strongly concentrated concerning set-hrebs, *La mijloc de codru* is extremely non-concentrated. Strong concentration means much synonymy and references, while small *KC* means different contents of lines, same name for the same thing, reduced references. Of course, poems of this sort must be scrutinized individually.

4.2. Text concentration

The concentration of the text can be measured in different ways. Here we use only the Repeat rate applied to data-hrebs. The data-hrebs contain all respective tokens of the text, e.g. for the poem *Lacul* we have (using Table 4)

9, 6, 4, 3, 3, 2, 2, 2, 2, 2, 2, 2, 39(1)

The *Repeat rate* is defined as

(4) $$R = \sum_{i=1}^{n} p_i^2 = \frac{1}{N^2}\sum_{i=1}^{n} f_i^2$$

where *n* is the number of hrebs, *N* is the number of tokens in all hrebs, and f_i is the size of the data-hreb. Computing the above expression for *Lacul* with *n* = 51 and *N* = 78 we obtain

$R(Lacul) = (9^2 + 6^2 + 4^2 + (2)3^2 + (7)2^2 + 39(1^2))/78^2 = 218/78^2 = 0.0358.$

R lies in the interval <1/*n*, 1>. In order to obtain the relative *R* in <0,1>, we use the MacIntosh version

$$(5)\qquad R_{rel} = \frac{1-\sqrt{R}}{1-1/\sqrt{n}}$$

and obtain for *Lacul*

R_{rel} = (1 - √0.0358)/(1 - 1/√51) = 0.9427

The sizes of data-hrebs of some individual poems are presented in Table 4, the values of *R* for these poems are presented in Table 7.

Table 7
Repeat rate of list hrebs in some poems (ordered by *n*)

Poem	***n***	***N***	***R***	R_{rel}	$Var(R_{rel})$
La mijloc de codru	25	33	0.0652	0.9308	0.00243
Peste vârfuri	26	38	0.0554	0.9512	0.00107
La steaua	30	55	0.0539	0.9394	0.00065
Ce te legeni?	33	66	0.0698	0.8909	0.00125
Trecut-au anii…	35	66	0.0565	0.9174	0.00091
Somnoroase păsărele	35	52	0.0407	0.9607	0.00046
Lacul	51	78	0.0358	0.9427	0.00056
Mai am un singur dor	55	101	0.0464	0.9070	0.00091
Atât de fragedă	78	134	0.0507	0.8738	0.00076
Pe lângă plopii fără soţ	82	148	0.0377	0.9058	0.00048

If a poem is concentrated to some few hrebs, the *Repeat rate R* is great and we can speak of semantic richness. However, the vocabulary richness is rather small. Hence if R_{rel} is great, the vocabulary richness is great and the text is not very concentrated semantically. It can be seen that the number of hrebs does not correlate with R_{rel}.

For the comparison of two poems for their relative *Repeat rate* repressenting text concentration and semantic richness one can use an asymptotic normal test. To this end we need the variance of R_{rel} which can be derived in form (cf. Altmann, Lehfeldt 1980: 160; Ziegler, Altmann 2002: 53)

$$(6)\qquad Var(R_{rel}) = \frac{n}{NR(\sqrt{n}-1)^2}\left(\sum_i p_i^3 - R^2\right).$$

For the computation one needs the sum of p^3. This can be done from using the sizes of list-hrebs, e.g. for *La mijloc de codru* we have

[6,3,2,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1] hence

$$\frac{1}{N^3}\sum_{i=1}^{n} f_i^3 = (6^3 + 3^3 + 2^3 + 22(1^3))/33^3 = 0.007596$$

and

$$Var(R_{rel}) = \frac{25}{33(0.0652)(\sqrt{25}-1)^2}(0.007596 - 0.0652^2) = 0.00243.$$

Using the variances one can set up the normal test

$$u = \frac{|R_{rel,1} - R_{rel,2}|}{\sqrt{Var(R_{rel,1}) + Var(R_{rel,2})}}.$$

For the sake of illustration let us compare the R_{rel}s of *La mijloc de codru* and *Atât de fragedă* using the numbers contained in Table 5. We obtain

$$u = \frac{|0.9308 - 0.8738|}{\sqrt{0.00243 + 0.00076}} = 1.01,$$

which is not significant. Using this procedure one could detect the semantic outliers.

4.3. Hreb and text diffuseness

If the elements of a hreb occur on different places in text, we may suppose that after a pause the poet returns to the same semantic entity. However, the elements can be placed in a short distance from one another, and the given hreb has only a local role. If the distance between the first and the last occurrence is great, we may speak of diffusity of the hreb.

In order to measure this property, we first localize the places of its elements in text. Then the diffuseness of the given data-hreb is defined using the maximal and minimal position of tokens occuring in it. If we denote the set of positions of tokens of a given hreb H_p by $<H_p>$ then the *diffuseness* of H_p is :

$$(7) \qquad D_{H_p} = \frac{\sup < H_p > - \inf < H_p >}{|H_p|},$$

i.e. the difference of the last and the first position divided by the cardinal number of the data-hreb. For example in the poem *Lacul* the positions of the hreb elements of the data-hreb LAC are:

LAC = (lacul (1), il (6), tresarind (8), el (11), cutremur-**ă** (12), lacul (57))

There are 6 elements in the data-hreb. Hence the diffuseness of LAC is given as

$$D(\text{LAC}) = (57 - 1)/6 = 9.33.$$

In this way one can compute the diffuseness of all data-hrebs whose cardinality is greater than 1. Adding all diffusenesses and dividing by the number of pertinent hrebs one can obtain the *Mean diffuseness of the text* as

$$(8) \qquad \bar{D}_{text} = \frac{1}{K}\sum_{i=1}^{K} D_j$$

where *K* is the number of data-hrebs with more than one element. For the poem *Lacul* we obtain the results in Table 8.

Table 8
Computing the mean diffuseness of *Lacul*

Hreb	Size	*D*
ALBASTRU = (albastru (3), cel albastru (58))	2	27.5000
NUFĂR = (nuferi (4), încarc-ă (7), de nufăr (61))	3	19.0000
TRESTIE = (din trestii (21), trestii (46))	2	12.5000
LIN = (lin (24),lin (47))	2	11.0000
LAC = (lacul (1), îl (6), tresărind (8), el (11), cutremur-ă (12), lacul (57))	6	9.3333
EA = (ea (20), să răsar-ă (22), să-mi cad-ă (24), nu vin-e (52))	4	8.0000
BARCĂ =(o barcă (13), în luntrea (28))	2	7.5000
APĂ = (de ape (32), apa (48), sun-e (51))	3	6.3333
NOI = (să săr-im (27), să plut-im (39))	2	6.0000
EU = (eu (14), trec (15), parc-ascult (18), parc-aştept (19), să scap (33), să-mi (23), să-mi (37), suspin (55), sufăr (56))	9	4.6666
A SCĂPA = (să scap (33), să-mi scap-e (38))	2	2.5000
A PĂREA =(parc-ascult (18), parc-aştept (19))	2	0.5000

The hrebs ordered according to decreasing diffuseness show that the strongest reminiscence in which the whole poem is wrapped has the hreb ALBASTRU (*blue*), but LAC (*lake*) which is the topic of the poem is dispersed over the whole text and, even if it displays the greatest distance, it has a smaller diffuseness. The meaning of diffuseness can further be interpreted psycholinguistically.

In order to obtain the mean diffuseness of the poem, we add the individual values (last column of Table 8) and divide the sum by the number of diffused hrebs (here 12). For *Lacul* we obtain

$$\bar{D}(Lacul) = (27.5000 + 19.0000 + 12.5000 + 11.0000 + 9.3333 + 8.0000 + \\ + 7.5000 + 6.3333 + 6.0000 + 4.6666 + 2.5000 + 0.5000)/12 = 9.5694$$

Other results are presented in Table 9.

Table 9
Text diffuseness in some poems

Poem	***n***	$\bar{D}$
La mijloc de codru	25	3.5533
Somnoroase păsărele	35	4.6755
Peste vârfuri	26	2.8757
La steaua	30	7.7100
Ce te legeni?	33	6.0966
Trecut-au anii…	35	7.4180
Lacul	51	9.5694
Mai am un singur dor	55	11.8523
Atât de fragedă	78	10.6121
Pe lângă plopii fără soț	82	8.8224

Ordering the texts according to the number of hrebs, one can easily see that mean $\bar{D}$ slowly increases. Though any statement would be preliminary, we found that the Zipf-Alekseev function given as

$$\bar{D} = 0.00000000102313{*}n^{11.2860469 - 1.37923368{*}\ln n}$$

is the first good approximation to this regularity ($R^2 = 0.88$), even if the first parameter is very strange.

Mean $\bar{D}$ is also an indicator of a kind of recall intensity and later on, when a number of different texts has been analyzed in this way, it will be possible to use it also for qualitative interpretations.

4.4. Text compactness

Another way of measuring semantic text concentration is the expression of its *compactness*. The smaller the number of hrebs, the more compact is the text. This fact can be expressed using the indicator

$$C = \frac{1 - \frac{n}{N}}{1 - \frac{1}{N}}$$

where *n* is the number of hrebs in the text and *N* is the number of tokens in all hrebs or better, hreb-tokens. Taking for example the poem *Lacul* containing 51 hrebs and 78 hreb-tokens we obtain

$C(Lacul) = (1 - 51/78)/(1-1/78) = 0.3506$

This indicator varies in interval <0,1>. Zero-compactness means at the same time small thematic concentration (all expressions belong to separate hrebs), great hreb-richness, and eo ipso great vocabulary richness. Maximal compactness (with $n = 1$) means that all expressions are elements of the same unique hreb. Though this situation is possible only in dada-poetry, we must take it into account. On the other hand, great compactness means at the same time great thematic concentration.

Using the results in Table 7 we compute *C* for the other texts and present them in Table 10.

Table 10
Text compactness of some poems

Poem	***n***	***N***	***C***
La mijloc de codru	25	33	0.25
Dintre sute de catarge	26	52	0.51
Peste vârfuri	26	38	0.32
Somnoroase păsărele	35	52	0.39
Lacul	51	78	0.35
Atât de fragedă	78	134	0.42
Pe lângă plopii fără soț	82	148	0.45
La steaua	30	55	0.46
Trecut-au anii…	35	66	0.48
Ce te legeni?	33	66	0.51
Mai am un singur dor	55	101	0.46

As can be seen, only two poems have $C > 0.5$, that is, they tend to a smaller compactness, smaller thematic concentration, greater hreb-richness (in all but one case $n > N/2$).

4.5. Thematic concentration

Finally, let us consider the concept of *thematic concentration* for hrebs by similar arguments as considered for word-forms (Popescu, Altmann, 2009: Chapter 6). The basic formula is

$$TC = 2\sum_{r'=1}^{T} \frac{(h-r')f(r')}{h(h-1)f(1)}$$

where *TC* is this time the thematic concentration of hrebs, *h* = the *h*-point of data-hrebs rounded to integer, *r´* = ranks of thematic data-hrebs (*r'* < rounded *h*), and *T* = total number of data-hreb elements in the pre-*h* domain. Table 11 displays the hrebs-*TC* numerical data for the same poems as considered above.

Table 11
Thematic concentration for hrebs of some poems

Poem (alphabetically)	**# Data-hrebs (*N*)**	***h***	**Pre-*h* elements**	***TC***	**Diffuseness** $\bar{D}$
Atât de fragedă	134	4	23,15,4,4	0.746	106.121
Ce te legeni?	66	4	11,10,4,3	0.864	60.966
La mijloc de codru	33	3	6,3,2	0.833	35.533
La steaua	55	4	7,4,4,4	0.786	77.100
Lacul	78	4	9,6,4,3	0.796	90.648
Mai am un singur dor	101	4	17,6,5,3	0.667	118.523
Pe lângă plopii fără soț	148	4	19,16,5,4	0.825	88.224
Peste vârfuri	38	3	5,4,2	0.933	28.757
Somnoroase păsărele	52	3	5,4,3	0.933	46.755
Trecut-au anii…	66	4	10,7,4,3	0.800	74.180

One may remark the high correlation between hrebs thematic concentration and text diffuseness, as illustrated in Figure 1 (where the outlier *La mijloc de codru* has been skipped by fitting).

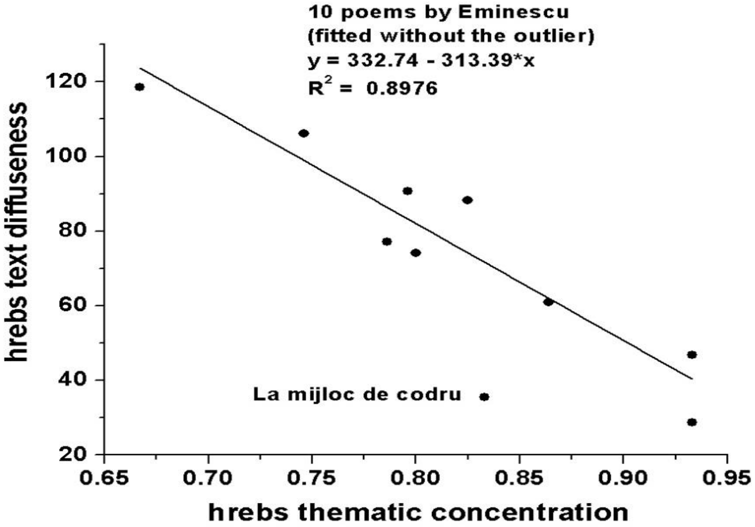


Figure 1. Hrebs thematic concentration and text diffuseness relationship

5. Hrebs vs. Lexical and Coreference Chains

In Tatar et al. (2013) the relationship between a Cohesion Chain (CC), defined as a Lexical Chain or a Coreference Chain, on one hand, and a hreb, on the other hand (more exactly a slightly modified kind of word-hrebs, quasi-hrebs) is presented. Lexical Chains are sequences of words which are in a lexical cohesion relation (synonymy, repetition, hypernymy, hyponymy, etc) with each other.

Lexical cohesion relationships between the words of Lexical Chains are established using an auxiliary knowledge source such as a dictionary or a thesaurus. Coreference Chains are chains of antecedents-anaphors. Lexical Chains and Coreference Chains (Cohesion Chains denoted by CCs) are intensively studied in Computational Linguistics, but few indicators are standard for them.

The indicators inspired from the hrebs could be studied and adopted for CCs, improving some application of CCs as for example Text segmentation and Text summarization. On the other hand, for Lexical Chains and Coreference Chains there exist at the moment numerous software tools. Thus, studying the relationship between CCs and hrebs could bring some benefits for both of these concepts.

In Tatar et al. (2013) it is shown how CCs could be obtained from the data-hrebs. Let us describe shortly the procedure consisting firstly in calculating a slightly modified version of hrebs, the quasi-hrebs. From the set of rules R1-R14, the rule R2 makes the difference when the quasi-hrebs are calculated. This rule is reproduced here:

R2. personal ending of the *verb* (noun or pronoun) $\in$ NOUN or PRONOUN

Applying all the rules, excepting R2, the quasi-hrebs (and the sizes of data quasi-hrebs) calculated from the poem *Lacul* are provided in Table 12.

Table 12
Quasi-hrebs of the poem *Lacul* (n = 51)

Quasi-hreb	**Elements**	**Size of data-quasi-hreb**	**Size of set-quasi-hreb**
EU	(eu 14, -mi 23, -mi 37)	3	2
LAC	(lacul 1, îl 6, el 11, lacul 57)	4	3
EA	(ea 20)	1	1
NUFĂR	(nuferi 4, nufăr 61)	2	1
APĂ	(de ape 32, apă 50)	2	1
BARCĂ	(barcă 13, luntrea 28)	2	2
TRESTIE	(trestii 21, trestii 46)	2	1
ALBASTRU	(albastru 3, albastru 58)	2	1
A PĂREA	(parc- 18, parc- 19)	2	1
LIN	(lin 25, lin 47)	2	1
A SCĂPA	(scap 33, scape 38)	2	1

As a remark, the hreb NOI has not a corresponding quasi-hreb, because both elements (sărim 27, plutim 39) are obtained by Rule R2.

Examining Table 12 of quasi-hrebs, we observe that: the quasi-hreb EU corresponds to a Coreference Chain (eu 14, -mi 23, -mi 37), the quasi-hreb LAC to a Coreference Chain (lacul 1, îl 6, el 11, lacul 57). The quasi-hreb EA is not a chain (it has only one element). The rest of quasi-hrebs represents Lexical Chains: (nuferi 4, nufăr 61), (ape 32, apă 50), (barcă 13, luntrea 28), (trestii 21, trestii 46), (albastru 3, albastru 58), (parc- 18, parc- 19), (lin 25, lin 47), (scap 33, scape 38). Thus, Lexical Chains and Coreference Chains of the poem are exactly the quasi-hrebs with the size of data representation equal or greater than 2.

For CCs the indicators inspired from the hrebs must be studied and adopted. For example, there is a large debate about how to select CCs to construct the summaries of a text: selecting long or short CCs is one of the questions. Using only kernel CCs (defined as CCs with the size bigger than a given coefficient), or CCs with a high topicality and /or high diffuseness could be a solution.

6. Conclusion

Since hreb is a unit staying one step higher than the word in the hierarchy of units, it has all properties of the word and in addition some other ones whose examination would fill a whole book. We restrict ourselves to the above ones and want to attract the attention of researchers to the fact that there are already several definitions of hrebs and that hrebs, just as any other language units, do not have a sharp boundary, they are fuzzy. One may continue the research with the degree of anaphora and cataphora within the hreb, the graph-theoretical connection of all hrebs, the properties of these graphs, and check whether the theory of "small world" holds also for texts decomposed in hrebs. Here we must dispense with these possibilities but hope that the research will take this way.

References

Agricola, E. (1969). *Semantische Relationen im Text und im System*. Halle: Niemeyer.

Altmann, G., Lehfeldt, W. (1980). *Einführung in die quantitative Phonologie*. Bochum: Brockmeyer.

Brown, C., Yule, G. (1983). *Discourse analysis*. Cabridge: Cabridge University Press.

Christmann, C. (2004). *Denotative Textanalyse am Beispiel von Zeitungsartikeln*. Seminararbeit, Trier.

Halliday, M.A.K., Hassan, R. (1976). *Cohesion in English*. London: Longman.

Hoffmannová, J. (1996). Analýza diskurzu (ve světle nových publikací). *Slovo a slovesnost 57(2), 109-115*.

Hřebíček, L. (1985). Text as a unit and co-references. In: Ballmer, T.T. (ed.), *Linguistic dynamics: 190-198*. Berlin-New York: de Gruyter.

Hřebíček, L. (1992). *Text in communication: Supra-sentence structure*. Bochum, Brockmeyer.

Hřebíček, L. (1993). Text as a construct of aggregations. In: Köhler, R., Rieger, B. (eds.), *Contributions to quantitative linguistics. Dordrecht: Kluwer: 33-39.*

Hřebíček, L. (1995). *Text levels. Language constructs, constituents and Menzerath-Altmann law*. Trier: WVT.

Hřebíček, L. (1996). Word associations and text. *Glottometrika 15, 12-17.*

Hřebíček, L. (1997). *Lectures on text theory*. Prague: Oriental Institute.

Köhler, R., Naumann, S. (2007). Quantitative analysis of co-reference structures in texts. In:Grzybek, P., Köhler, R. (eds.), Exact *methods in the study of language and text: 317-329.* Berlin-New York: Mouton de Gruyter.

Levinson, S.C. (1983). *Pragmatics*. Cambridge: Cambridge University Press.

Numan, D. (1993). *Introducing discourse analysis*. London: Penguin.

Palek, B. (1988). *Referenční výstavba textu*. Praha: Univerzita Karlova.

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

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

Schiffrin, D. (1987). *Discourse markers*. Cambridge: Cambridge University Press.

Schwarz, C. (1995). The distribution of aggregates in text. *ZET – Zeitschrift für empirischeTextforschung 2, 62-66.*

Stubbs, M. (1983). *Discourse analysis*. Oxford: Blackwell.

Tatar, D., Lupea, M., Kapetanios, E. (2013). Hrebs and cohesion chains as similar tools for semantic text analysis. *Studia Universitatis Babes-Bolyai, Informatica, 58(2) , 40-52.*

Vater, H. (1994). *Einführung in die Textliguistik. Struktur, Thema und Referenz in Texten.* München: Fink.

Viehweger, D. (1978). Struktur und Funktion nominativer Ketten im Text. *Studia Grammatica 17, 149-168.*

Ziegler, A., Altmann, G. (2002). *Denotative Textanalyse*. Wien: Praesens.

Ziegler, A. (2005). Denotative Textanalyse. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 423-447.* Berlin-New York: de Gruyter.

Glottometrics 28, 2014, 56-74

Script complexity: *A Case Study*

Tomi S. Melka
Gabriel Altmann

Abstract. In the article, complexity is defined as the combination of *form, joining* and *level of strokes*. Two scripts, the very simple Celtic *ogham* along with the very complex *rongorongo* of Easter Island, are analyzed. Given the restricted number of *ogham* letters, they are all accounted for, while a large-scale scrutiny of *rongorongo* glyphs is subject to future tests. Complexity is not identical with distinctivity.

***Key words**: complexity measurement, iconicity, ogham, rongorongo script, stick-like signs*

1. Introduction

The concept of *complexity* applied to scripts has as many aspects as we are able to conceive: it is not a single definition. There are subjective and objective factors making a script *more* or *less* complex. The subjective aspects are: (a) the ease of script learning; there is a difference between Chinese - English - Slavic languages (Sampson, 1985; Lyovin, 1997; Rogers, 2005, pp. 185-195; Su & Samuels, 2010). As a script gets more complex, more time is required to learn it (Coulmas, 2009, p. 13). (b) The complexity of reproducing it: the smaller the number of rules that must be obeyed when writing, the simpler is the script (see also E. Pulgram, 1976 [1966], p. 15); classical Latin or *ogham* writing are surely simpler to be produced than English, or Chinese, though English has almost the same number of letters as Latin (see e.g. S. Knight, 1996, p. 313, Figure 43; H. Rogers, 2005; F. Coulmas, 2009, p. 13-14). (c) The first objective complexity factor is the time necessary to produce a text using a certain script; the Russian script is constructed in such a way that any word can be handwritten without raising the arm but this is not possible in all Slavic languages; in Chinese there is a prescribed way in which the strokes (components) of a sign are written with one movement, or along with an order (M. A. French, 1976 [1966]; W. Boltz, 1994; J. Myers, 1996, & R. W. Sproat, 2000: 48-49 on “headedness”); supra-segmental and prosodic features of the speech, e.g. diacritical signs like hyphen, apostrophe, palatalization signs, accents, cedillas, vowels’ points, etc., plus contextual shaping, render the script even more complex (DeFrancis, 1989, pp. 170-171; Correll, 2013; and Wikipedia, 2013, see Figure 1).

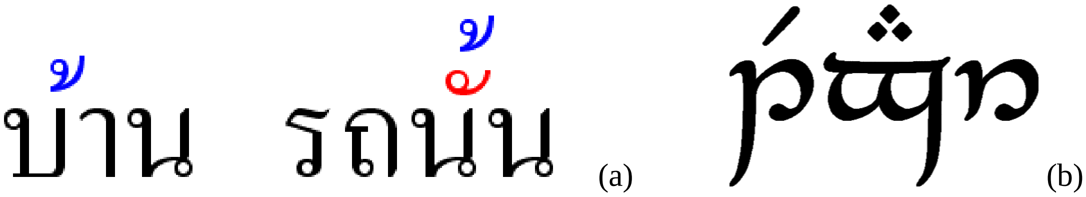


Figure 1. Stacking diacritics in Thai (a), after Correll (2013). The word “Tengwar” written in the artificial script Tengwar (b) invented by J.R.R. Tolkien contains script-like diacritic marks (Wikipedia, 2013; and Allan, 2002 [1978]).

(d) The sign number in a script: scripts using the same, say, letters, have the same degree of script complexity. As a general rule, the more iconic is the script, the more complex it is. But there are some further complications, e.g. in Japanese where one uses Chinese signs but the affixes and synsemantics are written in a syllabic script (*hiragana*), and, in addition, there is a second syllabic script (*katakana*) for children and for transcribing foreign words, e.g. erebētā (elevator). (e) However, the number of script components may be very small, e.g. in Assyrian or *ogham*, but their ordering may be very different: in cuneiform Assyrian one combines the arrows in different numbers, levels and directions rendering the script more complex than in *ogham* (cf. H. Arntz, 1935; J. Friedrich, 1971 [1957], pp. 35-40; G. Barthel, 1972; Walker, 1989; Lehmann, 1991 [1989]; S. Ziegler, 1994; A. Gaur, 2000, pp. 78-79). (f) In many scripts, the individual strokes (e.g. written with one movement, without changing the direction, etc.) are easily recognizable, in other ones, especially in printed form, one cannot speak of the number of movements, lifting the hand, etc., and must consider the form plus the joining of some lines. And just this is the point, at which objective measurement is possible, even if the scaling may be quite different. Some of the criteria may be inexact (e.g. length, angle, level, stroke's form) and relative, with scaling not depending on test persons. In this manner, the complexity of Maya or *rongorongo* glyphs (A. Robinson 2002: 134-136; T. Melka, 2012, p. 7, Footnote 3) can be estimated in a reproducible form. This last aspect lets us to also compute some other properties of the given script and examine the links between the properties.

If we consider only the quite evident properties of an alphabetic script, we obtain a preliminary control cycle presented in Altmann (2008) and shown in Figure 2.

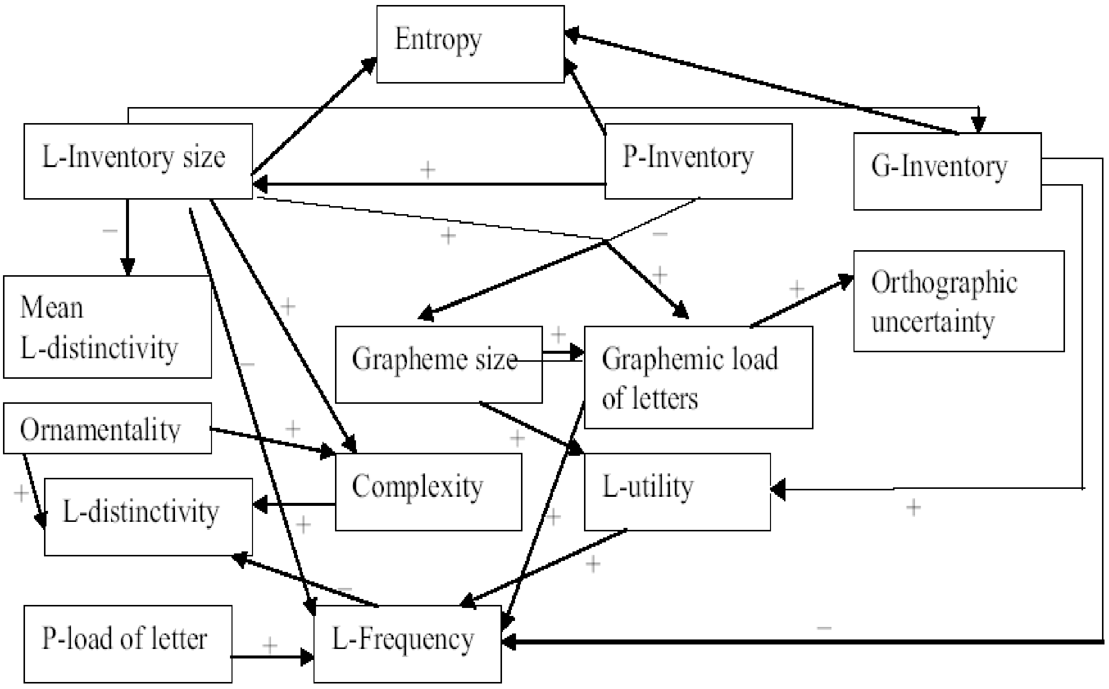
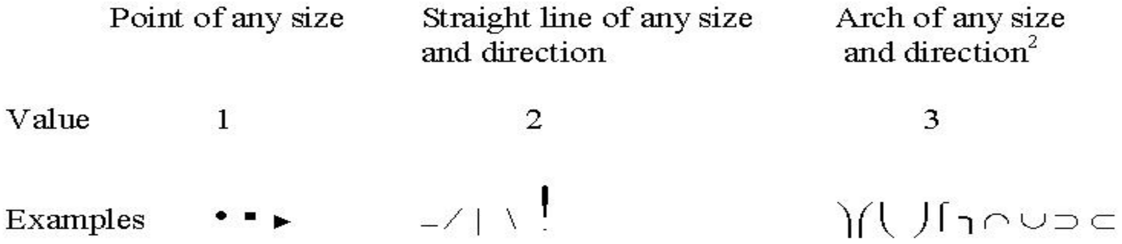


Figure 2. The control cycle of letter properties, based on G. Altmann (2008).

If one considers the complexity of any kind of signs, the above cycle may obtain a different form. Some of the properties must be omitted, some links formally expressed may take different parameters; the number of distinctive strokes will increase or decrease; one must introduce a finer scaling, etc. But we hope that after many script-types have been analyzed, some of the hypotheses mentioned in Altmann (2008) will be positively tested and obtain a very general form.

The simplest scaling can be performed by merely considering the form and the manner the graphical elements are joined. A more complex scaling takes also into account the position of a graphical element which may turn out to be distinctive, see e.g. the *ogham* script (Lehmann, 1991 [1989]; Ziegler, 1994; Ager, 1998-2013). According to a proposal of Altmann (2004), in the majority of cases the following scaling system is sufficient,



	Point of any size	Straight line of any size and direction	Arch of any size and direction[2]
Value	1	2	3
Examples	• ▪ ►	– / | \ !	)(( ) ∫ ┐ ⌒ ∪ ⊃ ⊂

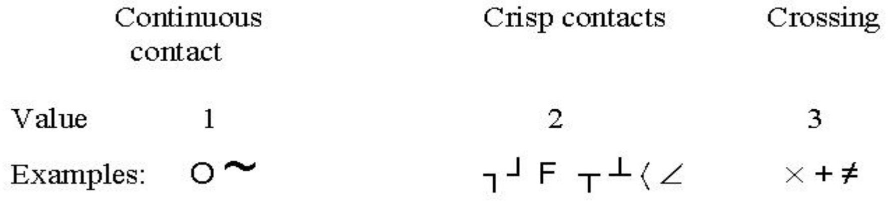
If the position is distinctive, one can add three vertical levels and three horizontal ones to obtain

Level	*Vertical*			*Horizontal*		
Value	1	2	3	1	2	3
Examples	___	___	‾‾‾	|…	…|…	…|

that is, the lowest (leftward) stroke in the sign obtains 1 score, the mid-point 2, and the upper (rightward) stroke 3. When the low and the upper strokes receive fewer or higher points is a matter of technical convenience (the order in which strokes should be retrieved), and assigned scores are interchangeable. Thus, the letter L has a left vertical stroke and a low horizontal stroke yielding 4 point for the two straight lines and 1 + 1 for their positions (+ one crisp contacts, see below). The letter E has 4 straight lines (8 points) and 1 + 2 + 3 for the vertical positions of three of them and 1 for the left position of the main stroke (+ 3 crisp contacts, see below).

Now, the strokes are not written separately but display some contact. There are three types of contacts: *continuous*, *crisp* and *crossing*. Using the continuous contact, the strokes' edges touch mutually in such a way that one of them continues in the "same" direction representing a turning point, or continues in the opposite direction. This can be represented by the sign "~" and by the "circle," or "O." The crisp contact means the existence of a point in which the strokes meet in a sharp manner, i.e. discontinuously, for instance in "E." The third possibility is crossing of two lines, e.g. in "X."

The examples of individual values and form are as follows:



	Continuous contact	Crisp contacts	Crossing
Value	1	2	3
Examples:	O ~	┐┘ F ┬ ┴ ⟨ ∠	× + ≠

If we look for the complexity of a sign, then, it can be computed as the sum of form, position and (all) contacts. In the *ogham* script there is always a horizontal middle line and the vertical lines can be either below it, above it, or crossing it. All this must be taken into account if the distinctivity of the entire script is computed. Recall that distinctivity refers to *the quality of being easily recognized and differentiated from other signs* (G. Antić & G. Altmann, 2005). Hence, in working out merely the complexity, some of the features may be omitted.

In Chinese, the stroke form, direction and position are not regarded, one counts only the strokes written without raising the writing instrument. We shall not adhere to this method. In more complex scripts having an iconic-like form (*rongorongo*, Zapotec and Maya hieroglyphs), all individual strokes and their connections must be taken into account. In the Assyrian script the position is also distinctive.

However, a *script* is not counted separately for each symbol; it only contains motifs in a structure, repeated in several signs. A vertical line can be combined (*touched*, *crossed*, *in parallel*) with another horizontal / vertical line or with an arc, etc., appearing combined in several signs. While in Chinese we count only the “one-movement” strokes, there is a possibility to identify more complex ones. A *motif* is a combination of strokes. Their occurrence in individual signs makes up the complexity of a script. For example, the Arial letter L contains three different motifs, “|”, “–”, and “L” itself. They all occur in other letters and their frequency yields a frequency distribution characterizing the given script. The motifs themselves have their own scaled complexity (cf. R. Čech & G. Altmann, 2011, p. 12). Seen from this viewpoint, the *ogham* script is much “simpler” than *rongorongo*. However, the evaluation of motifs in hand-written scripts may be very complex because any two strokes may be equal only in the abstract sense.

2. Sampling

Since the complexity spectrum covers dozens of examples, the focus is centred at both ends by selecting a fitting model, namely the *ogham* stick-like signs and *rongorongo* glyphs. The selective process is not a complete matter of convention, rather than it follows a scheme in R. Čech & G. Altmann (2011, p. 14; 1.7). On top of that, it does cannot exclude nor relegate other scripts known for their pictorial, Baroque-like or ultra-Baroque complexity, e.g. Mayan glyphs (cf. Lounsbury, 1991 [1989], pp. 232-233, Figure 10); monumental Egyptian hieroglyphs (cf. A. Sánchez Rodríguez, 2000); Zapotec glyphs (cf. G. Whittaker, 1992; J. S. Urcid, 2001, 2005); Brahmic scripts (Devanagari, Bengali, etc), or mixed letter-forms in Gray, 1982 [1971], pp. 68-69), or on the other side, the ones that have affinities for simplicity and equilibrium such as Libyco-Berber inscriptions, or Celtiberian writings (cf. W. Pichler, 2003, p. 197; J. Ferrer i Jané, 2005).

The *ogham* set and assorted RR glyphs are submitted to analysis. A brief overview is offered ahead.

2.1 *Ogham*

The ancient alphabet used by the Celtic-speaking people, mostly in Ireland, England, Wales and Scotland, dating from the 3rd to the 9th centuries CE, responds today to *ogham* – also spelled *ogam* (Rolleston, 1990 [1911]; Arntz, 1935; Lehmann, 1991 [1989]; McManus, 1991). The letters appear to be organized by horizontal or slanted notches along a central line. The set originally consisted of 20 characters arranged in four staves (*aicme*, i.e. group, class) of five letters each. A fifth set of five symbols, called in Irish tradition *forfeda* ("extra letters"), is seemingly a later development made by the Benedictine monks (EB, 2013). *Ogham* is also known as the *tree alphabet*, since the characters bear the name of specific trees. Presumably, it was the creation of a highly literate caste of priests, the Druids, guardians and transmitters of the sacred oral lore (Lehmann, 1991 [1989], pp. 159-160). The majority of the inscribed stone slabs appear to consist of personal names in the genitive (patronyms), usually meaning "*in memory of*," "*dedicated to*" or "X, *son / descendant of* Y," written in an Old form of Irish language, and perhaps in Pictish (Evans, 1967; McManus, 1991; EB, 2013). Such memorial markers do not appear to fix literary texts rather than brief linguistic material related to territory, family and tribal affinity, plus possible grave locations. The origins and age of *ogham* have been much discussed in many sources, for more details see R. P. M. Lehmann (1991 [1989], pp. 160-168).

2.2 Analysis - Complexity of *ogham* signs

Due to its minimal design, we begin with *ogham* so the analytical procedure is grasped with no perceived difficulty, which will assist afterward in tackling the "tougher" *rongorongo*.

Being of pure straight lines and their combinations, it is clear that *ogham* characters are as simple as they can get (see UC, 1991-2012). It stands to reason to propose that *ogham* letters were easily inscribed on stone with a hammer and chisel or on wood with any sharpened object, e.g. a bone, a knife or dagger. Considering the little variation *straight* or *slanted lines* can have, the Old trained readers were supposed to share proficiency and quickness in order to retrieve the written message. Similarly, as the *ogham* alphabet scrapped and eliminated the use of curved lines, it spared at the same time graphic convolution. The fact itself goes contrary to *rongorongo* of Easter Island. In terms of literacy, it may take quite a lot for a top-level scribe to not remember the *ogham* signs, but the thousands of glyphs in mixed writing systems can hardly be stored in one's memory at any given time.

Stone-working or -carving is quite time-consuming; hence the Celtic writers conveyed by relatively little efforts what was socio-culturally relevant in the times they lived: ancestry, affiliation, funerary memorials. A different picture appears in the case of several Old Egyptian hieroglyphics, Luwian (Anatolian) hieroglyphs[1] and of Maya glyphs written on monumental stone-works, where *artistry, calligraphy* and yet a formidable *complexity*, reached the apex.

In the *ogham* script, we have straight lines, crisp contacts and crossings, all of the same sort. Considering also punctuation marks we obtain the results presented in Table 1*a-f*.

[1] Although Luwian hieroglyphs generally appear to be non-calligraphic, even crudely carved, the time and painstaking efforts invested by the scribes justify them as an instance of complexity.

A crossing, even if it involves several straight lines is evaluated as an “*x* crossing” (*v. supra*, page 4), and has the value of 3.

Table 1
Complexity of *ogham* signs[2]

1*a*. The first group of ogham characters, or *aicme b* (first aicme).

Ogham character	ᚁ	ᚂ	ᚃ	ᚄ	ᚅ
Alphabetic letter	b	l	f	s	n
Complexity	**6**	**10**	**14**	**18**	**22**

1*b*. The second group of ogham characters, or *aicme h* (second aicme).

Ogham character	ᚆ	ᚇ	ᚈ	ᚉ	ᚊ
Alphabetic letter	h	d	t	c	q
Complexity	**6**	**10**	**14**	**18**	**22**

1*c*. The third group of ogham characters, or *aicme m* (third aicme).

Ogham character	ᚋ	ᚌ	ᚍ	ᚎ	ᚏ
Alphabetic letter	m	g	ng	z	r
Complexity	**7**	**12**	**17**	**22**	**27**

1*d*. The fourth group of ogham characters, or *aicme a* (fourth aicme)

Ogham character	ᚐ	ᚑ	ᚒ	ᚓ	ᚔ
Alphabetic letter	a	o	u	e	i
Complexity	**7**	**12**	**17**	**22**	**27**

1*e*. The fifth group of ogham characters, or *fifth aicme*. This *aicme*, or Forfeda, was added later on for use in literary and law manuscripts.

[2] *Ogham* character figures are extracted from S. Ager’s (1998-2013) webpage.

Ogham character	ᚕ	ᚖ	ᚗ	ᚘ	ᚙ
Alphabetic letter	ea	oi	ui	ia	ae
Complexity	**9**	**22**	**18**	**30**	**66**

1*f*. Extra-linguistic or “other” *ogham* symbols

Ogham symbol	᚛		᚜
Designation	Start of texts	space (boundary division)	End of texts
Complexity	**8**	**2**	**8**

Allowing for the complexity in the four original or pre-Christian *ogham* staves (alphabetic groups), the scale fluctuates between 6 and 26 in the first two ones (1*a*-*b*), and 7-27, in the two other groups (1*c*-*d*). At which point, the low complexity is related to the fully geometricized linear-like characters, having less crossings and zero arches / curves. Removal of such “unnecessary” scribal features, makes us think of a pre-conceived mathematical design or adaptation by the Druids, where economy of writing and reading were the main concern. The *forfeda* group characterized by later diphthong additions shows a different direction: increase in the value of complexity, from 9 to 66. The inclusion of extra letters encoding diphthongs may speak in favor of more literary variety (genres) in the expression of the early Irish Christian priests. Nonetheless, *ogham* appears to be a far cry from the complexity of the highly convoluted *rongorongo* glyphs (see ahead Table 4).

2. 3 ***Rongorongo***

Rongorongo is the classical script attested in Easter Island in 1864 (Eyraud, 1866, p. 71). It is nowadays extinct due to fateful circumstances, and what remains is but a scant number of texts (25), whose authentic condition may be reduced even further. The current corpus[3] is in stark contrast with the hundreds of texts assumed to have existed in pre-missionary times (before 1864). The writing order in tablets is the *inverted boustrophedon* (Fischer, 1997, p. 351; Sproat, 2000, p. 58; Robinson, 2002, p. 39), or “*shark-toothed*” which according to A. Gaur (1987, p. 54) *the writing material has to be turned upside down after completing a line*. The main reason for such an order might have been the “ease” and “improved legibility” of the tablets (see also Fischer, 1997, p. 353).

Another typical feature is the seamless linearity of the glyphs lacking clear boundary divisions in the best part of the corpus, in opposition to the spaced English words of the current paper. From a modern perspective, the unidentified figures and their meaning appear as all being waiting dormant; however, for the Old expert scribe textual recovery was

[3] S. Englert (1948, p. 322) points out, “*Las tabletas que existen actualmente son tan pocas que presentan un acervo escasísimo de textos*” [The tablets that currently exist are so few that represent a very scarce text legacy.]

uneventfully done, based on discerning skills and long, training and chanting practices. The point reflects R. Köhler's (2004, p. 6) observations on the Production and Decoding complexities of scripts, where one should take into account the Muscular / Nervous Effort and the Cognitive Effort. For that reason, we must bear in mind that modern epigraphers or other enthusiasts are not exactly the ancient scribes: raised in pre-missionary Rapanui to speak the local language, experts in managing obsidian flakes and a shark's tooth and comfortable with the operating system of *rongorongo*. Efforts to imitate the original process are told by F. Dederen (in Dederen & Fischer, 1993, pp. 183-184), describing the work of reproduction as "*…exceedingly tiresome, for penetration of the wood was very strenuous and caused one's finger joints, muscles, and arms to ache within a very short period of time*." Further experimental replications would be quite useful in the sense that they may help measuring *the Muscular / Nervous Effort* mentioned by Köhler (2004). In fact, given that the writing medium should be extraneous to any measurement concerning the script itself, the inclusion of the act of wood-carving as part of the "production complexity" may strike us as a bit odd. Otherwise, extrapolating, we could say that "stone-cut" English is more complex than "pen-and-ink" English. The argument is not supposed to perplex rather than to call attention to the *script complexity*, whose study and standard definition may suggest additional dimensions.

We are not discussing in details the history, script-related subtleties and the decipherment possibilities, as they have been elsewhere treated and handled at large (see K. Routledge, 1919; A. Métraux, 1940; S. Englert, 1948; N. Butinov & Y. Knorozov, 1957 [1956]; T. S. Barthel, 1958; J. Guy, 1982, 1985, 1990, 2006; J. Vignes, 1990; K. Pozdniakov, 1996; S. R. Fischer, 1997; A. Davletshin, 2002; R. W. Sproat, 2003) –, rather than talk about a few aspects related to the scope of the paper.

Hereby, we begin with a direct question, *how many signs does the script really have*? The answer to this, if not the most sought-after target in the RR studies, then, is the second in a list of inter-related controversies. Having that said, the first comes now: is it a genuine writing system and will it ever be fully deciphered and independently verified as such? Either of them is a non-trivial issue, as it conditions all and any effort in achieving concrete results. Specifically, F. Coulmas (2003, p. 69) points out that *the size of signary is just one of the several factors that account for the relative simplicity of a writing system*. As a rule, alphabets and syllabaries are less complex in their design as compared to pictographic or logo-syllabic writings. At present, if we discount the empty cells in T. Barthel's (1958) release of a 799-sign draft, then 605 basic shapes (*Grundtypus*) are listed (cf. R. Duranton, 1998, p. 43). Nevertheless, since there is much stylistic variation and compounded forms, the basic signary[4] of RR must be smaller. Upon the bare study only of Barthel (1958), we may decide on a 120-sign core, which can generate some 1500-2000 ligatures.

The point is moot, however. Insightful or anecdotal suggestions aside, one can take issue here with the fact that the nature of RR is *far from certain*. If every claim is to be believed, *rongorongo* may have massive semantic areas meshed in relational structures and falling in the proto-writing category[5]; similar, for example, to the Olmec clustering of elem-

[4] See for example A. Robinson's (2002, pp. 41-43) discussion on the basic signs of a given script.

[5] See J. Vignes (1990, p. 116), "*Nevertheless, the danger is to want to make the RR a more elaborate system than it was. A hieroglyphic writing* (*a semantic-phonetic script*, our note) *perhaps was not necessary for the Easter Islanders trained in the awful cerebral gymnastics which was required for the retention of oral tradition. Satisfied with the embryo-system they had created, they did not try to improve it*." If Vignes (1990) is right on his claim, it makes impossible, to all intents and purposes, to decipher *rongorongo*.

ents in Mesoamerica, cf. S. Houston (2004, pp. 284-286), or it may represent linguistic units at the level of the morpheme (expressing full words, or concepts via logograms as in much of the Zapotec writing, cf. Urcid Serrano, 2001), or phonemic segments (syllables and/or distinct phonemes; in line e.g. with the Linear B syllabary, see J. Chadwick, 2000 [1958]; E. Grumach, 1976 [1966], p. 47; Figs. 3.4-3.5; E. Bennett, 1996, pp. 125-126, Table 7.1), or a blend of all of them (close e. g. to Mayan glyphs; cf. Lounsbury, 1991 [1989], pp. 219-233). With no clear middle ground here and by reserved optimism we think it may have a sign list of 60-70 units, though that does not primarily suggest a syllable-template at play. In truth, since the remaining corpus is random and incomplete (in terms of sociolinguistic variables and in size; see Melka, 2009b), we cannot foretell with accuracy neither the type/ token ratio (TTR), nor the fact that RR inventory was *open-ended* and *always augmentable*. The *descriptors* are borrowed from E. Pulgram (1976 [1966], p. 15), with the statement implying that any Old Rapanui thought and neologisms could be expressed via expansion of syntax and morphology, i.e. combinations of "existing" glyphs and/or via the invention of new ligatures and new single glyphs along linear sequences (see also R. W. Sproat, 2000, p. 137). Another plausible premise is: providing the script has a large-scale symbolism and a metaphor-use, then, to be sure, a good number of signs benefit from polysemy or multivocality. Accordingly, while we do not predict a jumble of possibilities for a single sign, we cannot *a priori* exclude that it might have a semantic value as well as a phonetic value in different contexts.

We scrutinize the complexity of three RR glyph-subsets. To our knowledge, the proposed subject has not been examined on purpose in former studies. On the other side, it may be presumed that the commented feature of calligraphy and artistry in RR relates directly or not to our goal (see among others, A. Métraux, 1940, p. 393; 1957, p. 184; T. Barthel, 1958; T. Heyerdahl 1965, p. 372; R. Campbell, 1971, p. 374; S. R. Fischer, 1997[6]; P. Horley, 2009; & T. S. Melka, 2009c).

> "*From a purely technical, and even artistic, point of view we cannot but admire the quality of the incised work. In their masterly simplification, the designs have a vigour and lightness that makes one forget the heavy pressure the artist must have exerted on the wood in order to cut their grooved outlines with a shark's tooth or an obsidian graver. Graphic art has rarely reached such a level of perfection in any primitive*[7] *culture*" (A. Métraux, 1957, p. 184.)

This particular sign selection may provide ample room for discussion and / or confusion, e. g. the complexity from the point of view of the "writer" (Old Rapanui scribe) *vs*. the complexity from the point of view of the Old chanter (or of the modern observer), see R. Köhler (2004, p. 6). Most certainly, we may distinguish *simple* vs. *convoluted forms* based on their visual features. The widely referred "catalog" indexes series of hundreds, /1-700/ (cf. Barthel, 1958: *Formentaffel*; *Kennziffern 1-799* [Sign form plates; Reference numbers 1-799]; see also Fischer 1997, pp. 217-218; CEIPP, 2005), and mainly mirrors two trends. It follows somewhat principles of glyph frequency based on simplicity and on glyph-ordering, sharing

[6] Such a feature indicates that text carving/copying was not mechanically or morosely done. Quite the contrary: S. R. Fischer (1997, p. 559) sees in the RR script *an unparalleled artistic aptitude among all Pacific Islanders*.

[7] By "*primitive* culture" A. Métraux (1957) presumably means *Neolithic* culture. Otherwise, if it was meant *pre-Industrial* culture, there are several worldly scribal traditions that enjoy artistry to the highest degree.

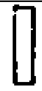
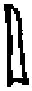
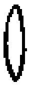
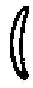
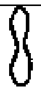
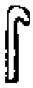
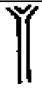
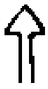
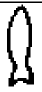
one or more design features. The outline of sign-forms /1-100/ apparently is less complex or relatively so as we climb the echelon, i.e. /200/-series, /300/-series, et cetera. Complexity in the upper strata often increases to pure iconicity. In case of distinct signs, the iconicity is quite recognizable. However, in case of compounds and of other conflated, "kinetic"-looking or aesthetic forms, there is a degree of opacity. These are convenience terms, anyway. We know that Barthel's (1958) transliteration is marred by inconsistencies, various aberrances and misidentifications, requiring serious intervention or complete replacement (see e. g. J. Guy, 1985, 2006; J. Vignes, 1990, p. 117; K. Pozdniakov, 1996; S. R. Fischer, 1997, pp. 218-219; R. Duranton, 1998; *Shortcomings* in CEIPP, 2005). Un-replaced to date by a better and an agreed catalog, T. Barthel (1958) is an imperfect, yet a necessary requisite in the RR studies.[8]

2.4 Analysis – RR complexity

In order to illustrate the three levels of complexity we present some computations on selected cases. We ignore the vertical placing and only consider the stroke type and line-joining. It must be remarked that in many cases it is not possible to distinguish between a crisp and a continuous (or a smooth) joining because the texts were not printed, rather than extracted from the personal tracings of Bodo Spranz (see Barthel, 1958: Foreword). We take the most probable joining. For example, for glyph /1/ we have four straight lines (i.e. 4 x 2 = 8) and four crisp contacts (i.e. 4 x 2 = 8). Hence the complexity is 8 + 8 = 16. The simplest signs, their catalog number and their complexity are presented in Table 2.

Table 2

Complexity featured in some simple *rongorongo* signs. As far as we can gather, quite many of them appear *highly conventionalized* as already said by A. Métraux (1940, p. 403).

Glyph										
T. Barthel's (1958) No.	1	4	21	22	40	45	63	64	86	710
Complexity	*16*	*14*	*8*	*10*	*10*	*16*	*17*	*18*	*22*	*23*

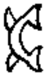
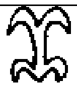
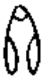
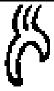
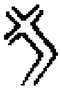
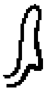
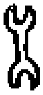
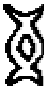
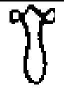
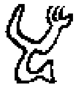
The simplest glyphs have a strong geometric feature in their contour, with their complexity ranging from *8* up to *23*. It has been mentioned that several of the simple-designed signs have *top frequency* in the RR corpus (Barthel, 1958, p. 165; Fischer, 1997, pp. 224-225; Harris & Melka, 2011a; Melka, 2013, p. 123). This piece of information fits well with observations regarding other types of scripts (G. Altmann, 2004). The *simplified forms* may be the "backbone" of the script; however they enter quite often in dependency associations, fusions and conflations with other signs (see e.g. Barthel 1958, p. 166; Fischer, 1997, p. 225; Horley,

[8] In R. Duranton's (1998, p. 43) choice of words, "*…still perhaps the only possible practical framework to thoroughly describe the corpus and to ease communication across experts*."

2005, p. 110), yielding at times quite complex and overworked sign forms, somewhat in collusion e. g. with Maya glyphs.

As it appears *simplicity* finds its way much faster and frequently to a RR context than *complexity* does. Are the principles of economy and efficiency at work for the most common Old Rapanui sounds and/or concepts?

Table 3
Complexity in different visually elaborated *rongorongo* signs: the middle group is made of not too simple or too complex samples.

Glyph										
Barthel's (1958) No.	7	39	51	56	68	76	83	86	90	756
Complexity	*63*	*48*	*32*	*22*	*28*	*20*	*38*	*50*	*46*	*60*

The complexity's scale varies between *20* (for glyph /76/) and *63* (for glyph /7/).

At this point there is something to consider. Intuitively, we would think that a glyph with any sort of symmetry (e.g. glyph /7/ has a bilateral symmetry if horizontally placed), should be rated less complex than a glyph of roughly the same number of strokes without any sort of symmetry, e.g. /756/. We, humans, use by and large *symmetry* as a diagnostic feature in the perception and assessment of form; hence it would be easier cognitively to recall /7/. So it seems quite strange that glyph /7/ should be more complex than /756/. Well, we must say for one that it amounts to a little more than that. In admitting that the recent complexity decisions are conventional (based on the discrete *points*, *lines*, *curves* and *contacts*), one needs to start and re-adjust all along the measuring process by reviewing the current method and by integrating more features.

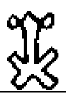
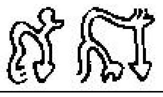
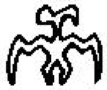
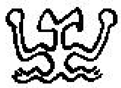
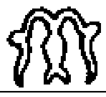
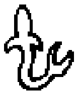
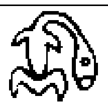
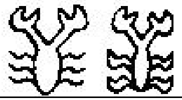
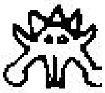
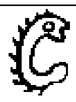
As for the real-life referent, different glyphs are identified with some certainty: glyph /7/ relates to *rei miro*, a crescent-shaped wooden gorget worn by ancient locals as a prominence and authority sign (T. Heyerdahl, 1975, pp. 203-204; M. Orliac & C. Orliac, 2008); glyph /51/ hints at the female pudenda (*komari*), and also equaled on the re-generative principle to *earth* or *land* (*henua*), see Melka (2009a, p. 52, Footnote 64); glyph /68/ is perhaps a variant of /67/, the "indigenous palm tree" (*Jubaea chilensis*) that once covered significant land portions on Easter Island (cf. Grau, 1998); /756/ suggests a "shark body with an affixed open right hand," in line with related glyphs of the /700/-series. As for the rest of glyphs, their referential identification is open to discussion. For sure, from a modern day viewpoint we may think of glyph /83/ standing for an *open-ended wrench* (*open-ended spanner*, in British English); of glyph /76/ for an *erect phallus* (Fischer, 1997), and of glyph /90/ embodying an *eared bowling pin*. Whatever is "concealed" in the last three signs, two of our guesses are certainly erroneous and misguided: neither metal wrenches/spanners, nor bowling games were ever reported or observed in the Neolithic society of Rapanui. As for shape resemblance of glyph /76/ this is undetermined, unless a total decipherment proves its phonetic and/or semantic value/s.

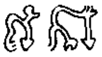
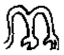
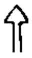
Subset No. 3 deals with some of the most complex signs, faring most of the time as narrative- and picture-like figures (see Table 4). While resolving the iconographic likeness, any educated person (not particularly well-versed in the Old Pascuan lore) can link the glyphs with a description, e.g. /618/ with a "*winged bird-man with a frigate head*"; /684/ with a "*double cormorant head-and-neck in a fish body*"; /720/ with a "*squalid*"; /761/ with a "*skink-like*" creature; /770/ with a "*double-headed crustacean*"; and /790/ with a "*hairy worm-like*" critter. A connoisseur of the Easter Island's ancient traditions may include a "much" different /761/ as a variation of "*lizard*"-glyph /760/ , therefore it may well be an out-of-standard "skink with a big tilted head," while we can glean the "face" of the "*ao*" paddle from glyph /781/. To put all the guesswork in perspective, pictorial-like scripts tend to be Old as in closer to the origin of writing for a particular culture (*Rapanui*, in our case), not in terms of absolute date.

The complexity of very convoluted signs (see Table 4) indicates that some variants may be strong simplifications, signaling a beginning in the transition from *iconicity* to *symbolism*. Consider at this juncture, Sumerian and other cuneiform-related scripts that had origins in pictography, but strayed from it significantly in later times (cf. J. Friedrich, 1971 [1957], pp. 34-51). In the same way, Chinese and Old Egyptian hieroglyphics appear to have evolutional stages toward more geometric and cursive forms (Jean, 1998 [1989]: 127; W. Boltz, 1994; F. Coulmas, 2003, pp. 50-52; Baines in Houston *et al*, 2003, pp. 439-445; M. A. Stadler, 2008, pp. 167-169), while with *rongorongo*'s chronology still unclear (Robinson, 2002, pp. 223-225), diachronic studies would set very different expectations on present evidence.

Some complexities, varying from *60* up to *178* (as per G. Altmann's 2004 *composition method*) are presented in Table 4.

Table 4
Complexity of lavishly designed *rongorongo* signs

Glyph					
Barthel's (1958) No.	99	491	618	642	684
Complexity	*86*	*63 86*	*108*	*94*	*90*
Glyph					
Barthel's (1958) No.	720	761	770	781	790
Complexity	*60*	*70*	*130 178*	*85*	*77*

We should say that in some cases it is not quite clear whether a contact is *crisp* or *continuous*, being a matter of interpretation. In written forms, even in the Latin script, we would be forced to interpret the handwriting of various persons in a different way. The entire occurrences of the same sign are different, thus we merely have a preliminary evaluation and by no means complete. *Rongorongo* variants can be seen e.g. in glyphs /491/ & /770/. Specifically, the composite glyph /491/ in Barthel's notation (1958) is related to dual variants . The second variant yields  /477/ +  /86/ after deconstruction, with the glyph /86/ in an upside down fashion (see Melka, 2009c, p. 84). Another coherent line of attack would be to measure /477/ and /86/ separately and cross-check their occurrence across the available texts.

We must realize by now that visually *rongorongo* is a very complex script. The fact suggests sumptuous pictorial-like elements in the glyphs, leading to remarkable individuality, plus imaginative and standout shapes. In a parallel manner, it brings about more chances for distraction and clutter (errors), with scribes normally failing to run a strict spell-check on their texts. Attempting to explain all signs given their visual representation is a potential roadblock at this moment. Despite the iconicity, the sequential and repetitive feature observed in a variety of RR contexts may involve areas of coded speech agreeing with the registered ancient folklore (Melka, 2012). One could also attain an approximate value taking the average complexity of the entire signs but this would be rather a Sisyphean work not warranting some definite result.

3. Conclusion

In principle, script complexity is measurable but it does not mean that our way of doing is more objective than other kinds of quantification may be (see Peust, 2006). It would also be possible to ask test-persons to order the signs according to complexity. However, talking from a strong position, that way of measurement in this domain would be both very subjective and we would be forced to take averages. In addition, the possible association of writing, reading, memorizing, etc. in *ogham* and *rongorongo* could add difficulty and distort the estimation. The concept of complexity in glyphs is intriguing, but there's clearly a lot more to do before it can be standardized and put to good use. In any case, we see that more iconic-like scripts are much more complex than purely symbolic (conventionalized) ones. It must be emphasized that this has nothing to do with script distinctivity which is based on the comparison of individual signs, commonality of strokes or even motifs. There is a possibility that high complexity or high ornamentality is associated with low distinctivity.

One reason to measure complexity for unknown writings (i.e. *rongorongo*) is that *complexity levels* could be closely related to other properties of the language, such as if higher script complexity is shown to reliably indicate logographs in known languages. Incidentally, our preliminary results endorse a diagnosis more and more accepted among mainstream research: *rongorongo* is *not* largely syllabic. At present, cases of early scripts factoring large numbers of glyphs such as Old Egyptian, cuneiform, Maya, Zapotec and Luwian speak in favor of this point, implying a degree of logography embedded in their systems. The level of *logography* vs. *phonography* as *a tool to classify entirely unknown writing systems to assist in attempts at archaeological decipherment* is referred in R. W. Sproat (2000, pp. 137-139); G. Penn & T. Choma (2006); and M. Harris (2010). At any rate, even the safest estimate here cannot be used of, until we discover and largely quantify these relationships for numerous known writing systems (or written phrases).

In our analysis, script complexity is dependent on specific characters/glyphs or compounds. Beyond this first approach, further steps are worthy of exploring, e.g. taking a reasonably long *rongorongo* text, *what is the distribution of complexities*? Or *what is the course of complexities*: *does complexity increase from the beginning to the end or vice versa*? Another interesting hypothesis regarding the future would be testing the complexity on a "word" and/or a "sentence" boundary. It appears that the formulation of a word or a sentence in a Latin-based script is at times non-complex (simple) and every so often pretty complex. Matters are all the more obscure in the case of non-fully-segmented and undeciphered scripts such as *rongorongo*. A habit of caution is most advised in such a case.

The study of texts –which are not available in sufficient number for all *known* or *unknown* scripts (cf. Robinson, 2002, pp. 34-37)– could show that high complexity is linked with small frequency distribution. Hence, all questions concerning the control cycle shown above (Figure 2) must be postponed until different scripts and multiple texts are analyzed. Logically, further research is pivotal in avoiding snap decisions as we examine the complexity in a long list of scripts, e.g. *Phaistos disc* printed characters, *Pictish* symbols, *Luwian* (*Anatolian*) hieroglyphs, *Maya* and *Zapotec* glyphs, *Linear B* signs, etc, looking for common and replicable patterns.

References

Ager, **Simon** (1998-2013). Ogham. In *Omniglot*: *The Online Encyclopedia of Writing Systems and Languages*.http://www.omniglot.com/writing/ogham.htm (accessed September 23, 2013).

Allan, **Jim** (Ed.) (2002 [1978]). *An Introduction to Elvish, other Tongues, Proper Names and Writing Systems of the Third Age of the Western Lands of Middle-Earth as set forth in the Published Writings of Professor John Ronald Reuel Tolkien.* As Authorized by the Mythopoeic Linguistic Fellowship, a Discussion Group of the Mythopoeic Society. Helios, Glenfinnan, Inverness-shire: Bran's Head Books.

Altmann, **Gabriel** (2004). Script Complexity. *Glottometrics*, 8: 68-74.

Altmann, **Gabriel** (2008). Toward a Theory of Script. In: Gabriel Altmann & Fenxiang Fan (Eds.), *Analyses of Script*: *Properties of Characters and Writing Systems* (Quantitative Linguistics, 63).. Berlin-New York: Mouton de Gruyter. pp. 149-164.

Antić, **Gordana** & **Gabriel Altmann** (2005). On Letter Distinctivity. *Glottometrics*, 9: 46-53.

Arntz, **Hermann** (1935). Das Ogom. *Beitrage zur Geschichte der Deutschen Sprache und Literatur*, 59: 321-413.

Barthel, **Gustav** (1972). *Konnte Adam schreiben*: *Weltgeschichte der Schrift. von der Keilschrift zum Komputersatz.* Bearbeitet und herausgegeben von Karl Gutbrod. Köln: Verlag M. DuMont Schauberg.

Barthel, **Thomas S**. (1958). *Grundlagen zur Entzifferung der Osterinselschrift*. (Abhandlungen aus dem Gebiet der Auslandskunde 64, Reihe B.). Hamburg: Cram, de Gruyter & Co.

Bennett, **Emmet** (1996). Aegean Scripts. Section 7. In: Peter T. Daniels & William Bright (Eds.). *The World's Writing Systems*. Oxford, NY: Oxford University Press. pp. 125-133.

Boltz, **William G**. (1994). *The Origin and early Development of the Chinese Writing System*. American Oriental Series, 78. New Haven, CT: American Oriental Society.

Brennan, **J**. **H**. (1994). *A Guide to Megalithic Ireland*. London: Aquarian / Thorsons.

Butinov, **Nikolai A**. & **Yuri V**. **Knorozov** (1957 [1956]). Preliminary Report on the Study of the Written Language of Easter Island. *Journal of the Polynesian Society*, 66(1): 5-17.

Campbell, **Ramón** (1971). *La Herencia Musical de Rapanui*: *Etnomusicología de la Isla de Pascua*. Santiago de Chile: Editorial Andrés Bello.

CEIPP (2005). *Thomas Barthel's Transliteration System*: *The Rongorongo of Easter Island*. http://web.archive.org/web/20071216102050/http://www.rongorongo.org/corpus /codes.html (accessed September 4, 2013).

Čech, **Radek** & **Gabriel Altmann** (2011). *Problems in Quantitative Linguistics. Vol*. *3*. Lüdenscheid: RAM-Verlag.

Chadwick, **John** (2000 [1958]). *The Decipherment of Linear B*. Cambridge: The Press Syndicate of the Cambridge University.

Coulmas, **Florian** (2003). *Writing Systems*: *An Introduction to Their Linguistic Analysis*. Cambridge: Cambridge University Press.

Coulmas, **Florian** (2009). Evaluating Merit - the Evolution of Writing Reconsidered. *Writing Systems Research*, 1(1): 5-17.

Correll, **Sharon** (2013). *NRSI*, *Computers* & *Writing Systems*: *Examples of Complex Rendering*. SIL International. http://scripts.sil.org/cms/scripts/page.php?site_id =nrsi&id=CmplxRndExamples (accessed October 1, 2013).

Daniels, **Peter T**. & **William Bright** (Eds.) (1996). *The World's Writing Systems*. New York-Oxford: Oxford University Press.

Davletshin, **Albert** (2002). Names in the *Kohau Rongorongo Script.* Paper presented as *From Kohau Rongorongo Tablets to Rapanui Social Organization*: *From Rapanui Social Organization to Kohau Rongorongo Script* at the 2nd International Conference "Hierarchy and Power in the History of Civilizations," Saint Petersburg, Russia, July 4-7, 2002.

Dederen, **François** & **Steven R**. **Fischer** (1993). The Traditional Production of the Rapanui Tablets. In *Easter Island Studies*: *Contributions to the History of Rapanui in Memory of William T. Mulloy*. *Oxbow Monograph 32*. Edited by S. R. Fischer. Oxford, UK: Oxbow Books. pp. 182-184.

DeFrancis, **John** (1989). *Visible Speech*: *The Diverse Oneness of Writing Systems*. Honolulu, HI: University of Hawai'i Press.

Duranton, **Raymond A**. (1998). Encoding and Imaging the *Rongorongo* Corpus. (A collective work of the Cercle d'Études sur l'Île de Pâques et la Polynésie (CEIPP) Paris, France.) In *Easter Island in Pacific Context South Seas Symposium*: *Proceedings of the Fourth International Conference on Easter Island and East Polynesia*, *University of New Mexico*, *Albuquerque*, *5-10 August 1997*. Edited by C. M. Stevenson, G. Lee & F. J. Morin. Los Osos, California: Bearsville and Cloud Mountain Presses (The Easter Island Foundation). pp. 42-48.

EB (Encyclopædia Britannica, Inc.) (2013). *Ogham Writing*. http://www.britannica. com/EBchecked/topic/425827/ogham-writing (accessed October 6, 2013).

Englert, **Sebastián** (1948). *La tierra de Hotu* Matu'*a*: *Historia*, *Etnología y Lengua de la Isla de Pascua* [The Land of Hotu Matu'a: History, Ethnology and Language of Easter Island]. Santiago de Chile, Chile: Padre las Casas.

Evans, **D**. **Ellis** (1967). *Gaulish Personal Names*: *a Study of some Continental Celtic Formations*. Oxford: Clarendon Press.

Eyraud, **Joseph-Eugène** (1866). Lettre du Frère Eugène Eyraud, au T. R. P. Supérieur Général de la Congrégation des Sacrés-Coeurs de Jesús et de Marie. Valparaíso, décembre 1864. *Annales de la Propagation de la Foi*. Lyon. 38:52-71, 124-138.

Ferrer i Jané, **Joan** (2005). Novetats sobre el sistema dual de diferenciació gràfica de les oclusives sordes i sonores. *Palaeohispanica*, 5: 957-982.

Fischer, **Steven R**. (1997). *RongoRongo, the Easter Island Script*: *History*, *Traditions*, *Texts*. Oxford, NY: Oxford University Press.

French, **M**. **A**. (1976 [1966]). Observations on the Chinese Script and the Classification of Writing-systems. In *Writing without Letters*. Edited by W. Haas. Mont Follick series. Vol. 4. Manchester, UK: Manchester University Press. pp. 101-131.

Friedrich, **Johannes** (1971 [1957]). *Extinct Languages*. Translated from the German, *Entzifferung Verschollener Schrifton und Sprachen* [Decipherment of Lost Scripts and Tongues] by Frank Gaynor. Westport, Connecticut: Greenwood Press, Publishers.

Gaur, **Albertine** (2000). *Literacy and the Politics of Writing*. Bristol, UK: Intellect Books. pp. 78-79.

Gray, **Nicolette** (1982 [1971]). *Lettering as Drawing*. New York: Taplinger Publishing Co., Inc.

Grumach, **Ernst** (1976 [1966]). The Cretan Scripts and the Greek Alphabet. In *Writing without Letters*. Edited by W. Haas. Mont Follick series. Vol. 4. Manchester, UK: Manchester University Press. pp. 45-70.

Guy, **Jacques B**.**M**. (1982). Fused Glyphs in the Easter Island Script. *Journal of the Polynesian Society*. 91: 445-447.

Guy, **Jacques B**.**M**. (1985). On a Fragment of the "*Tahua*" Tablet. *Journal of the Polynesian Society*, 94: 367-387.

Guy, **Jacques B**.**M**. (1990). On the Lunar Calendar of tablet '*Mamari*.' *Journal de la Société des Océanistes*, 91(2): 135-149.

Guy, **Jacques B**.**M**. (2006). General Properties of the *Rongorongo* Writing. *Rapa Nui Journal*, 20(1): 53-66.

Jean, **G**. (1998 [1989]). *Signs*, *symbols*, *and ciphers*. (S. Hawkes, Trans.). New York: Harry N. Abrams, Inc., Publishers.

Harris, **Martyn** (2010). *Corpus Linguistics as a Method for the Decipherment of* rongorongo. Mres in Applied Linguistics. London: Birkbeck University.

Harris, **Martyn** & **Tomi S**. **Melka** (2011a). The *Rongorongo* Script: on a Listed Sequence in the *recto* [*verso*, repaired] of Tablet 'Mamari.' *Journal of Quantitative Linguistics*, 18(2): 122-173.

Heyerdahl, **Thor** (1975). *The Art of Easter Island*. New York: Garden City, Doubleday & Company, Inc.

Horley, **Paul** (2005). Allographic Variations and Statistical Analysis of the *Rongorongo* Script. *Rapa Nui Journal*, 19(2): 107-116.

Horley, **Paul** (2009). *Rongorongo* Script: Carving Techniques and Scribal Corrections. *Journal de la Société des Océanistes*, 129(2): 249-261.

Houston, **Stephen**, **John Baines** & **Jerrold Cooper** (2003). Last Writing: Script Obsolescence in Egypt, Mesopotamia and Mesoamerica. *Comparative Studies in Society and History*, 45: 430-479.

Houston, **Stephen D**. (2004). Writing in Early Mesoamerica. In *The First Writing*: *Script Invention as History and Process*. Edited by Stephen D. Houston, Cambridge, UK: Cambridge University Press. pp. 274-309.

Knight, **Stan** (1996). The Roman Alphabet. In *The World's Writing Systems*. Edited by Peter T. Daniels & William Bright. Oxford, NY: Oxford University Press. pp. 312-332.

Köhler, **Reinhard** (2004). Quantitative Analysis of Writing Systems: an Introduction. In: A*nalyses of Script*: *Properties of Characters and Writing Systems*. *Quantitative Linguistics, 63*. Edited by Gabriel Altmann & Fan Fengxiang. Berlin-New York: Mouton de Gruyter. pp. 3-11.

Lehmann, **Ruth P**. **M**. (1989). Ogham: The Ancient Script of the Celts. In *The Origins of Writing*. Edited by Wayne M. Senner. Lincoln & London: University of Nebraska Press. pp. 159-171.

Lounsbury, **Floyd G**. (1991 [1989]). The Ancient Writing of Middle America. In *The Origins of Writing*. Edited by Wayne M. Senner. Lincoln. NE: University of Nebraska Press. pp. 203-237.

Lyovin, **Anatole V**. (1997). *An Introduction to the Languages of the World*. New York - Oxford: Oxford University Press.

McManus, **Damian** (1991). *A Guide to Ogam*. Maynooth Monographs. Volume 4. Maynooth: An Sagart.

Melka, **Tomi S**. (2009a). Some Considerations about the *kohau rongorongo* Script in the light of a Statistical Analysis of the Santiago Staff. *Cryptologia*, 33(1): 24-73.

Melka, **Tomi S**. (2009b). The Corpus Problem in the *Rongorongo* Studies. *Glottotheory*: *International Journal of Linguistics* (University of Saints Cyril and Methodius, Trnava), 2(1): 111-136.

Melka, **Tomi S**. (2009c). Linearity, Calligraphy and Syntax in the *Rongorongo* Script. *Glottotheory*: *International Journal of Linguistics* (University of Saints Cyril and Methodius, Trnava), 2(2): 70-97.

Melka, **Tomi S**. (2012). On a "Kinetic"-like Sequence in *rongorongo* Tablet "*Mamari*." *Writing Systems Research*. iFirst version at http://www.tandfonline.com/doi/full/10.1080/17586801.2012.742005

Melka, **Tomi S**. (2013). "*Harmonic*"-like Sequences in the *rongorongo* Script. *Glottotheory*: *International Journal of Linguistics* (Akademie-Verlag, Berlin), 4(2): 115-139.

Métraux, **Alfred** (1940). *Ethnology of Easter Island*. Bernice P. Bishop Museum Bulletin 160. Honolulu: Bernice P. Bishop Museum Press.

Métraux, **Alfred** (1957). *Easter Island*: *A Stone-Age Civilization of the Pacific*. Trans. from French by Michael Bullock. New York: Oxford University Press.

Myers, **James** (1996). *Prosodic Structure in Chinese Characters*. Presented as a Poster at the 5th International Conference on Chinese Linguistics, Tsing Hua University, Taiwan, June 1996. http://www.ccunix.ccu.edu.tw/~lngproc/Myers1996_ChinChar.pdf (accessed October 7, 2013).

Orliac, **Michel** & **Catherine Orliac** (2008). *Trésors de l'île de Pâques* / *Treasures of Easter Island*. Collection de la Congrégation des Sacrés-Coeurs de Jésus et de Marie SS.CC. Genève: Éditions D – Paris: éditions Louise Leiris.

Penn, **Gerald** & **Travis Choma** (2006). Quantitative Methods for Classifying Writing Systems. *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, pages 117-120, New York, June 2006. http://acl.ldc.upenn.edu/N/N06/N06-2030.pdf (accessed October 14, 2013).

Peust, **Carsten** (2006). Script Complexity Revisited. *Glottometrics*, 12: 11-15.

Pichler, **Werner** (2003). *Las Inscripciones Rupestres de Fuerteventura*. Traducción Marcos Sarmiento Pérez, Elena Alsó Juan. *Die Felsinschriften Fuerteventuras*. Puerto del Rosario: Cabildo de Fuerteventura, Servicio de Publicaciones.

Pozdniakov, **Konstantin** (1996). Les Bases du Déchiffrement de l'Écriture de l'Ile de Pâques. *Journal de la Societé des Océanistes*, 103(2): 289-303.

Pulgram, **Ernest** (1976) [1966]. The Typologies of Writing-systems. In *Writing without Letters*. Edited by W. Haas. Mont Follick series. Vol. 4. Manchester, UK: Manchester University Press. pp. 1-29.

Robinson, **Andrew** (2002). *Lost Languages*: *The Enigma of the World's Undeciphered Scripts*. New York: McGraw-Hill.

Rogers, **Henry** (2005). *Writing Systems*: *A Linguistic Approach*. Malden, MA / Oxford, UK: Blackwell Publishing Ltd.

Rolleston, **Thomas William Hazen** (1990 [1911]). *Celtic Myths and Legends*, also entitled *Myths & Legends of the Celtic Race*. Dover Publications.

Sampson, **Geoffrey** (1985). *Writing Systems*: *A Linguistic Introduction*. London: Hutchinson & Co. (Publishers) Ltd.

Sánchez Rodríguez, **Ángel** (2000). *Diccionario de Jeroglíficos Egipcios* (Spanish Edition). Madrid: Alderabán Ediciones.S.L.

Sproat, **Richard W**. (2000). *A Computational Theory of Writing Systems*. Stanford: Cambridge University Press.

Sproat, **Richard** (2003). *Approximate String Matches in the RR Corpus*. http://rws.xoba.com/ror/ (accessed September 5, 2013).

Stadler, **Martin Andreas** (2008). On the Demise of Egyptian Writing: Working with a Problematic Source Basis. In *Disappearance of Writing Systems*: *Perspectives on Literacy and Communication*. Edited by John Baines, John Bennet & Stephen Houston. London: Equinox Publishing Ltd. pp. 157-183.

Su, **Yi-Fen** & **S**. **Jay Samuels** (2010). Developmental Changes in Character-complexity and Word-length Effects when Reading Chinese Script. *Reading and Writing*, 23(9): 1085-1108.

Vignes, **Jaques** (1990). Is a New Approach to the Decipherment of *Rongorongo* Writing Necessary? In *State and Perspectives of Scientific Research in Easter Island Culture*. Edited by Heide-Margaret Esen-Baur. Courier Forschungsinstitute Senckeberg 125. Frankfurt am Mein: Senckenbergische Naturforschende Gesellschaft. pp. 15-19.

Walker, **C**.**B**.**E**. (1989). *Reading the Past*: *Cuneiform*. 2nd edition. London: The British Museum.

Wikipedia [in English] (2013). *Tengwar*. http://en.wikipedia.org/wiki/Tengwar (accessed September 22, 2013).

Whittaker, **Gordon** (1992). The Zapotec Writing System. In *Epigraphy*: *Supplement to the Handbook of Middle American Indians*. Edited by V. Reifler Bricker, with the assistance of P. A. Andrews (Vol. 5). Austin, TX: University of Texas Press. pp. 5-20.

Unicode Consortium [UC] (1991-2012). *Ogham*. Range 1680-168F. http://www.unicode.org/charts/PDF/U1680.pdf (accessed September 23, 2013).

Urcid Serrano, **Javier** (2001). *Zapotec Hieroglyphic Writing*. Dumbarton Oaks Pre-Columbian Art and Archaeology Studies Series. Studies in Pre-Columbian Art and Archaeology, Book 34. Washington, D.C.: Dumbarton Oaks Research Library and Collection.

Urcid, **Javier** (2005). *Zapotec Writing*: *Knowledge*, *Power and Memory in Ancient Oaxaca*. http://www.famsi.org/zapotecwriting/zapotec_text.pdf (accessed October 7, 2013).

Ziegler, **Sabine** (1994). *Die Sprache der altirischen Ogam-Inschriften*. Göttingen: Vandenhoeck and Ruprecht.

Glottometrics 28, 2014, 75-86

Towards a Theory of Compounding

Reinhard Köhler, Trier

Abstract. The paper attempts to explain the existence of compounds by its function as a means to reduce syntactic complexity in cases where a loss of semantic information is acceptable. A mathematical model is set up using Altmann's difference equation method. It is combined with a second model on the basis of a diversification approach for those compounds in a text which were no ad-hoc constructions but lexical elements. The result is a mixed Poisson distribution, which is successfully tested on data from a German text. Then, an alternative model is presented, which is based on the Popescu-Altmann function. It is assumed that the trend to reduce complexity can be considered as a continuous quantity. So, a differential equation can be justified as a model of compounding tendency. Finally, a perspective on a more complex model is presented, which could cover syntactic and morphological complexity as functional equivalents.

Keywords *compounding, morphological complexity, syntactic complexity, mixed Poisson distribution, Popescu-Altmann function*

1. Introduction

The most general meaning of the term compounding refers to a specific kind of word-forming mechanism, which combines two or more lexemes and results in a single word. The concrete implementation of this mechanism and the conditions which have to be met by the involved units to become elements of a compound differ from language to language. Linguists studying individual languages do, as a rule, not agree in what kinds of combinations of lexemes they recognise as compounds. Moreover, only few of all the linguistic concepts connected with word formation are agreed upon among researchers in the field; only few of them have been well-defined in linguistics. This fact aggravates the problem associated with the task to set up hypotheses and universal models of compounding. The corresponding difficulties can easily be illustrated by units such as *word* and *part of speech*. We have, however, to keep in mind that every definition of a unit, a property, a category etc. is based on conventions, not on empirical findings or philosophical truths.

Conventions belong to the basic elements of any science and of every theory. This is one of the reasons why definitions and other conventions are by no means arbitrary but should be determined in a way which corresponds as well as possible to the theoretical considerations, to the ideas behind the individual problems and hypotheses. Any unit and any property can be defined in various ways and may be derived from various aspects, purposes, and methods. We will therefore base our present considerations on a fundamental idea, analyse its central concepts, and try to find the best-possible definitions and operationalisations on the basis of the hypothetical interrelations between these concepts. This means that observational terms must be determined such that they, on the one hand, correspond to the theoretical notions involved in the hypothesis and, on the other hand, are maximally appropriate for the intended measurement of the phenomena under study.

2. A model

The highest level of any scientific activity is the explanation of the observed and described phenomena. We will therefore pose right at the beginning of our considerations the question *why* there are, in many languages, compounds, i.e. we ask for an explanation of this fact. An obvious answer to this question is the statement that compounding is, besides derivation, borrowing, and neologisms, a method to form new lexical units (Köhler 1990), i.e. one of the linguistic means to meet the lexical branch of the coding requirement as postulated in synergetic linguistics (Köhler 1986, 2005). This answer is not wrong but fails to take the background of communication processes into account. The characteristic function of compounds is not the increase of the lexical inventory but rather the *ad hoc* coding of meanings via syntagmatic means. This does not exclude, of course, that ad-hoc formed compounds are lexicalised, on the contrary (cf. also Köhler/Altmann 1993).

It is, as a consequence, a good idea to scrutinise our problem from a more general point of view and have also a look at syntagmatic and in particular at syntactic coding means and their properties. The following considerations are based on the problem to derive the theoretical probability distribution of syntactic constructions as presented in Köhler/Altmann (2000). We assume a general requirement of minimising the complexity of syntactic structures in analogy to the requirement of minimising production effort as known from the lexical sub-system in synergetic linguistics, where it has a decreasing effect on word length in dependence on frequency. Minimising syntactic complexity (another term is maximising compactness) works by shifting part of the code to another syntactic level. The requirement is abbreviated as *minX*. The relation between the sentences S_1 and S_2 in the following example can serve as an illustration of the principle:

S_1 $_{NP}$[The professor] was not prepared and so he could not explain the solution.
S_2 $_{NP}$[The unprepared professor] could not explain the solution.

Sentence S_1 consists of two clauses. Its syntactic complexity can be reduced by removing one of the clauses and shifting a part of its content into the nominal phrase of the first clause. The sentence becomes less complex, the NP, one level below, becomes more complex.

The requirement *minX* has its effect on each level of a sentence at the same time. Hence, while there is a tendency to shift complexity away from a given level, all the other levels have the same tendency and thus form resistance effects. These levels tend to get rid of complexity as well and are not ready to easily accept more complexity from other ones. This tendency will be abbreviated by *maxH*. We will have to take into account also the size of the inventory of syntactic constructions *I(K)* in the given language because less complexity is needed to express a given meaning if more constructions (coding means) are available, and a quantity *E,* which represents the degree of semantic explicitness: Not too much information should become lost when syntactic complexity is reduced. When we assume that *E* is a constant, i.e. that (almost) no information is lost, and adopting the general modelling approach proposed by Altmann (cf. Altmann & Köhler 1996), the following equation can be set up:

$$P_x = \frac{maxH + x}{minX + x}\frac{E}{I(K)}P_{x-1}\,. \qquad (1)$$

With *maxH* = *k*-1, *minX* = *m*-1, and *E*/*I*(*K*) = *q*, (1) can be written in the well-known form

$$P_x = \frac{k+x-1}{m+x-1} q P_{x-1} \qquad (2)$$

which yields the hyper-Pascal distribution (cf. Wimmer & Altmann 1999):

$$P_x = \frac{\binom{k+x-1}{x}}{\binom{m+x-1}{x}} q^x P_0 \qquad (3)$$

with $P_0^{-1} = {}_2F_1(k,1;m;q)$ – the hypergeometric function – as normalising constant. Tests of this hypothesis in the cited paper on data from English and German supported the model. Other studies (Naumann 2015) yielded deviating results and suggest other distributions (in particular the negative hypergeometric distribution, which is related to our model). Here, we are not interested in these details but rather in the general idea of a requirement *minX* with the tendency to reduce syntactic complexity.

Let us assume that there are situations in which the boundary condition that semantic explicitness must not decrease is relaxed. In this case, transformations become possible which allow a considerable reduction of complexity by removing a part of the expression without any compensation. An example is formed by S_3 and S_4:

S_3 This is the room where we use to have our dinner.
S_4 This is a room.

Here, a sentence was simplified by removing one of two clauses; a considerable amount of information was lost. In many languages, a compromise can be achieved when part of the information can be inferred from the context, the situation, or world knowledge. Compounds provide enough hints to infer the rest of the unexpressed meaning such as in S_5:

S_5 This is our dining room.

The information that "dining" specifies the usual function of the room and not any other property is lost in S_5 but this can be compensated for by means of world knowledge. Most people know that many flats and homes provide a special room where the families have their meals, and it helps to take the specific situation into account, e.g. that the sentence was not taken from a fantasy novel where a magic room has to be fed on a regular basis.

Compounds sometimes turn out to be useful in many situations, become more and more known and strongly associated with the intended meaning, which means that they are lexicalised. Such compounds can be investigated on material from dictionaries whereas the many ad-hoc compounds can be found only in texts, in particular in oral, everyday communication.

We modify equation (1) by removing the constants *E* and *I(K)*. As stated above, compounds tend to be used when the boundary condition that the information content, the semantic explicitness, be kept is relaxed. The inventory of syntactic constructions does not play any role here. The requirement *maxH* (maximising compactness = minimising complexity), which can be met by shifting a part of the construction to a lower level, can also be met by transforming a part of the syntactic construction into a compound when *E* is low or absent. The stronger *maxH* the more parts will be transformed into parts of a compound, i.e. the 'longer' resulting compounds will be in terms of lexematic elements. Therefore, the factor *maxH* will maintain its place in the model. The factor *minX*, the resistance effect from other syntactic levels, does not exist in the discussed case because other levels are not affected by the transformation into compounds. It is therefore removed from the model. Finally, a factor *x* will represent a repelling force: the probability that a compound of length *x* will be formed decreases with increasing *x*. The model takes the form (4).

$$P_x = \frac{maxH}{x} P_{x-1}. \qquad (4)$$

The solution to this difference equation is the Poisson distribution. Substituting *λ* = *maxH* yields the well-known formula (5):

$$P_x = \frac{e^{-\lambda}\lambda^x}{x!} \qquad (5)$$

This model is derived from the assumption that compounds are the result of a morphological way to reduce syntactic complexity. However, in a text, also lexicalised compounds are found. The occurrence of a compound such as "dining room" in a text is not likely to be the result of reducing the complexity of the construction "the room where we use to have dinner"; it is rather just the word which designates that kind of room. We will therefore have to expect a more complicated situation when we test our model on data from texts. We will have to find also a model of compounding as a kind of word-forming mechanism which is, at a given moment in time, independent of *ad-hoc* compounding during the process of text generation. Wimmer and Altmann (1995) presented a model of morphological productivity based on a birth-and-death process. For our purposes, this model seems to be too general as we need not take into account cases where compounds would lose some of their elements. It is naturally to assume that the mechanism is a simple diversification process (cf. Altmann 1991); the probability of a new compound which is formed on the basis of an existing compound of length *x* yielding a compound of length *x*+1 depends on the number of compounds of length *x* (a similar but also too complex model for our case here was presented in Wimmer et al. 1994). Assuming such a diversification process with a constant positive influence, which cares for a growth of more and more complex compounds, and a retarding effect which increases with increasing compound length, we arrive at exactly the model presented as formulas (4) and (5) above. We combine the two processes, the increase of compound length (1) caused by reduction of syntactic complexity and (2) caused by a regular word-formation process in form of a mixed Poisson distribution, where the two individual distributions have different parameters and are added and the sum normalised:

$$P_x = \frac{\alpha e^{-\lambda}\lambda^x}{x!} + \frac{(1-\alpha)e^{-\mu}\mu^x}{x!}, x = 0,1,2,... \qquad (6)$$

For our purposes, a displaced version will be applied because the minimum number of elements of a word is larger than 0.

3. Testing the model

For a first test of the model, the following text was selected: http://de.wikipedia.org/wiki/Grundstückverkehrsgesetz (June 12, 2014). This text is relatively short but on the other hand relatively rich with respect to compounds. We considered all kinds of compounded words irrespectively of their part of speech and age. The only criterion was the existence of an element of the segmented word as a lexeme. There are, of course, cases of doubt: "Nachmeldung" (approx. 'late registration') could be considered as a compound because "nach" exists as an adverb and as a preposition. This noun, however, was not considered as a compound because it is formed by derivation using the suffix "-ung" from the verb "nachmelden", in which "nach" is a (separable) prefix. The word "Nachkriegszeit", on the other hand, is formed from three lexical units: "nach" (preposition, 'after') "Krieg" (noun, 'war'), and "Zeit" (noun, 'time'). Similarly, also adjectives, adverbs, and verbs were included in the compound lists if they could be segmented into lexematic units. Abbreviations and acronyms were counted after expanding them: "z.B." was expanded into "zum" and "Beispiel" ('for example'). Proper nouns were ignored.

We decided that the design of the test data should be as similar as possible to the way how complexity was determined and evaluated in Köhler/Altmann (2000). Complexity of a syntactic construction was defined as the number of immediate constituents of a given construction, a number with the lower bound 1. We will therefore define complexity or length of a compound as the number of lexematic elements a word consists of – despite the fact that words with just one element are not called compounds but simplexes. The text contains words with 1..5 elements (cf. Tables 1 and 2a-2e). The result of fitting the mixed Poisson distribution to the data was successful (C = 0.0104) but not excellent.

Table 1
Fitting the mixed Poisson distribution to the data

x[i]	f[i]	NP[i]
1	522	516.68
2	95	96.22
3	18	24.65
4	14	8.83
5	1	3.62
Parameters:		
λ	1.2662	
μ	0.1300	
α	0.1395	

X²	P(X²)	DF	C
6.7473	0.0094	1	0.0104

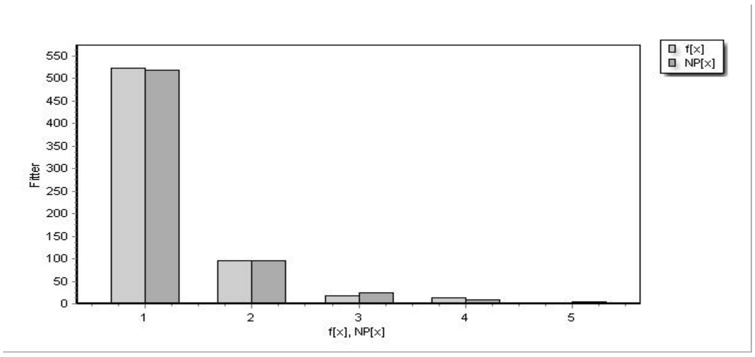


Fig. 1: Graph of the fitting result (cf. Table1)

4. An alternative model

Another way to set up a model of compounding tendency can be based on the assumption that the pressure to reduce complexity is a continuous quantity. This pressure may vary with infinitely small amounts. In some cases, the pressure reaches a threshold and part of the syntactic structure is replaced by a compound (under the boundary conditions discussed above) or, if a compound was already used or could have been used without sufficient complexity reduction, a more complex compound is formed.

We replace therefore the difference equation (4)

$$P_x = \frac{maxH}{x} P_{x-1}. \qquad (4)$$

where values of the quantity under consideration at subsequent discrete points it time are related to each other by a similar but continuous approach (7):

$$y' = maxH\, y \qquad (7)$$

which can be written as

$$y'/y = maxH, \qquad (8)$$

a very simple differential equation, which corresponds to a process with a constant relative change. The solution to this equation is

$$y = maxH\, e^{bx}. \qquad (9)$$

where *b* is an empirical parameter (the integration constant). This model was proposed by Popescu, Altmann, and Köhler (2010) and has been used successfully for several kinds of stratified data. As seen above, we expect in our data at least two strata, one consisting of lexicalised compounds, the other one of *ad-hoc* constructions. We furnish therefore our model with two terms of the form (9) and add the constant 1 because we will, of course, not observe a class with less than 1 compounds. The model has now the form (10):

$$y = 1 + maxH\, e^{bx} + Lex\, e^{cx} \qquad (10)$$

The parameters *maxH* and *Lex* represent the two sources of compounds; their values and those of the parameters *b* and *c* have to be estimated from the data.
Fitting this model to our data yields an extremely good fit with a determination coefficient R^2 = 0.9997. The values of the parameters were estimated as

maxH = 17.0059
b = -0.3398
Lex = 3012.78
c = -1.7802

Figure 2 shows the result in graphical form.

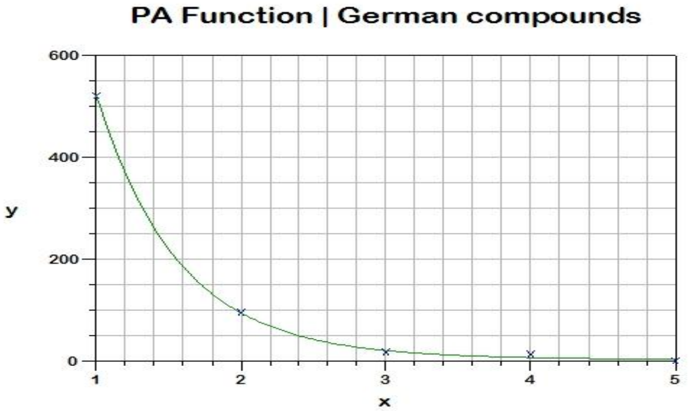


Fig. 2: Fitting the Popescu-Altmann function to the data.

Surprisingly, the version of the function with only one of the two terms, i.e.

$$y = 1 + a\,e^{bx} \qquad (11)$$

yielded a fitting result which is as good as the previous one: R^2 = 0.9995 with only two parameters: a = 282.1151 and b = -1.6919. The first parameter a represents both sources of compounds at the same time.

5. Conclusion

The test of the discrete model yielded a result, which indicates that it is compatible with the data. The goodness-of-fit values are not very good. A possible reason is that the large number of compounds with 4 elements in this text is likely to be an exception and due to the specific properties of the text type. Nevertheless, the attempt to derive a quantitative model of compounding by combining models of syntactic aspects and of morphological productivity was not rejected so far. On the other hand, we expect that a much more complex model will be needed to capture the boundary conditions imposed on the processes by differences between text types, syntactic rules and compounding mechanisms of individual languages.

As opposed to this result, we obtained an excellent fit with the continuous model, which is not affected by degrees of freedom and is much less sensitive against deviations such as the peak in the 4-elements class. This may, however, turn out to be a disadvantage because such stable results may fail to show the differences mentioned in the last paragraph.

It goes without saying that we will need much more data from as many languages as possible together with relevant information about the grammatical structures of the languages and meta data to the texts under analysis.

At the same time, the mathematical model should be extended according to theoretical considerations with respect to boundary conditions and functional alternatives. In section 1, we mentioned that Naumann obtained frequency distributions of syntactic complexity which did not confirm with the hypothesis expressed by formula (3). A plausible explanation of this fact is the following: The hyper-Pascal distribution can be expected only if the boundary condition E (full explicitness of the semantic information) holds. In this case, our model (6) may fail because syntactic complexity has no chance to be reduced by means of compounds. When we analyse texts in which context and world knowledge on the side of the readers compensate for loss of explicit information, more compounds may be expected, model (6) may be compatible with the data but the syntactic complexity may differ from predictions made by model (3). As a consequence, a complex model which covers both aspects and the corresponding boundary conditions must be developed.

Table 2a
The simplexes in the text *http://de.wikipedia.org/wiki/Grundstückverkehrsgesetz*

word	token frequency	word	token frequency	word	token frequency
der	39	verbessert	1	basis	1
die	31	bevölkerung	1	berechnen	1
in	17	makroökonomische	1	sei	1
und	15	aspekte	1	vermögen	1
ist	12	hierzu	1	sein	1
des	11	folgende	1	sonstigen	1
den	7	regelungen	1	zugelassen	1
vor	7	getroffen	1	sich	1
ein	7	behördlichen	1	anbetracht	1
im	7	hof	1	deutsche	1
dem	6	wege	1	prägenden	1
einem	6	gesetzlichen	1	schließlich	1
an	6	gerichtlichen	1	mildeste	1
zu	6	sowie	1	frage	1
das	5	bestellung	1	stehenden	1
wird	5	solchen	1	lösungen	1
von	5	erforderlich	1	entschieden	1
allem	5	antrag	1	dennoch	1
werden	5	erteilt	1	so	1
vom	5	laufe	1	alfred	1
genehmigung	4	jahre	1	pikalo	1
eines	4	immer	1	bernold	1
ob	4	liberaler	1	bendel	1
bei	4	geworden	1	ihrem	1
land	3	weil	1	kommentar	1
betriebe	3	erkenntnis	1	bleibt	1
besonders	3	gewonnen	1	unserer	1
hat	3	neben	1	da	1
veräußerung	3	betriebene	1	sie	1
bedarf	3	aus	1	nämlich	1
besonderen	3	anderen	1	steht	1
eine	3	gründen	1	inzwischen	1
kann	3	länder	1	besteht	1
zugewiesen	3	zum	1	mehr	1
auf	3	bestimmt	1	bisherige	1
dass	3	bestimmten	1	gilt	1
auch	3	größe	1	ländern	1
einer	3	keiner	1	fort	1
unter	3	dabei	1	es	1
keine	3	i	1	ersetzt	1

zuweisung	3	r	1	dies	1
durch	3	verstehen	1	bisher	1
mit	2	h	1	nur	1
genutzten	2	räumlich	1	erfolgt	1
sicherung	2	abgegrenzter	1	gehört	1
indem	2	ohne	1	zum	1
erhalten	2	art	1	beispiel	1
fällt	2	seiner	1	seltenen	1
nach	2	nutzung	1	südlichen	1
oder	2	nummer	1	noch	1
nicht	2	eingetragen	1	anzutreffen	1
boden	2	wirtschaftliche	1	er	1
bis	2	spielen	1	dann	1
d	2	rolle	1	gehen	1
teil	2	ferner	1	vorschriften	1
betrieb	2	13	1	deren	1
tod	2	geregelt	1	beendigung	1
wegen	2	wonach	1	geschichte	1
gesetzes	2	wenn	1	kennt	1
juli	2	entsprechende	1	folgenden	1
bereits	2	verfügung	1	stationen	1
15	2	todes	1	märz	1
war	2	vorliegt	1	1918	1
betriebs	2	voraussetzung	1	wollte	1
umstritten	2	interessant	1	notzeit	1
sollte	2	diesem	1	ersten	1
außer	2	lange	1	aufkauf	1
solle	2	gearbeitet	1	vermögens	1
für	2	worden	1	26	1
verhindern	2	erste	1	januar	1
kontrollierend	1	stammte	1	1937	1
eingreift	1	ihm	1	verwirklichung	1
verfolgt	1	möglichkeit	1	blut	1
vornehmlich	1	geschlossenen	1	ideologie	1
drei	1	einen	1	dienen	1
zwecke	1	bezeichnung	1	sollten	1
ausverkauf	1	vorgesehen	1	wiederum	1
ihres	1	dieser	1	verfolgten	1
bodens	1	am	1	ziel	1
geschützt	1	heftigsten	1	zerschlagen	1
mikroökonomischer	1	waren	1	will	1
aspekt	1	fragen	1	verbessern	1
betont	1	rechnung	1	bäuerliche	1
schutz	1	getragen	1	hand	1
natur	1	abfindung	1	familien	1
umwelt	1	weichenden	1	sichern	1

Table 2b
The compounds with two elements in the text
http://de.wikipedia.org /wiki/Grundstückverkehrsgesetz

word	token fre-quency	word	token fre-quency	word	token fre-quency
miterben	4	geschäftsverkehr	1	außerordentlich	1
gesetzgeber	3	landwirtschaftlich	1	fragwürdiges	1
landwirtschaft	3	fortbestandes	1	rechtsinstitut	1
grundstück	3	ernährungsvorsorge	1	fremdkörper	1
landwirtschaftlichen	3	rechtsgeschäftliche	1	rechtsordnung	1
grundstücken	2	genehmigungsverfahren	1	widerspruch	1
forstwirtschaftlicher	2	erbfolge	1	grundprinzipien	1
agrarstruktur	2	forstwirtschaftlich	1	vertragsfreiheit	1
erbengemeinschaft	2	grundstücks	1	eigentumsschutz	1
zuweisungsverfahren	2	nießbrauchs	1	bundeskompetenz	1
marktwirtschaft	2	genehmigungspraxis	1	landesgesetz	1
fortgesetzten	2	nebenberuflich	1	güterstand	1
gütergemeinschaft	2	agrarpolitischen	1	landwirten	1
volksernährung	2	volkswirtschaftlichen	1	landwirts	1
sicherstellen	2	erhaltungswürdig	1	auseinandersetzung	1

Table 2c
The compounds with three elements in the text
http://de.wikipedia.org /wiki/Grundstückverkehrsgesetz

word	token fre-quency	word	token fre-quency	word	token fre-quency
Landwirtschafts-behörde	1	grundbuchblattes	1	erblasserwillens	1
negativzeugnis	1	insoweit	1	gesamthandsgemeinsch aften	1
Landwirtschafts-behörden	1	landwirtschafts-gericht	1	grundstückslenkung	1
Landwirtschafts-gerichte	1	landwirtschafts-betrieb	1	bundesratsbekanntma-chung	1
Vollerwerbs-betrieben	1	betriebsübergabe	1	nachkriegszeit	1
erdoberfläche	1	gesetzgebungs-verfahren	1	großgrundbesitz	1

Table 2d
The compounds with four elements in the text
http://de.wikipedia.org /wiki/Grundstückverkehrsgesetz

word	**token frequency**
grdstvg	7
grundstückverkehrsgesetz	3
grundstücksverkehrsgesetz	2
grundstücksverkehrsgesetzes	1
asvg	1

Table 2e
The compound with five elements in the text
http://de.wikipedia.org /wiki/Grundstückverkehrsgesetz

word	**token frequency**
grundstückverkehrsbekanntmachung	1

References

Altmann, Gabriel (1991). Modelling diversification phenomena in language. In: Rothe, Ur sula (ed.), *Diversification processes in language: Grammar*. Hagen: Rottmann, 33-46.

Altmann, Gabriel; Köhler, Reinhard (1996). "Language Forces" and Synergetic Modelling of Language Phenomena. In: P. Schmidt (ed.), *Glottometrika 15. Issues in General Linguistic Theory and The Theory of Word Length.* Trier: WVT, 62-76.

Köhler, Reinhard (1986). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik.* Bochum: Brockmeyer. (= Quantitative Linguistics**;** 31)

Köhler, Reinhard (1990). Synergetik und sprachliche Dynamik. In: Walter A. Koch (Hg.): *Natürlichkeit der Sprache und der Kultur.* Bochum: Brockmeyer, 96-112.

Köhler, Reinhard (2005). Synergetic Linguistics. In: Köhler, Reinhard; Altmann, Gabriel; Piotrowski, Rajmund G. (eds.), *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook*. Berlin, New York: de Gruyter (= Handbücher zur Sprach- und Kommunikationswissenschaft; 27), 760-775.

Köhler, Reinhard; Altmann, Gabriel (1993). Begriffsdynamik und Lexikonstruktur. In: Frank Beckmann, Gerhard Heyer (eds.), *Theorie und Praxis des Lexikons.* Berlin, New York: de Gruyter, 173-190.

Köhler, Reinhard; Altmann, Gabriel (2000). Probability distributions of syntactic units and properties. *Journal of Quantitative Linguistics 7(3), 189–200.*

Naumann, Sven (2015). Structural versus morphological coding. A cross-linguistic study. To appear.

Popescu, Ioan-Iovitz; Altmann, Gabriel; Köhler, Reinhard (2010). Zipf's law – another view. *Quality & Quantity, 44(4), 713–731.*

Wimmer, Gejza; Altmann, Gabriel (1995). A model of morphological productivity. *Journal of Quantitative Linguistics* 2, 212-216.

Wimmer, Gejza; Altmann, Gabriel (1999). *Thesaurus of univariate discrete proba bility distributions.* Essen: Stamm.

Wimmer, Gejza; Köhler, Reinhard; Grotjahn, Rüdiger; Altmann, Gabriel (1994). Towards a theory of word length distribution. *Journal of Quantitative Linguistics* 1, 98-106.

Glottometrics 28, 2014, 87-91

Bibliography

Diversification

Diversification is one of the two main Zipfian processes (opposed to unification). It arises by self-organization and increases the number of variants or meanings of an entity, be it phonemic, morphological, syntactic, semantic, lexicological, idiolectal, dialectal, sociolectal, etc. This expansion is, however, not chaotic but controlled by the whole system of properties and by the requirements of language carriers (cf. Köhler 2005). It is based on the birth-and-death process. Semantic diversification is sometimes called "Beöthy Law", the dialectal version "Goebl Law".

General

Altmann, G. (1991). Modelling diversification phenomena in language. In: Rothe, U. (eds.): *33-46.*

Altmann, G. (1996). Diversification processes of the word. In: P. Schmidt, (ed.), *Glottometrika 15, 102-111.* Trier: WVT.

Altmann, G. (2005). Diversification processes. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 645-648.* Berlin: de Gruyter.

Best, K.-H. (2003). Slawische Entlehnungen im Deutschen. In: S. Kempgen, U. Schweier, T. Berger (eds.), *Rusistika - Slavistika - Lingvistika. Festschrift für Werner Lehfeldt: 464-473.* München: Sagner

Best, K.-H. (2006). *Quantitative Linguistik: Eine Annäherung.* 3., stark überarbeitete und ergänzte Auflage. Göttingen: Peust & Gutschmidt.

Čech, R., Altmann, G. (2011). *Problems in Quantitative Linguistics Vol. 3.* Lüdenscheid: RAM-Verlag.

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

Köhler, R., Altmann, G. (2009). *Problems in Quantitative Linguistics Vol. 2.* Lüdenscheid: RAM-Verlag.

Popescu, I.-I., Altmann, G. (2008). On the regularity of diversification in language. *Glottometrics 17, 97-111.*

Prün, C. (2005). Das Werk von G.K. Zipf. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 142-152.* Berlin: de Gruyter.

Rigo, G. (1994). L'entropie comme indice de diversification du vocabulaire dans les tragèdies de Sophocle. *Revue Informatique et Statistique dans les Sciences humaines 30, 109-125.*

Rothe, U. (1991). Diversification processes in grammar. An introduction. In: Rothe, U. (ed.) (1991).

Schweiger, F. (1987). Zu den Modellen der semantischen Diversifikation von G. Altmann. *Folia Linguistica 21, 191-194.*

Zipf, G.K. (1949). *Human behavior and the principle of least effort.* Cambridge, Mass.: Addison-Wesley.

Phonemics and script

Best, K.-H., (2009). Diversifikation des Phonems /r/ im Deutschen. *Glottometrics 18, 26-31.*

Best, K.-H., (2011). Diversification of a single sign of the Danube script. *Glottometrics 22, 1-4.*

Mačutek, J. (2008). On the distribution of graphemic representations. In: Altmann, G., Fan, F. (eds.), *Analyses of script. Properties and Characters of writing systems: 75-78.* Berlin-New York: Mouton de Gruyter.

Rothe, U. (1991). Distribution of spelling errors by Japanese English-Users. In: Rothe, U. (ed.) (1991), *168-171.*

Grammar

Altmann, G. (1991). Word class diversification of Arabic verbal roots**.** In: Rothe, U. (ed.) (1991), 5*7-59.*

Best, K.-H. (1994). Word class frequency in contemporary German short prose texts. *Journal of Quantitative Linguistics 1, 144-147.*

Best, K.-H. (1997). Zur Wortartenhäufigkeit in Texten deutscher Kurzprosa der Gegenwart, In: Best, K.-H. (ed.), *Glottometrika 16, 276-285.* Trier: WVT.

Best, K.-H. (2007). Kürzungstendenzen im Deutschen aus der Sicht der Quantitativen Linguistik. In: Bär, J.A., Roelcke, T., Steinhauer, A. (eds.), *Sprachliche Kürze. Konzeptuelle, strukturelle und pragmatische Aspekte: 45-62.* Berlin/NewYork: de Gruyter.

Best, K.-H. (2008; appeared 2010). Verteilungen von Fugenelementen im Deutschen. *Göttin ger Beiträge zur Sprachwissenschaft 16, 7-16.*

Best, K.-H. (2012). Diversifikation der starken Verben im Deutschen. *Glottometrics 24, 1-4.*

Boschtan, A., Best, K.-H. (2010). Diversification of simple attributes in German. *Glottotheory 3(2), 5-9.*

Brüers, N., Heeren, A. (2004). Plural-Allomorphe in Briefen Heinrich von Kleists. *Glotto metrics 7, 85-90.*

Junger, J. (1989). Diversification in the modern Hebrew verbal system. In: Hammerl, R. (ed.), *Glottometrika 10, 71-99.* Bochum: Brockmeyer.

Köhler, R. (1991). Diversification of coding methods in grammar. In: Rothe, U. (ed.) (1991)*, 47-55.*

Köhler, R. (2005). Quantitative Untersuchungen zur Valenz deutscher Verben. *Glottometrics 9, 13-20.*

Laufer, J., Nemcová, E. (2009). Diversifikation deutscher morphologischer Klassen in SMS. *Glottometrics 18, 13-35.*

Meuser, K., Schütte, J., Stremme, S. (2008). Pluralallomorphe in den Kurzgeschichten von Wolfdietrich Schnurre. *Glottometrics 17, 12-17.*

Popescu, I.-I., Kelih, E., Best, K.-H., Altmann, G. (2009). Diversification of the case. *Glottometrics 18, 32-39.*

Prün, C., Steiner, P. (2005). Quantitative Morphologie: Eigenschaften der morphologischen Einheiten und Systeme. In: Köhler, R., Altmann, G., Piotrowski, R.G.

(eds.), *Quantitative Linguistics. An International Handbook: 227-242.* Berlin: de Gruyter.

Raether, A., Rothe, U. (1991). Diversifikation der deutschen Komposita „Substantiv plus Substantiv". In: Rothe, U. (ed.) (1991), *85-91.*

Rothe, U. (ed.) (1991). *Diversification processes in language: grammar.* Hagen: Rottmann.

Schweers, A., Zhu, J. (1991). Wortartenklassifikation im Lateinischen, Deutschen und Chinesischen. In: Rothe, U (ed.) (1991), *157-165.*

Steiner, P. (2013). Diversification of English Valency Patterns. In: Köhler, R., Altmann, G. (eds.), *Issues in Quantitative Linguistics 3: 369-391.* Lüdenscheid: RAM-Verlag.

Steiner, P., Prün, C. (2007). The effects of diversification and unification on the inflectional paradigms of German nouns. In: Grzybek, P., Köhler, R. (eds.), *Exact Methods in the Study of Language and Text:* 623-632. Berlin/ New York: Mouton de Gruyter.

Tuzzi, A., Popescu, I.-I., Altmann, G. (2009). Parts-of-speech diversification in Italian texts. *Glottometrics 19, 42-48.*

Ziegler, A. (1998). Word class frequencies in Brasilian-Portuguese press texts. *Journal of Quantitative Linguistics 4, 269-280.*

Ziegler, A. (2001). Word class frequencies in Portuguese press texts. In: Uhlířová, L., et al. (eds.), *Text as a linguistic paradigm: levels, constituents, constructs: 295-312.* Trier: WVT.

Semantics

Altmann, G. (1985). Semantische Diversifikation. *Folia Linguistica 19, 177-200.*

Altmann, G. (1992). Two models for word association data. In: B. Rieger (ed.) *Glottometrika 13, 105-120.*

Altmann, G., Best, K.-H., Kind, B. (1987). Eine Verallgemeinerung des Gesetzes der semantischen Diversifikation. In: Fickermann, I. (ed.), *Glottometrika 8, 130-139.* Bochum: Brockmeyer.

Altmann, G., Kind, B. (1983). Ein semantisches Gesetz. In: Köhler, R., Boy, J. (eds.), *Glottometrika 5, 1-13.*Bochum: Brockmeyer.

Beöthy, E., Altmann, G. (1984). The diversification of meaning of Hungarian verbal prefixes. II. ki-. *Finnisch-Ugrische Mitteilungen 8, 29-37.*

Beöthy, E., Altmann, G. (1984). Semantic diversification of Hungarian verbal prefixes. III. "föl-", "el-", "be-". In: Rothe, U. (ed.), *Glottometrika 7, 45-56.* Bochum: Brockmeyer.

Beöthy, E., Altmann, G. (1991). The diversification of meaning of Hungarian verbal prefixes I. "meg-". In: Rothe U. (ed.) (1991), *60-66.*

Best, K.-H. (1990). Die semantische Diversifikation eines Wortbildungsmusters im Frühneuhochdeutschen. In: Hřebíček, L. (ed.), *Glottometrika 11, 107-110.* Bochum: Brockmeyer.

Best, K.-H. (1991). *Von:* Zur Diversifikation einer Partikel des Deutschen. In: Rothe, U. (ed.) (1991), 9*4-104.*

Best, K.-H., Boschtan, A. (2010). Diversification of simple attributes in German. *Glottotheory 3(2), 5-9.*

Fan, F., Altmann, G. (2008). On meaning diversification in English. *Glottometrics 17, 69-81.*

Fan, F., Popescu, I.-I., Altmann, G. (2008). Arc length and meaning diversification in English. *Glottometrics 17, 82-89.*

Fickermann, I., Markner-Jäger, B., & Rothe, U. (1984). Wortlänge und Bedeutungskomplexität. In: Boy, J., & Köhler, R. (Hrsg.) *Glottometrika 6: 115-126.* Bochum: Brockmeyer.

Fuchs, R. (1991). Diversifikation der Präposition *auf.* In: Rothe, U. (ed.) (1991),*105-115.*

Haight, F.A. (1966). Some statistical problems in connection with word association data. *Journal of Mathematical psychology 3, 217-233.*

Hammerl, R., Sambor, J. (1991). Untersuchungen zum Beöthy-Gesetz im Polnischen. In: Rothe, U. (ed.) (1991), *127-137.*

Hennern, A. (1991). Zur semantischen Diversifikation von "in" im Englischen. In: Rothe, U. (ed.) (1991), *116-126.*

Horvath, W.J. (1963). A stochastic model for word association tests. *Psychological Review 70, 361-354.*

Hřebíček, L. (1996). Word associations and text. In: Schmidt, P. (ed.), *Glottometrika 15, 96-101.* Trier: WVT.

Kuße, H. (1991). A und no in N.M. Karamzins Pis'ma Russkogo Petešestvennika. In: Rothe, U. (ed.) (1991), 173-182.

Nemcová, E. (1991). Semantic diversification of Slovak verbal prefixes. In: Rothe, U (ed.) (1991), 67-74.

Nemcová, E., Popescu, I.-I., Altmann, G. (2010). Word associations in French. In: Berndt, A., Böcker, J. (eds.), *Sprachlehrforschung: Theorie und Enpirie: 223-237.* Frankfurt: Lang.

Popescu, I.-I., Mačutek, J., Altmann, G. (2009). Aspects of word frequencies. Lüdenscheid: RAM-Verlag (esp. p. 86 ff.)

Roos, U. (1991). Diversifikation der japanischen Postposition "ni". In: Rothe, U. (ed.) (1991), 75-82.

Rothe, U. (1986). *Die Semantik des textuellen et.* Frankfurt: Lang.

Rothe, U. (1989). Semantische Beziehungen zwischen Präfixen deutscher denominaler Verben und den motivierenden Nomina. In: Hřebíček, L. (ed.), *Glottometrika 11: 111-121.* Bochum: Brockmeyer.

Rothe, U. (1990). Die Verteilung denominaler Verben nach ihren semantischen Wortbildungsmustern. In: Hammerl, R. (ed.), *Glottometrika 12:* 107-114. Bochum: Brockmeyer.

Rothe, U. (1991). Diversification of the case in German: genitive. In: Rothe, U. (ed.) (1991), 140-156.

Sanada, H., Altmann, G. (2009). Diversification of postpositions in Japanese. *Glottometrics 19, 70-79.*

Dialectology

Altmann, G. (1985). Die Entstehung diatopischer Varianten. Ein stochastisches Modell. *Zeitschrift für Sprachwissenschaft 4, 139-155.*

Etymological spectra

Best, K.-H. (2004). Das Fremdwort aus der Sicht der Quantitativen Linguistik. In*: Theorie, Steuerung und Medien des Wissenstransfers: 89-99.* Ed. by S. Wichter, O. Stenschke, M. Tants. Frankfurt: Lang.

Best, K.-H. (2005). Diversifikation der Fremd- und Lehnwörter im Türkischen. *Archív Orientální 73, 291-298.*

Best, K.-H. (2005). Ein Modell für das etymologische Spektrum des Wortschatzes. *Naukovyj Visnyk Černivec'koho Universytetu: Hermans'ka filolohija. Vypusk 266, 11-21.*

Best, K.-H. (2008). Das Fremdwortspektrum im Türkischen. *Glottometrics 17, 8-11.*

Best, K.-H. (2009). Zum etymologischen Spektrum des Hundeshagener Kochums. *Göttinger Beiträge zur Sprachwissenschaft 19, 25-29.* (Erschienen Ende 2010.)

Best, K.-H. (2010). Zum Fremdwortspektrum im Japanischen. *Glottotheory 3, 5-8.*

Names

Best, K.-H. (2005). Ernst Wilhelm Förstemann (1822-1906). *Glottometrics 12, 77-86.*

Best, K.-H. (2007). Diversifikation bei Eigennamen. In: Grzybek, P., Köhler, R. (eds.), *Exact Methods in the Study of Language and Text. Dedicated to Gabriel Altmann on the Occasion of his 75th Birthday: 21-31.* Berlin/ New York: Mouton de Gruyter.

Poetry

Best, K.-H. (2008). Zur Diversifikation lateinischer und griechischer Hexameter. *Glottometrics 17, 43-50.*

Best, K.-H. (2009). Zur Diversifikation deutscher Hexameter. *Naukovyj Visnyk Černivec'koho Universytetu: Hermans'ka filolohija.Vypusk 431, 172-180.*

Best, K.-H. (2008). Moritz Wilhelm Drobisch (1802-1896). *Glottometrics 17, 109-114.*

Karl-Heinz Best

Book Review

JI, Meng (2013). *Exploratory Statistical Techniques for the Study of Literary Translation.* Lüdenscheid: RAM-Verlag, IV+124 pp. ISBN 978-3-942303-17-0,

Reviewed by **Ruina Chen**

Despite the increasing use of corpus material and corpus methodologies in translation studies, there is a lack of systematic descriptions of quantitative methods that may be used for corpus translation studies. Such a situation poses serious hindrance for the theoretical development of the discipline as a whole. An important figure in this field in recent years is Meng Ji, a scholar and professor at Tokyo University, who is actively involved in the development of statistical concepts within the context of translation studies. Her new book *Exploratory Statistical Techniques for the Study of Literary Translation* (2013), adopts an essentially corpus-driven multivariate analysis of different sets of corpora, introducing principal component analysis and hierarchical clustering analysis from applied linguistics to translation studies. This book can be regarded as a parallel with *Quantitative Methods in Corpus-Based Translation Studies: A Practical Guide to Descriptive Translation Research*, edited by Oaks and Ji (2012), with the shared goal of introducing systematic quantitative methods of analysis to translation studies.

In corpus translation studies, the use of comparative statistics such as the chi-square test or log-likelihood test to compare data from two corpora based on the computation of comparative statistics is a well-established methodology. But this does not lend itself well to the comparison of multiple corpora and the subsequent visualization of the statistical result. In such circumstances, exploratory statistical techniques like principal component analysis and factor analysis gain their own weight. Both have been explored extensively in applied linguistics by Douglas Biber and his colleagues (Biber 1992; Conrad and Biber 2001, Biber 2006), whose multi-dimensional analysis methodology involves quantitatively and qualitatively analyzing large corpora of texts and identifying and describing linguistic variation contained in texts of academic speech and writing. But this methodology has been scarcely scrutinized in translation studies.

Translation studies requires the development of quantitative research materials and appropriate research methods to identify and analyze complex networks of relationships between the various social, cultural, stylistic, generic factors and the textual representation of a translation, with the central aim to "verify the existence of any statistically significant correlation between various textual variables under investigation" (Ji 2012: 56). But this remains under-explored in the past due to "the lack of relevant textual data purposely collected and annotated in the form of language corpora, and most importantly, the availability of advanced research methods to inquiry into the complex structure and changing nature of translation data" (Ji 2012: 55). With the construction and exploration of large-scale translation corpora, like Zhejiang University Corpus of Translational Chinese (ZCTC), and original language corpora like the Lancaster Corpus of Mandarin Chinese (LCMC), it becomes possible to extract, recognize, process and compare textual variables between different languages, in the current situation, that is, English and Chinese.

Corpus-based or corpus-driven translation studies are essentially experimental and exploratory, most studies investigated under this vein are with few presumptions made regarding the existence of any theoretical models and constructs. They "start from the deliberate construction of corpus resources which leads to the discovery of new textual and linguistic patterns. Textual patterns uncovered in translational corpora form the basis for the development of general conclusions. The generalizability of the conclusions made depend on the scale and size of the corpus materials used and analyzed" (Ji 2013: 72).

The methodological advantage of exploratory analysis over traditional corpus comparison is that "it can analyze and classify a large number of corpora as observational variables simultaneously" (Ji 2013: 63).The differences or variations among observational variables are measured by their linguistic and textual properties, for instance, the frequencies of occurrence of specific lexical and grammatical categories. The quantifying variables can be used for the construction of statistical models, composed of major dimensions or components extracted by principal component analysis. The similarities and dissimilarities among the different corpora may be gauged and detected by several statistical indicators, like the factor score (the larger is the factor score of an observational variable, the stronger is the correlation between the corpus and a specific principal component or dimension) and the Squared Euclidean distance (the smaller is the distance score, the more similar is a specific corpus with a principal component, dimension or a reference corpus). In this way, the comparative analysis of the observational variables, that is, the similarities and dissimilarities among them is streamlined. The result of exploratory empirical analysis may furnish important basis for the formulation and development of theoretical hypotheses for translation studies.

Exploratory techniques adopted in this book is principal component analysis and hierarchical clustering. The main purpose of principal component analysis is to identify the underlying patterns which can maximally explain variations and changes in the observational variables. Hierarchical clustering analysis aims to identify observational variables that are most similar to each other to form different levels of cluster, agglomerative clustering process continues until similar clusters merge together. These techniques have been used here to explore some unique and latent translational phenomenon that has been rarely discussed in translation studies, let alone from an empirical and quantitative perspective, like the issue of genre shifting or stylistic variation between the source and the target language, which has subverted one of the traditional view that the genre of a literary translation is always consistent with that of the source text; and also the issue of legitimate role of translated languages as a unique genre, which corroborates the existing hypothesis of translation universals.

The book comprises three independent empirical studies. The first one, "Stylistic Variation in Literary Translations: A Corpus Study of Two Chinese Translations of *One Hundred Years of Solitude*", explores the two Chinese translations of Garcia Marquez's *One Hundred Years of Solicitude.* Corpus comparison of a range of part-of-speech taggers in 19 corpora, including ZCTC, LCMC and two versions of translation of the novel, find that the version which is indirectly translated via English shows more resemblance with Chinese translations of detective and mystery fictions, the other one which is directly from Spanish is more close to Chinese translations of romance fictions, as is indicated by the dissimilarity matrix of squared Euclidean distance by the use of hierarchical clustering analysis.

The second study, "Genre Shifting in Literary Translation: A Corpus Study of Chinese

Translation of Confessions”, investigates key linguistic features and textual patterns in the Chinese version of Minato Kanae’s *Confessions*, a bestseller book in Japanese. Principal component analysis is first conducted on the Japanese-Chinese corpus of *Confessions*, sub-corpora of LCMC and ZCTC, and major statistical dimensions of individual corpus and their factor scores are derived and compared. Hierarchical clustering analysis is then used to detect the squared Euclidean distances among them. Comparison among the corpora results reveal that the Chinese translation has shifted from a detective and mystery fiction in the Japanese genre system to a religious text in the Chinese genre system.

The third study, “Enhanced Idiomaticity as a Potential Translation Universal: A Corpus Exploratory Analysis of Modern Chinese Translation”, identifies key linguistic and textual features of the translated Chinese. Principal component analysis between Chinese translational corpora (ZCTC) with the original Chinese corpora (LCMC) verifies the existence of common linguistic features in translations, generally described as translation universals or translationese. For example, the high frequencies of occurrence of some punctuation marks support the explication in translational Chinese; and the more frequent use of personal pronouns, object pronouns and location pronouns indicate enhanced cohesion of translational Chinese; and the increased use of idiomaticity supports the translation universal features such as normalization and standardization (Ji 2010).

It is quite obvious that the author here focuses on employing principal component analysis and hierarchical clustering analysis to conduct contrastive literary translation studies in terms of the number of factor scores and squared Euclidean distance scores of respective corpus, which may be conspicuously misleading to concentrate exclusively on specific, isolated linguistic markers without taking into account systematic variations which involve the co-occurrence of set of markers. Among the three case studies discussed in this book, two of them talk about the variation of the style or genre of a translation, which may be better approached by a set of co-occurring linguistic features rather than individually. In addition, the interpretation of each dimension in a particular genre or corpus should involve both linguistic and functional content, which remains the real essence of corpus-driven analysis. Thus, Ji’s illustration of the specific linguistic patterns and textual features in the translated Chinese in the third study may not just stop at using quantitative techniques; adding the co-occurrence patterns and the interpretation of their function in terms of the situational, social, and cognitive context may strengthen its conviction.

On the whole, this book reinforces the potential of quantitative methods in the exploration of intriguing textual and generic relationships bearing on the nature of literary translation. As the author states in the conclusion, the role of corpus-oriented approaches in the translation studies should be broadened to “the discovery of widely existent patterns in translations, instead of being relegated to the largely supplementary role of the verification and testing of presumed hypothesis” (Ji 2013:72). In this sense, this book succeeds in revealing the possible inconsistency in generic patterns between the source and target language of the same literary works, as well as the unique linguistic and textual patterns of the translational language, which still remain some of the crucial issues in translation studies. In addition, considering the possible mathematical and technical limitation of most translation researchers, the illustration of the statistical methods is concise and clear, the discussion of the relationship between the type of linguistic features and the research questions under exploration is through and detailed, with the appendix information attached in the book, the interested readers may have a hands-on experience with the exploratory empirical methods themselves. The statistical methods and research questions proposed in this book, plus those in the other counterpart (Oaks and Ji, 2012), will provide readers with the most up-dated development and innovation in the field of corpus-assisted contrastive translation studies.

References

Oaks, M.P., Ji, M. (eds) 2013. *Quantitative Methods in Corpus-Based Translation Studies: A Practical Guide to Descriptive Translation Research*, Amsterdam and Philadelphia, PA: John Benjamins.

Biber, D. (1992). The Multi-Dimensional Approach to Linguistic Analyses of Genre Variation: An Overview of Methodology and Findings. *Computers and the Humanities*, 26(5-6), pp. 331-345.

Biber, D. (2006). University Language: *A Corpus-based Study of Spoken and Written Registers*. Amsterdam: John Benjamins Publications.

Conrad, S., Biber, D. (eds.) (2001). *Variation in English: Multi-Dimensional Studies*. Essex: Pearson Education Limited.

Ji, M. (2010). *Phraseology in Corpus-Based Translation Studies.* Oxford and London: Peter Lang International Academic.

Ji, M. (2012). Hypothesis Testing in Corpus-Based Literary Translation Studies. In: M.P. Oakes and M. Ji (eds.), *Quantitative Methods in Corpus-Based Translation Studies*, John Benjamins. Pp. 53-74.