Glottometrics 14
2007

RAM-Verlag

ISSN 2625-8226

Glottometrics

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

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

Herausgeber – Editors

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

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

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

Die Deutsche Bibliothek – CIP-Einheitsaufnahme
Glottometrics. 14 (2007), Lüdenscheid: RAM-Verlag, 2007. 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 14 (2007) **ISSN 2625-8226**

Contents

Glottometrics 14, 2007, 1-21

Word frequency and position in sentence

Ludmila Uhlířová, Prague[1]

Abstract. The author hypothesizes that in sentences, there exist positional intervals in which some frequency rhythm is manifested, and applying various techniques and tests she finds relevant arguments in favour of this hypothesis. Data from Bulgarian texts are used.

1. Introduction

Every coherent text is structured in a complex, multifarious way. It is also reasonable to assume that the relationship between the word **frequency** in a text and the **position** which is occupied by that word in the linear (word order) arrangement of a sentence (and text) is also organized in a way: We assume that the distribution of word frequencies **across** the sentence (and text) is neither chaotic, nor uniform, but that it abides by a principle. We shall call the principle **frequency rhythm**. In the following we shall ask: How can this principle be captured?

Any rhythm, including the hypothesized frequency rhythm, or, Altmann's "Wiederholungen" of various kinds (Altmann: 1988), or Köhler's word length rhythm (Köhler 2006, 2007), is a matter of the syntagmatic axis of language. The results offered below represent several empirical statements. The author hopes that they are not contradictory to the philosophical framework of the fundamental work in this field by Altmann (1988) and, subsequently, by Wimmer et al. (2003), and other authors,. It is also hoped that these results may serve as a hint towards more sophisticated and more theoretical reasoning.

The frequency rhythm (defined as word frequency as conditioned by the word position in a sentence) will be tested here on ten texts from Bulgarian. The texts represent a single genre: they are private letters, written by and addressed to men and women with a university education, mostly concerning business topics. The owner of the letters consented to the use of these texts for linguistic analysis (some necessary steps were taken to anonymize them). The letters are numbered from 1 to 10 for the purpose of this article.

Let us take the following sentence:

	Blagodarja	*ti*	*za*	*pismoto*	*ti*	*ot*	*sedmi*	*noemvri.*
Lit.:	I-thank	you	for	letter-the	your	of	seventh	November
word position in the sentence:	1.	2.	3.	4.	5.	6.	7.	8.
word frequency in the text:	1	4	18	1	4	5	1	4

Seeking for the frequency rhythm, we assign serial numbers to the words in each sentence from left to right, and we provide each word with information about its text fre-

[1] Address correpsondence to: Uhlirova@ujc.cas.cz

quency. We shall not be interested in the concrete lexical meanings of words, but exclusively with word positions and the corresponding word frequencies, in other words, we replace the word by its frequency in the given text.

As a starting point, let us take the data from one of the Bulgarian texts (Letter 10). In Table 1 each word in Letter 10 is uniquely identified by its text frequency (given in the respective cell) and its position in sentence (given in the respective row). For technical reasons (for easier counting and better clearness) the sentences are presented in columns that have been ordered from left to right according to their increasing length (not according to the order in which they appear in the text).

Table 1
Word frequency in position (*Letter 10*)

1	2	2	2	25	9	2	9	1	3	2	9	2	1	1	1	4	2	18
1	1	1	3	6	4	1	1	4	1	1	1	1	1	1	2	1	7	1
	3	1	25	18	25	2	1	18	25	1	1	4	26	2	26	23	1	26
	1	1	1	1	18	9	1	1	4	1	18	1	2	4	1	1	1	3
			18	6	1	18	2	4	1	2	1	25	23	2	18	2	26	1
			1	2	1	1	25	5	2	4	1	1	1	23	2	25	1	25
						2	2	1	1	1	7	3	4	1	6	1	1	2
							2	1	5	1	1	3	3	1	1	26	8	1
									1	1	26	1	1	3	3	1	3	23
										2	2	6	1	2	26	1	1	1
											2	8	23	25	4	1	23	18
												3	1	4	1	1	1	4
														2	1	1	18	1
																1	1	5
																		2
1.00	**1.75**	**1.25**	**8.33**	**9.67**	**9.67**	**5.00**	**5.4**	**5.83**	**4.78**	**1.60**	**6.27**	**4.83**	**7.25**	**5.46**	**7.08**	**6.36**	**6.71**	**8.73**

rows continued:

1	9	1	2	3	23	1	5	9	3	1	1	6	1	2	2	1	1	4
3	4	1	6	18	3	18	18	1	2	1	18	1	18	4	6	1	8	1
5	1	1	1	2	5	1	1	3	7	1	1	26	1	2	8	25	25	26
6	2	3	1	2	2	6	4	1	1	26	2	2	2	4	1	18	1	2
1	5	1	23	1	1	18	1	25	26	3	1	2	26	1	2	1	26	3
1	23	4	1	6	6	2		2	1	1	25	9	4	2	25	1	1	9
1	1	1	18	8	4	1	1	1	1	2	18	18	6	18	1	26	1	4
1	4	23	1	2	25	4	26	7	4	26	1	1	1	1	1	1	1	1
4	6	1	2	1	18	1	2	1	1	1	2	1	1	25	1	25	1	25
2	8	1	2	4	1	1	1	7	1	3	8	2	1	1	1	18	2	1
1	1	1	3	1	7	6	1	1	1	1	18	23	1	1	26	1	1	23
26	18	2	1	23	1	1	2	26	3	6	1	2	3	1	1	23	1	1
4	1	1	1	2	16	1	5	1	2	1	1	26	1	2	1	25	1	3
4	1	26	23	3	1	1	1	1	1	1	23	2	1	1	18	18	23	2
1	23	1	1	1	26	23	26	18	1	18	18	9	1	25	25	1	26	6
1	1	1	2	3	2	4	1	1	1	4	1	18	6	1	1	1	1	1
						3	1	1	1	1	18	2	1	2	1	4	23	1
										1	1	3	1	1	1	25	26	1
												8	18	2	18	4	1	7
												1	3	5	1	2	1	3
													2	2	1	1	2	2
														5	23	23	2	23
														1	4	4	2	1

															3	3	1	18
																	1	3
																	25	1
																	4	6
																	2	1
																	1	1
																	4	2
																	1	1
																	4	3
																	1	18
																		1
																		26
																		2
3.88	**6.75**	**4.31**	**5.50**	**5.00**	**8.81**	**5.41**	**5.65**	**6.24**	**3.35**	**5.44**	**8.78**	**8.10**	**4.71**	**4.74**	**7.17**	**10.50**	**6.70**	**6.5**

For instance, in the sentence which consists of two words, the word in the first (= initial) position has $f = 1$ (first column, first row), and also the word in the second (= final) position has $f = 1$ (first column, second row), etc.

As can be seen from Table 1, our text consists of 38 sentences; the shortest one has only two words, whereas the longest one has 36 words. The range of sentence lengths is large, and the number of sentences of the same length is small. True, there is a group of six sentences which are 16 words long and there is another group of four sentences which are 17 words long, but, unfortunately, there is, e.g., just one sentence which is 9 words long, just one sentence which is 10 words long, one sentence which is 11 words long etc. Due to the considerable dispersion of sentence lengths in our text (and due to its limited size) any direct comparison of word frequencies in sentence positions would be awkward. That is the reason why we dare to propose the following experiment.

2. "Relative" position

Let us introduce the abstract notion of **relative position**. Let us divide **each** sentence in the text (Letter 10) into ten equal length intervals <0.0 - 0.1>, ..., <0.9 - 1.0>, regardless of its real length. For each interval, let us find all words which fall into it, separately for the word frequency f_1, f_2, f_3, f_4, f_5-f_{10}, f_{10}-f_{20} and f_{21} and more. For those sentences which are shorter than 10 words, the following convention is accepted: words are counted in the largest possible interval in which they can fall, e.g., in a sentence which consists of two words only, the first word is counted in the interval <0.4-0.5> and the second word is counted in the interval <0.9-1.0>. We have got Table 2. As can be seen, in the interval <0.0-0.1> there are 14 words with $f = 1$, there are 6 words with $f = 2$ etc.

Table 2
Word frequency in relative position (*Letter 10*)

Word frequency	Number of words with the given frequency in each interval									
	0.0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0
1	14	22	22	23	34	16	28	25	27	35
2	6	9	11	7	7	5	7	5	7	12
3	2	4	3	2	3	3	7	2	3	5
4	3	3	4	3	5	3	2	3	5	6
5-10	7	9	3	5	2	7	5	4	3	6
11-20	2	5	2	6	3	3	3	5	6	3
21-	3	5	10	9	14	6	4	8	6	6

Unfortunately, in the given text only words occurring once are represented with certain reliability, hence we shall use them for the analysis. Figure 1 shows the frequencies of hapax legomena in individual relative positions.

An oscillation can be seen also with other frequencies but since the numbers are small, it is not that conspicuous. It can easily be seen that the numbers of hapax legomena in the individual intervals are not homogeneous: the chi-square test for homogeneity yields X^2 = 30.44 with 9 *DF* and P = 0.0004. Hence there is no random oscillation, there is probably a rhythm. The other rows of Table 1 are homogeneous; the rhythm is not sufficiently expressed because of small numbers.

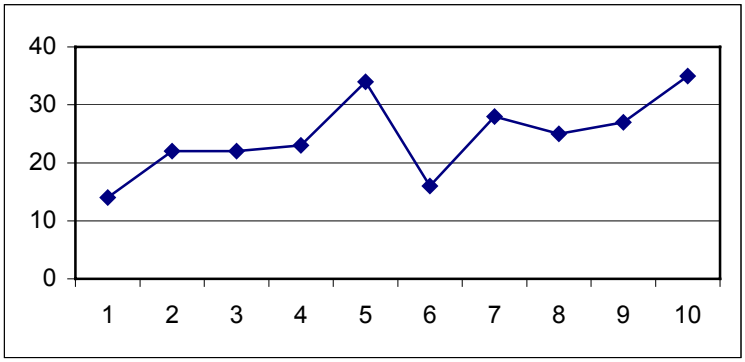


Figure 1. Occurrence of hapax legomena in "relative" positions

Looking at Figure 1 we can see that there are two "waves" dividing the series exactly in the middle. We can approach this situation in different ways. Either we use splines or wavelets or Fourier analysis or time series or a difference equation of higher order, etc. In any case, we may venture to express the following hypothesis: *In sentences, there exist positional intervals in which some frequency rhythm is manifested.*

For the sake of simplicity, we perform a Fourier analysis with our 10 points. Based on the computation of intensity yielding $I(f_1)$ = 29.79, $I(f_2)$ = 106.06, $I(f_3)$ = 1.61, $I(f_4)$ = 277.34, $I(f_5)$ = 0.8 we choose the frequencies 1, 2 and 4 and obtain the formula

$$y = 24.6 + 0.3382\cos(2\pi x/10) - 2.4172\sin(2\pi x/10) + 3.0493\cos(2\pi x\ 2/10) - \\ -3.4516\sin(2\pi x\ 2/10) + 6.8507\cos(2\pi x\ 4/10) - 2.9218\sin(2\pi x\ 4/10)$$

yielding the values in Table 3. The parameters are computed using standard methods.

Table 3
Fourier analysis of position-frequency data

x	y	y_t
1	14	13.85
2	22	22.81
3	22	21.10
4	23	23.31
5	34	34.16
6	16	16.15
7	28	27.19
8	25	25.90
9	27	26.69
10	35	34.83

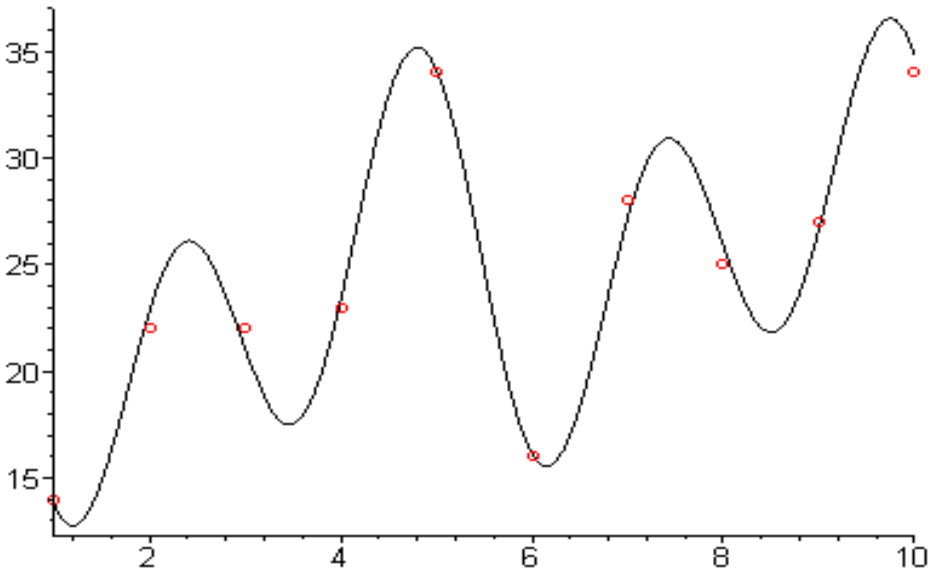


Figure 2. Fourier analysis of frequency rhythm

The hypothesis is reasonable. Autosemantic words with low frequencies are connected by synsemantics with high frequencies. Hence the frequency motion should be more or less regular. We suppose that in long texts more frequency classes will display a rhythm but in our text it is only the class x = 1. (For analogical cases of repetition in waves, namely for the repetition of stressed syllables in verse, see Altmann 1988:197ff, and Wimmer et al., 2003: 249ff.)

As far as the middle interval is concerned, much more data would be necessary so as to reveal at least some factors which may be at play.

As far as the final interval is concerned, its existence is strongly supported by the general principle of the communicative structure of sentence, well-known to linguists under various names, such as the principle of end weight, the principle of the increasing communicative dynamism, the theme-rheme structure or topic-comment structure, given and new information etc. This principle claims that the element(s) with the maximal weight / the highest degree of communicative dynamism /the rheme proper, the bearer of new information etc. display(s) a strong tendency to be placed to the sentence end. It is highly probable that an element which should push the information content of communication forward, will be expressed by a word with $f = 1$ or, less often, with $f > 1$, but still with quite a low frequency.

Let us test the relevance of the final sentence position for the frequency rhythm on more texts. The data are presented in the next paragraph.

3. "Absolute" position. Word frequency in the final position

Now let us take the data from ten Bulgarian letters and let us observe the frequency of the words which occupy the **final** position in **each** sentence in each text. The data are given in Table 4.

Table 4
Frequency of the word in the final position in each sentence (in ten texts)

Letter 1	1	1	1	3	3	2	2	2	2	1	1	2	1	1	1	1	3	5	1	1	2	1
Letter 2	1	1	1	2	1	1	1	2	1	1	1	1	1	3	1	2	1	2	2	1	1	1
Letter 3	1	9	1	1	2	1	1	1	2	2	3	1	1	1	1	1	2	2	1	1	1	2
Letter 4	1	1	1	1	1	1	2	1	1	1	1	13	1	1	1	2	2	1	1	1	1	1
Letter 5	1	1	3	1	1	2	1	1	1	1	2	1	2	2	1	1	7					
Letter 6	1	1	1	3	1	1	1	4	1	1	2	1	2	1	2	1	2	1	1	4	1	1
Letter 7	1	1	2	3	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	3	1
Letter 8	1	1	1	1	1	4	1	1	1	1	1	1	1	1	1	1	1	1				
Letter 9	1	1	1	1	1	1	1	2	2	2	1	2	2	1	1	1	1	1	3	1	1	1
Letter 10	1	1	1	1	2	1	2	2	1	1	2	2	3	1	2	1	1	1	2	1	1	1

rows continued:

1	2	1	1	1	1	4	6	1	1	1	1	1	1	3	1					**1.71**
1	3	1																		**1.36**
9	1	1	1	2	1	1	1	2	1	2	1	1	1	1	1	1	2	1	1	**1.67**
1																				**1.65**
																				1.71
1	4	1	1	2	1	1	1	1	1	2										**1.52**
4	1	1	1	1	1	1	2													**1.33**
																				1.17
1	1	1																		**1.28**
2	3	2	3	1	1	1	1	1	1	2	1	3	3	1	2					**1.55**

The texts have different lengths, and, consequently, the rows (= one row for each text) have different numbers of cells. In each cell the frequency of the word in the final position in each sentence is given (let us remind what was said above: at the beginning of counting the sentences were arranged according to their increasing length). E.g., in Letter 10, the last position in the shortest sentence is taken by a word with $f = 1$ (returning to Table 1 we can observe that this is the sentence consisting of 2 words), the last position in the two longer sentences is taken also by words with $f = 1$ (again: returning to Table 1, we can see that they are sentences consisting of four words), and so on. We can see now what was expected: Most sentences end with a word with $f = 1$; some sentences end with a word with $f > 1$, but only exceptionally with a word with ≥ 5.

In the last column of Table 4, the **average** word frequency in the final position is given for **each** text. The average word frequency in the final position falls within a very narrow frequency interval.

Thus, the importance of the final = informationally heavy position for the frequency rhythm is empirically supported by our data. The existence of a position (either final or – generally – any other one) marked for the frequency rhythm, if proved, could be important for text/speech recognition procedures: for example, in a continuous text/speech in which sentence boundaries are not formally/intonationally indicated, it could be the frequency rhythm that could – possibly – help. And, the word frequency in the final position in sentence may be considered as another case of the so-called "positional" repetition, dealt with by Altmann in detail (Altmann 1988:92ff).

4. "Absolute" position continued. Word frequency in the initial position

Now let us take the same ten texts and let us have a look at word frequencies in initial position. Let us ask whether also the initial position is relevant for the frequency rhythm.

The data, i.e. the frequencies of the words in initial position in each sentence, are given in Table 5, which is arranged along the same lines as Table 4 (see above for the details).

Table 5
Frequency of the word in the initial position in each sentence (in ten texts)

Letter 1	1	2	2	1	8	1	2	1	6	2	1	6	29	17	2	2	3	2	1	1
Letter 2	1	1	1	1	2	2	1	13	1	3	5	1	1	1	4	2	8	13	1	2
Letter 3	1	6	9	13	1	1	6	6	6	2	6	1	1	1	3	2	3	2	6	1
Letter 4	1	1	5	1	1	5	1	1	2	1	1	1	1	1	1	7	1	1	1	7
Letter 5	1	1	19	1	4	1	7	1	2	1	1	7	1	1	2	2	2			
Letter 6	1	1	2	1	2	2	24	1	1	2	4	27	3	1	1	1	12	1	27	1
Letter 7	1	8	1	1	1	1	1	3	16	3	1	1	1	1	2	4	1	1	1	4
Letter 8	1	2	5	1	3	3	1	2	1	1	2	1	1	5	1	1	1	1		
Letter 9	1	3	20	15	1	2	4	6	1	17	1	1	1	1	17	20	1	2	1	1
Letter 10	1	2	2	2	25	9	2	9	1	3	2	9	2	1	1	1	4	2	18	1

rows continued:

3	3	5	1	2	1	5	12	29	1	1	29	5	3	6	29	2	2					**6.03**
4	2	3	3	2																		**3.12**
1	2	4	5	1	1	3	5	1	1	1	14	1	8	1	2	1	2	1	13	3	6	**3.67**
3	1	1																				**2.00**
																						3.18
1	1	2	18	1	1	1	12	3	2	28	1	18										**6.18**
8	4	7	15	1	8	1	4	3	2													**3.53**
																						1.83
2	1	3	1	1																		**4.96**
9	1	2	3	23	1	5	9	3	1	1	6	1	2	2	1	1	4					**4.53**

This result is negative. The average frequencies (last column) are significantly higher than those in the final position. Without going into details, we suppose that this fact can be explained by the interplay of various textual factors, the typical cohesive function of words standing at the sentence beginning being among them. Nevertheless, there may be differences between languages, genres etc.

5. Average word frequency per position. "Absolute" position reconsidered

Let us return to the data from Letter 10 presented in Table 1. What can we say about the frequency rhythm on the basis of the **average** word frequency in each word order position? The average frequencies per position are given in the last row of Table 1. The lowest average frequency is $f = 1.0$ (for the first position), the highest amounts to $f = 10.5$ (for the thirty fourth position). Even if we disregarded these two extreme values, it would be evident from the data that there is a considerable oscillation above and below the mean value (mean =

5.92).Still, an empirical tendency shows: If the sentence length increases, the average word frequency per position slightly increases, too. See Fig. 3. This tendency may be modeled by the straight line equation $y = 0{,}605x + 4{,}728$. The coefficients are found in a least-squares sense.

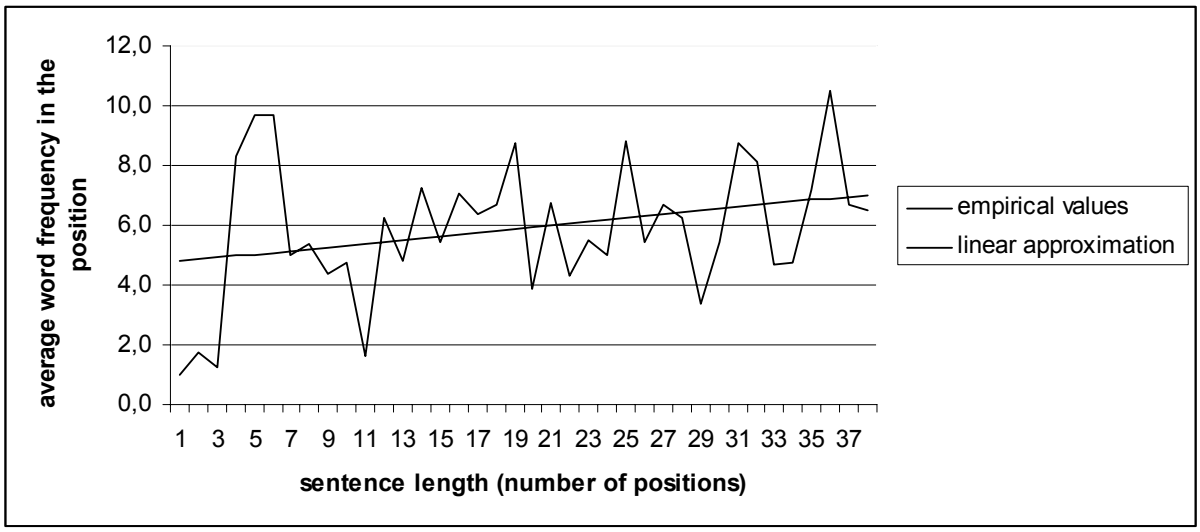


Figure 3. Average word frequency per position in *Letter 10*

Let us ask now whether this tendency is specific only for the given text, or whether it may be attested in the other texts, as well.

Let us take ten Bulgarian letters again. Let us calculate the **average** word frequency for each position, for each text. Let us create a table with ten rows (one row for each text), showing the average frequencies in the successive left-to-right positions in sentences. This data is given in Table 6. For example, the average word frequency in the first position in Letter 1 is $f = 6.7$ (first row, first column), the average word frequency in the first position in Letter 2 is $f = 4.5$ (second row, first column), etc. A fair number of points in Figure 4 below indicate, just as could be seen in Figure 3 above, that the data is considerably dispersed. However, obviously, if the sentence length increases, the average frequency per position increases from left to right. Thus, the same tendency which was found for Letter 10, is empirically observed in all ten texts.

Table 6
Average frequency per position in ten texts

Letter 1	6.7	1.3	4.8	2.8	5.2	7.4	5.6	6.8	5.2	5.3	6.4	5.5	6.8	5.3	4.0	5.6	5.2	3.8	6.2	5.9
Letter 2	4.5	2.8	4.0	2.8	3.8	3.4	7.1	3.8	3.4	4.8	4.0	4.1	4.7	2.6	5.0	5.2	4.2	4.4	3.8	3.8
Letter 3	6.8	3.0	6.0	1.0	3.4	3.8	4.6	3.1	4.8	5.3	2.9	4.0	1.4	5.0	2.3	4.4	5.8	3.4	3.8	3.8
Letter 4	7.9	3.9	3.6	3.2	3.8	5.3	5.6	4.0	4.3	4.5	4.8	4.7	4.0	4.0	4.4	3.1	5.2	4.3	5.1	5.1
Letter 5	4.4	5.8	4.2	5.8	3.1	4.5	5.4	4.5	4.6	5.2	4.3	3.9	4.1	5.2	4.6	4.5				
Letter 6	1.0	3.8	2.5	5.2	3.7	7.2	6.3	9.3	7.7	3.6	6.2	5.2	6.0	5.1	6.3	4.8	7.9	7.9	6.3	5.0
Letter 7	2.8	4.8	1.8	4.1	3.4	4.9	5.9	5.1	4.9	2.3	3.3	4.5	4.4	4.8	3.4	4.2	5.0	4.3	3.7	4.2
Letter 8	2.4	2.7	2.9	2.7	4.4	2.5	3.6	2.8	3.4	2.3	2.9	2.0	2.8	3.0	3.1	2.2	3.1			
Letter 9	1.7	5.4	4.0	5.3	3.7	5.2	5.6	6.2	4.9	4.3	4.6	4.7	5.3	4.5	5.6	5.0	2.9	5.1	5.7	6.7
Letter 10	1.8	1.3	8.3	9.7	9.7	5.0	5.4	4.4	4.8	1.6	6.3	4.8	7.3	5.5	7.1	6.4	6.7	8.7	3.9	6.8
Average	**4.0**	**4.0**	**4.0**	**4.3**	**4.4**	**4.9**	**5.5**	**5.0**	**4.8**	**3.9**	**4.6**	**4.3**	**4.7**	**4.5**	**4.6**	**4.5**	**5.1**	**5.3**	**4.8**	**5.2**

rows continued:

6.3	7.0	6.3	4.0	4.6	9.1	7.1	7.9	9.7	7.9	9.6	9.1	5.7	10.0	9.0	5.2	9.0				
4.7	4.1	4.0	4.1																	
4.7	4.7	4.5	4.6	5.9	5.2	4.0	2.7	6.3	5.1	4.3	6.1	4.5	6.1	4.0	5.0	4.0	3.7	4.3	3.8	4.9
3.4	5.0																			
6.1	7.2	5.8	5.9	6.9	7.6	6.9	8.6	6.5	7.8	7.5	6.9									
4.6	5.0	4.5	4.1	3.3	4.3	4.1	5.4	4.4												
5.6	4.5	5.5	5.8																	
4.3	5.5	5.0	8.8	5.4	6.7	6.2	3.4	5.4	8.8	8.1	4.7	4.7	7.2	10.5	6.7	6.5				
5.0	**5.4**	**5.1**	**5.3**	**5.2**	**6.6**	**5.7**	**5.6**	**6.5.**	**7.4**	**7.4**	**6.7**	**5.0**	**7.7**	**7.8**	**5.7**	**6.5**	**3.7**	**4.3**	**3.8**	**7.9**

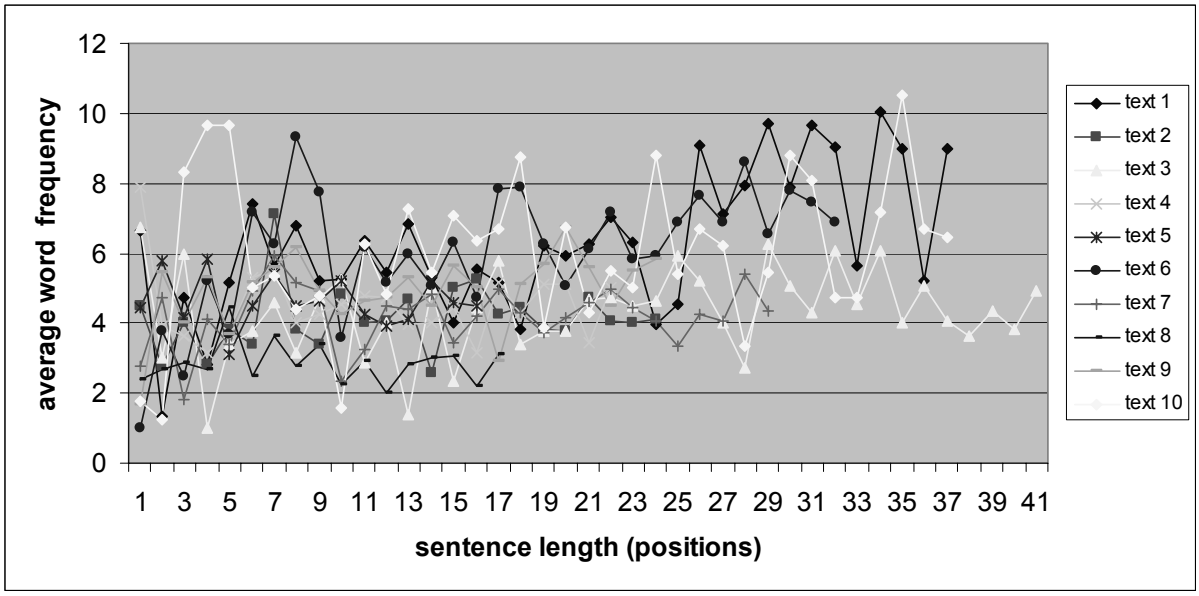


Figure 4. Average frequency per position in ten letters

Now, if we count the **average** value for each position **across ten average values**, as it is done in the last row of Table 6, we may use the same method of estimation as above. Now get the equation $y = 0.0439x + 4$. The least-square test demonstrates the same tendency as above: If the sentence length increases, the average word frequency also increases. For better illustration, see the graphic presentation (Fig. 5).

Seemingly, there is one exception, text 3, which is the longest one. A question arises: Is there a length limit, behind which the average frequency per position does not increase any more? This question remains open, for the time being.

This apart, it is reasonable to believe that if the average frequency per position depends on sentence length, then also the **sentence** length is very probably relevant as far as the frequency rhythm is concerned.

In addition, another question arises: Can we – on the basis of what we already know about the average frequency per position – hypothesize anything about the relationship between the **text** length and its frequency rhythm? We will pursue this line of investigation in the following section.

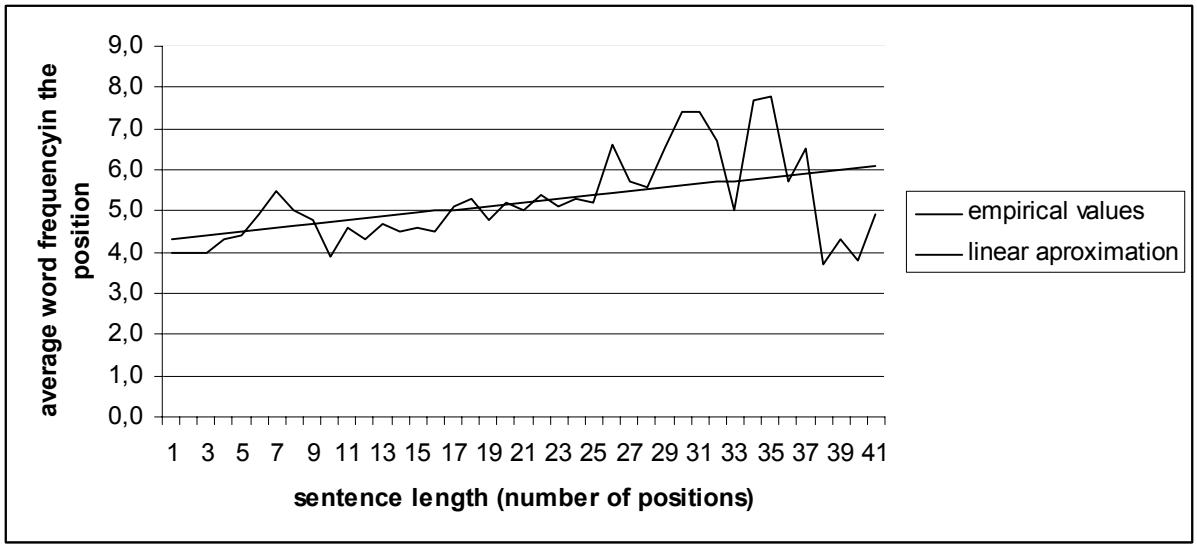


Figure 5. Average values calculated from the average frequencies per position in ten letters

6. Average frequency continued: Minimal–to–maximal range.

Let us pay attention to the range between the minimal and maximal average frequency in particular positions in each text. The <min–max> range may be easily read from the previous Table 6: for instance, for text 10, the minimal average $f = 1.3$, the maximal average $f = 10.5$, therefore the <min–max> range for this text is <1.3-10.5>. The <min–max> range for the ten Letters are presented in Table 7.

Table 7
Minimal–to–maximal range of the average frequency per position in ten Letters

	Min–max range of the average word frequency per position	Texts ordered by increasing min–max range	Texts ordered by the increasing text length
Letter 1	2.8-10.0	9.	8.-9.
Letter 2	2.8-7.1	8.	4.-5.
Letter 3	1.0-6.3	6.	10.
Letter 4	3.1-5.6	2.	3.
Letter 5	3.1-5.8	3.	1.
Letter 6	3.1-9.3	7.	7.
Letter 7	2.3-5.9	4.	6.
Letter 8	2.0-4.4	1.	2.
Letter 9	2.9-6.7	5.	4.-5.
Letter 10	1.3-10.5	10.	8.-9.

In Table 7 the <min–max> range is given for each text in column two; then the texts are ranked according to increasing <min–max> range (third column), so that the text with the smallest <min–max> range is assigned the ordinal number 1 (it is Letter 8), and so on. Comparing this ordering with the ordering of the texts according to increasing length, measured in number of sentences (last column: Letter 3 is longest, as we already know), we can see that there is a straightforward correlation between both orderings. Applying Spearman's rank correlation

coefficient we get $R = 0.7515$ with $DF = 8$, on the 0.05 significance level. We may say that the longer the text, the greater is the <min–max> range of the average frequency per position. If we interpret the <min–max> range as an indicator of the syntactic homogeneity of text – which is a natural and necessary property of any coherent text – then, obviously, the longer the text, the less homogeneous it is, at least in this particular respect. Furthermore, it is not unreasonable to suppose that the degree of syntactic homogeneity influences the frequency rhythm of the particular text.

7. Time series

Not only words, but also word sequences necessarily repeat in any coherent and reasonably long text. As was already mentioned above, the repetition of word sequences follows from the basic syntagmatic property of sentence as a text unit: autosemantic words are bound together, or, as Altmann puts it, they "associate with" various synsemantics, and the latter help to determine the syntactic functions of the former. If word sequences repeat, or if there is an "associative" repetition (Altmann 1988:115ff.) on the syntagmatic level, the following question is pertinent: Do also **frequency sequences** repeat? To answer this question, let us treat the frequency rhythm as a syntagmatic phenomenon of its own, as it is done below in this paragraph and, in more detail, in par. 8 and 10 below.

To begin with, let us take Letter 10 once again. Let us re-write it as a linear sequence of word frequencies, preserving the word order of the original text. We get the following sequence:

1,1,1,4,18,1,4,5,1,1,1,1,2,4,2,23,1,1, 3,2,25,4,2,1,3,5,6, ..., 1,1,26,2,23,1,4,3,1,1,23,1.

The text begins with a word which has $f = 1$, then two words with $f = 1$ follow, the fourth word has $f = 4$, the fifth word has $f = 28$, etc., and the text ends with a word with $f = 1$. The full sequence consists of $N = 555$ = total number of words in the text. Let us ask whether there is any regularity in the repetition of frequencies, and more concretely whether any frequency rhythm is apparent. Let us use a method of time series. As we already know from 2 and 3 above, the frequency class of $f = 1$ is prominent, hence it is reasonable to take the reversed frequency values for calculation.

Let $x(n)$ be a sequence of word frequencies f in the text, where n goes from 1 to N where $N = 555$ is the text length. Then $x'(n) = 1/x(n)$ is a sequence of the reversed values of f. Applying the autocorrelation function

$$R(k) = \frac{1}{N-k}\sum_{n=1}^{N-k} x'(n)x'(n-k)\text{, where k = 0,..., } k_{max},$$

we derive Table 8 and Fig. 6.

The frequency rhythm is quite clear; the waves repeat. Two rhythmic phenomena should be highlighted:

Firstly, hapax legomena, represented by the peaks, repeat rhythmically throughout the text, in regular waves. The first characteristic distance (= the first peak in Fig. 6) between two words with $f = 1$ is two words (two hapax legomena in the sequence are separated by one word with $f > 1$), another characteristic distance (another peak in Fig. 6) is four words (two hapax legomena in the sequence are separated by three words with $f > 1$). Other characteristic distances are eight words and eleven words. Greater distances are much less obvious. From a structural perspective – considering only Bulgarian: for example, we might note that

Table 8
Frequency rhythm in *Letter 10* – time series

k	*R(k)*
1	0.294
2	0.325
3	0.320
4	0.329
5	0.324
6	0.314
7	0.316
8	0.329
9	0.320
10	0.318
11	0.333
12	0.308
13	0.319
14	0.307

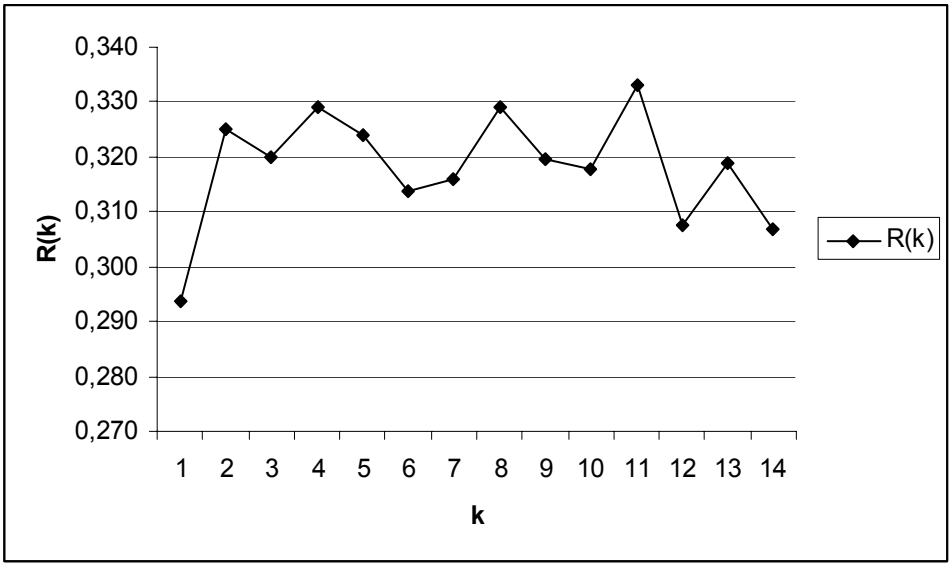


Fig. 6. Frequency rhythm in *Letter 10* – time series

Bulgarian is an analytical language with a practically empty inventory of case endings. Consequently, it is common in Bulgarian that prepositions serve as the main means of expressing syntagmatic relations; as such, they stand in the leftmost position of bare or compound nominal phrases.

Secondly, quite complementary to what was just said, let us note the first deep trough in Fig. 6. It demonstrates that it is very unusual for two words with very high frequencies to follow immediately one after another. From a structural perspective: sequences such as two prepositions or two conjunctions standing in contact positions are rare in Bulgarian.

Let us make a step further and generalize our result, still considering one language. Let us take the ten Bulgarian texts and repeat the time series procedure. Now the variable n goes from 1 to 5125, $N = 5125$. The results are offered in Table 9 and Fig. 7 below.

Table 9
Frequency rhythm in ten Bulgarian texts – time series

k	*R(k)*
1	0.303
2	0.323
3	0.330
4	0.326
5	0.325
6	0.328
7	0.324
8	0.325
9	0.328
10	0.323
11	0.326
12	0.325
13	0.327
14	0.322

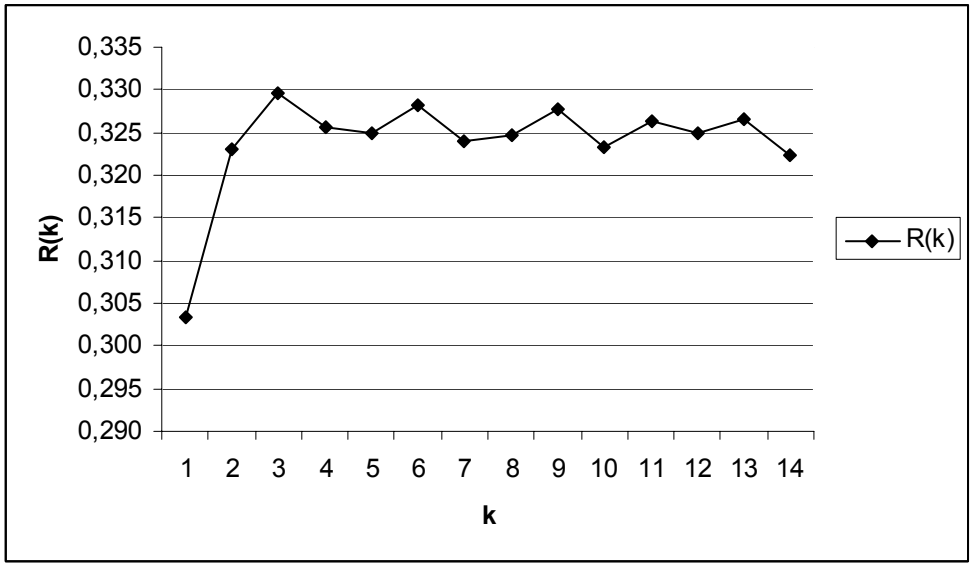


Figure 7. Frequency rhythm in ten Bulgarian texts – time series

Now the peaks are not the same as they appeared in the single short text Letter 10. But, it is notable that although we took ten **different** texts (albeit of the same genre) and although we did not respect the well-known methodological requirement for text homogeneity, the frequency rhythm is evident. The peaks repeat regularly, now with an interval of three. In our opinion, this observation supports the idea that the frequency rhythm is very probably a quite general quality of communication, a quality superimposed to "breaks" in text homogeneity.

8. An application of Köhler's syntagmatic approach

An approach to the syntagmatic dimension of text was undertaken by Köhler recently (Köhler, 2006, 2007), who studied word length sequences independently of all linguistic factors

proper (such as language, genre or text specificity, sentence boundary etc.). Köhler hypothesized that word length sequences in text are organized in "lawful patterns". He verified his hypothesis on extensive data so that he could answer it in positive. Below, Köhler's approach is applied, but instead of word length rhythm, word frequency rhythm is addressed.

Köhler departs from the notion of L-segment: It is "the text segment which, beginning with the first word in the given text, consists of word lengths which are greater or equal to the left neighbor". Applying Köhler's language independent criterion, we define **F-segment** as the text segment, which, beginning with the first word in the given text, consists of word **frequencies** which are greater or equal to the left neighbor. The F-segments are searched for across sentence boundaries, in the text as a whole.

Let us take the text Letter 10 again and let us segment it into F-sequences. The text is considerably shorter than Köhler's, but still sufficiently long to evidence different F-segments with different number of occurrences in it. The data are given in Table 10 below.

Table 10
Distribution of F-segments in *Letter 10*

F-segment	Number of occurrences	Length	F-segment	Number of occurrences	Length
1	1	1	1,5	2	2
1,1,1,1,2	1	5	1,6	2	2
1,1,1,1,1,8,25	1	7	1,6,8	1	3
1,1,1,2	1	4	1,7	3	2
1,1,1,2,4	1	5	1,9	1	2
1,1,1,3	2	4	1,16	1	2
1,1,1,18	1	4	1,18	4	2
1,1,1,23,26	1	5	1,23	3	2
1,1,1,1,26	1	5	1,23,25	1	3
1,1,1,2,25	1	5	1,23,26	1	3
1,1,2	1	3	1,25	5	2
1,1,2,2,2	1	5	1,26	5	2
1,1,4	1	3	2	9	1
1,1,4,25	1	4	2,2	1	2
1,1,8	1	3	2,2,3	1	3
1,1,18	3	3	2,3	1	2
1,1,23	3	3	2,3,19	1	3
1,1,25	2	3	2,7	1	2
1,1,26	3	3	2,23	1	2
1,2	3	2	3	5	1
1,2,7	1	3	3,23	1	2
1,2,8,18	1	4	3,5	1	2
1,2,26	1	3	4	7	1
1,3	3	2	4,25	1	2
1,3,9	1	3	18	6	1
1,4	4	2	**Total**	**106**	

What can we see? From the total number of F-segments in the text, which amounts to 106, most segments occur just once; they are 32 in number; four different F-segments occur twice, so that their frequency in the text makes $f = 8$ together, etc. On the other hand, only one segment repeats nine times in the text, i.e. its text frequency $f = 9$. If we order the F-segment frequencies by decreasing order (= by increasing rank) in the text, as shown in Table 11 below, we get a decreasing frequency spectrum. With the help of Altmann-Fitter, this distribution can be modelled by the hyperpascal probability model (Table 11, column three) with a good result. Note this is the same model as was applied by Köhler for L-segments. See also the illustration in Figure 8 below.

Table 11
Frequency spectrum of F-segments in *Letter 10*

Rank	Frequency F(i)	NP(i)
1	32	28.92
2	21	22.20
3	15	16.39
4	9	11.80
5	8	8.34
6	8	5.82
7	7	4.02
8	6	8.50
$k = 4.0537$ $m = 3.3770$ $q = 0.6394$ $X^2 = 4.93$ $P = 0.29$ $DF = 4$		

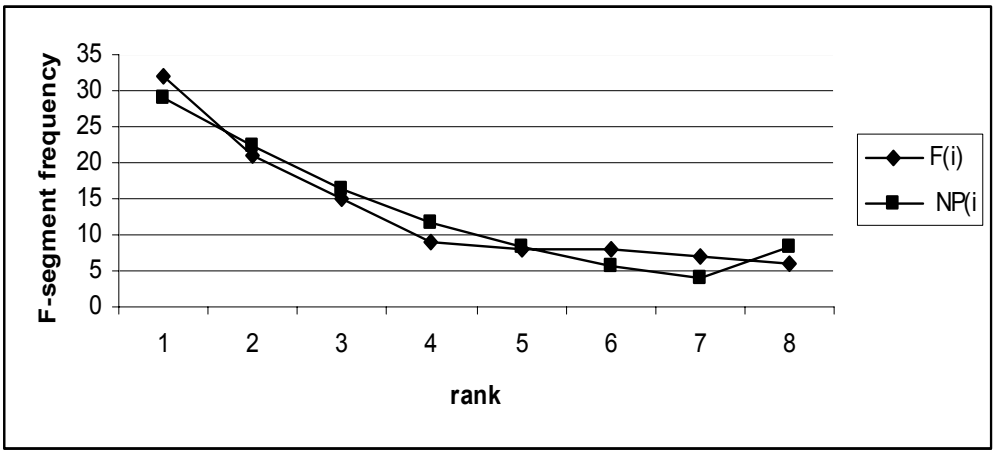


Figure 8. Frequency distribution of F-segments in *Letter 10*

This result is purely empirical, demonstrated on the data from one text, but it corroborates the hypothesis that the distribution of word frequency segments –as with Köhler's word lengths segments – abides by a law; as such, it may be taken as another argument in favour of the idea that the distribution of word frequency segments in texts is of a rhythmical nature.

Now let us arrange the F-segments according to their increasing **size,** i.e. according to increasing number of constituents (= according to increasing length), as shown in Table 10, last column. How often do the F-segments of different size occur in the text? After counting, we can see (Table 12) that there are 5 **different** F-segments which consist of a single

constituent each (actually, we already know this from the data in Table 10 above). Furthermore: there are 19 **different** F-segments which consist of two constituents; and so on. It is evident that the distribution of **different** F-segments of **different** size now differs substantially from what was found and tested in the previous paragraph. This time a different regularity is seen, and a different probability model fits: using Altmann-Fitter again, we find (Table 12, last column; see also Figure 9) that the size-frequency distribution of F-segments can be satisfactorily modeled by Poisson distribution, which is one of the most broadly applied and best-fitting probability models for modeling length distributions (Best, 1997; 2001).

Table 12
Size-frequency distribution of F-segments

Size of F-segment X(i)	Number of different F-segments F(i)	NP(i)
1	5	7.95
2	19	14.78
3	15	13.73
4	5	8.51
5	6	3.95
6	0	1.47
7	1	0.61
a = 1.8585, X^2 = 5.49, DF = 4 , P = 0.24		

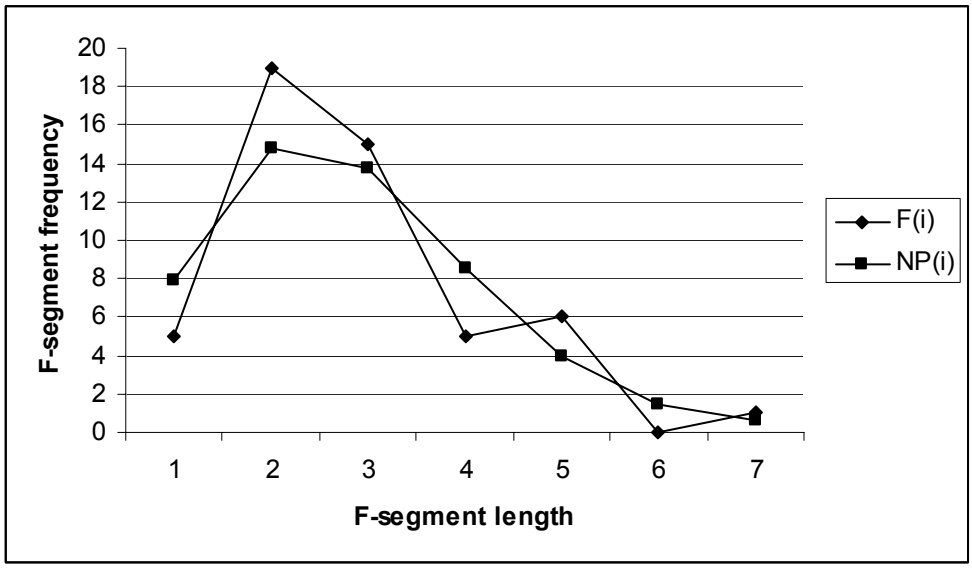


Figure 9. Size-frequency distribution of F-segments in *Letter 10*

This objective, "lawful" behaviour of F-segments, both of their rank–frequency distribution and of their size-frequency distribution, probides another argument in favour of considering frequency rhythm as one of the general organizing principles of the syntagmatic text dimension.

9. F-segments "from the inside"

Let us briefly remember that the inventory of F-segments in Letter 10 is represented by 51 different F-segments, and the total number of F-segments is 106; some F-segments repeat, the repetition index = 2.08.

Let us have a look into what we shall call the *inner structure* of F-segment. Let us characterize each F-segment by a single value – the **average frequency** *A(f)* counted from the **sum of the frequencies of its constituents**. If, for instance, there is an F-segment 1,1,1,2,25, then its *A(f)* = (1 + 1 + 1 +2 + 25)/5 = 6. Some F-segments with different numbers and different frequencies of constituents may happen to have an equal *A(f)*, e. g. the *A(f)* of an F-segment 1,1,1,3 equals (1 + 1+ 1+3)/4 = 1.5, and also the *A(f)* of an F-segment 1,2, equals (1,2)/2 = 1.5. Let us calculate *A(f)* for all 106 F-segments in the text. This is easily done from Table 10, so it is not necessary to give here the *A(f)* values in full. Now let us order *A(f)* according to their increasing values, beginning with the lowest value, which is, in our case, *A(f)* = 1.0, and ending with the highest value, which is, in our case, *A(f)* = 18, and let us ask whether there exists a probability model for our distribution. The result of fitting (using Altmann-Fitter again) shows a perfect fit (with $P \approx 1.0000$) of more than one probability model. As an illustration, let us choose one of them, the negative hypergeometric model, with the parameters $K = 2.5692$, $M = 1.8159$ and $n = 105$, with $X^2 = 29.8846$, $P \approx 1.000$, and $DF = 97$; see Fig.10.

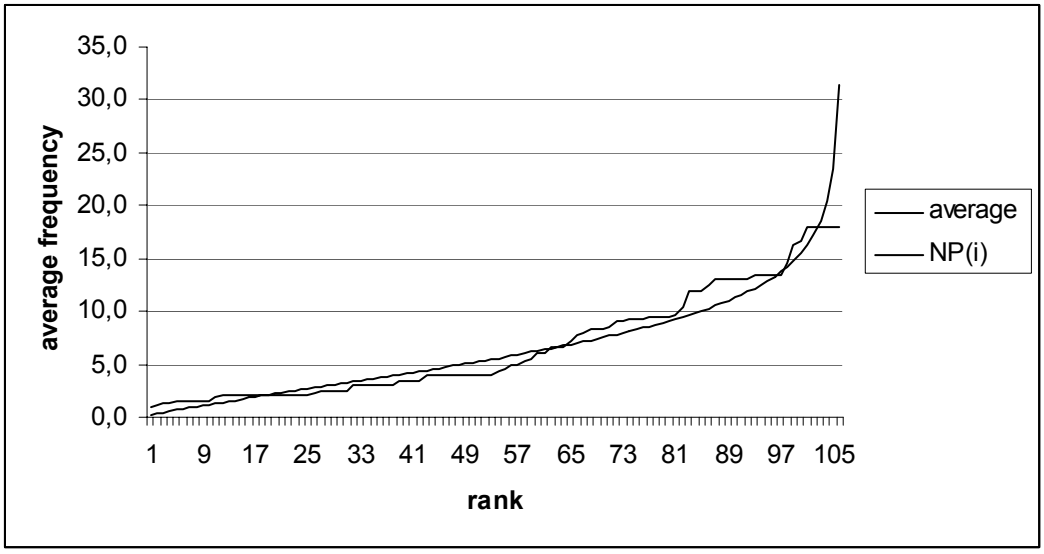


Fig. 10. Frequency distribution of F-segments in *Letter 10* in terms of the average frequency of their constituents *A(f)*

In this way, the "lawful" behaviour of F-segments (as pointed out by Köhler) is confirmed once more. Similar results were obtained when the *A(f)* values were ordered decreasingly, but we shall not go into full details here.

These results are in full accordance with what was expected intuitively.

Let us go a step further and consider the relationship between *A(f)* and the number of F-segments with a given *A(f)*. Is there any regularity, or perhaps, any rhythm? Since our text is relatively short, let us cumulate *A(f)* values into fifteen intervals according to the increasing value of *A(f)*, beginning with the interval *A(f)* = <1.0 – 2.0>and ending with the interval *A(f)* = <17.0 – 18.0>. How many F-segments fall into each *A(f)* interval? The following picture emerges (Table 13, Fig. 11):

Table 13
Number of segments in *A(f)* intervals in *Letter 10*

A(f) interval	Number of segments	NP(i)
1.0-2.0	25	22.48
2.0-3.0	13	11.70
3.0-4.0	13	8.81
4.0-5.0	4	7.31
5.0-6.0	4	6.34
6.0-7.0	3	5.64
7.0-8.0	3	5.09
8.0-9.0	6	4.65
9.0-10.0	8	4.28
10.0-11.0	0	3.95
11.0-12.0	3	3.66
12.0-13.0	7	3.40
13.0-14.0	5	3.15
14.0-15.0	1	2.91
15.0-16.0	0	2.68
16.0-17.0	2	2.44
17.0-18.0	6	5.52

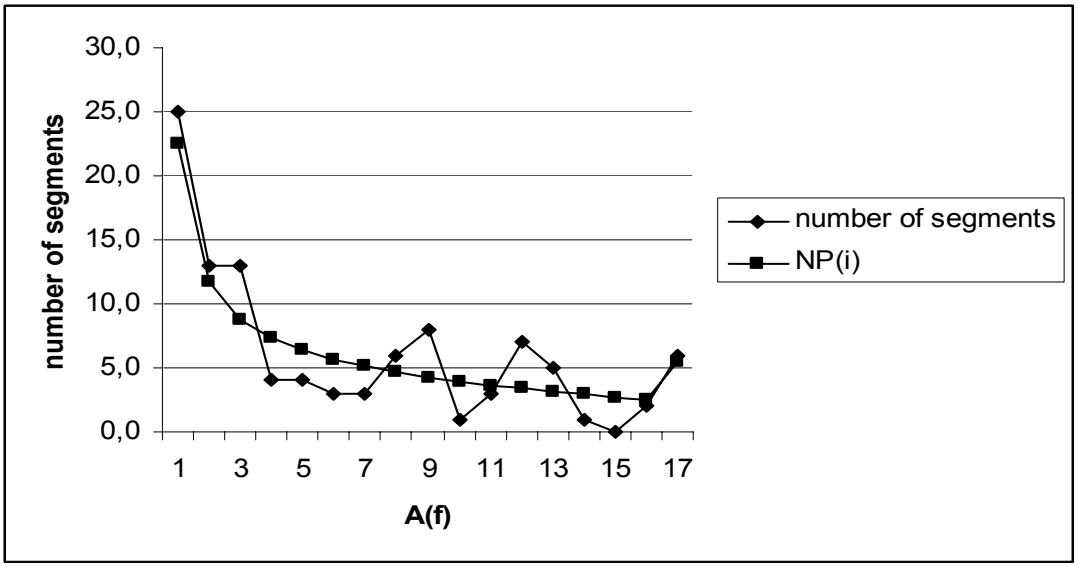


Figure 11. Number of segments with a given *A(f)* in *Letter 10*

Generally, if the average frequency *A(f)* of the F-segments increases, the number of occurrences of F-segments with the respective *A(f)* decreases. This tendency can be modeled by the Pólya distribution (third column of Table 13) with the parameters $s = 0.5608$, $p = 0.2961$, $n = 18$ and with $X^2 = 21.7884$, $P = 0.0587$, and with $DF = 13$. What does this result mean? If *A(f)* is an empirical variable which helps us to look into the inner structure of the F-segment, and if we interpret the *A(f)* distribution as an indicator of the inner complexity of F-segment, then we can say that the greater *A(f)* (or: the more complex a F-segment structure), the less often the F-segment occurs in the text, and, consequently, the less relevant it is for the

frequency rhythm. And vice versa: The lower *A(f)* (or: the less complex a F-segment structure), the more often the F-segment occurs in the text, and thus the more important it is for the frequency rhythm. This – again – supports existence of the frequency rhythm phenomenon that has been proposed and argued for throughout this article.

10. Distances between *f*1

So far we have asked about regularities in the distribution of word sequences and frequency sequences in sentences and texts. Is there any regularity in the distribution of **distances** between word frequencies? Let us notice that some words with *f*1 follow **immediately** one after another in the text, without any word with $f = \geq.1$ standing in-between. Quite often, not only two words with *f*1 follow immediately one after another, but also three, four, and (exceptionally) five words with *f*1 make clusters (for Letter 10 see Table 10 above).

Let us say that if two words with *f*1 follow immediately, then their distance is $d = 0$; if one word with $f \geq 1$ stands between them, their distance is $d = 1$, etc. Let us count all distances in Letter 10, following Altmann's procedure (Altmann 1988:146). Let us order the distances *d* increasingly, as shown in Table 14, column one, and give the frequency of their occurrence in our text in column two:

Table 14
Distances between *f*(1) in *Letter 10*

d	Number of occurrences	NP(x)
0	87	91.99
1	78	69.08
2	38	40.33
3	21	21.29
4	11	10.65
5	4	5.15
6	1	2.43
7	2	1.13
8	0	0.52
9	0	0.24
10	1	0.19
k = 1.8020 p = 0.5833 $X^2 = 3.09$ P = 0.69, DF = 5		

The most frequent distance between *f*1 words is $d = 0$. The numbers of occurrence of the increasing distances (second column) make a monotonously decreasing scale and the fitting of the negative binominal curve gives a good result, as is shown in Table 14, third column, and in Fig. 12. It is also worth noting that Altmann (1988:154) offers the negative binominal distribution as one of the suitable models for modeling distances.

The high frequency of $d = 0$ shows that making clusters is quite characteristic for the frequency class of *f*1. It may be considered another argument to support the hypothesis that the distribution of frequencies according to positions is not just accidental, but that there is something like a clustering regularity, a regularity which consists in a regular turn taking of the *f*1 clusters with sequences of $f \geq 1$ in the linear arrangement of sentence and text.

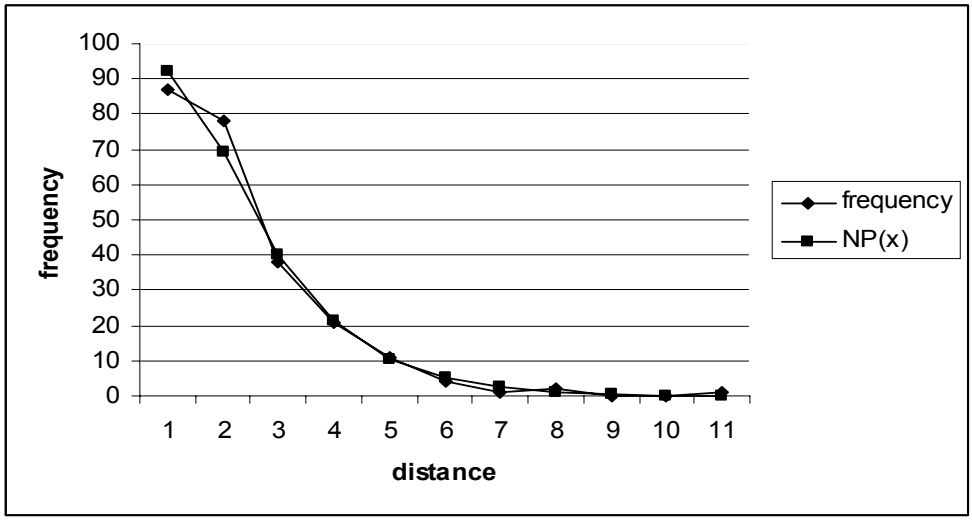


Figure 12. Distribution of distances between of *f*(1) in *Letter 10*

However, having tried to demonstrate a similar tendency for $f \geq 1$, we failed. The reason may be either that the text is too short, or that the pressure of syntax on the distribution of $f \geq 1$ word classes is too strong to allow any significant clustering; alternatively it may simply confirm a special role of hapax legomena in text structure.

In conclusion: If the final purpose of any quantitative analysis is the "Erforschung von Gesetzen, die die Konstruktion von Texten steuern und darauffolgender Aufbau einer Theorie der Texte" (Altmann 1988:3), then the study of frequency rhythm is also relevant here; and the existence of "lawful" behaviour of this phenomenon may be demonstrated.

Note
Supported by Project 1 ET 1011 20413 (Academy of Sciences of the Czech Republic) „Data a nástroje pro informační systémy".

Acknowledgment
This article was inspired by Prof. G. Altmann. I would like to thank him for a thousand and one points of advice, support and help (not only as far as a Fourier analysis is concerned...).

References

Altmann, G. (1988). *Wiederholungen in Texten*. Bochum: Brockmeyer.
Best, K.-H. (ed.) (2001). *Häufigkeitsverteilungen in Texten*. Göttingen: Peust & Gutschmidt.
Köhler, R. (2006). Word length in text. A study in the syntagmatic dimension. In: J. Genzor, M. Bucková (eds.), *Favete linguis. Studies in honour of Victor Krupa*, Bratislava: Slovak Academic Press, 145-152.
Köhler, R. (2007). The frequency distribution of the lengths of length sequences. In press.
Wimmer, G., Altmann, G., Hřebíček, L., Ondrejovič, S., Wimmerová, S. (2003). *Úvod do analýzy textov.* Bratislava: Veda.

Glottometrics 14, 2007, 21-31

Some properties of a metalinguistic verbal system

(in the metalanguage of the Macmillan English Dictionary's defining vocabulary)

Sergey Andreev, Smolensk[1]

Abstract. The study of metalanguage in linguistics has shifted from largely theoretical to more practical questions. In trying to achieve a more precise, universal and objective way of defining the meanings of lexical units, compilers of many dictionaries have started to work out 'defining vocabularies'. These are characterized by such features as a relatively short and closed list of elements, and the capability of expressing the various shades of meanings of all the lexical units in a language. Systems of 'defining vocabulary' actually represent empirically devised metalanguages which possess certain features of the natural languages from which they have been created, but at the same time differ from those natural languages, because they describe an already existing linguistic system.

The paper deals with (1) the predictive relevance of formal characteristics, singled out at different linguistic levels, for choosing words for the metalanguage and (2) a comparison of the relationship between the charateristics in a metalanguage and in the natural language from which that metalanguage was derived. The study is based on the verbal systems of the English language and the defining vocabulary in the Macmillan English Dictionary (2002). Correlation analysis (the coefficients of Cole for alternative characteristics and Jaccard) is used in the study.

Keywords: metalanguage, verbs, English

1. Introduction

One of the most important functions of a language – its meta-function, exemplified *inter alia* by the semantic description of linguistic units – has recently become the object of growing interest. In the course of trying to achieve more exact and accurate results in defining words, compilers of dictionaries have started to create what they often call as a 'defining vocabulary'. This is usually a limited vocabulary of the language, selected in such a way as to enable the compilers to define all the headwords of their dictionary without using any lexical units from outside that limited vocabulary.

Such defining vocabularies have been created and used in English language dictionaries including the Macmillan English Dictionary (2002, p. 1678-1689), the Cambridge International Dictionary of English (1995, p.1702-1707), the Longman Dictionary of Contemporary English (2005, p.1944-1949), and some others.

Defining vocabularies are devised by means of extensive research into empirical data, including such considerations as word frequency, and how these words can satisfy the demands of expressing the various shades of meaning of the lexical units of the language.

The criteria for the selection of words in such vocabularies may be summarized as follows:

– the words must be common and clear to the readers;

– the list of words must be relatively short and is not to be expanded;

[1] Address correspondence to: smolan@sci.smolensk.ru

– the words of this vocabulary should be such as to express any meaning of the headwords in the dictionary.

Based on this, it is clear that defining vocabularies represent empirically created metalanguages. A metalanguage is a specific integrated system. On the one hand, it must be able to act as a language and so to fulfil the conditions of sufficiency, adequacy, flexibility and diversity of linguistic means. On the other hand, the number of words in its vocabulary must be limited as far as possible, and the words must form a closed system.

Although the compilers of metalanguages are guided in their selection of words primarily by semantic factors, there are some formal features which possess predictive power for the selection of lexical units. These predictive tendencies originate independently of the semantic approach of the compilers and demonstrate the autonomous status of the formal structure of a language.

A natural language and its metalanguage exist on two different levels, since the former refers to the surrounding world and the latter refers to an existing language system. However, though it has been created on the basis of a very small number of the elements of a natural language, a metalanguage retains a considerable part of the system of relations between characteristics that is typical of the natural language.

2. Data and properties

This study is based on the verbal system in the defining vocabulary of the Macmillan dictionary. This vocabulary realizes fully the principles of a metalanguage as described above. Verbs were chosen for this analysis due to their role as the structural and semantic centre of the sentence in the predicative function.

The Macmillan defining vocabulary will henceforth in this paper be called metalanguage (Meta-L) and the verbal system of the English language will be referred to as the 'general verbal system' (GV-system).

Unfortunately, the list of words in Meta-L does not discriminate between verbs and nouns. This presents certain difficulties when dealing with conversion pairs, where a single form may be both noun and verb, because in such cases it is not clear whether Meta-L contains a verb or a noun.

To solve this problem, we checked how such words are used in dictionary definitions. It was found that in some conversion pairs, only a verb was included in Meta-L (*chase, catch, move, stay*, etc.); in other cases, only a noun (*frame, place, pump, reason, ship, view,* etc.). In most cases, however, the ambiguous word forms represent both members of the conversion pair at the same time (*act, change, fight, limit, pause, roll*, etc.).

According to our analysis the total number of verbs in Meta-L is 524. This list does not include several verbs which are found in definitions, but do not perform a defining function. Thus, for instance, *bag* is used as a verb in the following definition: "*if clothes bag, they become stretched and look wide*"; but this is not a defining unit, rather it simply repeats the headword of the entry. The same phenomenon is observed in the definition of the verb *taxi*: "*if a plane taxies, it moves on its wheels on the ground*".

Since the words of Meta-L are supposed to be used in definitions to describe the meanings of words of a natural language, in cases where their usage is different (such as repetition of the headword or illustration of the usage of the headword) we do not consider them as metalinguistic elements.

The other list of verbs, which represents the natural language (GV-system), was taken from The Concise Oxford Dictionary of Current English, 7 ed. 1987 (COD) and includes 8375 verbs.

The list of characteristics, used in this paper, corresponds in general to the list presented in some previous studies (Andreev, 1995; Sil'nitskij, Andreev, Kuzmin, Kuskov, 1990), and consists of the following parameters of different linguistic levels and aspects.

Phonetic characteristics

Length:
The verb has one syllable (SYL1): *ask, need, ride*;
the verb has two syllables (SYL2): *advise, copy, govern*;
the verb has three syllables (SYL3): *educate, remember*;
the verb has four syllables (SYL4): *communicate*.

Stress position:
The stress falls on the first syllable in a polysyllabic verb (STR1): *borrow, calculate, order*;
the stress falls on the second syllable (STR2): *commit, defend*;
the stress falls on the third syllable (STR3): *disagree, represent*.

Phoneme:
The first phoneme is a vowel ("VL): *invite, offer*,
the first phoneme is a sonorant ("SNR): *mean, refuse*;
the first phoneme is a obstruent consonant ("CON): *hire, pause*;
the last phoneme is a vowel (VL"): *care, hear*;
the last phoneme is a sonorant (SNR"): *happen, perform*;
the last phoneme is an obstruent consonant (CON"): *invent, read*.

Morphemic characteristics
The verb stem includes a prefix (Pf/): *remove, replace*;
the verb stem includes a suffix (Sf/): *concentrate, criticize*;
the verb stem is a compound form (CF): *telephone.*

Derivational characteristics
From the verb base a noun derivative can be formed (N): *organize - organization; recognize - recognition*;
from the verb base an adjective derivative can be formed (ADJ): *accept - acceptable, impress - impressive*.
Conversion: the verb has a correlated noun in a conversion pair. The direction of derivation is not taken into account (CONV): *cost, dance, mark.*

Syntactic characteristics
The verb is transitive (TR): *John built a house*.
The verb is intransitive (INTR): *John is walking*.
The verb generates a sentence structure with the syntactic position of indirect object (INDR): *John gave her a book*.
The verb generates a sentence structure with the syntactic position of a secondary predicate (2PRD): *The sun rose red*.
The verb generates a sentence structure with the syntactic position of an object clause (CL): *John said that he would go to London*.

Frequency ranks

First (highest) frequency rank (1FR): *achieve, open, pour, save*;
second frequency rank (2FR): *admire, flow, reward, shock*;
third frequency rank (3FR): *bake, curve, dream*;
fourth (lowest) frequency rank (4FR): *stitch, thumb*.
The information about frequency ranks is taken from Macmillan dictionary.

Chronological characteristics

These characteristics are based on the time of the origin of the verbs. The time when the verb first appeared in English was established according to COD.

Verb appeared in Old English period (OE): *hunt, help*, *shake*;
verb appeared in Middle English period (ME): *land, pull, save*;
verb appeared in New English period (NE): *joke, switch, suggest*.

Meta-L membership

This characteristic shows whether the verb is a member of the meta-language or not.

3. Analysis

All 8375 verbs of COD were examined and described in terms of possessing or not possessing each of the above-mentioned characteristics. As a result a database was formed in which lines are verbs and columns are characteristics which can take one of two values: 1 or 0. Table 1 presents an extract from the whole table.

In order to see what properties are decisive and how they influence the probability of a verb being included in Meta-L, correlation analysis was carried out. It is based on a frequency count of binary patterns of characteristics: [1,1] [1,0] [0,1] and [0,0].

Table 1
Extract from the verbal data-base (COD)

	Verb	SYL1	SYL2	SYL3	SYL4	STR1	STR2	STR3	“VL	“SNR	“CON	VL”	SNR”	CON”	Pf/	Sf/	CF
…																	
44	Account	0	1	0	0	0	1	0	1	0	0	0	0	1	1	0	0
45	Accredit	0	0	1	0	0	1	0	1	0	0	0	0	1	1	0	0
46	Accrete	0	1	0	0	0	1	0	1	0	0	0	0	1	0	0	0
47	Accrue	0	1	0	0	0	1	0	1	0	0	1	0	0	0	0	0
48	Acculturate	0	0	0	1	0	1	0	1	0	0	0	0	1	1	1	0
49	Accumulate	0	0	0	1	0	1	0	1	0	0	0	0	1	1	0	0
50	Accuse	0	1	0	0	0	1	0	1	0	0	0	0	1	0	0	0
…																	

Continued

		N	ADJ	CONV	TR	INTR	INDR	2PRD	CL	FR1	FR2	FR3	FR4	OE	ME	NE	Meta-L
…																	
44	Account	1	1	1	1	1	1	1	0	1	0	0	0	0	1	0	1
45	Accredit	1	1	0	1	0	1	0	0	0	0	0	1	0	0	1	0
46	Accrete	1	0	0	1	1	0	0	0	0	0	0	1	0	0	1	0
47	Accrue	1	1	0	0	1	1	0	0	0	0	0	1	0	1	0	0
48	Acculturate	1	1	0	1	1	0	0	0	0	0	0	1	0	0	1	0
49	Accumulate	1	1	0	1	1	0	0	0	0	0	1	0	0	0	1	0
50	Accuse	1	1	0	1	0	1	0	0	1	0	0	0	0	1	0	1
…																	

Thus, to analyze the relationship between any two characteristics **x** and **y**, the data is organized in the form of a two by two table:

	y	$\bar{y}$
x	a	b
$\bar{x}$	c	d

where **a** = the number of verbs in which both characteristics **x** and **y** are present; **b** = the number of verbs in which **x** is present and **y** is absent; **c** = the number of verbs in which **y** is present and **x** is absent; **d** = the number of verbs in which neither **x** nor **y** is present.

At this stage of analysis, we establish the degree of the strength and direction of a relationship between the characteristic "Meta-L membership" and other 31 characteristics.

The most widely used correlation coefficient for dichotomous (alternative) characteristics is the Pearson coefficient φ:

$$(1) \qquad \varphi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(d+b)}} \ .$$

This has a serious limitation: if one of the correlated characteristics has a much higher occurrence than the other, the Pearson coefficient cannot reach the values +1 or –1. To avoid this limitation, the results should be normalized: $\varphi_{norm} = \varphi/\varphi_{max}$ or $\varphi_{norm} = \varphi/\varphi_{min}$ (Tuldava, 1988).

L.C. Cole (Cole, 1949) proposes a coefficient (C) which makes it possible to arrive at normalized values of φ directly:

$$(2) \qquad C = \begin{cases} \dfrac{ad - bc}{(a+b)(b+d)}, & \text{if } ad \geq bc \text{ and } (a+b) \leq (a+c) \\ \dfrac{ad - bc}{(a+b)(a+c)}, & \text{if } ad < bc \text{ and } a \leq d \\ \dfrac{ad - bc}{(b+d)(c+d)}, & \text{if } ad < bc \text{ and } a > d \end{cases}$$

The Cole coefficient can reach the values +1 and – 1 for any frequency of occurrence of the correlated characteristics, thus having a wider range of variance and, consequently, reflecting better the differences in the force of the observed correlations.

The results were evaluated for statistical significance using a t-test. First the sample error σ was calculated using the following formulae, proposed for the coefficient by L.C. Cole (Cole, 1949):

$$(3) \qquad \sigma_C = \begin{cases} \pm\sqrt{\dfrac{(a+c)(c+d)}{n(a+b)(b+d)}}, & \text{if } ad \geq bc \text{ and } (a+b) \leq (a+c) \\ \pm\sqrt{\dfrac{(b+d)(c+d)}{n(a+b)(a+c)}}, & \text{if } ad < bc \text{ and } a \leq d \\ \pm\sqrt{\dfrac{(a+b)(a+c)}{n(b+d)(c+d)}}, & \text{if } ad < bc \text{ and } a > d \end{cases}$$

In this paper a *p*-value of 0.01 is treated as a "borderline acceptable" error. The value of Student's *t* at $p = 0.01$ for a two-sided test with infinite number of degrees of freedom is $t_{st} = 2.58$ which means that values of the coefficient will be considered as statistically significant if $|C| \geq 2.58\sigma_C$.

The results of the correlation analysis are given below. Statistically insignificant values are marked with asterisk.

Frequency rank prediction
- 1FR: 0.65;
- 2FR: 0.14;
- 3FR: –0.47;
- 4FR: –0.99.

Phoneme prediction
- "VL: 0.01*;
- "SNR: 0.02*;
- "CON: –0.04*;
- VL": 0.04*;
- SNR": –0.16*;
- CON": 0.01*.

Stress and syllable prediction
- STR1: –0.54;
- STR2: –0.10*;
- STR3: –0.81;
- SYL1: 0.37;
- SYL2: –0.08*;
- SYL3: –0.66;
- SYL4: –0.95;

Morphemic prediction
- Pf/: –0.73;
- Sf/: –0.73;
- CF: –0.67.

Derivational prediction
- N: 0.30;
- ADJ: 0.30;
- CONV: 0.19.

Syntactic prediction
- TR: 0.62;
- INTR: 0.63;
- INDR: 0.50;
- 2PRD: 0.34;
- CL: 0.14.

Chronological prediction
- OE: 0.28;
- ME: 0.27;
- NE: –0.67.

As can seen from the data above, the phonemic characteristics have no predictive power. Among the stress and syllable characteristics, there is a tendency for predictive force to shift to the beginning and the end of the stem of the verb. Thus characteristics STR1 and

STR3, SYL1 and SYL3, SYL4 are relevant in this respect, whereas STR2, SYL2 are not. The same tendency is also observed, to a certain extent, in frequency rank prediction, where 1FR and 3FR, 4FR possess strong predictive power and 2FR is much less relevant. Suffix in the stem of the verb and derivational suffix, which is added to the stem, are opposed to each other in influence: the former is a negative factor and the latter is positive.

The syntactic characteristics TR and INTR have the same high positive values of the Cole coefficient, which may seem illogical; we must however bear in mind that one and the same verb can be transitive in some of its meanings and intransitive in others. A further study of verbal syntactic characteristics with differentiation between (a) transitive verbs, (b) intransitive verbs and (c) verbs which combine both functions may resolve this problem.

Stress and syllable characteristics display only "limiting" influence (negative correlations) while the capability of verbs to be the motivating base for derivatives (nouns and adjectives) is a positive factor.

4. Similarity

At the next stage of analysis we establish the degree of similarity between the relations of characteristics (a) in the Meta-L and (b) in GV-system. To do this we compare the correlations existing between the "inner" characteristics (phonetic, morphemic and chronological) in Meta-L and in GV-system. Phonemic characteristics are reduced to the opposition "vowel/ consonant", only one member of which is used. The data for the GV-system is taken from our previous research (Andreev, 1995; Sil'nitskij, Andreev, Kuzmin, Kuskov, 1990).

Correlations between "inner" characteristics were established using the Cole coefficient; the results are presented in Tables 3 and 4. As in the previous case of correlation analysis, the level of significance is accepted at a value of 0.01. The values of the coefficient which do not satisfy the condition $|C| \leq 2{,}58\sigma_C$ are marked with an asterisk. Logically impossible combinations (SL1 + SL2, STR2 + SL1, etc.) are marked with 'x'.

These correlations form the basis for a comparison of the schemes of relations existing in both systems. To investigate the degree of similarity, the Jaccard coefficient was used (cf. Bartkov 1981):

$$(4) \qquad K = \frac{m}{S_1 + S_2 - m},$$

where m is the number of similar correlations between the same characteristics in Tables 3 and 4; S_1 is the number of all correlations in Table 3; S_2 is the number of all correlations in Table 4.

Table 3
Correlations of verbal characteristics in Meta-L

	"VL	VL"	STR 1	STR 2	STR 3	SYL1	SYL2	SYL3	SYL4	Pf/	Sf/	CF	OE	ME	NE
"VL	–	0,06	0,08	0,50	0,39	-0,77	0,48	0,38	0,39	0,22	-0,37*	-1,00	-0,64	0,11	0,20
VL"	0,06	–	0,21	-0,05*	0,16*	-0,21	0,17	0,05*	-1,00*	0,10*	-1,00	-0,01*	0,10	-0,09*	-0,10*
STR1	0,08	0,21	–	x	x	x	0,60	0,46	-1,00*	-0,73	0,74	1,00	-0,38	-0,10*	0,20
STR2	0,50	-0,05*	x	–	x	x	0,86	0,04*	1,00	0,67	-0,58	-1,00	-0,89	0,32	0,21
STR3	0,39	0,16*	x	x	–	x	x	1,00	-1,00*	0,47	-1,00*	-1,00*	-0,50*	0,08*	0,16*
SYL1	-0,77	-0,21	x	x	x	–	x	x	x	x	x	x	0,70	-0,17	-0,48

SYL2	0,48	0,17	0,60	0,86	x	x	–	x	x	0,38	0,14*	0,38	-0,66	0,23	0,23
SYL3	0,38	0,05*	0,46	0,04*	1,00	x	x	–	x	0,23	0,34	0,35	-0,92	-0,11*	0,42
SYL4	0,39	-1,00*	-1,00*	1,00	-1,00*	x	x	x	–	0,47	0,48	-1,00*	-1,00*	-1,00	1,00
Pf/	0,22	0,10*	-0,73	0,67	0,47	x	0,38	0,23	0,47	–	0,24	-0,12*	-0,76	0,12*	0,19
Sf/	-0,37*	-1,00	0,74	-0,58	-1,00*	x	0,14*	0,34	0,48	0,24	–	0,24	-0,33	-0,51	0,37
CF	-1,00	-0,01*	1,00	-1,00	-1,00*	x	0,38	0,35	-1,00*	-0,12*	0,24	–	-1,00*	-1,00*	-0,40*
OE	-0,64	0,10	-0,38	-0,89	-0,50*	0,70	-0,66	-0,92	-1,00*	-0,76	-0,33	-1,00*	–	x	x
ME	0,11	-0,09*	-0,10*	0,32	0,08*	-0,17	0,23	-0,11*	-1,00	0,12*	-0,51	-1,00*	x	–	x
NE	0,20	-0,10*	0,20	0,21	0,16*	-0,48	0,23	0,42	1,00	0,19	0,37	-0,40*	x	x	–

Table 4
Correlations of verbal characteristics in the system of English (GV-system)

	"VL	VL"	STR 1	STR 2	STR 3	SYL1	SYL2	SYL3	SYL4	Pf/	Sf/	CF	OE	ME	NE
"VL	–	-0,15	-0,39	0,48	0,27	-0,90	0,12	0,16	0,22	0,35	0,04	0,83	-0,57	-0,04*	0,18
VL"	-0,15	–	0,27	-0,14	-0,04*	-0,42	0,23	0,03*	-0,30	-0,07*	-0,32	0,06*	0,09	0,05	-0,10
STR1	-0,39	0,27	–	x	x	x	0,33	0,39	-0,52	-0,77	0,44	-0,26	-0,59	-0,21	0,28
STR2	0,48	-0,14	x	–	x	x	0,45	-0,29	0,59	0,56	0,01*	0,01*	-0,71	0,03*	0,10*
STR3	0,27	-0,04*	x	x	–	x	x	0,63	0,06	0,57	0,02*	0,38	-0,61	-0,56	0,52
SYL1	-0,90	-0,42	x	x	x	–	x	x	x	x	x	x	0,65	-0,27	-0,26
SYL2	0,12	0,23	0,33	0,45	x	x	–	x	x	0,12	-0,61	-0,26	-0,47	0,14	-0,04
SYL3	0,16	0,03*	0,39	-0,29	0,63	x	x	–	x	0,20	0,32	0,41	-0,86	-0,44	0,50
SYL4	0,22	-0,30	-0,52	0,59	0,06	x	x	x	–	0,20	0,45	0,06	-0,98	-0,85	0,81
Pf/	0,35	-0,07*	-0,77	0,56	0,57	x	0,12	0,20	0,20	–	-0,15*	0,99	-0,56	-0,21	0,27
Sf/	0,04	-0,32	0,44	0,01*	0,02*	x	-0,61	0,32	0,45	-0,15*	–	-0,71	-0,87	-0,55	0,60
CF	0,83	0,06*	-0,26	0,01*	0,38	x	-0,26	0,41	0,06	0,99	-0,71	–	-0,22*	-0,37	0,32
OE	-0,57	0,09	-0,59	-0,71	-0,61	0,65	-0,47	-0,86	-0,98	-0,56	-0,87	-0,22*	–	x	x
ME	-0,04*	0,05	-0,21	0,03*	-0,56	-0,27	0,14	-0,44	-0,85	-0,21	-0,55	-0,37	x	–	x
NE	0,18	-0,10	0,28	0,10*	0,52	-0,26	-0,04	0,50	0,81	0,27	0,60	0,32	x	x	–

To establish the statistical relevance of the results J. Tuldava (1974; see also Levitskij 2004, 129) suggests calculating σ, according to the following formula:

$$\sigma = \sqrt{\frac{K(1-K)}{S_1 + S_2}}. \tag{5}$$

We consider as similar those cases where in both tables the same pair of characteristics reveals the same type (either positive or negative) of statistically significant correlation. The strength of the correlations is not taken into account. Thus, for example, in both tables positive correlations are observed between STR2 and "VL (0.50 in Table 1 and 0.48 in Table 2), STR3 – "VL (0.39 and 0.27), Pf/ – STR2 (0.67 and 0.56), Pf/ – STR3 (0.47 and 0.57), OE – SYL1 (0.70 and 0.65).

On the other hand both in Meta-L and GV-system negative correlations were established between such characteristics as OE – SYL3 (–0.92 in Table 1 and –0.86 in Table 2), OE – STR2 (–0.89 and –0.71), SYL1 – "VL (–0.77 and –0.90).

The Jaccard coefficient is always positive and can take values in the range 0 to 1, but as explained above, the basis of comparison of the two tables is the concurrence of not only positive but also negative correlations.

Dissimilarities may be of two main types. Type 1 takes place when a positive correlation in one table corresponds to a negative correlation in the other. Thus STR1 and "VL in Meta-L (Table 3) are positively correlated (0.08), but in the general verbal system (Table 4) there is a negative correlation between them (–0.39). This corresponds to the notion of equipolent opposition. The second type of dissimilarity is observed when there is either a positive or a negative correlation in one table and no correlation at all in the other. Thus STR2 is negatively correlated with Sf/ in Table 3 (–0.58) and has no statistically significant correlation with this characteristic in Table 4 (0.01*). This type of dissimilarity corresponds to the notion of privative opposition.

Since Meta-L is a derivative based on the primarily existing language system, in cases of similarity we can also conventionally speak of 'stability' of relations, observed in Meta-L as compared to GV-system. This approach may be regarded as a dynamic vision of a metalanguage: in the process of transaction from the initial to the final stage, passing through semantic filters and losing over 99% of its lexics, the primary system retains or loses some of its features. This approach brings the present study closer to the problems of one of the fundamental questions of linguistics – self-regulation in language (Altmann, Köhler, 1996; Piotrowski, 2005).

The degree of similarity was established for different types of characteristics separately and then for all of them together. The results of the analysis are presented in Table 5 below.

Table 5
Similarity analysis data

Characteristics	m	S_1	S_2	S_1+S_2	K	σ	2,58σ (p=0,01)
Length:	26	37	37	74	0,542	0,058	0,149
Stress position	15	32	32	64	0,306	0,058	0,149
Phoneme	14	27	27	54	0,350	0,065	0,167
Morphemic characteristics	17	39	39	78	0,279	0,051	0,131
Chronological characteristics	20	36	36	72	0,385	0,057	0,148
All characteristics	46	86	86	172	0,365	0,037	0,095

Judging by the results, Meta-L retains certain features of the formal system of English verbs. All the values of the coefficient are statistically relevant at $p = 0.01$ ($C > 2.58\sigma_C$).

As Table 5 shows, characteristics which reflect verbal properties at different levels do not differ in maintaining the stability of the system of relations.

Stability is realized by both positive and negative correlations: there are 31 positive and 15 negative correlations that are similar in both systems.

The stability of positive and negative correlations for different groups of characteristics varies. They are distributed among the parameters as follows: phonemes (10 positive and 4 negative); stresses (12 and 3); syllables (19 and 7); morphemic structure (12 and 5), chronology (9 and 11).

All the characteristics fall into three groups depending on the number of positive and negative stable correlations, observed in our study. The first group includes 'positively oriented' characteristics, which possess mostly positive stable correlations: "VL, STR1, STR2, STR3, SYL2, SYL3, SYL4, Pf/, NE. The other group is formed of characteristics which have an equal or nearly equal number of positive and negative correlations: VL", Sf/, CF, ME. The last group is 'negatively orientated' as the characteristics in it have mostly negative stable correlations: SYL1, OE.

4. Conclusion

The results described above allow us to compose an image of a verb whose formal characteristics predict its inclusion into the Meta-L. The verb has frequency rank 1; consists of 1 syllable; has a simple stem; is characterized by affix derivation; has a syntactic function in which it collocates with an indirect object and second predicate; and originated either in Old or Middle English period. Transitive and intransitive functions are not taken into account because they have equal predictive force.

It is important to stress that this image is based on the predictive power of the characteristics, not on the number of verbs in Meta-L that possess the given characteristics. Among the above-mentioned characteristics some are found frequently in Meta-L: 1FR (69.1%), N (70.8%), INDR (60.5%), others are less widely spread: 2PR (38%), OE (33.2%). On the other hand conversion, found in 58.8% of the verbs in Meta-L, does not have high predictive power.

Comparison of the systematic relations that exist in GV-system and in Meta-L makes it possible to single out a set of parameters which support the stability of relations of the elements in the new system, as compared to the original system: "VL – STR1 – SYL2 – SYL3 – Pf/ – Sf/ – OE – NE.

As discussed at the outset, the guiding principle of the compilers of this metalanguage is the semantics of the words. But this study has demonstrated that there exist certain tendencies in selecting formal types of verbs, due to which the new system retains a number of features of the natural language. Stability of relationship was found both in positive and negative correlations.

References

Altmann, G., Köhler, R. (1996). "Language Forces" and Synergetic Modelling of Language Phenomena. *Glottometrika 15, 62-76.*

Andreev, S.N. (1995). *Mnogomernaja klassifikacija jazikovich edinic (na materiale anglijskich glagolov)* [Multivariate classification of language units (on the material of English verbs)]. Moscow: Inion RAN.

Bartkov, B.I. (1981). O koefficientach schodstva členov synonimičeskich rjadov [On the coefficients of similarity of synonymic groups]. In: *Strukturnaja i prikladnaja lingvistika: 6-13*. Kiev: KGU.

Cole, L.C. (1949). The measurement of Interspecific Association. *Ecology 30(4) 411-424.*

Levitskij, V.V. (2004). *Kvantitativnije metodi v lingvistike* [Quantitative methods in linguistics]. Černovci: Ruta.

Piotrowski, R.G. (2005). *Sinergetika teksta* [Synergetics of text]. Minsk: MGLU.

Sil'nitskij, G.G., Andreev, S.N., Kuzmin, L.A., Kuskov, M.I. (1990). *Sootnošenie glagol'nich priznakov različnich urovnej v anglijskom jazyke* [Relationship of verbal characteristics of different levels in English]. Minsk: Navuka i Technika.

Tuldava, J. (1974). Ob izmerenii leksičeskoj svjazi tekstov na urovne slovarja [On the measurement of lexical connection of texts]. In: *Voprosy statističeskoj stilistiki*: *35-42*. Kiev: Naukova dumka.

Tuldava, J. (1988). O primenenii koefficientov soprjažennosti v lingvistike [About the use of contingency coefficients in linguistics]. In: J.Tuldava (Ed.), *Prikladnaja lingvistika i avtomatičeskij analiz teksta: 83-84*. Tartu: TUP.

DICTIONARIES

Cambridge International Dictionary of English. (1995). Cambridge – New York – Melbourne: Cambridge University Press.

Longman Dictionary of Contemporary English. (2005). Fourth edition with Writing Assistant. Harlow: Pearson Education Limited.

Macmillan English Dictionary. (2002). Oxford: Bloomsbury Publishing Plc., Macmillan Publishers Limited.

Concise Oxford Dictionary of Current English. (1987). 7 ed. Bombay: Oxford Univ.Press.

Glottometrics 14, 2007, 32-45

Quantitative Untersuchungen zum deutschen Wörterbuch

Karl-Heinz Best, Göttingen[1]

Abstract. This paper deals with some properties of the German lexicon. The purpose of the paper is to present some further data concerning morphs, syllables, parts of speech, and borrowings and to show that entities and processes in the lexicon abide by language laws, too.

Keywords: German, word length, dictionary

0. Themen der Untersuchung

Gegenstand dieser Untersuchungen sind einige phonetische und morphologische Eigenschaften des deutschen Lexikons. Wörterbuchuntersuchungen sind im Göttinger *Projekt Quantitative Linguistik* zugunsten von Textanalysen etwas vernachlässigt worden. Um diese Lücke auszufüllen bedürfte es eines sehr großen Aufwandes, der von mir nicht zu leisten ist. Einige Aspekte sollen jedoch auf der Grundlage einer systematischen Stichprobenerhebung bearbeitet werden. Es handelt sich im Einzelnen um folgende Themen:

Wie sind Wortlängen – bestimmt nach der Zahl der Morphe oder Silben pro Wort – im Wörterbuch verteilt? Wie entwickelt sich die Zahl der Morphe, wenn man nach und nach die Zahl der Wörter erhöht? Wie verteilen sich die erhobenen Stichwörter auf die Wortarten? Welche Rolle spielen Entlehnungen aus anderen Sprachen im deutschen Wortschatz? In allen Fällen gilt es, die Hypothese zu überprüfen, dass sprachliche Erscheinungen immer theoretisch begründbaren Gesetzen unterliegen (Altmann 1985: 7). Zu den genannten Themen soll entsprechend jeweils ein Gesetzesvorschlag, der bereits in der Literatur für solche Fälle geprüft wurde, aufgegriffen und hier erneut überprüft werden. Es mag weiteren Untersuchungen vorbehalten bleiben, diese Vorschläge weiter zu überprüfen und ggfs. auch durch bessere zu ersetzen.

1. Datenerhebung

Grundlage der folgenden Darstellung sind einerseits neue Auswertungen bereits bekannter Daten; im Wesentlichen stützt sie sich jedoch auf eine neue Datenerhebung. Ausgewertet wurde zu diesem Zweck *Duden. Deutsches Universalwörterbuch* (^{4}2001), das in seinem Wörterbuchteil etwas über 1800 Seiten aufweist. Die Stichprobe erfasste in einem ersten Durchgang von jeder Doppelseite das letzte Wort der letzten Spalte; in einigen Fällen wurde ein vorhergehendes Stichwort gewählt, vor allem, wenn es ein Affix, Eigenname, Syntagma oder eine besondere Flexionsform war. In einem zweiten Durchgang wurde die Liste durch das letzte Wort der vorletzten Spalte ergänzt. Insgesamt wurden auf diese Weise 2710 Stichwörter aufgenommen. Eine größere Stichprobe wäre vielleicht wünschenswert; für die Zwecke dieser Arbeit sollten die Daten aber genügen.

Jedes Stichwort wurde aufgelistet; ihm wurden mehrere Informationen hinzugefügt: Wenn es sich bei einem Wort oder wenigstens bei einer seiner Konstituenten um eine Ent-

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

lehnung handelte, so wurde angegeben, aus welcher Sprache sie ins Deutsche übernommen wurde. Entscheidend war dabei der fremdsprachige Hintergrund. Es wurden also auch solche Konstituenten bzw. Wörter als Entlehnungen aufgefasst, die im Deutschen entstanden sind wie das allseits bekannte „Handy". Die Angaben folgten, wenn möglich, dem ausgewerteten Wörterbuch, wurden aber bei Bedarf um die Angaben in Kluge ([24]2002) und in Einzelfällen auch aus *Duden. Herkunftswörterbuch* ([3]2002; Datum des Vorworts) und Pfeifer ([2]1993/ 1995) ergänzt. „Fantasie" z.B. wurde als Entlehnung aus dem Lateinischen aufgenommen, der Sprache, über die das Wort ins Deutsche gelangte (Vermittlersprache), und nicht als Entlehnung aus dem Griechischen, woher das Wort letztlich stammt (Herkunftssprache). Eingetragen wurde also immer die Vermittlersprache, nicht die Herkunftssprache. Wörter wurden in der Regel als Ganze erfasst; nur wenn einzelne Konstituenten aus verschiedenen Sprachen stammten, wurden diese getrennt aufgelistet. Da also ggfs. die Wortkonstituenten einzeln erfasst wurden, gab es in etlichen Fällen für ein komplexes Wort mehrere Angaben zur Vermittlersprache, z.B. beim Wort „Ratingskala", dessen erste Konstituente aus dem Englischen stammt, während die zweite dem Italienischen entlehnt ist. Die Erhebung hat ergeben, dass von den 2710 aufgenommenen Wörtern 1583 (58.41%) keine entlehnte Konstituente enthalten. Das heißt, etwas mehr als 40% der Stichprobe enthalten mindestens eine entlehnte Konstituente.

Der zweite Eintrag galt der Wortartbestimmung; wenn *Duden. Deutsches Universalwörterbuch* ([4]2001) für ein Stichwort mehr als eine Wortart angab, wurde immer nur die erste übernommen. Nur bei Kontraktionen wie „im" wurden zwei Wortartangaben (Präposition + Artikel) eingetragen.

Der dritte Eintrag galt der Silbenstruktur: alle Wörter wurden in Silben segmentiert und in einer eigenen Spalte die Silbenzahl vermerkt. Das gleiche wurde dann auch für die Morphe der Wörter durchgeführt. Die Segmentierungsverfahren sind in Best (2001c,d, 2006a) beschrieben. Hier hat sich gezeigt, dass auf 1583 Wörter ohne entlehnte Konstituente 4488 Morphe und 4506 Silben kommen; bei der Gesamtstichprobe von 2710 Wörtern ergaben sich 7469 Morphe und 8242 Silben. Dabei machen sich vermutlich morphologische Segmentierungsprobleme bei Entlehnungen bemerkbar.

Alle Einträge erfolgten in einer Excel-Tabelle; man kann dann die Spalten oder Zeilen kopieren und in Word oder MS DOS umformatieren und damit Auswertungsprogramme nutzen, die auf diese Formatierungen angewiesen sind.[2]

2. Auswertungen

In den folgenden Abschnitten werden die Auswertungen der Datensätze vorgestellt. Zunächst geht es um die Wortlängenverteilungen im Wörterbuch.

2.1. Wortlängen (in Morphen gemessen) im Wörterbuch des Deutschen

2.1.1. Wortlängenverteilung (in Morphen) in Wahrigs Wörterbuch

In etlichen Untersuchungen konnte schon gezeigt werden, dass Wortlängen in Texten gesetzmäßig verteilt sind (Best 2001a; zuletzt Best 2006b); das gleiche gilt offenbar auch für Wortlängen in Wörterbüchern, wie einige Ergebnisse zeigen (Best [2]2003: 46, [3]2006: 42; Best & Zhu 2001: 103-105; Köhler 2002: 57-58; Rheinländer 2001: 151; Wimmer u.a. 1994: 102).

[2] In den angegebenen Daten stecken geringfügige Fehler (im Promillebereich), die u.a. beim Umformatieren auftreten und sich nicht alle vermeiden lassen.

Wortlängen verteilen sich in Wörterbüchern aber z.T. nach anderen Modellen, als dies bei Texten zu beobachten ist. Hier wird nun zum ersten Mal geprüft, ob auch Wortlängen, in Morphen gemessen, sich im Lexikon gesetzmäßig verteilen, wie dies Wimmer u.a. (1994) erwarten lassen. Zu diesem Zweck können Daten, die Gerlach (1982) erarbeitet hat, berücksichtigt werden. Es hat sich ergeben, dass sich in diesem Fall die 1-verschobene Conway-Maxwell-Poisson-Verteilung

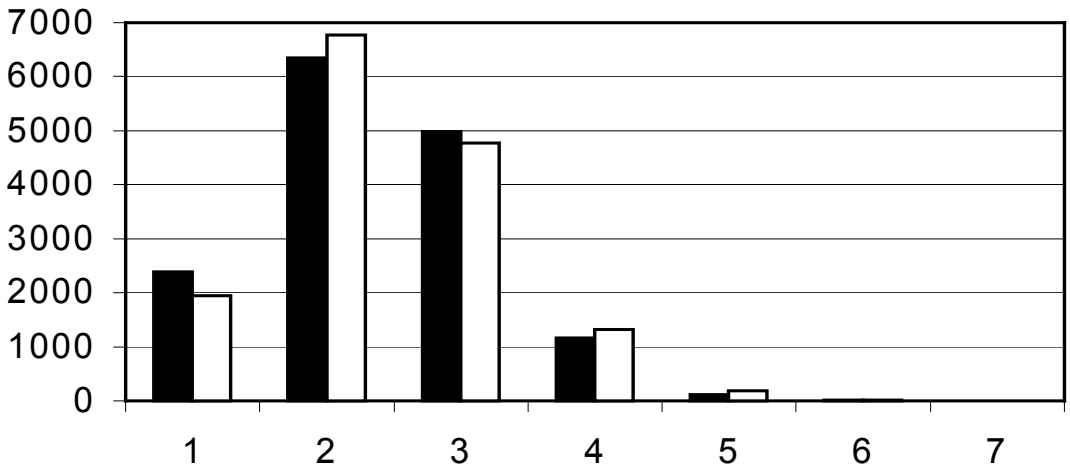
$$(1) \quad P_x = \frac{a^{x-1}}{[(x-1!)]^b \, T_1}, \quad x = 1, 2, ... \quad \text{mit} \quad T_1 = \sum_{j=0}^{\infty} \frac{a^j}{(j!)^b}$$

bewährt, wie Tabelle (1) und Graphik (1) zeigen.

Legende zu Tab. 1: (Chi² =) X^2 ist das Chiquadrat; *FG* die Zahl der Freiheitsgrade. *C* ist der Diskrepanzkoeffizient (Kontingenzkoeffizient) $C = X^2/N$, der dann verwendet wird, wenn wie hier eine sehr große Datei zu bearbeiten ist. Mit *a*, *b* werden die Parameter dieser Verteilung angegeben. *x* ist die Zahl der Morphe pro Wort, n_x die beobachtete, NP_x die berechnete Zahl der Morphe pro Wort. Das Testergebnis ist mit $C = 0.0128$ akzeptabel, wie auch Graphik 1 veranschaulicht, aber mit $C > 0.01$ nicht ganz ideal.

Tabelle 1
Anpassung der 1-verschobenen Conway-Maxwell-Poisson-Verteilung
an die Wortlängen (Morphe pro Wort) in Wahrigs *dtv-Wörterbuch*
(Gerlach 1982: 98; 15011 Wörter)

x	n_x	NP_x
1	2391	1939.58
2	6343	6773.95
3	4989	4773.88
4	1159	1319.14
5	112	187.59
6	13	15.94
7	4	0.92
$a = 3.4925$ $b = 2.3091$ $FG = 3$ $X^2 = 192.0775$ $C = 0.0128$		



Graphik 1: Verteilung der Wortlängen (in Morphen gemessen) in Wahrigs *dtv-Wörterbuch*

2.1.2. Wortlängenverteilung (in Morphen) in *Duden. Deutsches Universalwörterbuch*

Hier geht es noch einmal um die Wortlängenverteilung (nach der Zahl der Morphe gemessen). Der Versuch, die Conway-Maxwell-Poisson-Verteilung auf die Stichprobe aus *Duden. Deutsches Universalwörterbuch* ([4]2001) anzuwenden, hat sich weniger bewährt; stattdessen ließ sich in diesem Fall die 1-verschobene Hyperpoisson-Verteilung

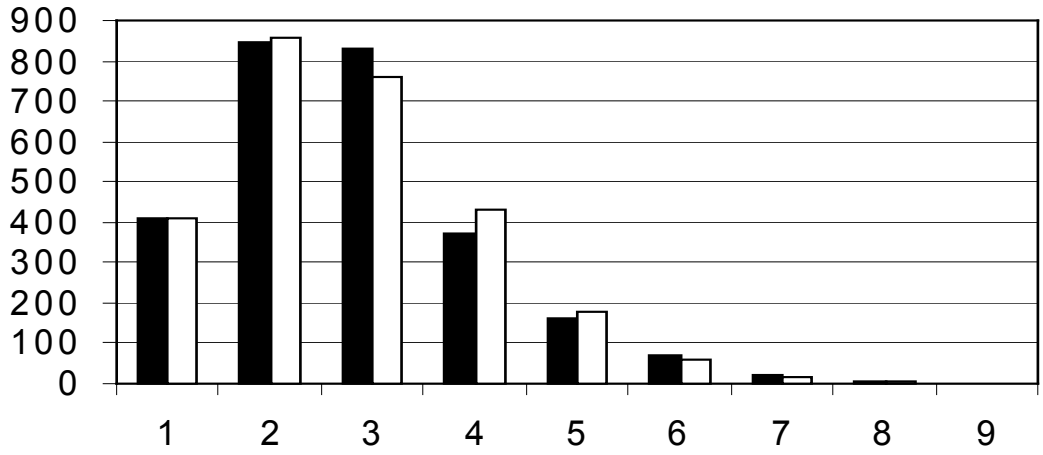
$$(2) \quad P_x = \frac{a^{x-1}}{b^{(x-1)}\,{}_1F_1(1;b;a)}, \quad x = 1, 2,...$$

mit gutem Ergebnis anpassen:

Tabelle 2
Anpassung der 1-verschobenen Hyperpoisson-Verteilung
an die Wortlängen (Morphe pro Wort) in einer Stichprobe aus
Duden. Deutsches Universalwörterbuch; 2710 Wörter)

x	n_x	NP_x
1	408	411.40
2	848	855.06
3	828	759.30
4	370	428.72
5	159	177.44
6	69	57.97
7	23	15.64
8	4	3.59|
9	1	0.87|
a = 1.5504 b = 0.7960 FG = 5 X^2 = 21.8864 C = 0.0081		

Die senkrechten Striche in der Tabelle zeigen an, dass die entsprechenden Längenklassen zusammengefasst wurden.



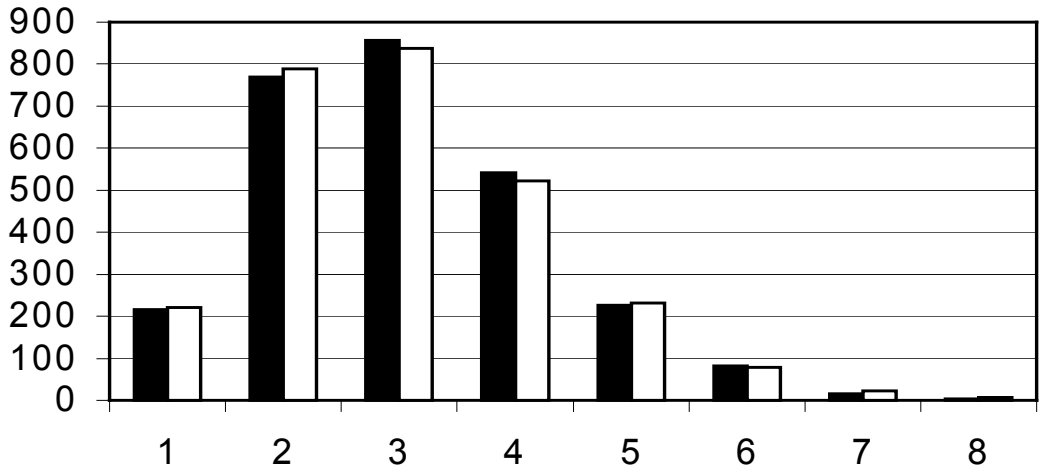
Graphik 2: Verteilung der Wortlängen (in Morphen gemessen) in *Duden. Deutsches Universalwörterbuch*

2.2 Wortlängenverteilung (in Silben) in *Duden. Deutsches Universalwörterbuch*

Die Wortlängenverteilung für die Zahl der Silben pro Wort wurde bereits in Best (²2003: 46) für ein alphabetisches Wörterbuch des Deutschen behandelt; es zeigte sich, dass die Conway-Maxwell-Poisson-Verteilung (Modell 1) mit gutem Ergebnis (C = 0.0094) an die Daten angepasst werden konnte. Dieses Modell lässt sich mit C = 0.0027 auch auf die Stichprobe aus *Duden. Deutsches Universalwörterbuch* anwenden. (Zur Anwendung dieses Modells auf ein ungarisches Wörterbuch siehe Anhang.) Noch besser ist das Ergebnis allerdings, wenn man stattdessen die 1-verschobene Hyperpoisson-Verteilung anpasst, wie die folgende Tabelle 3 und die Graphik zeigen:

Tabelle 3
Anpassung der 1-verschobenen Hyperpoisson-Verteilung
an die Wortlängen (Silben pro Wort) in einer Stichprobe aus
Duden. Deutsches Universalwörterbuch; 2710 Wörter)

x	n_x	NP_x
1	216	221.36
2	769	788.08
3	856	837.72
4	541	523.37
5	227	231.53
6	82	79.28
7	15	22.14
8	4	6.52
a = 1.5155 b = 0.4257 FG = 5 X^2 = 5.0431 C = 0.0019		



Graphik 3: Verteilung der Wortlängen (in Silben gemessen) in *Duden. Deutsches Universal wörterbuch*

2.3. Verteilung des Wortschatzes auf die Wortarten

Wortartenverteilungen waren schon häufig Gegenstand quantitativer Untersuchungen und wurden dabei u.a. als Diversifikationsphänomen betrachtet (Altmann 2005). Die Untersuchungen galten dem Vorkommen der Wortarten in Texten. Im vorliegenden Fall wird die

Möglichkeit genutzt, Wortarten einmal in einem Lexikon zu untersuchen. Es wird geprüft, ob die nach Häufigkeit geordneten Wortarten Altmanns Modell für beliebige Rangordnungen

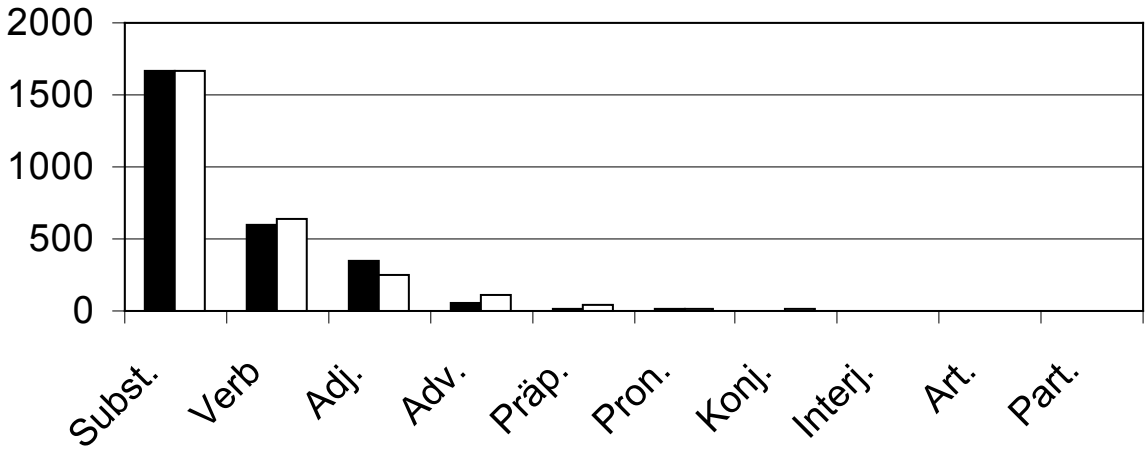
$$(3) \quad y_x = \frac{\binom{b+x}{x-1}}{\binom{a+x}{x-1}} c, \qquad x = 1, 2, 3, ...$$

(Altmann 1993: 61f., Gesetzesvorschlag 11) folgen. Das Ergebnis (vgl. Tabelle 4):

Tabelle 4
Wortarten im Lexikon (2710 Wörter)

Wortart	beobachtet	berechnet	Wortart	beobachtet	berechnet
Substantiv	1667	1667.00	Pronomen	12	19.10
Verb	602	641.64	Konjunktion	4	8.48
Adjektiv	347	255.16	Interjektion	3	3.85
Adverb	56	104.60	Artikel	3	1.79
Präposition	14	44.11	Partikel	3	0.85
$a = 45.1592$ $b = 16.1518$ $D = 0.9947$					

Anmerkung: Die abweichende Summe aufgrund von Kontraktion.



Graphik 4: Wortarten im Lexikon

2.4. Zunahme der Morphe bei wachsender Wortzahl

In diesem Fall geht es nun um die Frage, wie die Zahl der Morphtypes sich entwickelt, wenn man immer mehr Wörter in Betracht zieht. Zu diesem Zweck wurden Abschnitte von zunächst 10, 50, dann 100, 200, 300 usw. Wörtern gebildet und für jede so entstandene Wortmenge mit Hilfe einer geeigneten Software bestimmt, wie viele Morphtypes jeweils benötigt wurden, um diese Wörter zu bilden. Es hat sich herausgestellt, dass die Zunahme der Morphtypes der gleichen Gesetzmäßigkeit folgen, die bereits beim Wortschatzwachstum in Texten zu beobachten war (Best 2004a,b, 2006c), nämlich dem Potenzgesetz

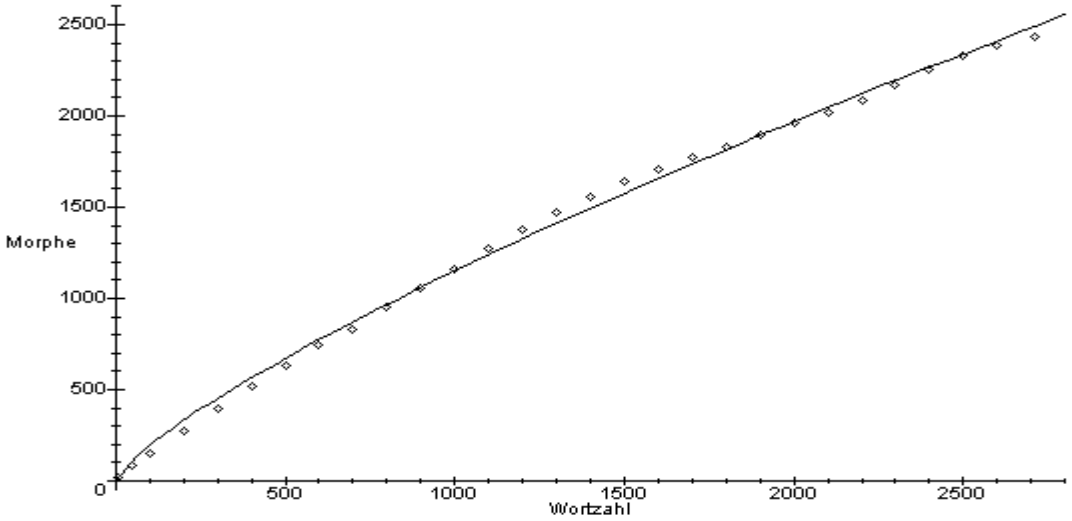
$$(4) \quad y = a\, x^{-b},$$

das in der Linguistik vielfach als „Menzerath-Altmann-Gesetz“ Verwendung findet. Die Dynamik dieses Prozesses wird auf zweierlei Weise vorgestellt: einmal in absoluten, danach in relativen Zahlen. Der Zuwachs in absoluten Zahlen stellt sich wie folgt dar (vgl. Tabelle 5):

Tabelle 5
Anpassung des Potenzgesetzes an die Zunahme der Morphtypes bei wachsender Wortzahl in einer Stichprobe aus *Duden. Deutsches Universalwörterbuch*; 2710 Wörter

Textabschnitt Wortzahl	Morphtypes beobachtet	Morphtypes berechnet	Textabschnitt Wortzahl	Morphtypes beobachtet	Morphtypes berechnet
10	19	32.27	1400	1555	1492.21
50	82	112.49	1500	1644	1574.26
100	151	192.59	1600	1708	1655.09
200	270	329.75	1700	1769	1734.79
300	398	451.65	1800	1833	1813.45
400	522	564.59	1900	1896	1891.14
500	631	671.30	2000	1957	1967.91
600	747	773.30	2100	2021	2043.83
700	832	871.54	2200	2088	2118.94
800	951	966.67	2300	2172	2193.29
900	1054	1059.16	2400	2258	2266.92
1000	1162	1149.37	2500	2333	2339.86
1100	1269	1237.58	2600	2391	2412.16
1200	1378	1324.01	2710	2430	2490.96
1300	1467	1408.84			
$a = 5.4076$ $b = -0.7758$ $D = 0.9974$					

a, *b*: Parameter des Modells. *D*: Determinationskoeffizient, der mit $D = 0.9974$ eine sehr gute Übereinstimmung zwischen Modell und Beobachtung anzeigt.

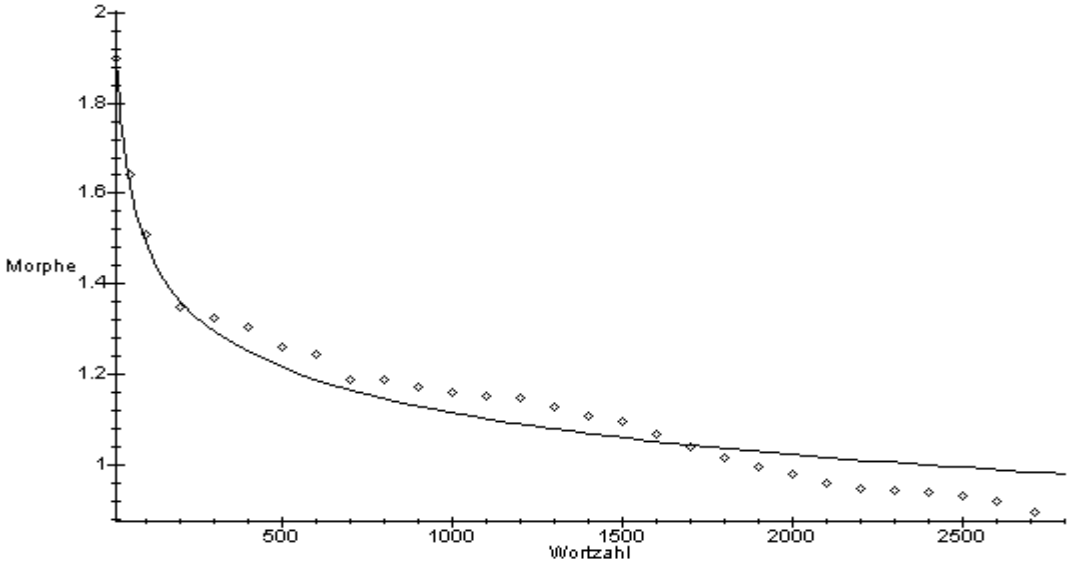


Graphik 5: Zunahme der Morphtypes bei wachsender Wortzahl (2710 Wörter)

Betrachtet man statt der absoluten Werte die Relation Morphtypes/Wörter, so ergeben sich Resultate, die in Tabelle 6 dargestellt sind.

Tabelle 6
Anpassung des Potenzgesetzes an die Relation Morphtypes/ Wörter in einer Stichprobe aus *Duden. Deutsches Universalwörterbuch*; 2710 Wörter

Textabschnitt Wortzahl	Morphtypes/ Wörter beobachtet	Morphtypes/ Wörter berechnet	Textabschnitt Wortzahl	Morphtypes/ Wörter beobachtet	Morphtypes/ Wörter berechnet
10	1.9000	1.9819	1400	1.1107	1.0795
50	1.6400	1.6212	1500	1.0960	1.0696
100	1.5100	1.4868	1600	1.0675	1.0604
200	1.3500	1.3636	1700	1.0406	1.0519
300	1.3267	1.2963	1800	1.0183	1.0439
400	1.3050	1.2506	1900	0.9979	1.0365
500	1.2620	1.2162	2000	0.9785	1.0296
600	1.2450	1.1889	2100	0.9624	1.0168
700	1.1886	1.1662	2200	0.9491	1.0109
800	1.1888	1.1470	2300	0.9443	1.0053
900	1.1711	1.1302	2400	0.9408	1.0000
1000	1.1620	1.1154	2500	0.9332	0.9949
1100	1.1536	1.1009	2600	0.9196	0.9900
1200	1.1483	1.1023	2710	0.8967	0.9849
1300	1.1285	1.0903			
$a = 2.6418$ $b = 0.1248$ $D = 0.9534$					



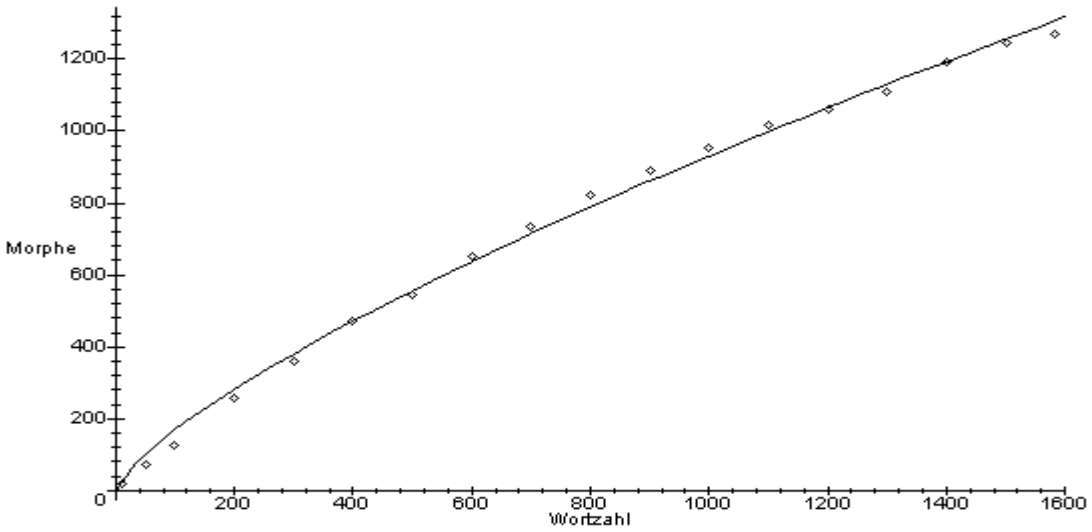
Graphik 6: Relation Morphtypes/ Wortzahl (2710 Wörter)

Die Anpassung des Modell ist mit $D = 0.95$ sehr gut; dennoch sieht man, dass mit wachsendem Wortschatz die Beobachtungswerte unterhalb der Trendlinie bleiben. Um heraus-

zufinden, welchen Anteil daran evt. der recht hohe Anteil an Entlehnungen hat, wurden aus der Liste der 2710 Wörter alle diejenigen entfernt, die mindestens eine Wortkonstituente enthalten, die aus einer anderen Sprache übernommen wurde. Es handelt sich also keineswegs immer um Wörter, die insgesamt entlehnt wurden. Viele dieser Wörter sind auch schon vor so langer Zeit ins Deutsche gekommen, dass ihr fremder Charakter nicht mehr erkennbar ist. Wieder andere sind gelehrte Bildungen, die morphologisch einen fremdsprachigen Hintergrund haben, aber in der verwendeten Form nicht entlehnt, sondern im Deutschen selbst gebildet wurden. Die 1583 Wörter ohne entlehnte Konstituente zeigen folgenden Trend (s. Tabelle 7):

Tabelle 7
Anpassung des Potenzgesetzes an die Zunahme der Morphtypes
bei wachsender Wortzahl in einer Stichprobe aus
Duden. Deutsches Universalwörterbuch; 1583 Wörter ohne Entlehnungen

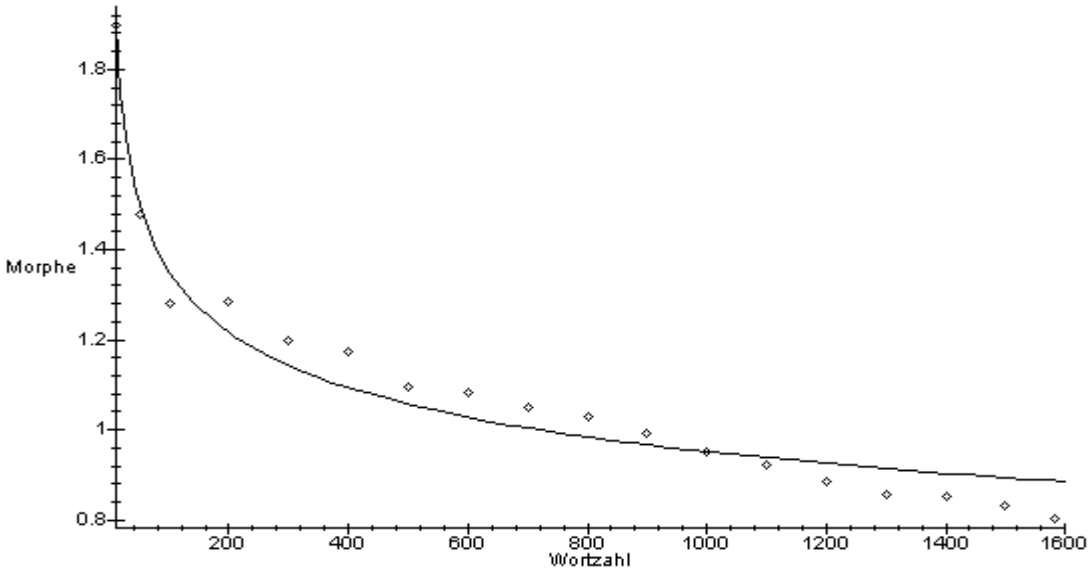
Textabschnitt Wortzahl	Morphtypes beobachtet	Morphtypes berechnet	Textabschnitt Wortzahl	Morphtypes beobachtet	Morphtypes berechnet
10	19	30.55	800	822	788.09
50	74	100.80	900	892	860.03
100	128	168.55	1000	952	929.94
200	257	281.85	1100	1015	998.06
300	359	380.74	1200	1063	1064.59
400	470	471.30	1300	1111	1129.71
500	547	556.13	1400	1192	1193.55
600	651	636.66	1500	1245	1256.21
700	735	713.77	1583	1272	1307.41
a = 5.5374 b = -0.7417 D = 0.9970					



Graphik 7: Zunahme der Morphtypes bei wachsender Wortzahl (1583 Wörter)

Tabelle 8
Anpassung des Potenzgesetzes an die Relation Morphtypes/ Wörter in einer Stichprobe aus *Duden. Deutsches Universalwörterbuch*; 1583 Wörter ohne Entlehnungen

Textabschnitt Wortzahl	Morphtypes/ Wörter beobachtet	Morphtypes/ Wörter berechnet	Textabschnitt Wortzahl	Morphtypes/ Wörter beobachtet	Morphtypes/ Wörter berechnet
10	1.9000	1.9179	800	1.0275	0.9842
50	1.4800	1.5011	900	0.9911	0.9667
100	1.2800	1.3507	1000	0.9520	0.9513
200	1.2850	1.2154	1100	0.9227	0.9376
300	1.1967	1.1427	1200	0.8858	0.9253
400	1.1750	1.0937	1300	0.8546	0.9140
500	1.0940	1.0572	1400	0.8514	0.9038
600	1.0850	1.0282	1500	0.8300	0.8943
700	1.0500	1.0044	1583	0.8035	0.8870
$a = 2.7232$ $b = 0.1523$ $D = 0.9618$					



Graphik 8: Relation Morphtypes/ Wortzahl (1583 Wörter ohne Entlehnungen)

Man kann bei diesen beiden Auswertungen erkennen, dass die Abweichungen zwischen den Beobachtungen und den Anpassungen des Modells gegen Ende der Kurven deutlich größer werden. Die Kurvenverläufe zeigen außerdem, dass mit wachsender Wortzahl die Zahl der hinzukommenden Morphe abnimmt. Die Trends sind also gleich, ob nun mit oder ohne Berücksichtigung der Entlehnungen.

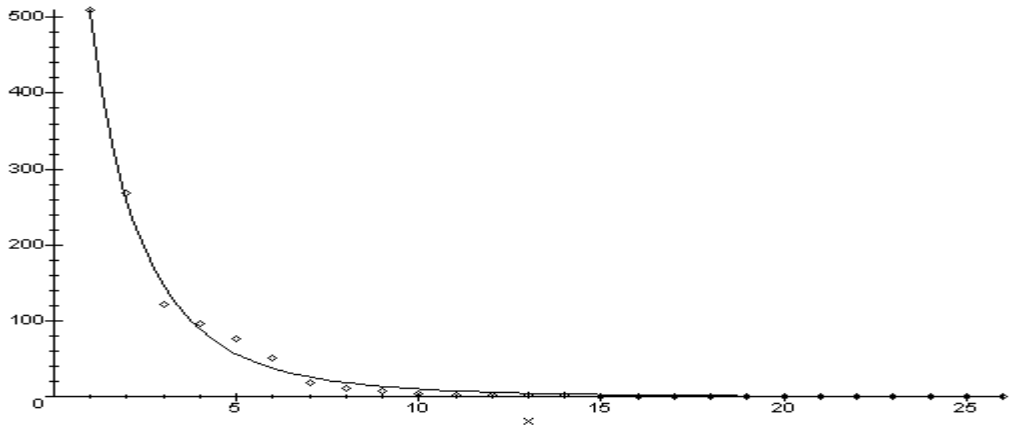
Trotz der vergleichsweise geringen Stichprobe darf nun angenommen werden, dass die Zahl der benötigten Morphtypes bei zunehmendem Wortschatz ebenfalls zunimmt. Allerdings schwächt sich diese Zunahme der Morphtypes anfangs sehr stark, danach weniger deutlich ab. Einstweilen muss offen bleiben, ob bei wesentlich größeren Stichproben oder gar Gesamterhebungen das gleiche Modell beibehalten werden kann oder vielleicht doch durch ein anderes ersetzt werden muss.

2.5. Das etymologische Spektrum

Mehrfach wurden für das Deutsche etymologische Spektra entwickelt (Best 2001b, 2005, Körner 2004). In Best (2005) konnte gezeigt werden, dass das etymologische Spektrum des Deutschen Altmanns Modell für beliebige Rangordnungen (Modell 3) folgt. Dies gilt auch für die hier erhobenen Daten. Es hat sich ja gezeigt, dass von den ausgewerteten 2710 Wörtern der Stichprobe 1127 (knapp 40 %) entlehnt sind oder wenigstens eine entlehnte Konstituente enthalten. Die Anpassung von Modell (3) an diese Daten ergibt (s. Tabelle 9):

Tabelle 9
Das etymologische Spektrum der entlehnten Wörter/ Konstituenten des Deutschen
(1127 von 2710 Wörtern)

x	Herkunft	beobachtet	berechnet	*x*	Herkunft	beobachtet	berechnet
1	Latein	510	510.00	17	Rotwelsch	2	3.73
2	Französisch	268	260.35	18	Afrikaans	1	3.01
3	Englisch	121	146.28	19	Arabisch	1	2.45
4	Griechisch	97	88.27	20	Gotisch	1	2.02
5	Niederdeutsch	76	56.30	21	Hebräisch	1	1.67
6	Italienisch	52	37.54	22	Hindi	1	1.40
7	Niederländisch	18	25.95	23	Jiddisch	1	1.18
8	Spanisch	12	18.49	24	Sanskrit	1	1.00
9	Russisch	8	13.52	25	Slowakisch	1	0.85
10	Slawisch	5	10.10	26	Sumer./ Skyt.	1	0.73
11	Japanisch	2	7.70	27	Tschechisch	1	0.63
12	Norwegisch	2	5.96	28	Türkisch	1	0.55
13	Persisch	2	4.69	29	Ungarisch	1	0.48
$a = 6.5333$ $b = 2.3562$ $D = 0.9949$							



Graphik 9: Das etymologische Spektrum der entlehnten Wörter/ Konstituenten

Anmerkungen: „Slawisch“ bedeutet, dass die genaue Herkunft eines Wortes/einer Konstituente nicht bestimmt werden kann. Dasselbe gilt für „Sumer./ Skyt.“ (= Sumerisch, Skytisch). Da manche Wörter Konstituenten enthalten, die aus verschiedenen Sprachen stammen, ergibt sich mit 1187 Herkunftsangaben eine höhere Summe als bei der Zahl der Wörter mit fremdsprachigem Hintergrund.

3. Zusammenfassung und Perspektive

Die Untersuchung vermittelt einige Perspektiven auf den Wortschatz des Deutschen. Bei allen erhobenen Daten kann gezeigt werden, dass sie einem der für entsprechende Fälle bereits entwickelten Sprachgesetze folgen. Da diese Sprachgesetze bisher überwiegend an Texten erprobt wurden, sind Überprüfungen am Beispiel von Lexikon-Daten weiterhin sinnvoll. Eine Stichprobe von nur 2710 Wörtern aus einem Lexikon, das nach Eigenauskunft „rund 140000 Wörter[...] und Wendungen" (*Duden. Deutsches Universalwörterbuch,* Titelseite) enthält, ist nicht gerade groß, sollte aber genügen, um die Hypothese, dass Sprache generell Gesetzen unterliegt, zu stützen.

Im Anschluss an die Untersuchung zur Zunahme der Morphtypes drängt sich die Frage auf, ob sich dieses Ergebnis auf Morpheme übertragen lässt. Dies war eine der Motivationen für die vorliegende Untersuchung. Um diesen Schluss von Morphtypes auf Morpheme vorzunehmen, muss man deren Verhältnis kennen; also z.B.: Wie viele Morpheme entsprechen einer bestimmten Zahl von Morphtypes? Die Frage lässt sich nicht einfach beantworten: Einander widersprechende Aspekte, Homonymie einerseits und Allomorphie andererseits, beeinflussen das Verhältnis. Es ist auch zu prüfen, ob es sich dabei um ein festes oder doch um ein variables Verhältnis handelt, verschieden womöglich bei unterschiedlichen Textsorten. Bevor solche Fragen geklärt sind, verbietet sich einstweilen ein Schluss von den Morphtypes auf Morpheme.

Anhang

Die Anpassung der Conway-Maxwell-Poisson-Verteilung an die Daten eines ungarischen Wörterbuchs haben zuerst Wimmer u.a. (1994: 102) dargestellt. In Best (2005a: 49) sollte der Test mit der inzwischen überarbeiteten Software wiederholt werden, wobei aber die Tabelle versehentlich gegen die eines Textes ausgetauscht wurde; die Graphik ist dagegen die richtige und nicht verwechselt worden. Als Richtigstellung hier die korrekte Tabelle.

Anpassung der 1-verschobenen Conway-Maxwell-Poisson-Verteilung an die Wortlängen (gemessen nach der Zahl der Silben pro Wort) in einem ungarischen Wörterbuch (Wimmer u.a. 1994: 102)

x	n_x	NP_x
1	1421	1494.64
2	12333	11893.35
3	20711	20991.89
4	15590	15353.99
5	5544	6011.01
6	1510	1449.19
7	289	235.11
8	60	27.29
$a = 7.9573$	$b = 2.1726$	$X^2 = 118.60$ $FG = 6$ $C = 0.0021$

Obwohl ein Freiheitsgrad mehr gegeben ist, ist das Testergebnis fast genau so gut wie mit $C = 0.0019$ bei Wimmer u.a. (1994).

Der Vollständigkeit halber sei darauf verwiesen, dass Köhler (2002: 58) offenbar beim gleichen ungarischen Wörterbuch bei etwas größeren Daten, die mir nicht zur Verfügung standen, mit $C = 0.0016$ ein noch etwas besseres Ergebnis erzielte.

Literatur

Altmann, Gabriel (1985). Sprachtheorie und mathematische Modelle. *Christian-Albrechts-Universität Kiel, SAIS [= Seminar für Allgemeine und Indogermanische Sprachwissenschaft] Arbeitsberichte. H. 8, 1-13.*

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

Altmann, Gabriel (2005). Diversification processes. In: Köhler, Reinhard, Altmann, Gabriel, & Piotrowski, Rajmund G. (Hrsg.), *Quantitative Linguistik - Quantitative Linguistics. Ein internationales Handbuch: 646-658.* Berlin/ N.Y.: de Gruyter.

Best, Karl-Heinz (2001a). Kommentierte Bibliographie zum Göttinger Projekt. In: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten: 284-310.* Göttingen: Peust & Gutschmidt.

Best, Karl-Heinz (2001b). Wo kommen die deutschen Fremdwörter her? *Göttinger Beiträge zur Sprachwissenschaft 5, 7-20.*

Best, Karl-Heinz (2001c). Zur Länge von Morphen in deutschen Texten. In: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten:* 1-14. Göttingen: Peust & Gutschmidt.

Best, Karl-Heinz (2001d). Silbenlängen in Meldungen der Tagespresse. In: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten: 15-32.* Göttingen: Peust & Gutschmidt.

Best, Karl-Heinz (²2003, ³2006). *Quantitative Linguistik: Eine Annäherung.* 3., stark überarbeitete und ergänzte Aufl. Göttingen: Peust & Gutschmidt

Best, Karl-Heinz (2004a). Wortschatzwachstum. In: *Wissenstransfer und gesellschaftliche Kommunikation. Festschrift für Sigurd Wichter zum 60. Geburtstag:.333-342.* Hrsg. v. Albert Busch & Oliver Stenschke. Frankfurt u.a.: Peter Lang.

Best, Karl-Heinz (2004b). Zum Wortschatzwachstum und -umfang in Texten. *Naukovyj Visnyk Černivec'koho Universytetu: Hermans'ka filolohija. Vypusk 206-207, 31-43.*

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

Best, Karl-Heinz (2005a). Wortlängen im Ungarischen (und anderswo). Ein Nachtrag. In: *Lihkkun lehkos! Beiträge zur Finnougristik aus Anlaß des sechszigsten Geburtstages von Hans-Hermann Bartens* (S. 43-56). Hrsg. von Cornelius Hasselblatt, Eino Koponen & Anna Widmer. Wiesbaden: Harrassowitz.

Best, Karl-Heinz (2006a). Wie viele Morphe enthalten Wörter in deutschen Pressetexten? *Glottometrics 13, 47-58*.

Best, Karl-Heinz (2006b). Wortlängen im Deutschen. *Göttinger Beiträge zur Sprachwissenschaft 23-49*.

Best, Karl-Heinz (2006c). Wortschatzwachstum und -umfang in Texten und Textkorpora. *Festschrift für Slavomír Ondrejović* (eingereicht).

Best, Karl-Heinz, & Zhu, Jinyang (2001). Wortlängenverteilungen in chinesischen Texten und Wörterbüchern. In: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten: 101-114*. Göttingen: Peust & Gutschmidt.

Duden. Das Herkunftswörterbuch. 3., völlig neu berabeitetete und erweiterte Auflage. Mannheim/ Leipzig/ Wien/ Zürich: Dudenverlag 2001 (Datum des Vorworts).

Duden. Die deutsche Rechtschreibung. 22., völlig neu berabeitetete und erweiterte Auflage. Mannheim/ Leipzig/ Wien/ Zürich: Dudenverlag 2000.

Duden. Deutsches Universalwörterbuch. 4., neu berabeitetete und erweiterte Auflage. Mannheim/ Leipzig/ Wien/ Zürich: Dudenverlag 2001.

Gerlach, Rainer (1982). Zur Überprüfung des Menzerath'schen Gesetzes in der Morphologie. In: Lehfeldt, Werner, & Strauss, Udo (Hrsg.), *Glottometrika 4,* 95-113. Bochum: Brockmeyer.

Kluge. Etymologisches Wörterbuch der deutschen Sprache. Bearbeitet von Elmar Seebold. 24., durchgesehene und erweiterte Auflage. Berlin/ New York: de Gruyter.

Köhler, Reinhard (2002). Power Law Models in Linguistics: Hungarian. *Glottometrics 5, 51-61.*

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

Menzerath, Paul (1954). *Die Architektonik des deutschen Wortschatzes*. Bonn: Dümmler.

Pfeifer, Wolfgang (Ltg.) (1995). *Etymologisches Wörterbuch des Deutschen.* 2. Auflage, ungekürzte durchgesehene Ausgabe. München: Deutscher Taschenbuch Verlag.

Rheinländer, Nicole (2001). Die Wortlängenhäufigkeit im Niederländischen. In: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten: 142-152*. Göttingen: Peust & Gutschmidt.

Wahrig, Gerhard [Hrsg.] (1978). *dtv-Wörterbuch der deutschen Sprache.* München: Deutscher Taschenbuch Verlag.

Wimmer, Gejza, Köhler, Reinhard, Grotjahn, Rüdiger, & Altmann, Gabriel (1994). Towards a Theory of Word Length Distribution. *Journal of Quantitative Linguistics 1, 98-106.*

Software

Altmann-fitter (1997). *Iterative Fitting of Probability Distributions.* Lüdenscheid: RAM-Verlag.

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

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

Glottometrics 14, 2006, 46-57

Wortlängen im Weißrussischen

Svitlana Kiyko[1]

Abstract. The aim of this paper is to show that word lengths in Belorussian texts follow the hyper-Poisson distribution.

Keywords: word length, Belorussian

Vorbemerkung

Die bisherigen Untersuchungen zu Wortlängenverteilungen in rund 50 Sprachen im Rahmen des Göttinger *Projekt Quantitative Linguistik* (Best 2001) haben im wesentlichen die theoretische Annahme bestätigt, dass die Verteilung von Wortlängenhäufigkeiten in der Tat nicht zufällig ist, sondern einen gesetzmäßigen Charakter trägt: es ließen sich immer Modelle finden, die an jeweilige Textgruppe angepasst werden konnten. Bisher haben sich die Hyperpoisson-Verteilung, die Poisson-Verteilung und die negative Binominalverteilung als die wichtigsten Verteilungen bei Wortlängen in deutschen Texten bewährt (Best 1997); über alle Sprachen hinweg scheint immer wieder die Hyperpoisson-Verteilung ein gutes Modell zu sein. Es müssen aber oft auch andere Verteilungen angepasst werden.

In einigen westslawischen Sprachen, z.B. im Polnischen und Tschechischen, konnte an die Wortlängen die erweiterte positive Binominalverteilung angepasst werden (vgl. Uhlířová 1995, 1996). Die Untersuchung von P. Girzig anhand 31 russischen Gedichte und sieben Kurzgeschichten hat gezeigt, dass in fast allen Fällen die Anpassung der erweiterten positiven Binominalverteilung an das untersuchte Material gelungen ist (Girzig 1997). Im Unterschied hierzu erwiesen sich in der Untersuchung von 52 Briefen A. S. Puškins die erweiterte positive Poisson-Verteilung, die Hyperpoisson-Verteilung, in einem Einzelfall die positive Poisson-Verteilung und die 1-verschobene Poisson-Verteilung als geeignete Modelle (Stitz 1994). Ungeachtet dessen sind als Ergebnis der bisherigen Untersuchungen zwei das Russische vermeintlich charakterisierende Verteilungsmodelle diskutiert worden (die sog. Hyperpoisson-Verteilung und die erweiterte positive Binomialverteilung) (Culp 1994, Best / Zinenko 1998, Best / Zinenko 2001).

Zu ostslawischen Sprachen wurden auch die Wortlängen in Briefen und Gedichten des ukrainischen Autors Ivan Franko (Best, Zinenko 1999) untersucht. Bei allen Texten konnte die Hyperpoisson-Verteilung angepasst werden.

Zum Weißrussischen (Belarussischen nach neuerer Bezeichnung) liegen bisher noch keine Daten vor. Das Ziel dieser Arbeit ist deshalb, an Bespielen weißrussischer Lyrik und Kurzgeschichten zu ermitteln, wie die Verteilung der Wortlängenhäufigkeit ausfällt. Ferner ist es festzustellen, ob und inwieweit die genetisch und typologisch eng miteinander verwandten Sprachen Russisch, Ukrainisch und Weißrussisch identische Wortlängenverteilungen aufweisen.

Wortlänge wird in dieser Arbeit wie auch in den meisten anderen des Göttinger Projekts nach der Zahl der Silben pro Wort bestimmt. Als „Wort“ gilt das orthographische Wort, also die ununterbochene Zeichenkette, die durch Leerstellen oder Interpunktionszeichen begrenzt

[1] Address correspondence to: jurkij@sacura.net

ist. Der Bindestrich wird nicht als worttrennendes Interpunktionszeichen gewertet. Die Zahl der Silben pro Wort wird danach bestimmt, wie viele Vokale oder Diphthonge das Wort aufweist.

Theoretischer Hintergrund

Wie bei allen Arbeiten im Göttinger Projekt bilden auch hier die Arbeiten von Wimmer u.a. (1994) sowie Wimmer & Altmann (1996) den theoretischen Hintergrund. In diesen Untersuchungen wurde in Auseinandersetzung mit früheren Vorschlägen ein Gesetz der Wortlängenverteilung entwickelt, das als eine der wichtigsten Varianten auch die Hyperpoisson-Verteilung enthält.

Zu einigen relevanten Besonderheiten des Weißrussischen

Im Weißrussischen gibt es wie im Russischen und Ukrainischen silbenlose Wörter, d.h. Einheiten, die nur aus Konsonanten bestehen. Dazu gehören Präpositionen weißruss. ***к*** *(k),* ***в*** *(v),* ***з*** *(z),* unsilbiges ***ў*** *(ŭ)* und nullsilbige Partikelvarianten ***б*** *(b),* ***жс*** *(ž).* Dadurch erklärt sich die Erweiterung der Tabellen auf die Klasse $x = 0$.

In dieser Arbeit wurden auch Daten ohne nullsilbige Wörter untersucht. Das Argument für die Version ohne Nullsilbige ist, dass diese - phonetisch gesehen - nur mit den umgebenden Wörtern zusammen als phonetisches Wort aufgefasst werden können. Die Nullsilbigen erscheinen dann mehr als Enklitika, auch wenn sie orthographisch selbständig sind[2].

Im Weißrussischen sind alle Diphthonge mit dem ***ŭ*** (kurzes *i*, vergleichbar mit dt. *j*) zusammengesetzt: ***оў*** (betont entspricht dem dt. *eu*), ***оў*** (unbetont entspricht dem dt. *ei* bzw. *ai*) und ***аў*** (dt. *ei* bzw. *ai*). Auch werden die Verbindungen **ao** und **ay** in Fremdwörtern gelegentlich zu einem Diphthong: *фрау* (*Frau* als Anrede einer dt. Staatsbürgerin), *хаос* (dt. *Chaos*).

Das Weißrussische gehört zu den Sprachen mit einem stark entwickelten Flexionssystem und zeigt im Unterschied zu südslawischen Sprachen wenig analytische Tendenzen. Die Zahl der mehrsilbigen Wörter ist hier durchschnittlich höher als in den anderen Sprachen. Außerdem zeigt das Weißrussische auf phonetisch-phonologischer Ebene eine starke Tendenz zur Akkomodation der Laute, was historisch das Vorhandensein von Füllvokalen zwischen Konsonantenverbindungen des Typs *бл, бр, вл, рн* u.ä. am Wortende bedingt hat, vgl. weißruss. *карабель* und russ. *корабль* „Schiff“, weißruss. *журавель* und russ. *журавль* „Kranich“ usw. Darum kommen Wörter mit 6, 7 oder auch 8 Silben im Weißrussischen nicht selten vor.

Kriterien der Textauswahl

Die Textauswahl richtete sich nach folgenden Kriterien:

1. Die Anzahl der Wörter sollte nicht allzu unterschiedlich sein, so dass Texte mehr oder weniger gleicher Länge untersucht wurden. Außerdem liegen alle Texteinheiten unter

[2] Antić, Kelih & Grzybek (2006) behandeln das Problem der nullsilbigen Wörter, ohne sich eindeutig für oder gegen sie zu entscheiden. In dieser Untersuchung hat sich erwiesen, dass an die Texte mit nullsilbigen Wörtern ebenso wie an die ohne diese immer die Hyperpoisson-Verteilung angepasst werden kann. Die Testergebnisse sind bei den Dateien ohne die Nullsilbigen etwas besser.

dem von Hammerl (1990:155) empfohlenen kritischen Wert von 2000 Wörtern, um die Homogenität zu wahren.

2. Es sollten sowohl Kurzgeschichten, als auch Gedichte und Prosagedichte bedeutender weißrussischer Autoren berücksichtigt werden. In 20 Fällen wurden Sachtexte, in denen es sich um Beschreibung der Sitten und Bräuche des weißrussischen Volkes handelt, untersucht. In einem Fall wurde eine Autobiographie nachgeprüft.
3. Die untersuchten Texte stammen aus dem gleichen Zeitabschnitt (1920 bis 1970), um den Einfluss des Faktors „Zeit" auf Ergebnisse zu reduzieren.

Genaue Angaben zu den Texten finden sich im Anhang.

Das Modell der Wortlängenverteilungen im Weißrussischen

An die Dateien der Texte mit nullsilbigen Wörtern wurde die Hyperpoisson-Verteilung, an die ohne die nullsilbigen wurde die Hyperpoisson-Verteilung in 1-verschobener Form angepasst. Die Formel der Verteilung in 1-verschobener Form lautet wie folgt:

$$P_x = \frac{a^{x-1}}{b^{(x-1)}\ {}_1F_1\,(1;b;a)}, \quad x = 1,2,...$$

Dabei sind *a* und *b* Parameter; ${}_1F_1(1; b; a)$ ist die konfluente hypergeometrische Funktion, d.h.

$${}_1F_1\,(1;b;a) = 1 + \frac{a}{b} + \frac{a(a+1)}{b(b+1)} + ...$$

und $b^{(x-1)} = b(b+1)(b+2)...(b+x-2)$. (Für die unverschobene Form der Verteilung muss man nur *x-1* durch *x* ersetzen.)

Die Ergebnisse der Anpassung der (1-verschobenen) Hyperpoisson-Verteilung an die weißrussischen Texte finden sich in den folgenden Tabellen. Die Anpassungen wurden mit einer geeigneten Software, dem Altmann-Fitter (1997) durchgeführt[3].

In den Tabellen sind folgende Angaben enthalten: *x* ist die Silbenzahl der Wörter, n_x die Zahl der Wörter mit der entsprechenden Silbenzahl in dem jeweiligen Text, NP_x die aufgrund der 1-verschobenen Hyperpoisson-Verteilung zu erwartende Anzahl der Wörter dieser Länge; *a* und *b* sind die Parameter der Hyperpoisson-Verteilung; X^2 ist das Chiquadrat, *P* die Überschreitungswahrscheinlichkeit für das berechnete Chiquadrat; *FG* gibt die Zahl der Freiheitsgrade an. In einigen Fällen mussten die Wortlängenklassen so zusammengefasst werden, dass mangels Freiheitsgraden kein *P* bestimmt werden kann; in diesen Fällen gilt der Diskrepanzkoeffizient $C = X^2/N$ als Prüfgröße für die Güte der Anpassung. Eine Anpassung mit $P \geq 0.05$ bzw. $C \leq 0.01$ gilt als zufriedenstellend. Diese Bedingungen sind in allen Fällen erfüllt.

Die Ergebnisse der Anpassung stellen sich wie folgt dar[4]:

[3] Für Berechnungen, Anregungen und Korrekturvorschläge bedanke ich mich herzlich bei Dr. K.-H. Best.

[4] In den Tabellen sind jeweils Daten mit Nullsilbigen (Dateinamen ohne Buchstabenzusatz) und ohne Nullsilbigen (Dateinamen mit Buchstabenzusatz) angegeben. Die senkrechten Striche in den Tabellen zeigen an, dass die betreffenden Längenklassen zusammengefasst wurden.

Sachtexte

	Z 91-92		Z 91-92a		Z 108		Z 108a	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
0	12	12.41|			9	6.31		
1	69	71.38|	69	71.27	21	24.03	21	19.17
2	82	88.55	82	84.70	30	32.35	30	32.98
3	78	61.56 |	78	61.02 |	28	26.27	28	28.08
4	30	29.73 |	30	31.54 |	18	15.36	18	15.89
5	8	15.37 |	8	18.47 |	7	7.00	7	6.73 |
6					2	3.78	2	3.15 |
a =	1.6707		1.8299		2.0731		1.6870	
b =	0.2906		1.5398		0.5449		0.9809	
X^2 =	0.397		0.382		3.061		1.134	
FG =	0		0		4		3	
P =	-		-		0.55		0.77	
C =	0.0014		0.0014					

	Z 123		Z 123a		Z 139		Z 139 a	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
0	9	0.35|			5	12.80|		
1	58	69.08|	58	58.57	66	54.09|	66	62.45
2	71	66.34	71	68.51	71	68.82	71	65.44
3	30	31.93	30	32.62	40	51.53	40	48.74
4	10	10.25	10	9.74	25	27.34|	25	28.16
5	3	3.05	3	2.56	21	11.23|	21	13.29|
6					3	5.19|	3	7.92|
a =	0.9650		0.8028		1.8204		2.5757	
b =	0.0049		0.6862		0.4310		2.4580	
X^2 =	0.536		0.394		3.534		2.973	
FG =	2		2		1		2	
P =	0.76		0.82		0.06		0.23	

	Z 265-266		Z 265-266a		Z 210		Z 210a	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
0	10	0.07|			12	2.56|		
1	62	71.48|	62	60.26	35	46.37|	35	34.90
2	86	92.52	86	93.54	58	51.31	58	53.58
3	72	59.92	72	61.45	23	29.28	23	30.51
4	21	25.87	21	25.60	14	11.26|	14	10.67
5	9	8.38	9	7.81	3	4.22|	3	3.34
6	0	2.17|	0	1.87|			
7	1	0.59|	1	0.47|			
a =	1.2960		1.1389		1.1786		0.9055	
b =	0.0013		0.7337		0.0651		0.5898	
X^2 =	4.971		4.234		2.450		3.287	
FG =	3		3		1		2	
P =	0.17		0.24		0.12		0.19	

	Z 266		Z 266a		Z 268		Z 268a	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
0	10	10.98			12	12.32		
1	61	67.03	61	58.35	39	40.05	39	38.65
2	82	77.30	82	87.88	50	43.93	50	45.03
3	57	49.22	57	52.94	23	28.98	23	29.80
4	21	21.64	21	19.93	14	13.67	14	13.77
5	5	9.83	5	6.90	8	7.05	8	6.75
a =	1.4220		1.0043		1.6556		1.5324	
b =	0.2331		0.6669		0.5094		1.3152	
X^2 =	4.521		1.394		2.255		2.347	
FG =	3		2		3		2	
P =	0.21		0.50		0.52		0.31	

	Z 272		Z 272a		Z 234		Z 234a	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
0	10	0.43|			14	12.92		
1	46	57.26|	46	46.43	62	70.52	62	60.11
2	62	61.03	62	62.59	88	83.88	88	93.00
3	35	32.65	35	33.48	65	55.99	65	60.40
4	11	11.66	11	11.17	24	25.97	24	24.82
5	3	3.97	3	3.33	7	9.23	7	7.46
6					2	3.49	2	2.21
a =	1.0745		0.8869		1.5210		1.1195	
b =	0.0081		0.6580		0.2786		0.7237	
X^2 =	0.495		0.112		4.083		0.749	
FG =	2		2		4		3	
P =	0.78		0.95		0.39		0.86	

	Z 142		Z 142a		Z 267		Z 267a	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
0	12	10.23			6	0.45|		
1	40	46.93	40	40.54	37	43.27|	37	36.71
2	57	46.10	57	50.46	49	48.95	49	49.79
3	19	25.36	19	26.97	29	27.85	29	28.59
4	10	9.68	10	9.18	11	10.58	11	10.41
5	4	3.70	4	2.85	2	3.02|	2	2.77|
6					1	0.88|	1	0.73|
a =	1.2501		0.9368		1.1448		0.9962	
b =	0.2726		0.7526		0.0120		0.7346	
X^2 =	5.539		3.769		0.270		0.120	
FG =	3		2		2		2	
P =	0.14		0.15		0.87		0.94	

	Z 169-170		Z 169-170a		Z 132-133		Z 132-133a	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
0	23	24.54			15	6.49|		
1	88	93.91	88	83.53	95	103.68|	95	93.20
2	115	110.87	115	121.54	133	128.90	133	130.48
3	84	77.39	84	83.53	77	83.37	77	85.11
4	41	38.34	41	37.57	48	36.44|	48	36.19|
5	13	14.72	13	12.56	7	16.12|	7	15.02|
6	2	6.23	2	4.27				
a =	1.7076		1.3025		1.3483		1.2215	
b =	0.4463		0.8952		0.0845		0.8725	
X^2 =	4.408		2.110		0.734		1.140	
FG =	4		3		1		1	
P =	0.35		0.55		0.39		0.29	

	Z 265		Z 265a		Z 267-268		Z 267-268a	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
0	12	12.08			11	10.95|		
1	39	47.11	39	39.31	57	56.77|	57	56.98
2	63	52.71	63	63.50	74	67.19	74	73.98
3	39	34.42	39	38.26	34	44.89	34	44.75|
4	15	15.87	15	14.17	32	20.89|	32	17.64|
5	4	7.81	4	4.76	3	10.31|	3	6.65|
a =	1.5688		0.9611		1.5339		1.1325	
b =	0.4023		0.5950		0.2960		0.8724	
X^2 =	6.277		0.184		3.802		0.000	
FG =	3		2		1		-	
P =	0.10		0.91		0.05		-	
C =							0.0000	

	Z 280		Z 280a		Z 233-234		Z 233-234a	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
0	15	19.49			18	0.65|		
1	80	69.04	80	78.81	99	118.25|	99	98.36
2	72	77.30	72	73.27	156	147.45	156	148.58
3	45	51.40	45	47.94	82	92.25	82	94.92
4	29	24.30	29	24.20	43	38.52	43	38.45
5	9	8.91	9	9.94	14	12.07	14	11.40
6	3	2.67|	3	3.44	1	3.81	1	3.28
7	1	0.89|	1	1.40				
a =	1.6372		2.2084		1.2556		1.1070	
b =	0.4623		2.3752		0.0070		0.7329	
X^2 =	4.904		1.426		4.552		4.849	
FG =	4		4		3		3	
P =	0.30		0.84		0.21		0.18	

	Z 263-264		Z 263-264a		Z 126-127		Z 126-127a	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
0	17	17.11			13	11.06|		
1	148	148.95	148	144.52	75	84.47|	75	84.85
2	163	161.79	163	166.32	131	123.18	131	129.90
3	92	93.71	92	95.37	76	99.28|	76	100.24|
4	44	37.01	44	36.41	71	55.29|	71	51.70|
5	7	11.08	7	10.42	32	23.52|	32	20.03|
6	2	3.35	2	2.96	5	8.09|	5	6.21|
7					5	3.11|	5	2.07|
a =	1.2410		1.1429		1.8022		1.5559	
b =	0.1426		0.9931		0.2360		1.0163	
X^2 =	3.404		3.266		1.091		1.579	
FG =	4		3		0		0	
P =	0.49		0.35		-		-	
C =					0.0027		0.0040	

In allen Fällen kann die Hyperpoisson-Verteilung (mit nullsilbigen Wörtern) oder die 1-verschobene Hyperpoisson-Verteilung (ohne Nullsilbige) angepasst werden. Allerdings müssen in fünf Fällen (beim Text Z 91-92 und Text Z 126-127 mit und ohne Nullsilbige sowie Text Z 267-268 ohne Nullsilbige) die Wortlängen so stark zusammengefasst werden, dass 0 Freiheitsgrade übrig bleiben, so dass in diesen Fällen nur C als Kriterium bleibt. Versuche mit anderen Verteilungen haben zu keinen besseren Ergebnissen geführt.

Die Anpassung an kurze Erzählungen:

	Bj 62-64		Bj 62-64a		Bj 75-76		Bj 75-76a	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
0	17	3.00|			5	5.67		
1	172	198.79|	172	177.24	83	94.14	83	84.30
2	252	221.72	252	239.33	139	121.23	139	131.90
3	118	124.70	118	125.68	77	81.20	77	85.85
4	42	46.89	42	40.97	37	36.75	37	35.28
5	7	13.24	7	9.68	13	12.56	13	10.59
6	3	2.99|	3	1.79|	2	4.45	2	3.09
7	1	0.67|	1	0.31|				
a =	1.1344		0.8593		1.3959		1.1146	
b =	0.0171		0.6364		0.0841		0.7324	
X^2 =	8.787		3.763		5.570		2.331	
FG =	3		3		4		3	
P =	0.03		0.29		0.23		0.51	

	Č 284-286		Č 284-286a		Č 327-329		Č 327-329a	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
0	25	25.24			17	0.31|		
1	207	209.02	207	204.99	176	189.86|	176	172.46
2	225	217.15	225	220.84	213	220.15	213	219.48
3	115	120.35	115	121.88	130	127.76	130	129.40
4	44	45.48	44	45.21	56	49.44	56	49.64
5	17	13.03	17	12.63	12	14.35	12	14.12
6	1	3.73	1	3.46	1	3.33|	1	3.19|
7					1	0.80|	1	0.71|
a =	1.1879		1.1315		1.1618		1.0984	
b =	0.1435		1.0503		0.0019		0.8631	
X^2 =	3.771		3.777		2.647		2.321	
FG =	4		3		3		3	
P =	0.44		0.29		0.45		0.51	

Autobiographie:

	Č 391		Č 391a	
x	n_x	NP_x	n_x	NP_x
0	11	13.04|		
1	72	64.00|	72	68.23
2	88	86.07	88	83.39
3	58	67.07	58	64.86
4	32	36.79	32	37.00
5	27	15.57	27	16.66|
6	1	5.36|	1	6.20|
7	1	2.10|	1	2.67|
a =	1.8529		2.1391	
b =	0.3777		1.7502	
X^2 =	4.783		2.335	
FG =	3		2	
P =	0.19		0.31	

In allen fünf Fällen ist eine zufrieden stellende Anpassung an die Hyperpoisson-Verteilung zu beobachten.

Die Anpassungen an Gedichte:

	Br 113		Br 113a		Br 115		Br 115a	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
0	2	9.66|			1	1.05		
1	66	55.59|	66	67.71	41	43.21	41	40.74
2	61	65.35	61	62.58	48	47.29	48	51.27
3	57	42.78|	57	39.70|	32	26.23	32	26.67
4	13	19.41|	13	19.17|	7	9.74	7	8.74
5	3	9.21|	3	10.84|	1	2.72|	1	2.09|
6					1	0.76|	1	0.49|
a =	1.4774		2.0231		1.1245		0.8868	
b =	0.2567		2.1889		0.0274		0.7047	
X^2 =	0.442		0.239		2.786		1.745	
FG =	0		0		3		2	
P =	C = 0.0022		C = 0.0012		0.43		0.42	

	Br 128		Br 128a		Br 133		Br 133a	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
0	5	0.54|			9	9.59|		
1	34	40.88|	34	37.35	62	66.07|	62	64.25
2	53	53.30	53	54.15	89	81.79	89	92.24
3	46	35.05	46	35.05	76	55.62|	76	56.50|
4	9	15.41	9	14.60	14	26.07|	14	21.99|
5	3	5.09	3	4.48	2	12.86|	2	8.02|
6	1	1.34|	1	1.09|				
7	0	0.29|	0	0.22|				
8	1	0.10|	1	0.06|				
a =	1.3269		1.1689		1.5091		1.0685	
b =	0.0176		0.8062		0.2191		0.7443	
X^2 =	7.137		6.694		0.991		0.543	
FG =	3		3		0		0	
P =	0.07		0.08		*C* = 0.0039		*C* = 0.0022	

	Br 154		Br 154a	
x	n_x	NP_x	n_x	NP_x
0	5	3.89		
1	17	19.45	17	16.77
2	27	25.44	27	26.63
3	18	19.13	18	20.24
4	13	10.10	13	10.11
5	3	4.11	3	3.76
6	0	1.36|	0	1.11|
7	1	0.52|	1	0.38|
a =	1.7713		1.4580	
b =	0.3545		0.9180	
X^2 =	2.310		1.380	
FG =	4		3	
P =	0.68		0.71	

Die Anpassung der Hyperpoisson-Verteilung an die Texte mit nullsilbigen Wörtern gelingt bei 3 von 5 Texten. In vier Fällen (bei den Texten Br 113 und Br 133 jeweils mit und ohne Nullsilbige) sind nur schwache Anpassungen festzustellen, so dass in diesen Fällen nur C als Kriterium bleibt**.**

Prosagedichte:

	Bj 5-6		Bj 5-6a		Bj 6		Bj 6a	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
0	2	2.20			1	5.97|		
1	41	45.18	41	39.97	33	28.62|	33	32.98
2	61	59.08	61	66.24	39	38.07	39	38.98
3	48	39.90	48	42.16	42	29.41|	42	29.06|
4	19	18.16	19	16.61	9	16.00|	9	15.82|
5	2	8.48	2	6.02	4	9.93|	4	10.15|
a =	1.3967		1.0334		1.8423		2.0196	
b =	0.0681		0.6236		0.3848		1.7089	
X^2 =	7.086		4.283		0.035		0.000	
FG =	3		2		0		0	
P =	0.07		0.12		-		-	
C =					0.0003		0.0000	

	Bj 8-9		Bj 8-9a		Bj 9-10		Bj 9-10a	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
0	9	10.66			4	4.96		
1	73	86.49	73	71.63	37	45.94	37	36.93
2	124	103.37	124	121.68	65	53.01	65	64.88
3	65	66.69	65	72.75	35	32.62	35	35.49
4	33	29.46	33	26.40	12	13.68	12	11.50
5	6	9.90	6	8.54	2	4.35	2	2.65|
6	0	3.43			1	1.44	1	0.55|
a =	1.4017		0.9227		1.3184		0.7946	
b =	0.1728		0.5432		0.1425		0.4523	
X^2 =	11.916		3.304		6.412		0.041	
FG =	4		2		4		2	
P =	0.018		0.19		0.17		0.98	

	Bj 26		Bj 26a	
x	n_x	NP_x	n_x	NP_x
0	7	7.74		
1	50	55.35	50	49.13
2	68	60.26	68	69.10
3	37	35.51	37	37.45
4	15	14.34	15	12.57
5	1	4.40	1	3.06|
6	1	1.40	1	0.69|
a =	1.2846		0.8818	
b =	0.1798		0.6270	
X^2 =	4.410		1.320	
FG =	4		2	
P =	0.35		0.52	

Die Anpassung der Hyperpoisson-Verteilung gelingt bei 4 von 5 Texten. In einem Fall (Text Bj 6 mit und ohne Nullsilbige) ist nur eine schwache Anpassung festzustellen.

Ergebnis

An alle 31 Texte mit und ohne nullsilbige Wörter kann die Hyperpoisson-Verteilung angepasst werden. Sie erweist sich damit wiederum als ein geeignetes Modell für die Wortlängenverteilungen in einer weiteren slawischen Sprache. Die Unterschiede zwischen den verschiedenen Textsorten sind nur gering. Diese Stichprobe ist aber nicht genug groß, um endgültige Schlussfolgerungen zu ziehen. Man kann nur vermuten, dass sich „neben den sprachspezifischen Bedingungen auch historische, autorenspezifische und textsortenspezifische Einflüsse bemerkbar machen“ (Girzig 1997). Dazu bedarf es jedoch der Untersuchung eines wesentlich größeren Textkorpus. Einstweilen darf festgestellt werden, dass die Hypothese, Wortlängen folgten einer begründbaren Verteilung, erneut bestätigt wurde.

Literatur

Antić, Gordana, Kelih, Emmerich, & Grzybek, Peter (2006). Zero-Syllable Words in Texts. In: Grzybek, Peter (ed.), *Contributions to the Science of Text and Language: Word length studies and related issues: 117-156.* Dordrecht: Springer

Best, Karl-Heinz (1997). Warum nur: Wortlänge? Nicht nur ein Vorwort. In: Best, K.-H. (Hrsg.), *Glottometrika 16, 5-12.* Trier: Wissenschaftlicher Verlag Trier.

Best, Karl-Heinz (2001). Kommentierte Bibliographie zum Göttinger Projekt. In: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten: 284-310.* Göttingen: Peust & Gutschmidt.

Best, Karl-Heinz, & Zinenko, Svetlana (1998). Wortlängenverteilungen in Briefen A. T. Twardowskis. *Göttinger Beitraege zur Sprachwissenschaft 1, 7-19.*

Best, Karl-Heinz, & Zinenko Svitlana (1999). Wortlängen in Gedichten des ukrainischen Autors Ivan Franko. In: Jozef Genzor, & Slavomír Ondrejovi≐ (ed.), *Pange Lingua. Zborník na Počesť Viktora Krupu* (S. 201-213). Bratislava: Veda, Vydavatel'stvo SAV.

Best, Karl-Heinz, & Zinenko, Svetlana (2001). Wortlängen in Gedichten A.T. Twardowskis. In: Uhlířová, Ludmila, Wimmer, Gejza, Gabriel Altmann & Reinhard Koehler (eds.), *Text as a Linguistic Paradigm: Levels, Constituents, Constructs. Festschrift in Honour of Luděk Hřebíček : 21-28.* Trier: Wissenschaftlicher Verlag Trier.

Culp, Christine (1994). *Untersuchung zur Häufigkeit von Wortlängen in ausgewählten Briefen Majakovskijs.* Seminararbeit, Göttingen.

Girzig, Patricia (1997). Untersuchung zur Häufigkeit von Wortlängen in russischen Texten. In: Best, Karl-Heinz (Hrsg.), *Glottometrika 16. The Distribution of Word and Sentence Length: 152-162.* Trier: Wissenschaftlicher Verlag Trier.

Hammerl, Rolf (1990). Untersuchungen zur Verteilung der Wortarten im Text. In: *Luděk Hřebíček* (Hrsg.), *Glottometrika 1, 142-156.* Bochum: Brockmeyer.

Stitz, Katrin (1994). *Untersuchungen zu den Wortlängen in deutschen und russischen Briefen des 19. Jahrhunderts.* Staatsexamensarbeit, Göttingen.

Uhlířová, Ludmila (1995). O jednom modelu rozložení délky slov. (= On a model of word-length distribution) *Slovo a slovesnost 56, 8-14.* (engl. summ.)

Uhlířová, Ludmila (1996). How long are words in Czech? In: Schmidt, Peter (Hrsg.), *Glottometrika 15, 134-146.* Trier: Wissenschaftlicher Verlag Trier.

Wimmer, Gejza, & Altmann, Gabriel (1996). The Theory of Word Length Distribution: Some Results and Generalizations. In: Schmidt, Peter (Hrsg.), *Glottometrika 15, 112-133*. Trier: Wissenschaftlicher Verlag Trier.

Wimmer, Gejza, Köhler, Reinhard, Grotjahn, Rüdiger, & Altmann, Gabriel (1994). Towards a Theory of Word Length Distribution. *Journal of Quantitative Linguistics 1, 98-106.*

Software

Altmann-Fitter (1997). *Iterative Fitting of Probability Distributions.* Lüdenscheid: RAM-Verlag.

Informationen zum *Göttinger Projekt:* Homepage: http://wwwuser.gwdg.de/~kbest/

Angaben zu den untersuchten Texten *(Transliteriert nach ISO)*:

Z Zemljarobčy kaljandar (abrady i zvyčai). Hg. v. V.K. Bandarčyk, K.P. Kabašnikav, A.S. Fjadosik, Minsk: Navuka i technika, 1990. - 403 S.
Insgesamt 20 Dateien. Das sind Beschreibungen der Sitten und Bräuche des weißrussischen Volkes. Die Texte sind künstlich zusammengestellt.

Br Broŭka, Pjatrus': Veršy. Paėmy. Peraklady (1926-1940). Zbor tvoraŭ u sjami tamach, T. 1. - Minsk: Mastackaja literatura, 1975. – 382 S.
Insgesamt 5 Dateien, Textsorte: reimgebundene Gedichte.

Bj Bjadulja, Zmitrok: Veršy ŭ proze. Liryčnyja imprėsii. Apavjadanni. Zbor tvoraŭ u pjaci tamach, T. 1. - Minsk: Mastazkaja literatura, 1986. - 352 S.
Insgesamt 7 Dateien. Dabei handelt es sich um kleine Erzählungen und Prosagedichte.

Č Čorny, Kuz'ma: Apavjadanni. Publicystyka. Zbor tvoraŭ u šasci tamach, T. 1. - Minsk: Mastazkaja literatura, 1988. - 432 S.
Insgesamt drei Dateien, Textsorte: zwei kurze Erzählungen, eine Autobiographie.

Glottometrics 14, 2007, 58-71

On the dynamics of word classes in text

Ioan-Iovitz Popescu, Bucharest[1]
Karl-Heinz Best, Göttingen
Gabriel Altmann, Lüdenscheid

Abstract: In this study, the distributions of certain parts of speech, especially auxiliaries, is investigated. Using the definition of the *h*-point, we define the thematic concentration of the text and introduce the concentration unit *tcu*.

Keywords: Word classes, auxiliaries, thematic concentration, distributions

1. Introduction

The dynamics of word classes in text can be evaluated in different ways. Firstly, according to their frequency distribution, which follows a special probability distribution (cf. Hammerl 1990; Best 1994, 1997, 2000, 2001; Schweers, Zhu 1991; Ziegler 1998, 2001); secondly, according to sequences of individual classes in text building a special distance pattern or a time series (cf. Pawlowski 2005; Ziegler, Best, Altmann 2001, 2002) and thirdly, according to the participation of a word class in a frequency class, an aspect which can be characteristic both of texts and languages. This last mode of evaluation will be scrutinized in this paper.

Consider the frequency distribution of words in a text, ranked according to decreasing frequency. For the sake of simplicity we shall differentiate only auxiliaries (function words, synsemantics) and autosemantics (nouns, verbs, adjectives, adverbs, numerals). This differentiation is not crisp and there are different classifications even in one language because language does not care for classes; they are no more than our conceptual creations, mostly based on Latin. If we look at words occurring exactly once in the text (g(1)), we state that the proportion of auxiliaries in this frequency class is very small. Taking class g(2) of words occurring twice, the proportion of auxiliaries increases. Continuing to the most frequent class we can observe that the proportion of auxiliaries grows monotonously and in some class it attains 1.00. That means that in long texts the curve of proportions of auxiliaries begins with zero (or a number very near zero) and in the class of the most frequent word – which is usually an auxiliary – it must be 1.00. Now, since the curve increases slowly at the beginning and approaches 1.00 long before it attains it, its form must change from convex to concave, i.e. at least for auxiliaries it must have an inflection point. This assumption can easily be translated in mathematics. Let y_x be the proportion of auxiliaries in the frequency class x. Then the rate of change of y_x is proportional to its own height and to its distance to 1, i.e.

$$\frac{dy}{dx} = ay(1-y)\,, \tag{1}$$

whose solution yields

[1] Address correspondence to: iovitzu@gmail.com

(2) $$y = \frac{1}{1 + be^{-ax}}$$

known as the logistic curve and from historical linguistics as Piotrowski law. Up to now the curve has been used mostly in historical linguistics where a class of changes has this form (there are also incomplete and reversible changes); its appearance in text structuring is a surprise. In the present article we shall analyze only the two fuzzy classes of auxiliaries and autosemantics. Many words vary in their class membership. Consider the German morpheme "auf". It can be a preposition ("auf dem Haus" – on the house), a fixed prefix ("Auftrag" – assignment), a prefix followed by another prefix ("aufgewendet" – spent), or a verbal particle with the status of an adverb ("stehe auf!" – stand up!), although not all grammarians would agree with this last. If one uses a PoS-tagger, the program represents the grammatical philosophy of the programmer/linguist. It is not our aim to emphasize a particular grammar but to show that under the given conditions "grammar words" have a certain dynamic.

2. Application

Consider first the distribution of word forms in A. v. Droste-Hülshoff´s "Der Geiergriff" as shown in the first two columns of Table 1. Here $g(x)$ is the number of words occurring exactly x-times; *Aux* is the number of auxiliaries in $g(x)$, $p(Aux)$ is the proportion of auxiliaries and $p(Aux)_{theor}$ is the computed value from the fitting logistic curve, Eq.2.

.

Table 1
Frequency distribution of word forms in Droste-Hülshoff

x	***g(x)***	***Aux***	***p(Aux)***	***p(Aux)$_{theor}$***
1	380	23	0.061	0.114
2	71	7	0.099	0.178
3	14	4	0.286	0.265
4	11	4	0.364	0.376
5	6	4	0.667	0.501
6	8	5	0.625	0.626
7	4	3	0.75	0.737
8	2	4	0.5	0.824
9	1	9	1	0.886
10	1	10	1	0.929
11	2	11	1	0.956
12	2	12	1	0.973
16	2	16	1	0.996
17	1	17	1	0.998
20	1	20	1	1
26	1	26	1	1
36	1	36	1	1
39	1	39	1	1

In Figure 1 a trend can be clearly seen. The formula obtained by optimization is

$p_{theor} = 1/(1 + 12.9149 \exp(-0.5125\, x))$

and the determination coefficient is $R^2 = 0.92$. Hence we can conclude that at least in this text the auxiliaries have the frequency structure (2). The inflection point of (2) is $x = (\ln b)/a$, i.e. in the above case $x = (\ln 12.9149)/ 0.5125 = 4.99 \approx 5$.

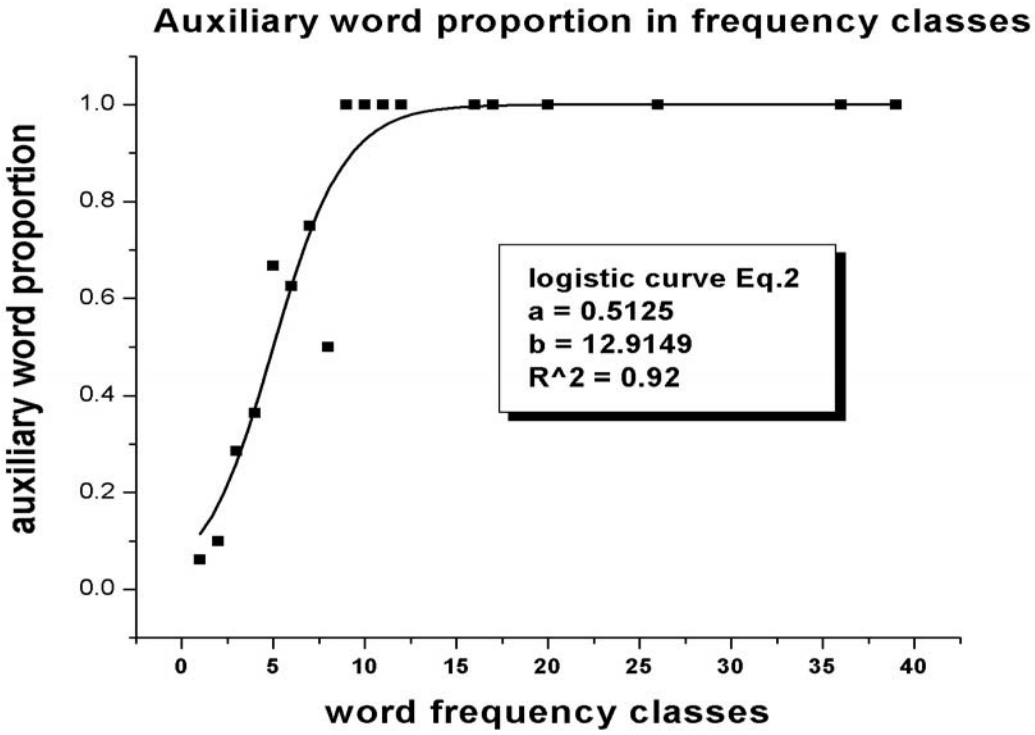


Figure 1. The frequency structure of auxiliaries in A. Droste-Hülshoff´s text

In the same way, we analysed some other short German texts and obtained the results in Table 2. In some cases, when the subject of the text is named very frequently (e.g. "Becher" in Schiller's "Taucher"), the given noun gets under the auxiliaries and if there are no other words of the same frequency, we obtain an observed $y = 0$ in that point. Such points are left out because they were "reserved" for non-auxiliaries. This happens in rather shorter texts. The curve has in all cases the same course and displays an inflection point (*IP*) which can be considered a characteristic of the text. Since the inflection point is a function of the two parameters ($x = (\ln b)/a$), it is perhaps sufficient to use them to characterize the text. In the texts of Table 2, the *IP* points are relatively stable.

Table 2
Frequency structure of auxiliaries in some short German texts

Author	Text	*b*	*a*	R^2	*IP*
Goethe	Elegie 19	5.0590	0.5776	0.63	2.81
Anonym	Mäuschen	5.0833	0.3261	0.85	4.99
Goethe	Der Gott und die Bajadere	10.0648	0.7103	0.94	3.25
Droste-Hülshoff	Der Geiergriff	12.9121	0.5125	0.92	4.99
Anonym	Zaunbär	18.4856	0.5314	0.86	5.49
Möricke	Peregrina	19.2843	0.8214	0.95	3.60
Schiller	Der Taucher	34.2635	0.8656	0.95	4.08

The same procedure can be performed not only for formal classes, whether isolated or pooled, but also for semantic classes like "processual expressions" containing verbs expressing some activity (no verbs like "sleep", "be", "have" etc.) and also nouns derived from activity verbs or even adjectives (e.g. "donnernder Huf" - thundering hoof). In English a number of conversions belong to this class. According to our hypothesis, the frequency distribution of words (in this case, rather, word forms) is semantically structured; and even individual frequency classes have their particular semantic spectrum, changing from class to class.

Let us now consider an English text, namely Rutherford's Nobel lecture. The PoS-tagging of the text has been performed with the aid of the CLAWS tagger (http://www.comp.lancs.ac.uk/ucrel/trial.html) and the frequency count with the aid of a very reliable counter that can be found at http://www.georgetown.edu/faculty/ballc/webtools/web_freqs.html. Some small uncertainties of the tagger have been corrected by hand. The following "parts of speech" have been pooled in one class: articles (ATO), adverbs (AVO, AVP, AVQ), conjunctions (CJC, CJS, CJT), determiners (DPS, DTO, DTQ), existential "there" (EXO), pronouns (PNI, PNP, PNQ, PNX, POS), prepositions (PRF, PRP), the infinitive marker "to" (TOO), the negation (XXO). In the same way as above, the participation of this class in frequency classes has been ascertained and we obtained the result given in Table 3. The classes x = 1,..,20 have been left separated, classes 21-100 have been pooled and the mean x = 60.5 has been taken; for the pooled classes 101-464 the mean 282.5 has been taken into account. In Figure 2 the trend is displayed graphically.

Table 3
The observed and computed proportion of "auxiliaries" in Rutherford's Nobel lecture

x	$p(Aux)$	$p(Aux)_{theor}$
1	0.0373	0.0693
2	0.0513	0.0794
3	0.0875	0.0908
4	0.0227	0.1037
5	0.0976	0.1182
6	0.1250	0.1344
7	0.2273	0.1524
8	0.2500	0.1944
9	0.3333	0.2185
10	0.1667	0.2446
11	0.1667	0.2728
13	0.4000	0.3030
14	0.2500	0.3349
16	0.5000	0.4032
20	0.5000	0.5489
60.5	0.9286	0.9979
282.5	1.0000	1.0000
$y = 1/(1 + 15.5663 \exp(-0.1471x))$, $R^2 = 0.94$, $IP = 18.66$		

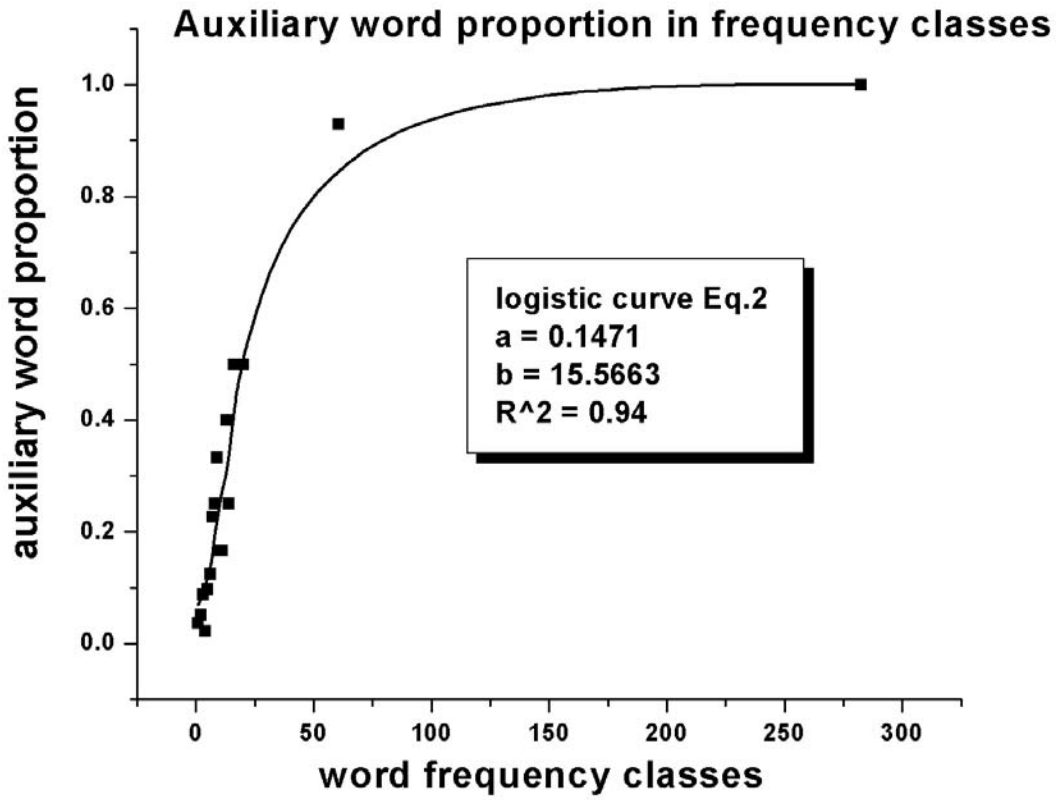


Figure 2. The dynamics of auxiliaries in an English text (Rutherford´s Nobel lecture)

The fit is very good, evidently we are dealing with a tendency that holds at least in some European languages. Of course, a number of further tests are necessary to corroborate this. Comparing the parameters *a* in German and English, we can preliminarily say that the greater *a*, the smaller is the analyticity (= the greater is syntheticity) in language; but this is a very preliminary statement requiring a lot of testing. We wish only to draw attention to this possibility.

3. The curves of word classes

One natural question which arises now is whether other word classes abide by this (or the opposite) trend; but we entrust this problem to the interested reader. Here we shall consider another question. As is generally known, the rank-frequency distributions of word classes follow the Zipf-law which – in the language of distributions – is represented by the (usually right-truncated) zeta distribution or some of its manifold generalisations and modifications (cf. Zörnig, Altmann 1995; Wimmer, Altmann 1999; Baayen 2001). Though the majority of these can be subsumed under a very general model (cf. Wimmer, Altmann 2005), we can consider the rank-frequency distribution of word forms from another point of view, namely as a mixture of word classes, each of which has its own distribution. The question is, what kind of distribution word classes follow, if any. The problem cannot be solved without first solving the linguistic problem of word classes. The PoS-taggers analysing word forms go to different depths and a false attribution can distort the form of the distribution. But even a "hand made" analysis involves decisions based on two hundred (or two thousand) years of discussion. Thus

we rely on taggers and perform two experiments. First, we consider the frequency spectrum of word classes (for nouns, verbs, adjectives and adverbs): a quite usual distribution in which some values of the variable are not realized. For the Rutherford text, we obtain the result in Table 4.

Table 4
Frequency spectra of word classes (Rutherford's Nobel lecture)

Nouns		Verbs		Adjectives		Adverbs	
x	g(x)	x	g(x)	x	g(x)	x	g(x)
1	205	1	114	1	105	1	38
2	85	2	40	2	30	2	28
3	31	3	28	3	13	3	11
4	21	4	9	4	7	4	4
5	16	5	7	5	11	5	3
6	13	6	3	6	2	6	1
7	7	7	4	7	2	7	3
8	8	8	4	8	4	8	2
9	6	9	1	9	1	11	1
10	1	10	3	12	1	12	1
11	4	12	3	16	1		
12	4	13	1	20	1		
13	2	14	1	22	1		
14	2	15	3				
15	1	17	1				
19	2	40	1				
28	1	42	1				
38	1	51	1				
44	1	85	1				
48	1						
51	1						
60	1						

It can easily be shown that any of the current distributions (Waring, Yule, Zipf-Mandelbrot, Good, zeta etc.) can be adequately fitted to these data. For the sake of illustration, we present in Table 5 the fit of the right truncated zeta distributions (Zipf) to the distribution of adjectives. The software inserts automatically zero for $g(x)$ if x is not realized, because the probability must be computed. It automatically pools frequency classes beginning from below to get the theoretical frequencies > 1. The chi-square test and the parameters of the zeta distribution are in the last row of the table. R is the truncation parameter at the right side of the distribution. In Figure 3 and 4 the fit is shown graphically. Similar results can be obtained also for nouns, verbs, adverbs and auxiliaries.

Table 5
Fitting the zeta distribution to adjectives in Rutherford´s Nobel lecture

x	g(x)	zeta
1	105	106,07
2	30	28,54
3	13	13,24
4	7	7,68
5	11	5,03
6	2	3,56
7	2	2,66
8	4	2,07
9	1	1,65
10	0	1,35
11	0	1,13
12	1	0,96
13	0	0,82
14	0	0,72
15	0	0,63
16	1	0,56
17	0	0,50
18	0	0,44
19	0	0,40
20	1	0,36
21	0	0,33
22	1	0,30
a = 1.89419; R = 22; DF =12; X^2 = 14,74; P = 0,26		

The right- truncated zeta distribution is defined as

$$(3) \qquad P_x = \frac{x^{-a}}{T}, \quad x = 1, 2, ..., R$$

where $T = \sum_{j=1}^{R} j^{-a}$.

This result is not surprising, but its consequences are rather strange. First, the resulting overall distribution looks (mathematically) like a superposition of identical distributions with differing parameters and different weights. But this is not true. The generating mechanism is the valency of the word classes. The appearance of a class in text leads to the appearance of another class which lies in the domain of its valency. Though the realization of a special class from this domain is controlled probabilistically (e.g. a noun admits an adjective but does not need to have it in any case, but on the contrary, an adjective requires a noun; a verb presupposes a noun or a pronoun but the pronoun can be an inflection, e.g. Latin "vocamus", etc.; this is realized differently in different languages), the associated classes follow the distribution of the main class.

Secondly, considering the frequencies of individual word classes as given in Table 4 and fitting a probability distribution to the data we have two problems: (i) the variable does not contain all the values; there are a number of *x*-values whose *f*(*x*) is zero. But we must compute the probability of these *x*, too, and use them for testing. (ii) Some expected frequencies are less

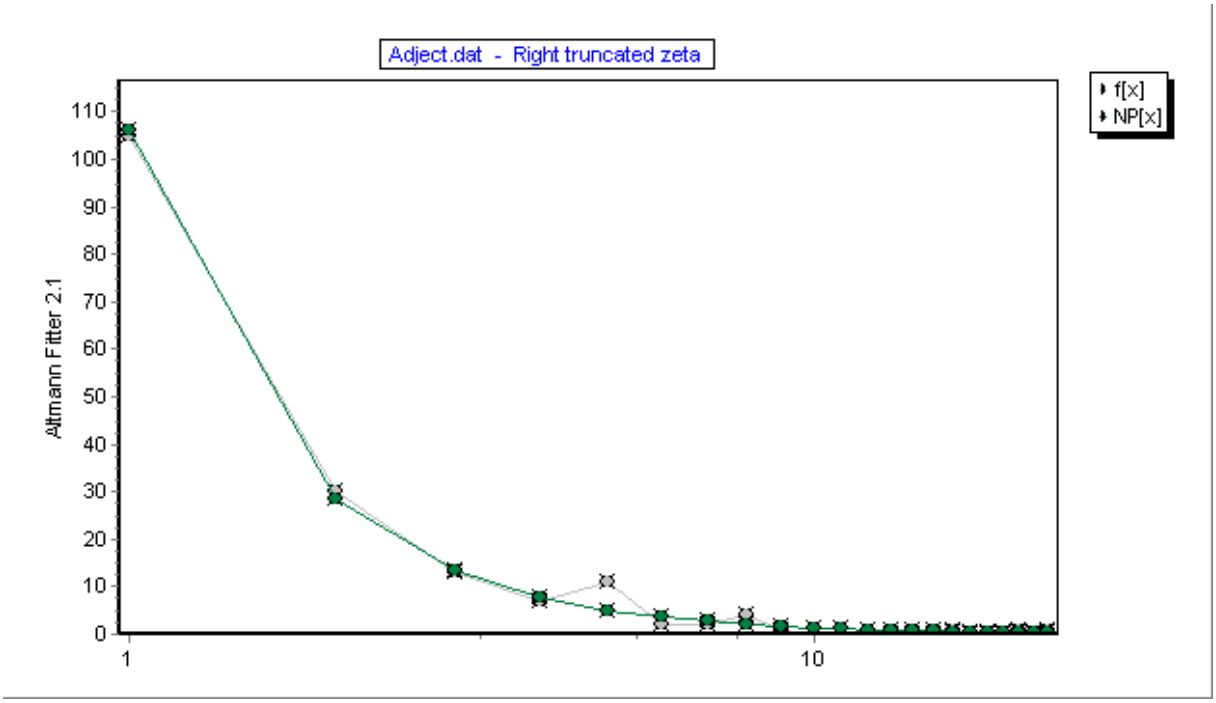


Figure 3. Fitting the right truncated zeta distribution to adjectives in Rutherford (half logarithmic presentation)

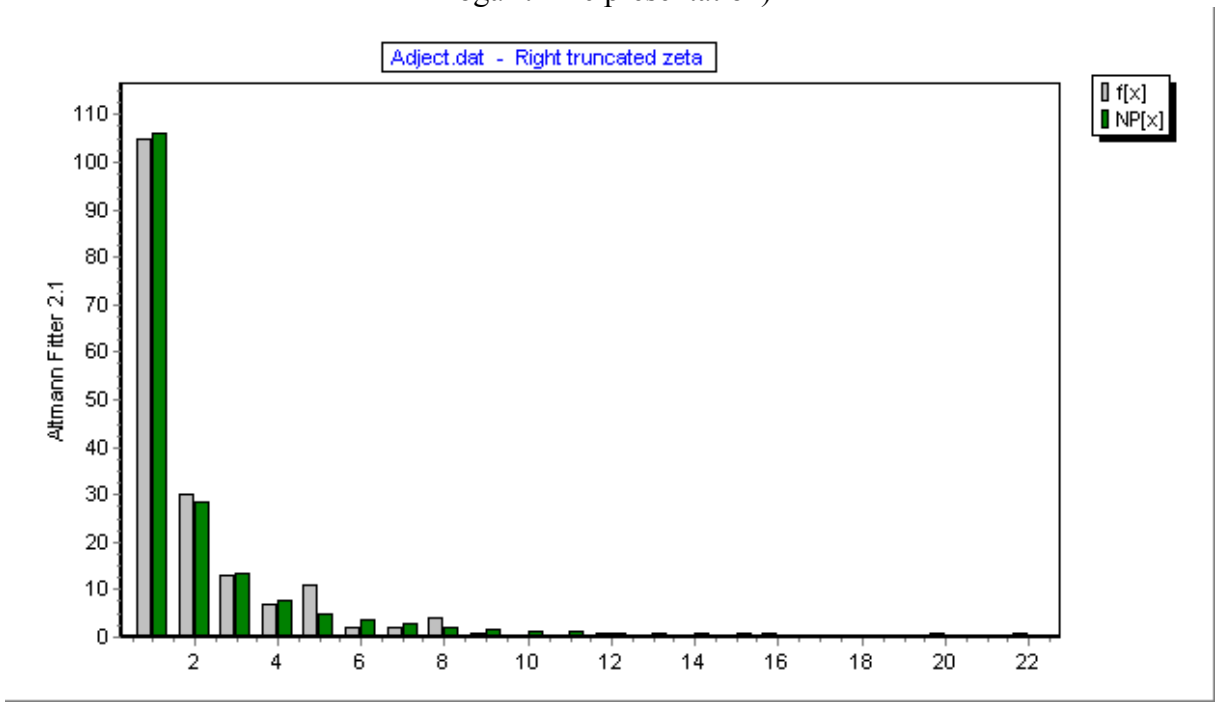


Figure 4. Fitting the right truncated zeta distribution to adjectives in Rutherford (bar presentation)

than 1, a circumstance which is not possible in reality unless it is zero, i.e. case (i). A part of the quantities in these classes has been, as a matter of fact, "shifted" to non-existing classes. In order to avoid these problems, we can set up a model of word-class distributions using an adequate curve. Using a curve or a sequence instead of a probability distribution, we can avoid the shortcomings above, but ignore the normalization. Real phenomena can be modelled by curves or sequences, probability distributions, discrete or continuous. There is no more truth in any of them; they are merely steps in the approximation of reality. Recalling

that Zipf used the zeta-function (not the zeta-distribution), we apply the same series, taking into account only the realized frequencies. Since the sequence alone converges to zero, we add a constant 1, i.e. we apply

(4) $y = bx^{-a} + 1$.

where y are the frequencies. From the data in Table 4 we obtain the results in Table 6.

Table 6
Fitting the modified zeta function to word-class distributions

Nouns			Verbs			Adjectives			Adverbs		
x	y	y_{theor}	x	y	y_{theor}	x	y	y_{theor}	x	y	y_{theor}
1	205	207.63	1	114	114.86	1	105	105.13	1	38	40.62
2	85	70.39	2	40	38.01	2	30	29.01	2	28	17.36
3	31	37.65	3	28	20.18	3	13	14.00	3	11	10.75
4	21	24.30	4	9	13.03	4	7	8.54	4	4	7.75
5	16	17.40	5	7	9.38	5	11	5.94	5	3	6.08
6	13	13.31	6	3	7.23	6	2	4.50	6	1	5.07
7	7	10.66	7	4	5.85	7	2	3.61	7	3	4.31
8	8	8.82	8	4	4.91	8	4	3.03	8	2	3.79
9	6	7.50	9	1	4.23	9	1	2.62	11	1	2.86
10	1	6.51	10	3	3.72	12	1	1.94	12	1	2.66
11	4	5.74	12	3	3.03	16	1	1.55			
12	4	5.13	13	1	2.78	20	1	1.36			
13	2	4.64	14	1	2.58	22	1	1.30			
14	2	4.24	15	3	2.41						
15	1	3.91	17	1	2.15						
19	2	3.00	40	1	1.29						
28	1	2.09	42	1	1.27						
38	1	1.67	51	1	1.19						
44	1	1.53	85	1	1.08						
48	1	1.47									
51	1	1.42									
60	1	1.33									

Nouns : $y = 206.6252x^{-1.5743} + 1$, $R^2 = 0.99$
Verbs : $y = 113.6624x^{-0.1563} + 1$, $R^2 = 0.99$
Adjectives : $y = 104.1267x^{-1.8942} + 1$, $R^2 = 0.996$
Adverbs : $y = 39.6188x^{-1.2761} + 1$, $R^2 = 0.89$

Note that for adjectives, both the zeta distribution and the above power function have the same parameter $a = 1.8942$. The above modified zeta-function for adjectives is shown in Figure 5.

We can conclude that not only the overall spectrum is zeta-like (Zipf-like) but also the individual main classes of words. Even mixing two classes (of the same text), e.g. verbs and adjectives, yields a zeta-like result. This picture can change if we take very long or very short texts. Since very long texts are necessarily mixtures whose parts abide by the same regime, however with different parameters, the mixture can cause model modifications. On the other hand, in very short texts the classes cannot take shape sufficiently. Nevertheless, in general,

both the overall spectrum and that of individual word classes abide by the power law or its modifications.

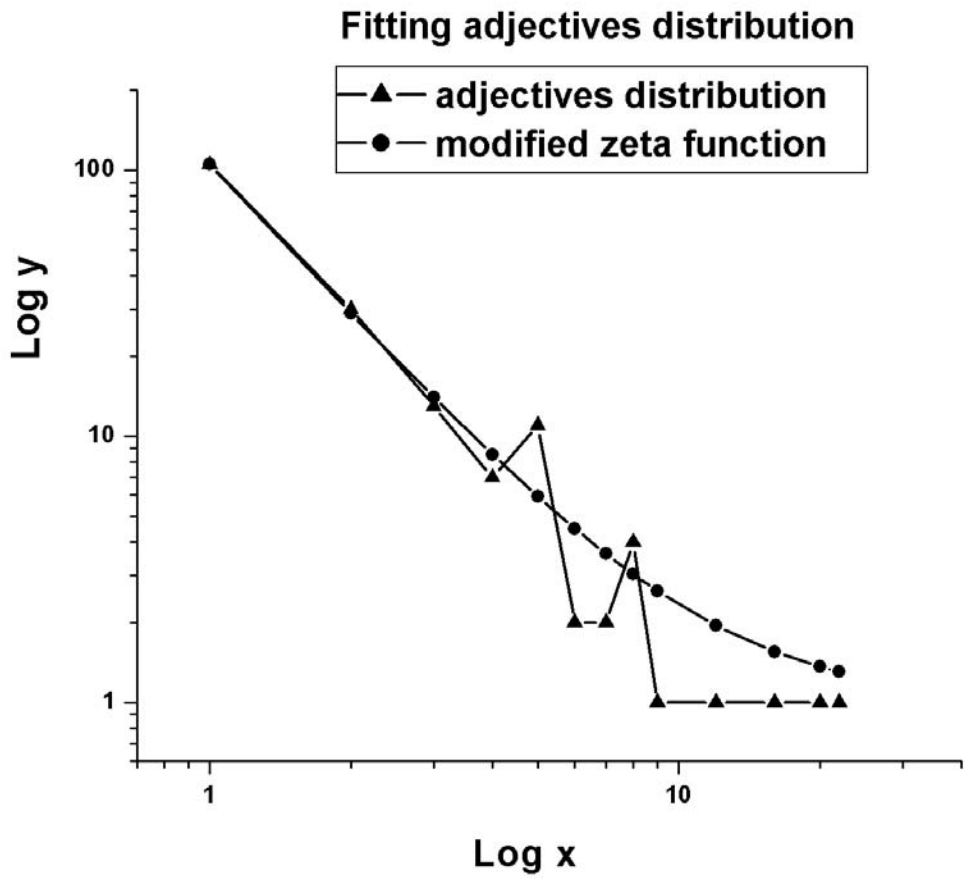


Figure 5. Fitting the modified zeta-function (4) to the distribution of adjectives

If we believe in the existence of laws in language, then from the above facts we can draw the conclusion that if an established word class in any language does not follow this law, it has been established incorrectly. This concerns even the well analysed languages. Ambiguities are mostly resolved syntactically. But even in that case many uncertainties remain which can be resolved using the frequency distribution.

Further research possibilities in this domain are: (i) Has parameter *a* of the power function (distribution or curve) characteristic values for individual word classes? (ii) Does the above regularity hold only for the main classes or for all? Which classes do not abide by this regularity? (iii) If the frequency spectrum deviates from the power law – and this has been shown in many cases – why does this happen, at which places, to what extent and in what direction? That is, we can begin to search for the boundary conditions of the rise of a word-frequency distribution.

4. Thematic concentration

The thematic concentration of a text can be evaluated intuitively, i.e. by reading it and expressing an expert opinion. Another possibility is to let a program search for key concepts in the text and to then try to interpret the results in terms of thematic concentration. Some tasks

of content analysis concern this aim. A third method is denotative analysis, also taking into account references and setting up a text-core (cf. Ziegler, Altmann 2002). A mechanization of this task is not yet fully developed. However, easy methods, which could not only help to find the core of the text but also measure its conspicuousness, are always highly desirable. In the following such a method will be proposed.

In a previous work (Popescu 2006) it has been shown that in the (monotonously decreasing) rank-frequency distribution of words or word forms there is a point in which $r = f(r)$, i.e. the rank of a word is equal to its frequency. This point is called h-point and it is the nearest point to the origin [0,0]. This point separates in a fuzzy way the great class of auxiliaries from autosemantic words. Of course, auxiliaries are found throughout the lower part (tail) of the distribution (below the *h*-point) but the point itself can be used for different purposes (cf. Popescu, Altmann 2006, 2007). Since we consider the pre-*h* domain as that of auxiliaries, autosemantics occurring in this domain must be extremely emphasized in the text, i.e. they must be part of the text theme, either its subject or its properties or activities. Sometimes they are proper names. This holds for any language, even extremely synthetic ones.

Consider the first thirty most frequent words in Rutherford´s Nobel lecture in Table 7. We

Table 7
The first thirty most frequent words in Rutherford´s Nobel lecture

r	f(r)	Word	r	f(r)	Word	r	f(r)	Word
1	466	the	11	60	it	21	39	atom
2	382	of	12	58	from	22	36	with
3	140	a	13	56	particles	23	30	for
4	121	that	14	53	helium	24	29	rays
5	116	and	15	51	is	25	28	an
6	113	in	16	45	particle	26	27	as
7	87	to	17	42	be	27	25	its
8	85	was	18	42	this	28	24	on
9	64	radium	19	41	at	29	24	or
10	63	by	20	40	were	30	23	radioactive

can see that the h-point is located at $r = 26$, or more exactly, at $h = 26.5$. There are only a few autosemantics with a rank smaller than *h*, namely *radium, particles, helium, particle, atom, rays*. From these words one can easily reconstruct what Rutherford spoke about. They build the primary theme of his lecture. In order to characterize quantitatively the concentration on these thematic words, we propose the following index of thematic concentration:

$$TC = \frac{2}{h}\sum_{r'=1}^{T}\frac{(h-r')f(r')}{(h-1)f(1)}. \qquad (5)$$

Here *h* is the *h*-point, $f(1)$ is the frequency of the most frequent word, r' are those ranks which point to thematic autosemantics, $f(r')$ is the frequency at these ranks and *T* is the number of those ranks. If there are no autosemantics in this domain, then $f(r')$ is in each case zero, *T* is zero, hence $TC = 0$. In the other extreme case – when all these words are autosemantics with frequency theoretically equal to $f(1)$ – we obtain $T = h$ and adding

$$\sum_{r=1}^{h}(h-r)=h(h)-\sum_{r=1}^{h}r = h^2 - \frac{h(h+1)}{2} = \frac{h(h-1)}{2}.$$

Dividing the sum by this constant we obtain formula (5) which is normalized and cannot surpass 1. Hence $TC \in <0, 1>$. Let us illustrate the computation using the above data from Rutherford. There are 6 autosemantics in the pre-h-domain, i.e. $T = 6$; $f(1) = 466$ and $h = 26$, hence

$$\frac{2}{h(h-1)f(1)} = \frac{2}{26(25)466} = 0.0000066028 = K$$

by which the sum will be multiplied. For the individual words we obtain:

word	rank	frequency	$(h - r')f(r')K$
radium	9	64	0.00718
particles	13	56	0.00481
helium	14	53	0.00420
particle	16	45	0.00297
atom	21	39	0.00129
rays	24	29	0.00038.

The sum of the last column yields $TC = 0.02083$. As can be seen, the realized non-zero values are of the order of ten of thousandth. Thus, defining the value of 1 *tcu* (thematic concentration unit) as a thematic concentration having $TC = 1/1000$, the thematic concentration of Rutherford's Nobel lecture has a value of 20.83 *tcu*. Needless to say, a lemmatised text would yield different values.

For comparative purposes we evaluated some other Nobel lectures and obtained the results as given in Table 8. The thematic words show automatically the field of the author.

Table 8
The TC-values of some Nobel lectures
(r_{min} is the smallest thematic rank in the pre-*h* domain)

Author	***h***	***f*(1)**	**pre-*h* words**	***TC***
Banting, F.G.	32	622	insulin, sugar, diet, patient, blood, carbohydrate	0.01692
Bellow, S.	26	297	art	0.00027
Buchanan,J.M.Jr.	23	366	political, politics, individual, individuals, rules	0.00985
Buck, P.	39	617	people, novel, Chinese, novels, China	0.01034
Feynman,R.P.	41	780	time, theory, quantum, electrodynamics	0.02222
Lewis, S.	25	237	American, America	0.00494
Mcleod, J.	24	460	insulin, sugar, pancreas, blood, symptoms	0.00604
Marshall,G.C.	19	229	peace	0.00506
Pauling, L.	28	546	nuclear, world, weapons, war, great, nations, human	0.01935
Russell, B.	29	342	power	0.00022
Rutherford, E.	26	466	radium, particles, helium, particle, atom, rays	0.02083

Multiplying *TC* by 1000 we obtain the following order (Table 9):

Table 9

Author	**Field**	***tcu***
Rutherford, E.	Chem	20.83
Pauling, L.	Peace	19.35
Banting, F.G.	Med	16.92
Buck, P.	Lit	10.34
Buchanan, J.M. Jr.	Econ	9.85
Mcleod, J.	Med	6.04
Marshall, G.C.	Peace	5.06
Lewis, S.	Lit	4.94
Feynman, R.P.	Phys	2.22
Bellow, S.	Lit	0.27
Russell, B.	Lit	0.22

In general, the representatives of "hard sciences" write texts with greatest thematic concentration. Linus Pauling was also an outstanding chemist, P. Buck and R. Feynman are "exceptions". But if we examine different genres, the picture will possibly be different. In any case art and social sciences will follow "hard science" and within art, poetry will have the smallest thematic concentration. Further research must be made on a very broad basis.

The above results lead automatically to the consideration of following possibilities: (i) Setting up a *post-h domain* consisting of ranks ($h,\ 2h - r_{min}$), r_{min} being the smallest thematic rank in the pre-*h* domain, e.g. with Rutherford $h = 26$, $r_{min} = 9$, from which $h - r_{min} = 17$, i.e. ($h,\ 2h - r_{min}$) = (26, 43) This domain can be called secondary thematic concentration domain. (ii) Using the pace of $h - r_{min}$ (which is 17 for Rutherford), one could partition the whole rank-frequency distribution in subsequent domains and study their contribution to thematic concentration. Preliminary computations have, however, shown that even the secondary domain contributes very little to the thematic concentration. Nevertheless, the series of domains could display some regularity concerning not only the theme of the text but also the dynamics of word classes. Here we shall not follow this possible direction of research.

Conclusions

The present article shows the first steps in scrutinizing the dynamics of word classes in text. Some trends are very regular and need a broad investigation in different languages. The whole rank-frequency domain can be partitioned in intervals based on the *h*-point and the behaviour of word classes can be studied within these intervals. The classes can be defined formally (e.g. length classes), grammatically (e.g. parts of speech) or semantically (e.g. words expressing activity). The dynamics can serve for the characterization of individual texts, of genres, epochs or even languages. The research can be performed with very simple mathematical apparatus.

References

Baayen, R.H. (2001). *Word frequency distributions*. Dordrecht: Kluwer**.**

Best, K.-H. (1994). Word class frequencies 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. *Glottometrika 16, 276-285.*

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

Best, K.-H. (2001). Zur Gesetzmäßigkeit der Wortartenverteilungen in deutsche Pressetexten. *Glottometrics 1, 1-26.*

Hammerl, R. (1990). Untersuchungen zur Verteilung der Wortarten im Text. *Glottometrika 11, 142-156.*

Pawlowski, A. (2005). Modelling the sequential structures in text. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 738-750.* Berlin: de Gruyter.

Popescu, I.-I. (2006). The ranking by the weight of highly frequent words. In: Grzybek, P., Köhler, R. (eds.), *Exact methods in the study of language and texts: 553-562*. Berlin: Mouton-de Gruyter

Popescu, I.-I., Altmann, G. (2006). Some aspects of word frequencies. *Glottometrics 13, 23-46.*

Popescu, I.-I., Altmann, G. (2007). Some geometric properties of word frequency distributions. *Göttinger Beiträge zur Sprachwissenschaft (in print).*

Schweers, A., Zhu, J. (1991). Wortartenklassifizierung im Lateinischen, Deutschen und Chinesischen. In: Rothe, U. (ed.), *Diversification processes in language: grammar: 157-165.* Hagen: Rottmann.

Wimmer, G., Altmann, G. (1999). *Thesaurus of univariate discrete probability distributions.* Essen: Stamm

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: 191-208.* Berlin: de Gruyter.

Ziegler, A. (1998). Word class frequencies in Brazilian-Portuguese press texts. *Journal of Quantitative Linguistics 5, 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. Festschrift in Honour of Luděk Hřebíček: 295-312.* Trier: WVT.

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

Ziegler, A., Best, K.H., Altmann, G. (2001). A contribution to text spectra. *Glottometrics 1, 97-108.*

Ziegler, A., Best, K.H., Altmann, G. (2002). Nominalstil. *ETC – Empirische Text- und Kulturforschung 2, 72-85.*

Zörnig, P., Altmann, G. (1995). Unified representation of Zipf distributions. *Computational Statistics & Data Analysis 19, 461-473.*

Glottometrics 14, 2007, 72-98

History of Quantitative Linguistics

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

XXV. Lorenzo Bianchi (1889-1960)

Geb. 20.10.1889 in Porto Maurizio, Italien, gest. 6.7.1960 in Bologna. Gymnasium 1900-1907 in Carcare und Bologna; Studium der Klassischen Philologie in Bologna (1907-1911) und zusätzlich der deutschen Literaturgeschichte in Heidelberg (1913-14), Promotion 1911 in Bologna, ab 1911 Arbeit als Gymnasiallehrer und Lektor, 1915-1954 Mitarbeiter und Übersetzer im Verlagshaus Zanichelli, ab 1919 Lehrbeauftragter und später Prof. für deutsche Sprache und Literatur an verschiedenen Hochschulen in Bologna und Mailand, 1959 Ruhestand. Werke: Viele literaturwissenschaftliche Arbeiten, Übersetzungen, eine italienisch verfasste Grammatik des Deutschen. (Nach: Mignardi 2003; dort wesentlich ausführlichere Informationen.) Soweit ich bisher sehen kann, ist Bianchi in der Quantitativen Linguistik bisher völlig unbekannt geblieben.

Obwohl kein Sprachwissenschaftler, ist Lorenzo Bianchi für die Quantitative Linguistik von Bedeutung, weil er unter Berufung auf Marbe und seine Schule (Forschungsüberblick: Bianchi 1922: 5-12) Untersuchungen zu rhythmischen Einheiten bei einigen deutschen Autoren durchgeführt hat (Bianchi 1922: 12ff.). Er ist damit ein früher Vertreter der „Quantitativen Literaturwissenschaft“ (Fucks 1968: 88). Wie Marbe (1904) nimmt er die ersten 1000 Wörter der ausgewählten Texte, akzentuiert sie und stellt in Tabellen zusammen, wie häufig die rhythmischen Einheiten verschiedener Länge zu beobachten sind. Es werden also künstliche Textblöcke und nicht natürliche Texte oder Textabschnitte bearbeitet. In Best (2005) wurde bereits darüber berichtet, dass diese Entscheidung bisweilen zu Problemen führt, wenn man versucht, Modelle für die Verteilung rhythmischer Einheiten zu testen. Die Arbeit mit vollständigen Texten erwies sich als wesentlich erfolgreicher (Best 2002, 2005a: 210f.).

Ausgehend von der Hypothese, dass rhythmische Einheiten in Texten gesetzmäßig verteilt sein sollten, so wie andere Einheiten auch (Wimmer u.a. 1994; Wimmer & Altmann 1996), wurde an die Daten, die Bianchi für deutsche Autoren erarbeitet hat, die 1-verschobene Hyperpascal-Verteilung

$$P_x = \frac{\binom{k+x-2}{x-1}}{\binom{m+x-2}{x-1}} q^{x-1} P_0 , \quad x = 1, 2, ...$$

angepasst, die sich in allen Fällen als geeignet erweist, wie die folgenden Tabellen zeigen. Dabei wurden nur die jeweiligen Gesamtdateien (Bianchi 1922: 36) berücksichtigt.

Legende zu den Tabellen:

x: Länge der rhythmischen Einheiten, beginnend mit $x = 1$ für 0 unbetonte zwischen zwei betonten Silben; $x = 2$: 1 unbetonte Silbe zwischen zwei betonten; etc.

f_x: beobachtete Zahl der rhythmischen Gruppen mit x Silben;

NP_x: aufgrund der Hyperpascal-Verteilung berechnete Zahl rhythmischer Einheiten mit x Silben;

q: Parameter der Hyperpascal-Verteilung;

X^2: Wert des Chiquadrats;

FG: Freiheitsgrade;

P: Überschreitungswahrscheinlichkeit des Chiquadrats, der mit $P \geq 0.05$ eine gute Übereinstimmung zwischen Beobachtung und Modell anzeigt. Diese Bedingung ist in allen Fällen erfüllt.

Die senkrechten Striche in der Tabelle zeigen an, dass die entsprechenden Längenklassen zusammengefasst wurden.

	Kleist, Michael Kohlhaas		J. Grimm, Selbstbiographie		Hebel, Schatzkästlein	
x	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	17	18.86	26	27.80	26	30.49
2	126	132.98	162	171.45	148	152.22
3	125	122.42	124	129.51	125	121.50
4	98	83.77	107	83.88	76	74.63
5	47	50.37	56	51.23	36	41.00
6	33	28.16	33	30.34	36	21.16|
7	16	15.02	14	17.62	9	10.49|
8	2	7.76	5	10.11	3	5.06|
9	2	3.91	6	5.74	2	4.44|
10	0	1.93	1	3.24		
11	1	1.80	0	1.82		
12			1	2.28		
	$k = 1.3185$	$m = 0.0802$	$k = 0.4678$	$m = 0.0406$	$k = 1.1072$	$m = 0.0917$
	$q = 0.4289$	$X^2 = 11.639$	$q = 0.5355$	$X^2 = 15.337$	$q = 0.4135$	$X^2 = 3.412$
	$FG = 7$	$P = 0.11$	$FG = 8$	$P = 0.05$	$FG = 2$	$P = 0.18$

	Hebel, Briefe		Hebel, Biblische Geschichten		Grimm, Märchen	
x	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	39	43.39	36	44.27	28	30.36
2	170	173.51	161	170.91	171	177.03
3	140	133.27	184	153.79	125	121.52
4	86	78.76	84	90.31	79	68.40
5	40	41.56	40	42.26	42	35.59
6	24	20.57	16	17.11	11	17.75
7	8	9.77	5	6.26	10	8.63
8	1	4.51	1	2.13|	1	7.73
9	1	3.65	0	0.68|		
10			1	0.29|		
	$k = 1.1786$	$m = 0.1160$	$k = 4.5342$	$m = 0.2360$	$k = 0.6625$	$m = 0.0492$
	$q = 0.3934$	$X^2 = 7.133$	$q = 0.2010$	$X^2 = 9.331$	$q = 0.4332$	$X^2 = 7.608$
	$FG = 5$	$P = 0.21$	$FG = 4$	$P = 0.05$	$FG = 3$	$P = 0.05$

Damit hat sich erwiesen, dass die rhythmischen Einheiten der von Bianchi untersuchten Texte gesetzmäßig verteilt sind. Das Modell, das sich sonst bei deutschen Texten bewährt hat, die 1-verschobene Hyperpoisson-Verteilung, konnte nicht verwendet werden. Dies kann damit zu tun haben, dass Bianchi willkürlich gebildete Textabschnitte bearbeitet hat. Da es sich um verschiedene Autoren und auch um verschiedene Textsorten handelt, wäre es auch sinnvoll, jeweils nach einem besonders geeigneten Modell zu suchen, wenn man in den einzelnen Fällen nicht nur einen Textabschnitt hätte, sondern möglichst mehrere, dafür aber natürliche Texte oder Textteile wie z.B. Kapitel. Dass bei verschiedenen Sprachen auch mit verschiedenen Modellen für die Verteilung rhythmischer Einheiten zu rechnen ist, zeigt die Untersuchung altgriechischer Textabschnitte (Best (2006), bei denen sich die geometrische Verteilung bewährt. Dafür, dass natürliche Texte bessere Ergebnisse erwarten lassen, spricht Kaßel (2002), die an deutsche (15 Briefe Kleists und 15 Pressetexte verschiedener Presseorgane) sowie englische Texte (15 Briefe Jane Austens und ebenfalls 15 Pressetexte verschiedener Presseorgane) die 1-verschobene Hyperpoisson-Verteilung anpassen konnte.

In dieser Vorstellung Bianchis habe ich darauf verzichtet, auch seine Analysen zu 100-Wort-Blöcken zu berücksichtigen, die ja wesentlich weniger rhythmische Einheiten aufweisen und damit nur wenig Daten aufweisen.

Literatur

Best, Karl-Heinz (2002). The Distribution of Rhythmic Units in German Short Prose. *Glottometrics 3, 136-142.*

Best, Karl-Heinz (2005). Karl Marbe (1869-1953). *Glottometrics 9, 74-76.*

Best, Karl-Heinz (2005a). Längen rhythmischer Einheiten. In: Köhler, Reinhard, Altmann, Gabriel, & Piotrowski, Rajmund G. (Hrsg.), *Quantitative Linguistik - Quantitative Linguistics. Ein internationales Handbuch* (S. 208-214). Berlin/ N.Y.: de Gruyter.

Best, Karl-Heinz (2006). Rhythmische Einheiten im Altgriechischen. *Göttinger Beiträge zur Sprachwissenschaft* (eingereicht).

Bianchi, Lorenzo (1922). *Untersuchungen zum Prosa-Rhythmus Johann Peter Hebels, Heinrich von Kleists und der Brüder Grimm.* Heidelberg: Weiss'sche Universitätsbuchhandlung.

Fucks, Wilhelm (1968). *Nach allen Regeln der Kunst*. Stuttgart: Deutsche Verlags-Anstalt.

Kaßel, Anja (2002). *Zur Verteilung rhythmischer Einheiten in deutschen und englischen Texten.* Staatsexamensarbeit; Göttingen.

Marbe, Karl (1904). *Über den Rhythmus der Prosa*. Giessen: J. Ricker'sche Verlagsbuchhandlung.

Mignardi, Guilia (2003). Bianchi, Lorenzo. In: *Internationales Germanistenlexikon 1800-1950. Band 1: A-G.* Herausgegeben und eingeleitet von Christoph König. Berlin/ New York: de Gruyter.

Karl-Heinz Best

XXVI. Manfred Faust (1936-1997)

Geb. in Chemnitz 6.12.1936, 1943-49 Volksschule Wilsdruff und Pfofeld, 1949-57 Gymnasium Oettingen und Würzburg, 1957/58-1962/63 Studium (Vergleichene Sprachwissenschaft, Altphilologie, Germanistik) in Marburg und Tübingen, Promotion Tübingen 20.6.1963 und Habilitation Tübingen 7.6.1975 in Allgemeine und Vergleichende Sprachwissenschaft,

15.1.1976 venia legendi (Allgemeine und vergleichende Sprachwissenschaft, Tübingen); 1963-1980 in verschiedenen Funktionen/ Stellen (wiss. Hilfskraft, Verwalter einer Assistentenstelle, Assistent, wissenschaftlicher Angestellter, Stipendiat der DFG) in Tübingen; 1976-80 Privat-Dozent in Tübingen; ab 1.4.1980 Prof. für Germanistische Sprachwissenschaft in Konstanz. Seit dem 5.11.1997 galt Faust als vermisst; er wurde am 25.1.98 entdeckt, verstorben an einer Medikamentenvergiftung (Andreas Plecko, *Abendzeitung*, 31.1./ 1.2.1998). Seine Hauptarbeitsgebiete waren: „Varietätenlinguistik, Textlinguistik, historische Sprachwissenschaft – Deutsch, Sprachen des antiken Mittelmeerraums“ (Kürschner 1994: 219).

Manfred Faust war ein Sprachwissenschaftler mit vielfältigen Interessen, auch solchen, die außerhalb der Hauptströmungen der Linguistik seiner Zeit lagen. Besonders deutlich wird das dadurch, dass er zusammen mit Helmut Bachmaier Herausgeber von Karl Valentin, Sämtliche Werke in acht Bänden (München: Piper 1991-1997) war. Seine Bedeutung für die Quantitative Linguistik erwächst aus dem Umstand, dass er seine Argumentation zu und Darstellung von sprachlichen Sachverhalten mehrfach auf statistische Erhebungen stützte. Er gehört damit zu den Philologen, die uns immer wieder im Bestreben nach präziser Information Daten liefern, anhand derer man Gesetzeshypothesen testen und damit theoretische Annahmen überprüfen kann. Einer dieser Fälle findet sich in Faust (1972: 100f.), wo er die Länge der rund 9000 Bildtitel von Paul Klee in 8 Lebensphasen ermittelt. Aus seinen Angaben wurde die folgende Übersicht entwickelt und daraufhin geprüft, ob sie als ein Prozess im Sinne des Modells für den vollständigen Sprachwandel (Altmann 1983: 60) erwiesen werden kann. Die folgende Tabelle zeigt das Ergebnis (Best 2003b):

Tabelle1
Die durchschnittliche Länge der Bildtitel (in Wörtern) von Paul Klee

festgesetzter Zeitpunkt	t	Länge der Titel (beob.)	Länge der Titel (berechn.)	festgesetzter Zeitpunkt	t	Länge der Titel (beob.)	Länge der Titel (berechn.)
1898.5	1	3.61	3.65	1930.5	32	2.47	2.54
1917	18.5	2.97	2.90	1934.5	36	2.30	2.45
1922	23.5	2.83	2.75	1938	39.5	2.44	2.38
1926.5	28	2.55	2.63	1939.5	41	2.47	2.35
a = -0.73 b = 0.0058 D = 0.95							

Legende:

a, *b*: Parameter des Modells.
beob.: beobachtet, d.h. die aufgrund der Angaben von Faust ermittelten durchschnittlichen Längen der Bildtitel je Zeitabschnitt.
berechn.: berechnet, d.h. die aufgrund der folgenden Formel berechneten Werte für den Wandel des Idiolekts.
t: Zeitpunkt, für die Berechnung transformiert.
D: Determinationskoeffizient, der mit $D \geq 0.80$ eine gute Übereinstimmung des Modells mit den beobachteten Daten anzeigt.

Dieser Sprachwandel folgt also mit sehr gutem Determinationskoeffizient $D = 0.95$ der Formel

$$p_t = \frac{1}{1 - 0.73\, e^{-0.0058\, t}},$$

wie auch die Abb. 1 zeigt:

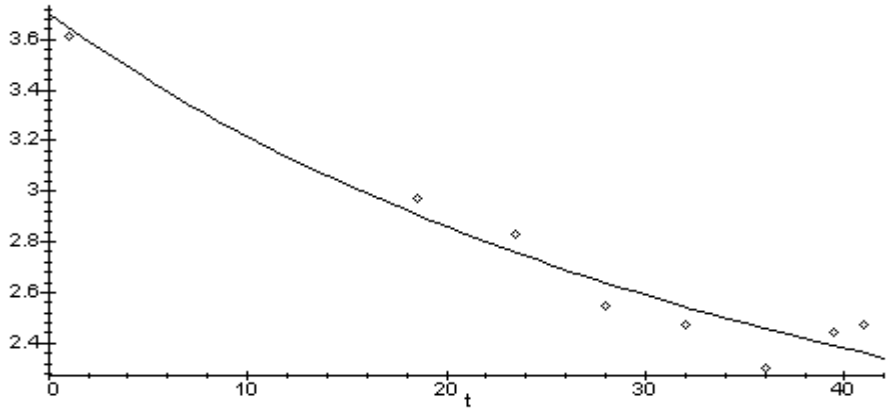


Abb. 1: Die durchschnittliche Länge der Bildtitel (in Wörtern) von Paul Klee (aufgefasst als vollständiger Sprachwandel)

Da am Ende die Bildtitellänge wieder zunimmt, kann man deren Wandel auch als reversiblen Prozess verstehen (vgl. dazu Best 2003b).

Ein weiterer Fall von Sprachwandel, für den Faust (1980: 400-404) Daten erhoben hat, ist der Übergang ehemals starker deutscher Verben in die Klasse der schwachen; er stellt sich (Best 2003a: 13; korrigiert) wie folgt dar:

Tabelle 2
Letztes Auftreten starker Formen ehemals starker Verben

Jhd.	*t*	*x* (beobachtet)	*x* (kumulativ)	*x* (berechnet)
15.	1	3	3	5.5645
16.	2	16	19	17.0253
17.	3	13	32	32.3567
18.	4	7	39	40.8113
19.	5	5	44	43.4214
20.	6	1	45	44.0698
	a = 30.2339	*c* = 44.2674	*b* = 1.4695	*D* = 0.99

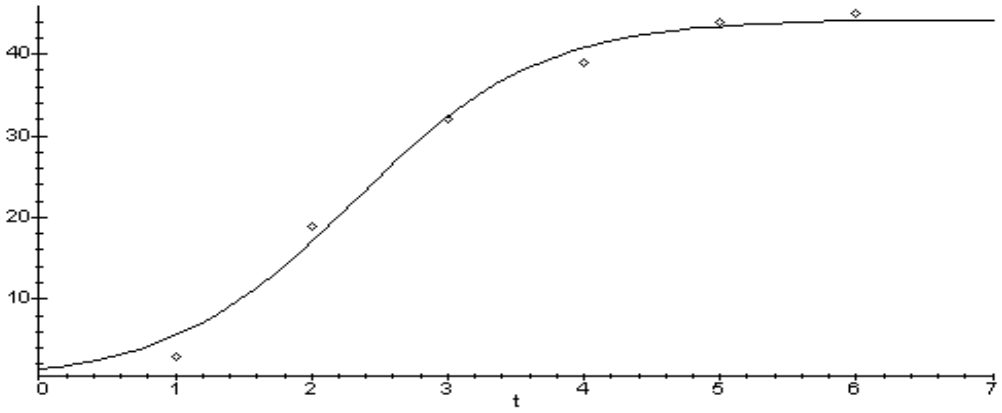


Abb. 2. Letztes Auftreten starker Formen ehemals starker Verben

Dieser Prozess folgt damit dem Modell für den unvollständigen Sprachwandel (Altmann 1983: 61):

$$p_t = \frac{44.2674}{1+30.2339\, e^{-1.4695t}}.$$

Auch wenn man die Jahrhunderte betrachtet, in denen schwache Formen ehemals starker Verben erstmals auftreten, erhält man ebenso gute Ergebnisse (Best 2003a: 12).

Diese beiden Sprachprozesse, deren Daten Faust erhoben hat, zeigen, dass er für die Quantitative Linguistik nicht ohne Bedeutung ist. Weitere Hinweise dazu: In (Faust 1983: 240) weist er auf ein quantitatives Argument von Jean Paul im Hinblick auf die s-Fuge hin (Best 2006); in einer Untersuchung der Sprachkenntnisse einiger griechischer Schulkinder in Deutschland stieß er auf erhebliche lexikalische Defizite, die sich beispielsweise in der Vereinfachung des lexikalischen Feldes der Sitz- und Liegemöbel auf z.T. nur zwei Wörter (*Stuhl* und *Bett* und die griechischen Entsprechungen dazu) bemerkbar machten; er erklärt dies mit Hinweis auf die hohe Gebrauchsfrequenz gerade dieser Wörter (Faust 1984: 122). Für sein Interesse an sprachstatistischen Themen sprechen auch Rezensionen, so die zu Best (1973), in der er auch auf die quantitativen Aspekte der Analogie eingeht (Faust 1977), ebenso wie auch die zu Ruoff (1981), in der er u.a. eine Geschichte der Häufigkeitswörterbücher des Deutschen skizziert (Faust 1983).

Die Quantitative Linguistik verdankt Forscherkollegen wie Manfred Faust viele Erkenntnisse, die sie in ihre eigenen theoretischen Konzepte einbeziehen kann bzw. an denen sie ihre Annahmen überprüfen kann. Wir haben allen Grund, diese Wissenschaftler in unser kollektives Gedächtnis aufzunehmen.

Für Auskunft zu Bildungsgang und beruflicher Laufbahn danke ich Dr. Wischnath, Universitätsarchiv Tübingen. Versuche (E-Mail und Brief), zusätzliche Auskünfte von der Universität Konstanz zu erhalten, blieben bis zur Abgabe des Manuskripts viele Wochen unbeantwortet.

Literatur

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

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

Best, Karl-Heinz (2003b). Zum Wandel von Idiolekten. *Naukovyj Visnyk Černivec'koho Universytetu: Hermans'ka filolohija. Vypusk 165-166, 36-43.*

Best, Karl-Heinz (2006). Jean Paul (1763-1825). *Glottometrics 12* (eingereicht).

Faust, Manfred (1972). Diachronie eines Idiolekts: Syntaktische Typen in den Bildtiteln von Klee. In: Gunzenhäuser, Rul (Hrsg.), *Mathematisch orientierte Textwissenschaft. = Zeitschrift für Literaturwissenschaft und Linguistik 2(8), 97-109.*

Faust, Manfred (1977). Rez. zu Karl-Heinz Best (1973). Probleme der Analogieforschung. München: Hueber. *Zeitschrift für Dialektologie und Linguistik XLIV, 183-187.*

Faust, Manfred (1980). Morphologische Regularisierung in Sprachwandel und Spracherwerb. *Folia Linguistica XIV, 387-411.*

Faust, Manfred (1983). Jean Paul's essay on word formation. In: Faust, Manfred, Harweg, Roland, Lehfeldt, Werner, & Wienold, Götz (Hrsg.); *Allgemeine Sprachwissenschaft,*

Sprachtypologie und Textlinguistik. Festschrift für Peter Hartmann: 237-248. Tübingen: Narr.

Faust, Manfred (1983). Rez. zu: Arne Ruoff (1981). Häufigkeitswörterbuch gesprochener Sprache. Tübingen: Niemeyer. *Zeitschrift für Dialektologie und Linguistik L, 242-246.*

Faust, Manfred (1984). On the Bilingual Lexicon of Greek School Children in the Federal Republic of Germany. In: *Navicula Tubingensis. Studia in honorem Antonii Tovar* (S. 115-126). Hrsg. von Francisco J. Oroz Arizcuren unter Mitarbeit von Eugenio Coseriu und Carlo de Simone. Tübingen: Narr.

Kürschners deutscher Gelehrtenkalender 1992. Bio-bibliographisches Verzeichnis deutschsprachiger Wissenschaftler. 16. Ausgabe. Bd. 1. Berlin/ New York: de Gruyter 1992.

Kürschner, Wilfried (Hrsg.) (1994). *Linguisten-Handbuch. Biographische und bibliographische Daten deutschsprachiger Sprachwissenschaftlerinnen und Sprachwissenschaftler der Gegenwart. Band 1: A-L.* Tübingen: Narr.

Anmerkung: Das Literaturverzeichnis nennt nur einschlägige Arbeiten. Weitere Angaben zum Werk Fausts finden sich in Kürschner (1994).

Karl-Heinz Best

XXVII. Erwin Kunath (1899-1983)

Vollständiger Name: Karl Erwin Kunath. Geb. am 26.10.1899 in Dresden, gest. am 26.12.1983, ebenfalls in Dresden. 1906-1916/17 Schulbesuch in Dresden. 1917-19 Militärdienst, unterbrochen durch einen zehnwöchigen Kurs, der Anfang 1918 zur Matur führte. 1919 - 1922 Studium (Germanistik, Geschichte, Englisch, Philosophie und Pädagogik) in Leipzig, Promotion 1922 mit einer Dissertation über den Barock-Lyriker David Schirmer. 1923 Staatsexamen für das höhere Lehramt und Eintritt in den Schuldienst bei den Technischen Lehranstalten der Stadt Dresden (Fächer: Deutsch, Englisch, Reichsbürgerkunde, Buchführung, niedere Mathematik); nebenamtliche Lehrtätigkeit am Vorbereitungsinstitut Laue (Deutsch, Englisch, Geschichte). Militärdienst zu Beginn des zweiten Weltkrieges; im Dezember 1939 uk gestellt und mit dem Aufbau der Städtischen Ingenieur- und Techniker-Vorschule beauftragt, die den Technischen Lehranstalten angegliedert war. 1939 zum pädagogischen Studiendirektor ernannt.

Kunath war Mitglied in mehreren nationalsozialistischen Organisationen und wurde im November 1945 aus dem Schuldienst entlassen. 1947 wurde er von der Entnazifizierungskommission im Stadtkreis Dresden rehabilitiert (Protokoll der Sitzung v. 10.12.1947). Ende 1945 Tätigkeit als Hilfsarbeiter, dann bis in die 60er Jahre Lehrkraft an der Volkshochschule Dresden (Sarfert 2004) und Lehrer für Privatstunden.

Die Angaben beruhen auf Dokumenten und Informationen, die mir das Sächsische Staatsarchiv - Hauptstaatsarchiv Dresden, das Stadtarchiv der Landeshauptstadt Dresden und das Universitätsarchiv der Technischen Universität Dresden dankenswerterweise zur Verfügung gestellt haben. Mein ganz besonderer Dank gilt Frau Angela Buchwald vom Universitätsarchiv der TU Dresden, ohne deren engagierte Unterstützung ich wesentliche Daten nicht erhalten hätte.

Kunath spielt für die Quantitative Linguistik eine gewisse Rolle, da er früher als die meisten anderen Forscher Wortlängenverteilungen verschiedener Autoren erhoben hat; sie

dienen ihm dazu, Argumente für stilistische Empfehlungen für die verständliche Gestaltung von Texten zu entwickeln. Zusammen mit seiner Aufforderung, kurze Sätze zu bevorzugen, sieht man sich bereits an die Lesbarkeitsforschung verwiesen (Best 2005; Groeben 1982). Auf Kunath (1937: 18) bezieht sich Esser (1960: 76), der daran anknüpft und die Datenbasis erweitert. Kunath (1937, ²1941: 18) stellt in seiner Stil- und Regelkunde Prozentwerte für Texte von 11 Autoren zusammen. (Die Ausgabe von 1937 war mir trotz aller Bemühungen bisher nicht zugänglich; die von 1941 nur als Zufallstreffer über ein Antiquariat. Die Seitenangabe durch Esser für die 1. Auflage stimmt mit der der 2. Auflage überein.) Da Kunaths Buch nach meinen Erfahrungen nicht ganz leicht zu bekommen ist, habe ich die entsprechenden Passagen in (Best 2006a) zitiert; dieses Zitat sei hier wiederholt:

„**7. Von der Schlichtheit des Wortes**

Dritte Stilregel: Sprich schlicht!

Es ist ein törichter Aberglaube, wenn einer denkt, daß er sich möglichst seltsam und gespreizt ausdrücken muß, sobald er die Feder zur Hand nimmt und schreibt. Die Forderung, schlicht zu schreiben, gilt schon rein äußerlich. Die deutsche Sprache drängt ihrem Wesen nach zum kurzen Wort. Die langatmigen Wortgebilde des Gelehrtendeutsch sind nicht aus dem Gefühl für die Eigenart unserer Muttersprache erwachsen. Die bedeutendsten Dichter, die immer unsere Stilvorbilder sein werden, haben schon rein äußerlich das schlichte, das kurzsilbige Wort gesucht. Unter hundert zusammenhängenden Wörtern, die an verschiedenen Stellen ihrer Werke durchgezählt wurden, verwendeten im Mittel:

	Goethe	Keller	Storm	Raabe	Löns	Flex	Blunk
einsilbige Wörter	60%	52%	52%	52%	64%	48%	53%
zweisilbige “	27%	26%	35%	32%	33%	36%	27%
dreisilbige “	7%	14%	9%	14%	3%	9%	14%
viersilbige “	6%	4%	3%	2%	0%	7%	5%
fünfsilbige “	0%	2%	1%	0%	0%	0%	1%
vielsilbige “	0%	0%	0%	0%	0%	0%	0%

Die verwendeten Wortlängen zeigen eine auffallende Übereinstimmung. Stellt man daneben Reden oder Schriften wissenschaftlichen oder politischen Inhalts, bei denen die stofflich notwendigen Fachwörter einer sprachlich glatten Form häufiger zuwiderlaufen, so erhält man beim guten Stilisten noch immer annähernd das gleiche Bild. So finden sich bei:

	Ranke	List	Bismarck	Hitler
einsilbige Wörter	45%	53%	44%	53%
zweisilbige “	32%	23%	32%	28%
dreisilbige “	17%	12%	14%	6%
viersilbige “	6%	9%	6%	11%
fünfsilbige “	0%	2%	4%	2%
vielsilbige “	0%	1%	0%	0%“

(Kunath ²1941: 18).

Weitere Werke mit linguistischem Inhalt hat Kunath offenbar nicht verfasst. Außer Esser (1960) ist mir bisher auch keine weitere Bezugnahme eines Linguisten auf Kunath bekannt.

Literatur

Best, Karl-Heinz (2006). Sind Wort- und Satzlänge brauchbare Kriterien der Lesbarkeit von Texten? In: Wichter, Sigurd, & Busch, Albert (Hrsg.), *Wissenstransfer – Erfolgskontrolle und Rückmeldungen aus der Praxis* (S. 21-31). Frankfurt/ M. u.a.: Lang.

Best, Karl-Heinz (2006a). Wortlängen im Deutschen. *Göttinger Beiträge zur Sprachwissenschaft 13* (erscheint).

Esser, Wilhelm M. (1960). Zum konsonantischen Element im deutschen Sprachlautkörper. Ein Beitrag zur vergleichenden Stilistik. *Wirkendes Wort 10, 68-78*.

Groeben, Norbert (1982). *Leserpsychologie: Textverständnis - Textverständlichkeit*. Münster: Aschendorff.

Kunath, Erwin (1922). *David Schirmer als Dichter und als Bibliothekar. Ein Beitrag zur Geistesgeschichte Kursachsens im 19. Jahrhundert.* Diss., Leipzig (liegt nur handschriftlich vor).

Kunath, Erwin (1937, ²1941). *Klares Deutsch. Stil- und Regelkunde für berufsbildende Schulen.* 2, verbesserte Auflage. Leipzig: Klinkhardt.

Sarfert, Hans-Jürgen (2004). David Schirmer - Poet und Bibliothekar am kurfüstlichen Hof. *SLUB-Kurier 2004/1, 9-10.*

Karl-Heinz Best

XXVIII. Otto Behaghel (1854-1936)

Geboren am 3.5.1854 in Karlsruhe, Gymnasium Karlsruhe, Studium der klassischen und neueren Philologie 1873-76 in Heidelberg, Göttingen und Paris; Promotion 1876 in Heidelberg, Habilitation 1878 ebenfalls in Heidelberg. 1882 ao. Professor, 1883 o. Professor Basel, 1888 Professor für deutsche Philologie in Gießen. Gestorben 9.10.1936 in München. Seine thematischen Schwerpunkte sind: deutsche Sprachgeschichte und deutsche Syntax; Mitwirkung im deutschen Sprachverein.

In Bibliographien zur Quantitativen Linguistik wird gelegentlich auf Behaghel Bezug genommen (auf die Untersuchung zum Dativ-{e} [Behaghel 1900] bei Guiraud 1954: 66, Nr. 2021 bei Köhler 1995; vgl. auch Meier ²1967: 387). Er behandelt jedoch nicht nur dieses Thema auf eine für die Quantitative Linguistik einschlägige Weise:

Fremdwörter: Es geht um eine Diskussion der Frage, ob das Deutsche für Fremdwörter besonders anfällig sei oder nicht. Behaghel (1918) erörtert dabei linguistische und statistische Probleme, die entstehen, wenn man die Fremdwortbestände in den Nachbarsprachen mit denen des Deutschen vergleichen will (Meier ²1967: 23).

Attributive Adjektive: Untersuchung der Häufigkeit von Adjektiven („Beiwort“) in Versdramen von Schiller. Behaghel (1905) zählt dazu aus, wie viele attributive Adjektive in den ersten 1000 Versen der zeitlich geordneten Werke vorkommen und stellt fest: „Also zunächst ein Anschwellen der Zahl, dann wieder ein Rückgang“ (Behaghel 1905: 181). Diese Schwankungen seien auch nicht mit der unterschiedlichen Zahl der Substantive zu erklären. Es handelt sich also um einen reversiblen Wandel im Stil Schillers, der dem Modell

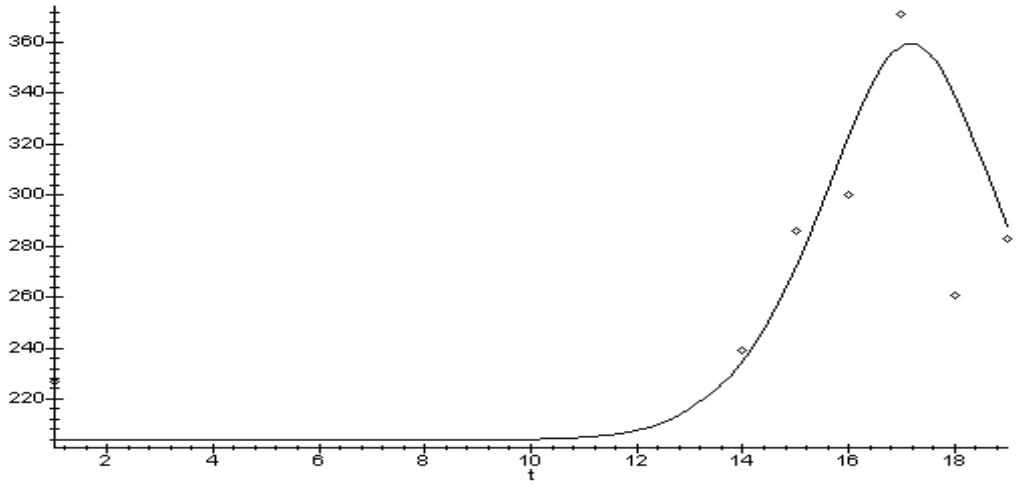
$$p = \frac{d}{1 + ae^{-bt + ct^2}}$$

folgen sollte (Altmann 1983: 62). Die von Behaghel bereitgestellten Daten sind problematisch, da sie aus willkürlich festgelegten Textabschnitten stammen und außerdem eine ungünstige zeitliche Verteilung aufweisen; hinzu kommt, dass es sich auch nur um 7 Werke handelt. So ist es nicht verwunderlich, wenn die Anpassung des Modells zunächst nicht gelingt. Betrachtet man aber die Werte von *Wilhelm Tell* einmal als Ausreißer, dann lässt sich der Trend mit Hilfe von NLREG sehr gut modellieren:

Tabelle 1
Häufigkeit der attributiven Adjektive in Versdramen von Schiller

Drama	Jahr	*t*	$Adj._{beobachtet}$	$Adj._{berechnet}$
Don Carlos	1787	1	227	204.00
Piccolomini	1800	14	239	234.94
Maria Stuart	1801	15	286	271.17
Jungfrau v. Orleans	1802	16	300	322.67
Braut v. Messina	1803	17	371	358.59
Wilhelm Tell	1804	18	261	-
Demetrius	1805	19	283	287.95
a = -0.0000000000000003		b = -4.0586	c = -0.1182	d = 204 D = 0.89

Anmerkung: Die Datierung wurde nach Wilpert (1997) vorgenommen. *a*, *b*, *c* und *d* sind die Parameter des Modells; *t* sind die Zeitabschnitte in Jahren, beginnend mit $t = 1$ für 1887; *D* ist der Determinationskoeffizient, der mit $D \geq 0.80$ eine gelungene Anpassung des Modells anzeigt, wie auch die Graphik verdeutlicht:



Graphik zu Tab. 1: Häufigkeit der attributiven Adjektive in Versdramen von Schiller

In der Graphik ist der Wert für *Wilhelm Tell* eingetragen, der bei der Berechnung aber nicht berücksichtigt wurde.

Gesetz der wachsenden Glieder: Behaghel (1909) untersucht koordinierte Wortgruppen daraufhin, ob eher eine längere vor einer kürzeren erscheint oder umgekehrt. Seine Beobachtungen zum Griechischen, Lateinischen und Deutschen in Versliteratur und Prosa führen ihn dazu, ein „G e s e t z d e r w a c h s e n d e n G l i e d e r" zu postulieren (Behaghel 1909: 139), das er mit einem "Gesetz [...] von der Späterstellung des Wichtigsten" (Behaghel 1930: 86) ergänzt. Nur bei Plautus beobachtet er eine Ausnahme von dieser allgemeinen Tendenz, die er auf den Stil volkstümlicher, gesprochener Sprache zurückführt. Sonst gilt aber: „Im allgemeinen wirkt unser Gesetz bei den von mir geprüften Schriftstellern ungefähr in derselben Stärke" (Behaghel 1909: 141). Behaghel (1909: 138) stellt fest, dass dieses Gesetz generell in germanischen Sprachen gilt, und zwar auch bei anderen syntaktischen Konstruktionen, und führt es auf die Sprachverarbeitungsmechanismen bei Sprecher und Hörer zurück. Auch rhythmische Aspekte spielen dabei eine Rolle (Behaghel 1912). Behaghel verweist ferner auf Beobachtungen Eheloffs, der dieses Gesetz auch im Assyrisch-Babylonischen nachweise (Behaghel 1930: 86). Fenk-Oczlon & Fenk (2002: 26) versuchen eine Teil-Erklärung: „The prevailing of the (very frequent and therefore) rather short function words in the first part of sentences might contribute to or even account for Behaghel's (1909) ‚Gesetz der wachsenden Glieder'." Ähnliche Beobachtungen wie Behaghel hat Fenk-Oczlon (1989: 517) am Beispiel von „freezes", idiomatisierten Redewendungen wie „Lust und Laune", „Kind und Kegel" etc., gemacht: „the more frequent and therefore informationally poorer elements tend to occupy initial position." Man muss sich dann nur noch daran erinnern, dass die häufigeren Elemente in der Sprache ja auch die kürzeren sind. Auch Beobachtungen zum Tschechischen bestätigen Behaghels Gesetz, differenzieren es aber auch etwas (Uhlířová 1997).

Gegenwärtig werden entsprechende Befunde u.a. unter dem EIC-Prinzip (EIC: Early Immediate Constituents) behandelt, teils mit Bezugnahme auf Behaghels Entdeckung (Vulanović & Köhler 2005: 279ff.). Hoffmann (1999: 110f.) erläutert dazu: „Since with the head of a last phrase the IC-structure of the whole phrase is known to the processor and thus the whole structure containing this last element may be cleared from working memory, it is advantageous to order constituents in that way that the longer follows the shorter."

Das Allomorph -{e} im Dat.Sg. bei deutschen Maskulina und Neutra: Dieses Thema greift Behaghel mehrfach auf (Behaghel 1900, 1909b, 1919). In einer tabellarischen Übersicht, geordnet nach dem Geburtsjahr der Autoren, zeigt er für die Zeit von Luther bis zum Ende des 19. Jahrhunderts, dass die Verwendung dieses Allomorphs von Autor zu Autor schwankt, wobei sich Unterschiede zeigen bei einsilbigen Wörtern, Wörtern mit Vorsilbe, Wörtern mit Nachsilbe und vor nachfolgendem <e> (Behaghel 1900: 174). Ordnet man die Übersicht nach dem Erscheinungsjahr der Werke, ergibt sich das Resultat in Tabelle 2.

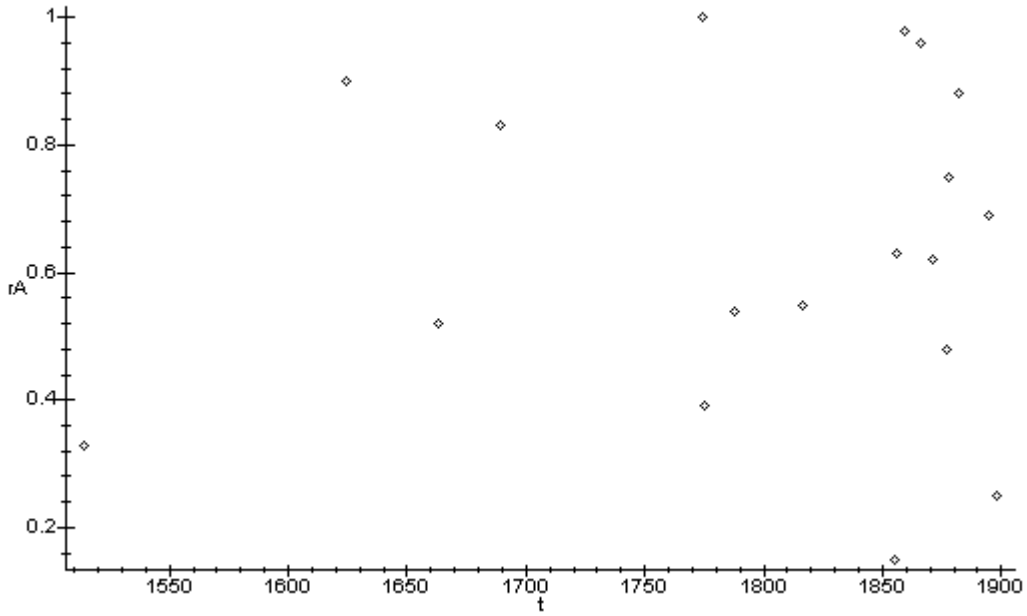
Die Tabelle enthält die meisten von Behaghel angegebenen Titel. Einige wurden ausgelassen, wenn ihre Datierung nicht bestimmt werden konnte. Bei Werken, deren Erscheinen sich über Jahre erstreckte, wurde als Messpunkt die mittlere Jahreszahl bestimmt. Die Graphik zeigt, dass die Nutzung des dat.sg. mit -{e} nach Behaghels Auszählungen eine Punktwolke ohne klaren Trend ergibt.

(Behaghels Tabelle ist differenzierter; die einzelnen Kategorien sind aber oft nur schwach belegt. Deshalb wird hier für jeden Autor nur ein Gesamtwert angegeben.)

Tabelle 2
Das Vorkommen von -{e} im Dat. Sg. Mask./ Neutrum bei deutschen Substantiven
(n. Behaghel 1900)

Autor	Werk	Erscheinungs-jahr	∑ Dat.Sg.	mit -{e}	rel. Anteil
Luther	Bibel	1541	220	72	0.33
Opitz	Buch von der dt. Poeterey	1624	82	74	0.90
Gryphius	Horribilicribrifax	1663	132	68	0.52
Lohenstein	Arminius...	1689f.	109	91	0.83
Lessing	Laokoon	1766	nur mit -{e}		1.00
Wieland	Abderiten	1774 (1781 Neufassung)	135	84	0.62
Geßner	Schriften	1775	56	22	0.39
Schiller	Abfall der Niederlande	1788	84	45	0.54
Goethe	Dichtung und Wahrheit	1811-22	58	32	0.55
Scheffel	Ekkehard	1855	93	14	0.15
Riehl	Kulturgesch. Novellen	1856	142	89	0.63
Fichte	Reden an die dt. Nation	1859	47	46	0.98
Süpfle	Aufgaben zu lateinischen Stilübungen	[12]1866	49	47	0.96
Leander = Richard v. Volkmann	Träumereien an franz. Kaminen	1871	81	50	0.62
Freytag	Markus König (Die Ahnen)	1873-81	121	58	0.48
Keller	Zürcher Novellen	1878	95	71	0.75
Gindely	Gesch. d. dreißigj. Krieges	1882	69	61	0.88
Tovote	Heißes Blut	1895	42	29	0.69
Fulda	Jugendfreunde	1898	64	16	0.25

Nach Wegera (1987: 115ff.) ist in frühneuhochdeutscher Zeit zu beobachten, dass die Formen auf -{e} zwar in den meisten Dialekten in unterschiedlichem Maße, aber doch insgesamt deutlich abnehmen. Dazu Hartweg & Wegera (1989: 118): „Das Dat.-*e* unterliegt dem umfassenden Prozeß der *e*-Apokope ...Es wird bis zum 16. Jh. im Obd. [= Oberdeutschen, Verf.] nahezu ganz, im Wmd. [= Westmitteldeutsch] weitestgehend und im Omd. [= Ostmitteldeutsch] ansatzweise getilgt. Seit dem 16. Jh. nimmt die Verwendung des Dat.-*e* vom Omd. ausgehend wieder zu, wird jedoch nicht obligatorisch.“ In der Gegenwart tritt das Dativ-e nur noch zuweilen und unter bestimmten Bedingungen auf (Duden. Die Grammatik [7]2005: 210ff.). Daraus ergibt sich, dass es einen Trend geben muss, bei dem das Dativ-{e} an Boden verliert. Die Punktwolke der Graphik lässt kaum etwas von diesem langfristigen Trend der allmählichen Aufgabe des -{e}-Allomorphs erkennen.



Graphik: Anteil der Formen des dat.sg. mit -{e} (rA: relativer Anteil von -{e} an allen dat.sg.)

Statistische Erhebungen im Dienst der Sprachforschung: Nimmt man alles zusammen, so kann man feststellen, dass Behaghel der Gedanke nahe lag, linguistische Befunde statistisch abzusichern. Er hat dies mehrfach getan. Speziell seine Idee, die Wortstellung nach dem Prinzip „kurz vor lang" könne Gesetzescharakter haben, hat in der Quantitativen Linguistik wiederholt Beachtung gefunden. Seine Untersuchung zu den Adjektiven in Dramen von Schiller sollte aber ebenfalls aufgegriffen werden, da sie einen der seltenen Fälle darstellt, in denen die sprachliche Entwicklung eines Idiolekts verfolgt werden kann. (Zur Entwicklung von Idiolekten vgl. Best 2003; Kohlhase 1983.) Beim Dativ-{e} ist es ihm gelungen, verschiedene Faktoren nachzuweisen, die Einfluss darauf hatten, ob dieses Allomorph gewählt wurde oder statt seiner das Null-Allomorph.

Literatur

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

Behaghel, Otto (1900). Das -e im Dativ der Einzahl männlicher und sächlicher Hauptwörter. *Zeitschrift des allgemeinen deutschen Sprachvereins, Wissenschaftliche Beihefte 17-18, 251-277.*

Behaghel, Otto (1905). Zum Gebrauch des Beiworts bei Schiller. *Zeitschrift des Allgemeinen Deutschen Sprachvereins, Wissenschaftliche Beihefte, Vierte Reihe, Heft 26, 180-198.* (Auch in: Behaghel 1927: 108-130.)

Behaghel, Otto (1909). Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanische Forschungen 25, 110-142.*

Behaghel, Otto (1909b). Der Dativ der Einzahl männlicher und sächlicher Hauptwörter. *Zeitschrift des Allgemeinen Deutschen Sprachvereins 24, Nr. 2, 33-39* (Auch in: Behaghel 1927: 305-314).

Behaghel, Otto (1912). Wortstellung und Rhythmus. *Magyar Nyelvör XLI, 18-21* (Auch in Behaghel 1927: 281-284).

Behaghel, Otto (1918). Die Verdeutschungsbestrebungen und die Preußische Akademie der Wissenschaften. *Deutscher Wille (Kunstwart) XXXI, 13-16* (Auch in: Behaghel 1927: 353-358).

Behaghel, Otto (1919). Wieder einmal vom e. *Zeitschrift des Allgemeinen Deutschen Sprachvereins 54, Nr. 6, 97-100* (Auch in Behaghel 1927: 314-318).

Behaghel, Otto (1927). *Von deutscher Sprache. Aufsätze, Vorträge und Plaudereien.* Lahr in Baden: Druck und Verlag Moritz Schauenburg.

Behaghel, Otto (1930). Von deutscher Wortstellung. *Zeitschrift für Deutschkunde 1930, 81-89.*

Behaghel, Otto (1953). *Neue deutsche Biographie.* Hrsg. von der Historischen Kommission bei der Bayerischen Akademie der Wissenschaften. Erster Band: Aachen – Behaim (S. 747-748). Berlin: Duncker & Humblot.

Best, Karl-Heinz (2003). Zum Wandel von Idiolekten. *Naukovyj Visnyk ≠ernivec'koho Universytetu: Hermans'ka filolohija. Vypusk 165-166, 36-43.*

Duden. Die Grammatik. 7., völlig neu erarbeitete und erweiterte Auflage. Mannheim/ Leipzig/ Wien/ Zürich: Dudenverlag 2005.

Fenk-Oczlon, Gertraud (1989). Word frequency and word order in freezes. *Linguistics 27, 517-556.*

Fenk-Oczlon, Gertraud, & Fenk, August (2002). Zipf's Tool Analogy and Word Order. *Glottometrics 5, 22-28.*

Guiraud, Pierre (1954). *Bibliographie critique de la statistique linguistique.* Utrecht/ Anvers: Editions Spectrum.

Hartweg, Frédéric, & Wegera, Klaus-Peter (1989). *Frühneuhochdeutsch. Eine Einführung in die deutsche Sprache des Spätmittelalters und der frühen Neuzeit.* Tübingen: Niemeyer.

Hoffmann, Christine (1999). Word Order and the Principle of „Early Immediate Constituents" (EIC). *Journal of Quantitative Linguistics 6, 108-116.*

Karstien, C. (1924). Verzeichnis der Schriften von Otto Behaghel 1876-1923. In: Horn, Wilhelm (Hrsg.), *Beiträge zur Germanischen Sprachwissenschaft. Festschrift für Otto Behaghel* (S. 1-34). Heidelberg: Winter.

Köhler, Reinhard 1995). *Bibliography of quantitative linguistics*. Amsterdam: John Benjamins.

Kohlhase, Jörg (1983). Die Entwicklung von *ward* zu *wurde* beim Nürnberger Chronisten Heinrich Deichsler. In: Best, Karl-Heinz, & Kohlhase, Jörg (Hrsg.), *Exakte Sprachwandelforschung* (S. 103-106). Göttingen: edition herodot.

Meier, Helmut (²1967). *Deutsche Sprachstatistik*. 2., erw. u. verb. Aufl. Hildesheim: Olms.

Stroh, Fritz (1934). Otto Behaghels Schriften. Bücher, Abhandlungen, Aufsätze, Vorträge und Besprechungen 1924-1933. In: Goetze, Alfred, Horn, Wilhelm, & Maurer, Friedrich (Hrsg.), *Germanische Philologie. Ergebnisse und Aufgaben. Festschrift für Otto Behaghel* (S. 531-541). Heidelberg: Winter.

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

Vulanović, Relja, & Köhler, Reinhard (2005). Syntactic units and structures. In: Köhler, Reinhard, Altmann, Gabriel, & Piotrowski, Raimund (Hrsg.) (2005), *Quantitative Linguistik - Quantitative Linguistics. Ein internationales Handbuch* (S. 274-291). Berlin/ N.Y.: de Gruyter.

Wegera, Klaus-Peter (1987). *Flexion der Substantive.* Heidelberg: Winter (= Grammatik des Frühneuhochdeutschen. Beiträge zur Laut- und Formenlehre, Bd. 3. Hrsg. v. Hugo Moser, Hugo Stopp und Werner Besch.).

Wilpert, Gero von (Hrsg.) (1997). *Lexikon der Weltliteratur. Band 2. Biographisch-bibliographisches Handwörterbuch nach Autoren und anonymen Werken. L-Z.* München: Deutscher Taschenbuch Verlag.

Software

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

Karl-Heinz Best

XXIX. Paul Menzerath (1883-1954)

Geb. 1.1.1883 in Düren, gest. 8.4.1954. Besuch verschiedener Schulen in Düren bis 1903. Studium in Freiburg, Berlin, Marburg, Würzburg und Kiel bis 1908. Menzerath war u.a. in Marburg Schüler des Indogermanisten und Neogräzisten Albert Thumb (Thumb 1911, 10) und des Psychologen N. Ach. Promotion in Würzburg 1906 in Philosophie, vergleichender Sprachwissenschaft und klassischer Philologie. Nach Aufenthalten in Kiel (1907), Genf und Paris (1908) und Brüssel (1908: Institut de Sociologie Solvay) wurde er 1908 Leiter des neugegründeten Psychologischen Instituts Fort Jacco bei Uccle (Belgien) und a.o. Prof. an der Universität Brüssel, 1914 aus Belgien ausgewiesen. 1914 Militärdienst, 1915 Lektor für französische Sprache in Bonn, 1916 a.o. Prof. für Psychologie in Gent, 1917 o. Prof. in Gent, 1918 wieder ausgewiesen. 1918 Lektor für französische Sprache in Bonn, 1920 Habilitation für Psychologie und Phonetik in Bonn, 1921 Gründung des „Phonetischen Laboratoriums“ in Bonn, dessen erster Leiter als a.o. Prof. Menzerath wurde (später umbenannt in „Institut für Phonetik“, 1951 erneute Umbenennung in „Institut für Phonetik und Kommunikationsforschung“). 1930 Vorsitzender der 1. Tagung der Internationalen Gesellschaft für experimentelle Phonetik in Bonn, 10.-14.6. Das Institut wurde am 18.10.1944 bei einem Bombenangriff völlig zerstört. 1946 Diäten-Dozent in Bonn. Wiederaufbau des Instituts. 1949 vom Entnazifizierungs-Hauptausschuss als „entlastet“ eingestuft. 1951 Emeritierung als o.Prof.
Menzeraths Leistungen als Phonetiker werden in der Literatur hinreichend gewürdigt (Meyer-Eppler, Wodarz u.a.). Im Folgenden wird der Schwerpunkt auf die Aspekte seiner Arbeit gelegt, die für die Quantitative Linguistik interessant sind.

Thumb (1911, 10) verweist auf Menzeraths Untersuchungen zu Assoziationen, die von ihm, Thumb, angeregt wurden (1911, 68). Menzerath (1908) knüpft an Thumb & Marbes Untersuchung zur Analogie an (1901) und erforscht den Einfluss der Geläufigkeit auf die Reaktionszeit von Versuchspersonen in Assoziationsexperimenten. Thumb (1911: 11) definiert das „Geläufigkeitsgesetz“: „Je geläufiger eine Assoziation ist, desto kürzer ist die durchschnittliche Zeit, in der sie hervorgerufen wird.“ An Menzeraths Ergebnisse zu diesem Problem kann man als Modell

$$y = ae^{-bx}$$

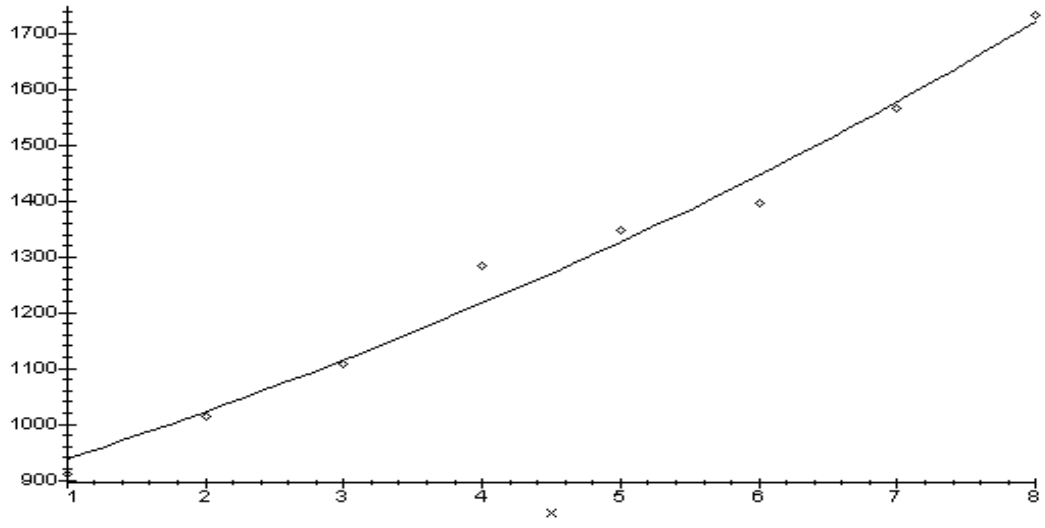
anpassen; dies entspricht einer der Versionen des sog. Menzerath-Altmann-Gesetzes (Altmann 1980: 3):

Tabelle 1
Abhängigkeit der Reaktionszeit von der Geläufigkeit

Geläufigkeitsgrad	Reaktionszeit beobachtet	Reaktionszeit berechnet	Geläufigkeitsgrad	Reaktionszeit beobachtet	Reaktionszeit berechnet
1	911.2	937.24	5	1348.5	1326.39
2	1015.2	1022.24	6	1397.0	1446.70
3	1107.1	1114.96	7	1565.5	1577.91
4	1284.6	1216.09	8	1732.2	1721.03
a = 859.2959 b = -0.0868 D = 0.98					

Die Geläufigkeitsgrade nehmen ab; Geläufigkeitsgrad 1 betrifft also die geläufigsten Einheiten, Geläufigkeitsgrad 2 die zweitgeläufigsten Einheiten, etc. Die Reaktionszeit wurde in 1/1000 Sekunden gemessen. a, b sind die Parameter des Modells; D ist der Determinationskoeffizient, der die Testbedingung $D \geq 0.80$ sehr gut erfüllt.

Thumb (1911: 11) weist darauf hin, dass die Assoziationsexperimente, die er mit Marbe durchgeführt hat, einen zunächst stärker, dann immer schwächer werdenden Anstieg der Reaktionszeit aufwies; bei Menzerath dagegen sei es ein linearer Anstieg gewesen. Eine mögliche Lösung dieses Widerspruchs lässt er offen. Mein Test zeigt eher eine kontinuierlich steigende Tendenz des Zusammenhangs. Allerdings kann man mit guten Ergebnissen auch noch weitere Varianten des Modells verwenden. Insofern ist hier nur ein Modell unter mehreren gefunden, das jedenfalls das Testkriterium erfüllt; es gibt aber noch andere Möglichkeiten. Eine theoretische Begründung, die eine Entscheidung zwischen den verschiedenen Modellen nahelegt, ist bisher noch nicht entwickelt.



Graphik zu Tabelle 1: Abhängigkeit der Reaktionszeit von der Geläufigkeit

1921 übernimmt Menzerath die Leitung des neugegründeten „Phonetischen Laboratoriums“ in Bonn und widmet sich entsprechend von da an der Phonetik. Wesentlich ist hier die Suche nach „phonischen Gesetzen“, die so genannt wurden, um eine Verwechslung mit dem „Laut-

gesetz“ der Sprachhistoriker zu vermeiden. Als entscheidend ist die Untersuchung von Menzerath & de Oleza (1928) anzusehen, eine Arbeit, in der die Autoren den „phonischen Gesetzen“ auf die Spur kommen: „Den Wirkungsbereich solcher Gesetze innerhalb der spanischen Lautquantität konnten wir in ganz ungeahnter Weise feststellen und damit eine Regelmäßigkeit entdecken, die einer nichtexperimentellen Phonetik niemals zugänglich wäre“ (Menzerath & de Oleza 1928: 9). Gegenstand ihrer Untersuchung sind – aus den Tabellen ermittelt – 1432 echte Wörter, die so ausgewählt wurden, dass die spanischen Laute mit allen möglichen Nachbarn in der Liste vertreten sind; sie enthalten 3888 Silben und 8440 Laute. (Menzerath & de Oleza 1928: 36 beziffern nur 7883 Laute als Gesamtbestand; die Differenz erklärt sich vermutlich dadurch, dass während der Untersuchung einige Wörter ergänzt werden mussten, da doch nicht alle phonetischen Erscheinungen in der Ausgangsliste der Testwörter vorhanden waren. Die Differenz lässt sich nicht aufklären, da die Wortlisten im Text nicht vollständig sind, wie die Zahl der einsilbigen Wörter zeigt: Es wurden 73 ausgewertet, aber in der Liste sind nur 41 aufgeführt.)

Die Wörter wurden nach der Zahl ihrer Silben in Worttypen I – VII eingeteilt, also 1-7silbige Wörter. Subklassifizierung der Worttypen danach, welche Silbe den Wortton trägt. IV,2 ist also die Gruppe der viersilbigen Wörter mit Wortton auf der zweiten Silbe. Die Wortlisten wurden nur von einer Versuchsperson, de Oleza (37 Jahre, stud.phil. an der Universität Bonn, Muttersprache: Kastilisch) gelesen (Menzerath & de Oleza 1928: 10f.). Zuerst wurde die Dauer sämtlicher 7883 Laute addiert und dann mittels Division durch diese Zahl der „absolute[1] [...] Durchschnittswert“ mit 13.03 Hundertstelsekunden für die Dauer spanischer Laute bestimmt. Anschließend wurde die relative Dauer der Laute mit 13 = 1 festgelegt (Menzerath & de Oleza 1928: 36). Man kann also die absolute Dauer durch Multiplikation mit 13 jederzeit berechnen.

Menzerath & de Oleza (1928: 91) resümieren:

„Wir fanden, daß allgemein ein Laut um so kürzer ist, je länger das Wort ist oder je mehr Silben das Wort hat, zu dem er gehört. Diese Feststellung erhält die Bezeichnung: **phonisches Quantitätsgesetz**. Diesem 1. allgemeinen Gesetz treten zwei weitere phonische Quantitätsgesetze an die Seite, nämlich 1. die Feststellung, daß das Quantitätsgesetz ebenfalls Geltung hat im einzelnen Worttypus, d.h. steigt die Lautzahl in Wörtern gleicher Silbenzahl, so nimmt die Dauer des Lautes ab; 2. die Feststellung, daß das Quantitätsgesetz ebenfalls gilt für die Silbe, d.h. innerhalb jedes beliebigen Wortes oder Worttypus ist die lautreichere Silbe auch die relativ kürzere.“

Die folgenden Tabellen zeigen die Ergebnisse der Untersuchung von Menzerath & de Oleza (1928), einschließlich einer Anpassung der einfachsten Form des Menzerath-Altmann-Gesetzes

$$y = ax^{-b}$$

(Altmann 1980: 3), die in allen Fällen mit sehr guten Ergebnissen gelingt:

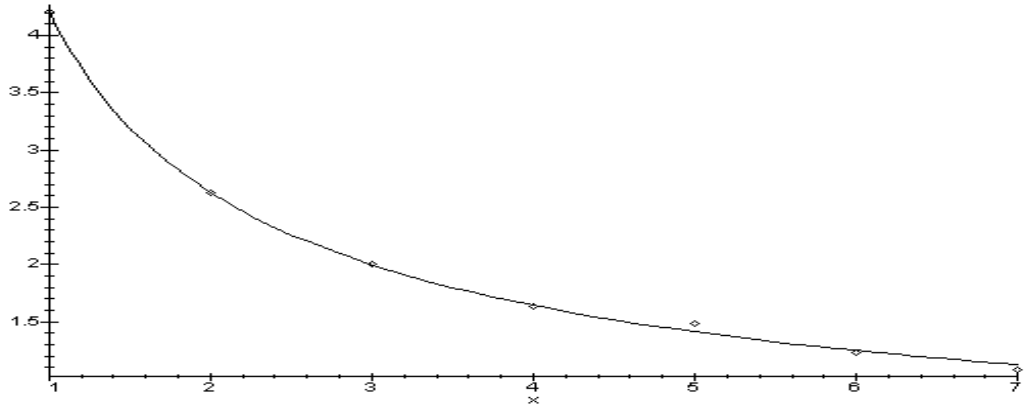
[1] Unterstrichen bedeutet hier und im Folgenden: im Original gesperrt gedruckt.

Tabelle 2
Abhängigkeit der Silbendauer von der Zahl der Silben pro Wort

Worttyp	Silbendauer beobachtet	Silbendauer berechnet	Worttyp	Silbendauer beobachtet	Silbendauer berechnet
I	4.2011	4.2039	V	1.4842	1.4115
II	2.6261	2.6274	VI	1.2248	1.2473
III	2.0012	1.9958	VII	1.0809	1.1235
IV	1.6283	1.6421			
$a = 4.2039$ $b = 0.6781$ $D = 0.9989$					

(Quelle: Tab. XXIX, Figur V.)

a, *b*: Parameter des Modells; *D*: Determinationskoeffizient, der mit $D > 0.80$ in allen Fällen ein gutes Ergebnis anzeigt. „Dauer“ bedeutet immer 1/13 der tatsächlich gemessenen Dauer in Hunderstelsekunden.



Graphik zu Tabelle 2: Abhängigkeit der Silbendauer von der Zahl der Silben pro Wort

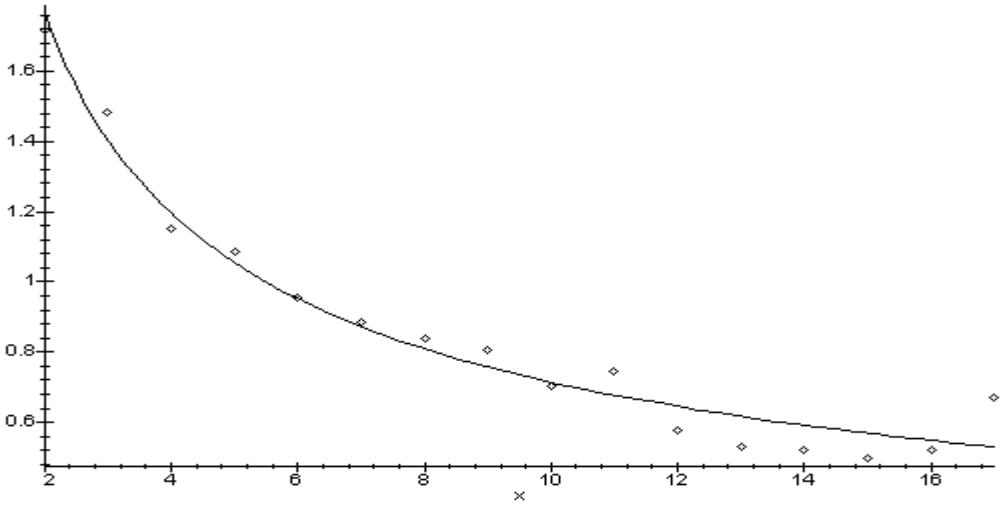
Tabelle 3
Abhängigkeit der Lautdauer von der Zahl der Laute pro Wort

Wortlänge in Lauten	Lautdauer beobachtet	Lautdauer berechnet	Worttyp	Lautdauer beobachtet	Lautdauer berechnet
2	1.7138	1.7617	10	0.7014	0.7148
3	1.4810[2]	1.4036	11	0.7456	0.6776
4	1.1502	1.1946	12	0.5795	0.6453
5	1.0879	1.0541	13	0.5328	0.6170
6	0.9549	0.9517	14	0.5232	0.5919
7	0.8836	0.8730	15	0.4987	0.5695
8	0.8409	0.8100	16	0.5195	0.5492
9	0.8054	0.7583	17	0.6728	0.5309
$a = 2.5981$ $b = 0.5605$ $D = 0.9672$					

(Quelle: Tab. XXV, Figur I.)

[2] Letzte Ziffer fehlt im Original; <0> von mir ergänzt.

Obwohl nur ein 17lautiges Wort vorkommt, ist die Anpassung des Modells insgesamt sehr gut.

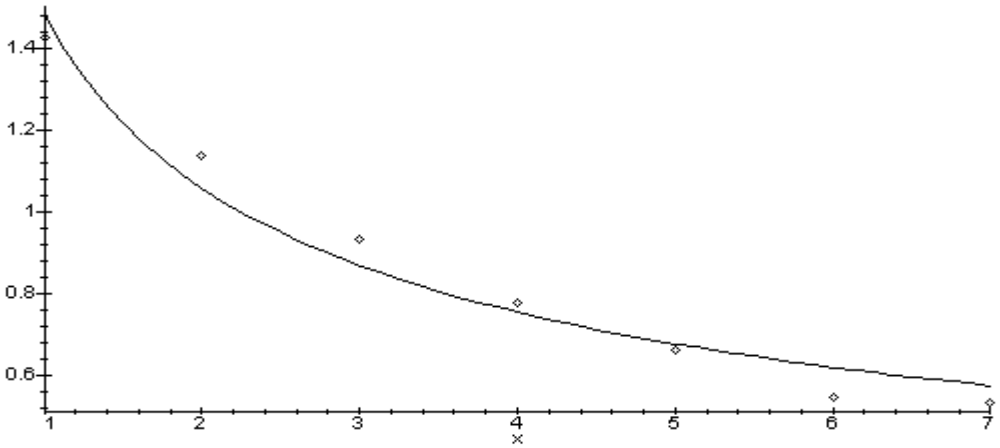


Graphik zu Tabelle 3: Abhängigkeit der Lautdauer von der Zahl der Laute pro Wort

Tabelle 4
Abhängigkeit der Lautdauer von der Zahl der Silben pro Wort

Worttyp	Lautdauer beobachtet	Lautdauer berechnet	Worttyp	Lautdauer beobachtet	Lautdauer berechnet
I	1.4264	1.4797	V	0.6636	0.6788
II	1.1388	1.0579	VI	0.5499[3]	0.6215
III	0.9319[4]	0.8693	VII	0.5335	0.5768
IV	0.7809	0.7563			
$a = 1.4797$ $b = 0.4842$ $D = 0.9675$					

(Quelle: Tab. XXVIII, Figur III.)



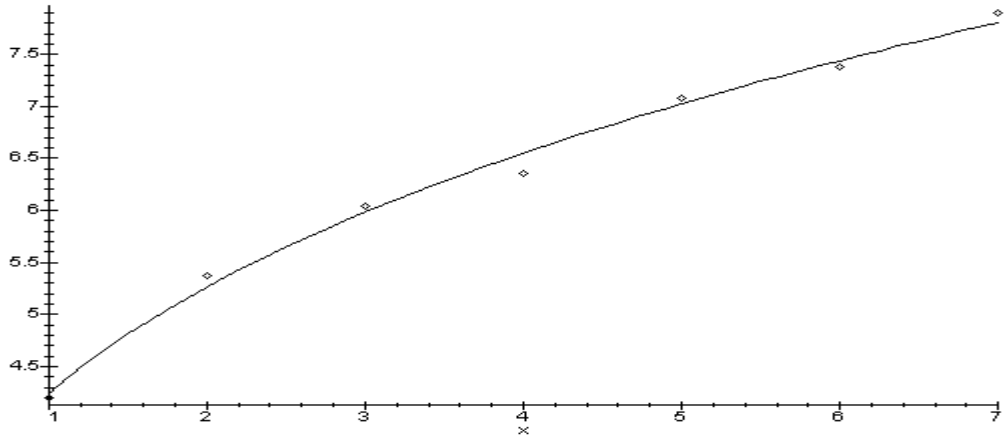
Graphik zu Tabelle 4: Abhängigkeit der Lautdauer von der Zahl der Silben pro Wort

[3] Auf vier Stellen von mir gerundet.
[4] Auf vier Stellen von mir gerundet.

Tabelle 5
Abhängigkeit der Wortdauer von der Zahl der Silben pro Wort

Worttyp	Wortdauer beobachtet	Wortdauer berechnet	Worttyp	Wortdauer beobachtet	Wortdauer berechnet
I	4.2011	4.2367	V	7.0761	7.0252
II	5.3723	5.2676	VI	7.3742	7.4394
III	6.0368	5.9834	VII	7.8961	7.8087
IV	6.3548	6.5495			
$a = 4.2367$		$b = -0.3142$		$D = 0.9930$	

(Quelle: Tab. XXXII, Figur X.)

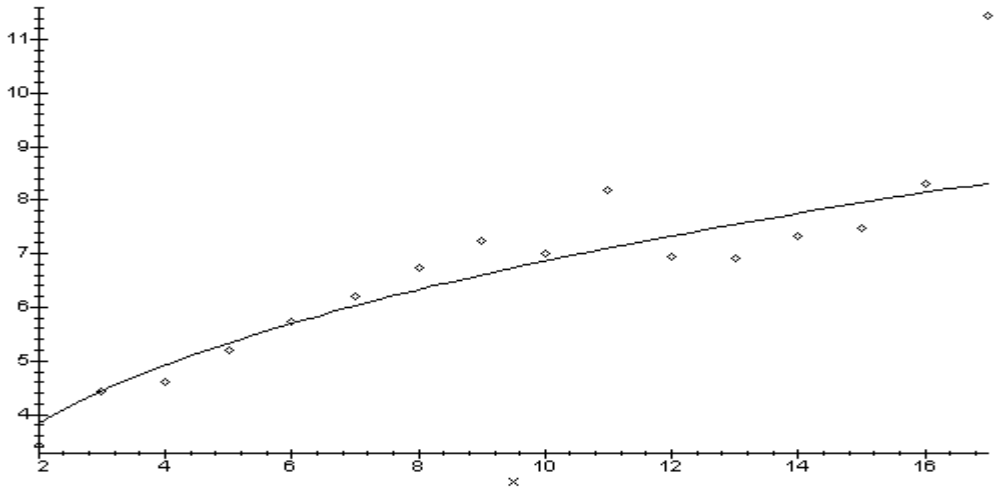


Graphik zu Tabelle 5: Abhängigkeit der Wortdauer von der Zahl der Silben pro Wort

Tabelle 6
Abhängigkeit der Wortdauer von der Zahl der Laute pro Wort

Wortlänge in Lauten	Wortdauer beobachtet	Wortdauer berechnet	Worttyp	Wortdauer beobachtet	Wortdauer berechnet
2	3.4277	3.8291	10	7.0141	6.8623
3	4.4426	4.4353	11	8.2025	7.1035
4	4.6012	4.9228	12	6.9551	7.3311
5	5.1951	5.3376	13	6.9275	7.5470
6	5.7259	5.7023	14	7.3254	7.7525
7	6.2094	6.0300	15	7.4807	7.9488
8	6.7337	6.3290	16	8.3122	8.1369
9	7.2499	6.6051	17	11.4384	-
$a = 2.9783$		$b = -0.3625$		$D = 0.8920$	

(Quelle: Tab. XXX, Figur VIII.)



Graphik zu Tabelle 6: Abhängigkeit der Wortdauer von der Zahl der Laute pro Wort

Menzerath & de Oleza (1928: 73) erklären, die 10-13lautigen Wörter seien zu selten, um als repräsentativ gelten zu können. Dies gilt aber vielmehr für das einzige 17lautige Wort; es wurde als Ausreißer betrachtet und bei der Berechnung nicht berücksichtigt. Die berechneten Werte beruhen auf der gesamten Tabelle mit Ausnahme des einen 17lautigen Wortes. Lässt man zusätzlich auch die 10-13lautigen Wörter weg, erhält man mit $D = 0.9429$ allerdings ein deutlich besseres Ergebnis.

Menzerath (1928) erklärt die Suche nach „phonischen Gesetzen" als eine „aussichtsreiche und zudem reizvolle Aufgabe". Dazu gehört das „Quantitätsgesetz": „In der Untersuchung über ‚Spanische Lautdauer' ...konnten wir nachweisen, dass ein Laut um so kürzer wird, je grösser das mit ihm verbundene Lautganze ist (Quantitätsgesetz). Es liess sich sogar zeigen, dass dies Gesetz Geltung noch innerhalb desselben Worttypus hat, so z.B. dass ein dreisilbiges Wort von 8 Lauten relativ länger dauert als ein gleiches Wort von 9 oder mehr Lauten; ja selbst für die Silbe trifft das Gesetz zu, insofern die lautreichere Silbe auch die relativ kürzere ist, und umgekehrt" (Menzerath 1928: 104). Im gleichen Artikel heißt es zu den deutschen Vibranten: „Das Quantitätsgesetz trifft ausnahmslos zu: je größer die Schlagzahl, um so kürzer ist die relative Schlagdauer" (Menzerath 1928: 105).

Das Gesetz hat eine Vorgeschichte (Menzerath & de Oleza 1928: 3ff.), die in die Phonetik des 19. Jahrhunderts zurückreicht (frühe Vorstellungen dazu: Sievers 1876: 122; als allgemeineres Prinzip: Sievers 1893: 240f.; Grégoire 1899; Altmann & Schwibbe 1989: 60; Best 2006); auch in der Literaturwissenschaft des frühen 20. Jahrhunderts stößt man bei Siegfried Behn auf ganz ähnliche Aussagen (Behn 1912: 97; Best 2006b). Die Verbesserung gegenüber diesen frühen Versuchen charakterisiert Menzerath (1936: 246f.) wie folgt: „Hatten ältere Forscher auf Grund experimenteller Befunde nur behaupten können, dass der *Vokal* mit wachsender Lautgruppe (Wortlänge z.B.) verkürzt wird, so wissen wir jetzt, dass unter diesen Umständen *jeder* Laut schlechthin kürzer wird. Ein zwölflautiges Wort dauert also nicht viermal so lange wie ein dreilautiges, sondern nur etwa doppelt so lange. Das soll als *‚allgemeines Laut-Quantitätsgesetz*' bezeichnet werden."

Menzerath (1942a) befasst sich mit französischer Dichtung. Er bearbeitet die Ballade *Mort de Jeanne d'Arc* von Casimir Delavigne (1793-1843), indem er Hebungen und Senkungen markiert, die Silbenzahl je Zeile feststellt und für jede Zeile den Reim bestimmt. Fasst man einmal die Verteilung der Verslängen als Diversifikationsphänomen (Altmann 1991) auf,

dann kann man an die Daten, die sich aus Menzeraths Bearbeitung gewinnen lassen, die 1-verschobene Poisson-Binomial-Verteilung

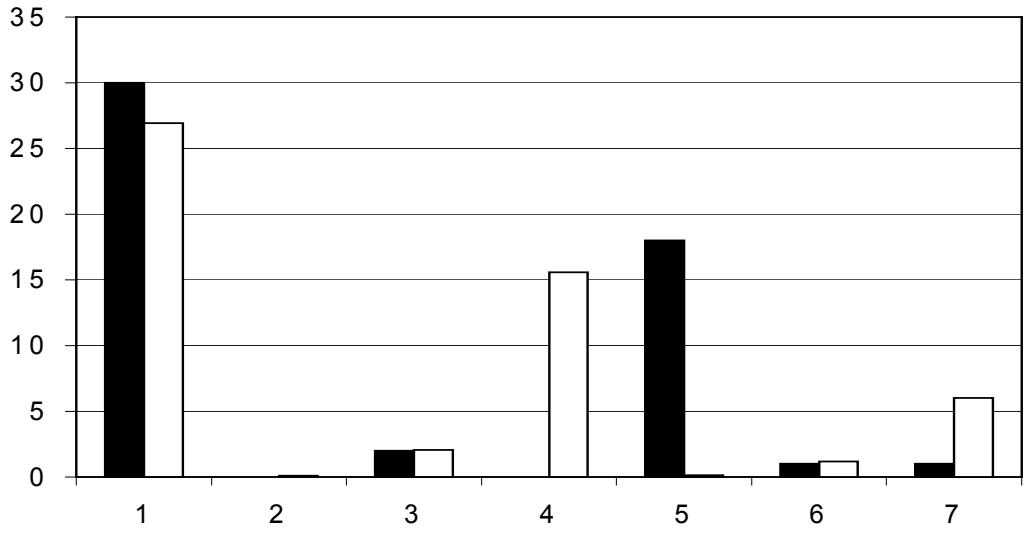
$$P_x = e^{-a} \sum_{j=0}^{\infty} \binom{nj}{x-1} p^{x-1} q^{nj-x+1} \frac{a^j}{j!}, \quad x = 1,2,...$$

anpassen (s. Tab. 7).

Tabelle 7
Silbenzahl je Verszeile in einer französischen Ballade

x	Silben pro Zeile	n_x	NP_x
1	12	30	26.93
2	11	0	0.09
3	10	2	2.05
4	9	0	15.59
5	8	18	0.13
6	7	1	1.19
7	6	1	6.03
$a = 0.6579$	$n = 3$	$p = 0.9580$	$X^2 = 4.8813$ $FG = 1$	$P = 0.0271$

Legende zu Tabelle 7: x: Verslängenklasse; n_x: beobachtete Zahl der Verse der Länge x; NP_x: berechnete Zahl der Verse der Länge x; X^2: Chiquadrat; FG: Freiheitsgrade; a, n, p: Parameter der Verteilung; P: Überschreitungswahrscheinlichkeit des Chiquadrats. Die Anpassung der Poisson-Binomial-Verteilung ist mit $P = 0.0271$ schwach, aber nicht ganz als gescheitert anzusehen. Da die Ballade nur 52 Verszeilen hat, ist sie auch für solche Tests nicht als besonders geeignet anzusehen.



Graphik zu Tabelle 7: Silbenzahl je Verszeile in einer französischen Ballade

In der Quantitativen Linguistik ist bisher Menzeraths *Die Architektonik des deutschen Wortschatzes* (1954) am ausführlichsten rezipiert worden. Das Buch wurde, wie Menzerath selbst im Vorwort des Buches berichtet, 1943 in Schweden konzipiert und auch im Wesentlichen in diesem Jahr fertiggestellt, konnte aber erst 1954 veröffentlicht werden, wobei in der Zwi-

schenzeit erfolgte Kritik und Paralleluntersuchungen berücksichtigt wurden. Das Buch ist das Ergebnis der sprachtypologischen Bemühungen, deren Aufgabe „die Untersuchung des strukturellen Aufbaues der Wörter und des Wortschatzes einer Sprache" ist, wobei sich herausstellt, „dass jede Sprache eine charakteristische Struktur besitzt (Menzerath & Meyer-Eppler 1950: 54). Das Hauptziel dieser Sprachtypologie besteht darin, „aus den statistischen Daten sprachtypologische Gesetze abzuleiten" (Menzerath & Meyer-Eppler 1950: 56). Schon in Menzerath (1944: 77) nennt er als wichtiges Ergebnis: „Stellt man dann aber für die mehrsilbigen Wörter die entsprechenden Lautzahlen tabellarisch zusammen, so findet man, daß die relative Lautzahl mit steigender Silbenzahl abnimmt." Außerdem erklärt er: „Bezogen auf die Lautzahl ist das Siebenlautwort das häufigste; alle anderen Werte ordnen sich im Verhältnis dazu nach dem bekannten Gaußschen ‚Verteilungsgesetz' an. Dazu zeigen sich weitere erstaunliche Gesetzmäßigkeiten, beispielsweise im Verhältnis von Formtyp und Lautzahl. In allen Gruppen bestätigt sich übereinstimmend das vorgenannte Verteilungsgesetz" (Menzerath 1944: 77). Zugrunde liegt die statistische Auswertung der Stichwörter eines Lexikons im Hinblick auf ihre Silbenstrukturen. Menzerath (1954: 101) formuliert die allgemeine Hypothese: *„Je größer das Ganze, um so kleiner die Teile"*. Zwischen 1948 und 1954 wurden in Bonn in diesem Zusammenhang 6 Dissertationen erarbeitet, die Wodarz (1974: 202) auflistet, ergänzt um den Hinweis, dass die Arbeiten keinem Druckzwang unterlagen und daher fast unbekannt blieben; allerdings findet man einige Ergebnisse dieser Arbeiten in Menzerath (1954: 112-121, passim):

Schönhage, A. (1948). *Zur Struktur des französischen Wortschatzes. Der französische Einsilber.*

Feuser, Margot (1948). *Das einsilbige Wort im Englischen. Eine sprachstatistische Strukturuntersuchung.*

Rosić, M.S. (1950). *Zur Struktur des serbokroatischen Wortschatzes. Eine typologische Untersuchung der einsilbigen Wörter.*

Gajić, Dragomir M. (1950). *Zur Struktur des serbokroatischen Wortschatzes. Eine Typologie der serbokroatischen mehrsilbigen Wörter.*

Rettweiler, Hildegard (1950). *Die Stichprobenentnahme bei sprachtypologischen Untersuchungen, als Problem nachgeprüft an der italienischen Sprache.*

Miron, Paul (1954) *Zur typologischen Struktur des Rumänischen.*

In einer ausführlichen Rezension kritisiert Vértes (1955), dass Menzerath (1954) den Wortschatz ohne Rücksicht auf Wortbildungstyp und Herkunft behandelt und steuert eigene Untersuchungen zum Ungarischen bei.

Altmann (1980) gab der Menzerathschen Hypothese eine mathematische Form, wendete sie generell auf die Sprache an und führte sie in seiner allgemeineren Formulierung als Menzerath-Gesetz wie folgt aus: *„Je größer ein sprachliches Konstrukt, desto kleiner seine Konstituenten"* (Altmann & Schwibbe 1989: 5). Dieses Sprachgesetz ist daher auch verbreitet als Menzerath-Altmann-Gesetz bekannt (Aichele 2005; Cramer 2005) und hat sich in vielfacher Weise bewährt, im Sprachsystem ebenso wie in seiner Verwendung (Altmann & Schwibbe 1989; Asleh & Best 2005). Hřebíček (1997) konnte am Beispiel türkischer Texte zeigen, dass dieses Gesetz die Strukturierung der Sprache von den größten bis zu den kleinsten Einheiten organisiert; selbst die Schrift unterliegt diesem Gesetz (Bohn 1998: 8ff.; Prün 1994). Es bildet daher die ‚vertikale' Achse eines Modells der Sprache (Best 2003: 128). Dabei ist zu beachten, dass das Gesetz für das Verhältnis zwischen sprachlichen Konstrukten und ihren direkten Konstituenten gilt. Betrachtet man den Zusammenhang zwischen Konstrukten und ihren indirekten Konstituenten, dann muss dies bei der Modellbildung berücksichtigt werden, so wie das im Falle des sog. Arens-Gesetzes geschehen ist (Altmann 1983; Best 2006a).

Nach der frühen Phase, in der Menzerath einen Beitrag zum Geläufigkeitsgesetz geliefert hat, widmet er sich die längere Zeit seines Forscherlebens den phonischen Gesetzen. Beide Phasen sind von der Suche nach Gesetzmäßigkeiten geprägt. Die frühe Phase ist für die Quantitative Linguistik erst noch zu entdecken; aus der späteren verdienen die Arbeiten ab ca. 1928 verstärkte Beachtung. Ihr eigentliches Gewicht erhalten Menzeraths Forschungen m.E. aber vor allem dadurch, dass es Altmann (1980) gelungen ist zu zeigen, dass seine Ergebnisse theoretisch begründet und mit den Mitteln der Statistik überprüft werden können.

Das folgende Literaturverzeichnis nimmt auch Arbeiten auf, die nicht zitiert wurden, um einen Überblick über seine Forschungen zu geben. Menzeraths Werke sind oft nicht ohne Mühe erreichbar, was z.T. an unzulänglichen, manchmal auch falschen bibliographischen Angaben liegt. Das gilt besonders für einige der frühen Publikationen. Mit Stern* gekennzeichnete Arbeiten konnten bisher nicht direkt eingesehen werden. (Wichtige, aber sehr unvollständige Quellen: Miron 1956; Simon 2006.)

Literatur

Aichele, Dieter (2005). Quantitative Linguistik in Deutschland und Österreich. In: Köhler, Reinhard, Altmann, Gabriel, & Piotrowski, Rajmund G. (Hrsg.), *Quantitative Linguistik - Quantitative Linguistics. Ein internationales Handbuch* (S. 16-23). Berlin/ N.Y.: de Gruyter.

Altmann, Gabriel (1980). Prolegomena to Menzerath's law. In: Grotjahn, Rüdiger (Ed.). *Glottometrika 2* (S. 1-10). Bochum: Brockmeyer.

Altmann, Gabriel (1983). H. Arens' „Verborgene Ordnung" und das Menzerathsche Gesetz. In: Faust, Manfred, Harweg, Roland, Lehfeldt, Werner, & Wienold, Götz (Hrsg.); *Allgemeine Sprachwissenschaft, Sprachtypologie und Textlinguistik. Festschrift für Peter Hartmann* (S. 31-39). Tübingen: Narr.

Altmann, Gabriel (1991). Modelling diversification phenomena in language. In: Rothe, Ursula (Hrsg.), *Diversification Processes in Language: Grammar* (S. 33-46). Hagen: Margit Rottmann Medienverlag.

Altmann, Gabriel, & Schwibbe, Michael H. (1989). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen.* Hildesheim/ Zürich/ New York: Olms.

Asleh, Laila, & Best, Karl-Heinz (2005). Zur Überprüfung des Menzerath-Altmann-Gesetzes am Beispiel deutscher (und italienischer) Wörter. *Göttinger Beiträge zur Sprachwissenschaft 10/11, 9-19.*

Behn, Siegfried (1912). *Der deutsche Rhythmus und sein eigenes Gesetz. Eine experimentelle Untersuchung.* Straßburg: Trübner.

Best, Karl-Heinz (²2003b). *Quantitative Linguistik. Eine Annäherung*. 2., überarbeitete und erweiterte Auflage. Göttingen: Peust & Gutschmidt.

Best, Karl-Heinz (2006). Eduard Sievers (1850-1932). In Arbeit.

Best, Karl-Heinz (2006a). Hans Arens (1911-2003). *Glottometrics 13, 75-79.*

Best, Karl-Heinz (2006b). Siegfried Behn (1884-1970). *Glottometrics 13, 85-88.*

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

Cramer, Irene M. (2005). Das Menzerathsche Gesetz. In: Köhler, Reinhard, Altmann, Gabriel, & Piotrowski, Rajmund G. (Hrsg.), *Quantitative Linguistik - Quantitative Linguistics. Ein internationales Handbuch* (S. 659-688). Berlin/ N.Y.: de Gruyter.

Grégoire, A. (1899). Variations de durée de la syllabe française suivant sa place dans les groupements phonétiques. *La parole I, Heft 3, 161-176; Heft 4, 263-280; Heft 6, 418-433.*

Hřebíček, Luděk (1997). *Lectures on Text Theory.* Prag: Oriental Institute of the Academy of Sciences of the Czech Republic.

Hug, Marc (oJ., 2003 oder später). La loi de Menzerath appliquée à un ensemble de textes. cavi. univ-paris3.fr/lexicometrica/article/numero5/lexicometrica-hug.pdf.

***Ley, August, & Menzerath, Paul** (1911). *L'Ètude expérimentale de l'association des idées dans les maladies mentales: VIe Congrès belge de Neurologie et de Psychiatrie, Bruges 1911: Rapport de Psychologie.* Gant: Imprimerie A. van der Haeghen.

***Ley, August, & Menzerath, Paul** (1913). *Le témoignage des normaux et des aliénés. Rapport.* Bruxelles: Imprimerie médicale et scientifique Severeyns.

Menzerath, Paul (1908). Die Bedeutung der sprachlichen Geläufigkeit oder der formalen sprachlichen Beziehung für die Reproduktion. *Zeitschrift für Psychologie und Physiologie der Sinnesorgane. I. Abteilung: Zeitschrift für Psychologie, Bd. 48, 1-94* (= Diss., Würzburg).

Menzerath, Paul (1909). Psychologische Untersuchungen über die sprachliche Kontamination. *Zeitschrift für angewandte Psychologie und psychologische Sammelforschung 2, 280-290.*

***Menzerath, Paul** (1912). Contribution à la psycho-analyse. *Archives de psychologie 12, 372-389.*

***Menzerath, Paul** (1912). Le genre grammatical. *Bulletin de la Société d'Anthropologie de Bruxelles XXXI, 163-184.*

***Menzerath, Paul** (1912). A propos des calculateurs prodiges. *Bulletin de la Société d'Anthropologie de Bruxelles XXXI, 229-234.*

Menzerath, Paul (1913). The Association Method in Criminal Procedure. *Journal of the American Institute of Criminal Law and Criminology 4, 58-66.* (Übersetzt von William S. Forster.)

***Menzerath, Paul** (1913). Les illusions optiques - Discussion. *Bulletin de la Société d'Anthropologie de Bruxelles XXXII, 37-89.*

***Menzerath, Paul** (1913). Les légendes étiologiques. *Bulletin de la Société d'Anthropologie de Bruxelles XXXII, 222-242.*

***Menzerath, Paul** (1913). Un phénomène d'optique paradoxal. *Bulletin de la Société d'Anthropologie de Bruxelles XXXII, 35-37.*

***Menzerath, Paul** [Hrsg.] (1926*). Beiheft zur deutschen Lauttafel: mit verkleinerter Lauttafel und 1 Abbildung.* Bonn: Marcus & Weber.

Menzerath, Paul, & Evertz, Erich (1927/ 28). Atem und Lautdauer. *Teuthonista 4, 114-124, 204-214.*

Menzerath, Paul (1928). Über einige phonetische Probleme. *Actes du premier congrès international de linguistes à La Haye, du 10 – 15 avril 1928* (S. 104-105). Leiden: Sijthoff.

Menzerath, Paul, & de Oleza, Joseph M. (1928). *Spanische Lautdauer. Eine experimentelle Untersuchung.* Berlin/ Leipzig: de Gruyter.

Menzerath, Paul (1928/29). Vokalquantität und Dialektgeographie. *Teuthonista 5, 208-212.*

Menzerath, Paul (1929). Assimilation und Nasalierung. Ein experimenteller Versuch. In: *Donum Natalicium Schrijnen. Verzameling van Opstellen door Out-leerlingen en bevriende Vakgenooten opgedragen aan Mgr. Prof. Dr. Jos. Schrijnen. Bij Gelegenheid van zijn zestigsten Verjaardag, 3 Mei 1929* (S. 63-68). Nijmegen – Utrecht: N.V. Dekker & VAN DE VEGT.

Menzerath, Paul [Hrsg.] (1930). *Bericht über die I. Tagung der Internationalen Gesellschaft für experimentelle Phonetik in Bonn vom 10. bis 14. Juni 1930.* Bonn: Bonner Universitäts-Buchdruckerei Gebr. Scheur.

Menzerath, Paul (1933). Was ist Akzent? *Le Maître Phonétique, Bd. 48, Nr. 41, 2-3.*[5]

***Menzerath, Paul** (1933). Zur deutschen Umschrift. *Le Maître Phonétique, Bd. 48, Nr. 41.*

Menzerath, Paul, & de Lacerda, A. (1933). *Koartikulation, Steuerung und Lautabgrenzung. Eine experimentelle Untersuchung.* Berlin/ Bonn: Dümmler. (= Phonetische Studien, 1)

***Menzerath, Paul** (1934). Beobachtungen zur deutschen Lautquantität. *Le Maître Phonétique, Bd. 49, Nr. 47-48.*

Menzerath, Paul (1934). Zur deutschen Lautquantität. *Teuthonista 10, 238-248.*

***Menzerath, Paul** (1935). Die Chromographie, eine neue Registriermethode. *Geistige Arbeit (5. Jan. 1935), Nr. 1, S. 6.*

Menzerath, Paul (1935). Lautabgrenzung und Wortstruktur. In: *Actes du 3ème congrès international de linguistes (Rome, 19-26 septembre 1933 – XI)/ Atti del III congresso internazionale dei linguisti (Roma, 19-26 settembre 1933 – XI)* (S. 59-66). Rédigés par Bruno Migliorini et Vittore Pisani. Florence: Felice Le Monnier.

***Menzerath, Paul** (1935). Phonetik im Sprachunterricht. *Geistige Arbeit 2, 18, (20.IX.1935), 5.*

Menzerath, Paul (1935). Der Stand der heutigen Lautwissenschaft. Betrachtungen zum II. Kongress für Phonetik und Phonologie (London 22.-26. Juli 135). *Wärbel. Dolgozatok a Debreceni Tudományegyetem Nyelvatlasz és Fonetikai Intézetéböl I. Kötet, I. Szám, 5-18.*

***Menzerath, Paul** (1935). Die „Stimmhaftigkeit". *Le Maître Phonétique, Nr. 50, S. 24-25.*

***Menzerath, Paul** (1935). Nochmals zur deutschen Umschrift. *Le Maître Phonétique, Bd. 50, Nr. 49, 2-5.*

***Menzerath, Paul** (1936). Eine anomale Artikulation des Zungen-*r*. *Archives néerlandaises de phonétique expérimentale 12, 69-70.*

Menzerath, Paul (1936). Die phonetische Struktur. Eine grundsätzliche Betrachtung. *Acta Psychologica I, 241-262.*

***Menzerath, Paul** (1936). Neue Untersuchungen zur Steuerung und Koartikulation. In: D. Jones & D.B. Firth (eds.), *Proceedings of the 2d international Congress of Phonetic Sciences, London 22-26 July 1935* (S. 220-225). Cambridge: University Press.

Menzerath, Paul (1937). Die Sprechartikulation als Struktur. *Forschungen und Fortschritte 13, 364-366.*

Menzerath, Paul (1937). Neue Untersuchungen zur Lautabgrenzung und Wortsynthese mit Hilfe von Tonfilmaufnahmen. In: *Mélanges de Linguistique et de Philologie offerts à Jacq. van Ginneken à l'Occasion du soixantième anniversaire de sa naissance (21 avril 1937)* (S. 35-41). Paris: Klincksieck.

Menzerath, Paul (1938). Neue Untersuchungen zur Wortartikulation. *Actes du 4e congrès international des linguistes, Copenhague 1936* (S. 67-75). Kopenhagen: Munksgaard.

Menzerath, Paul (1940). Der Diphthong, sein Wesen und sein Aufbau. *Forschungen und Fortschritte 16, 209-210.*

Menzerath, Paul (1941). *Der Diphthong. Eine kritische und experimentelle Untersuchung.* Bonn/ Berlin: Dümmler. (= Phonetische Studien, 2)

Menzerath, Paul (1942). Gedanken über Kern- und Wendepunkt in der Phonetik. Aus Anlaß von J. Forchhammer: „Die Sprachlaute in Wort und Bild" (Heidelberg, Winter 1942). *Archiv für Vergleichende Phonetik 6, 89-102.*

Menzerath, Paul (1942a). Die Polyrhythmie des französischen Verses. *Archiv für Vergleichende Phonetik 6, 1-15.*

Menzerath, Paul (1944). Zum Aufbau des deutschen Wortschatzes. *Forschungen und Fortschritte 20, 76-77.*

[5] Für die Überprüfung einiger Beiträge in *Le Maître Phonétique* danke ich Reinhard Köhler; weitere Angaben nach *Indogermanisches Jahrbuch.*.

Menzerath, Paul (1948). Zur Reform der deutschen Orthographie. *Zeitschrift für Phonetik und Allgemeine Sprachwissenschaft 2, 38-43.*

Menzerath, Paul (1950). Typology of languages. *Journal of the Acoustical Society of America 22, 698-701.*

Menzerath, Paul (1951). Bemerkungen zu Lauri Posti: On Quantity in Estonian. *Zeitschrift für Phonetik und allgemeine Sprachwissenschaft 5, 247-252.*

Menzerath, Paul (1954). *Die Architektonik des deutschen Wortschatzes.* Bonn/ Hannover/ Stuttgart: Dümmler. (= Phonetische Studien, 3)

Menzerath, Paul, & Meyer-Eppler, Werner (1950). Sprachtypologische Untersuchungen. I. Teil: Allgemeine Einführung und Theorie der Wortbildung. *Studia Linguistica IV, 54-93.*

Meyer-Eppler, Werner (1953). Paul Menzerath 70 Jahre. *Zeitschrift für Phonetik und allgemeine Sprachwissenschaft 7, 146-149.*

Miron, Paul (1956). Paul Menzerath (1 janvier 1883 – 8 avril 1954). *Orbis V, 290-294.*

Prün, Claudia (1994). Validity of Menzerath-Altmann's Law: Graphic Representation of Language, Information Processing Systems and Synergetic Linguistics. *Journal of Quantitative Linguistics 1, 148-155.*

Sievers, Eduard (1876). *Grundzüge der Lautphysiologie zur Einführung in das Studium der Lautlehre der indogermanischen Sprachen.* Leipzig: Breitkopf & Härtel.

Sievers, Eduard (1893). *Grundzüge der Phonetik zur Einführung in das Studium der Lautlehre der indogermanischen Sprachen.* 4., verbesserte Auflage. Leipzig: Breitkopf & Härtel.

Simon, Gerd, unter Mitwirkung von Dagny Guhr und Ulrich Schermaul (2006). *Chronologie Menzerath, Paul.* http://homepages.uni-tuebingen.de/gerd.simon/ChrMenzerath.pdf

Thumb, Albert (1911). Experimentelle Psychologie und Sprachwissenschaft. Ein Beitrag zur Methodenlehre der Philologie. *Germanisch-Romanische Monatsschrift 3, 1-15; 65-74.*

Thumb, Albert, & Marbe, Karl (1901). *Experimentelle Untersuchungen über die psychologischen Grundlagen der sprachlichen Analogiebildung.* Leipzig: Engelmann (Neuausgabe: David D. Murray. Amsterdam: John Benjamins 1978).

Vértes, E. (1955). Rez. zu: Paul Menzerath, Die Architektonik des deutschen Wortschatzes. *Acta Linguistica Academiae Scientiarum Hungaricae V, 415-431.*

Wenig, Otto (Hrsg.) (1968). *Verzeichnis der Professoren und Dozenten der Rheinischen Friedrich-Wilhelms-Universität zu Bonn 1818-1968.* Bonn: Bouvier u.a.

Wodarz, Hans Walter (1972). Phonetik und Phonologie bei Paul Menzerath. *Phonetica 25, 65-71.*

Wodarz, Hans Walter (1974). Zur Entwicklung der Phonetik in Deutschland: Panconcelli-Calzia und Menzerath. In: *Kommunikationsforschung und Phonetik. Festschrift zum fünfzigjährigen Bestehen des Instituts für Kommunikationsforschung und Phonetik der Universität Bonn* (S. 183-206). Hamburg: Buske.

Wodarz, Hans Walter (1974). Das Institut für Kommunikationsforschung und Phonetik in Vergangenheit und Gegenwart. In: *Kommunikationsforschung und Phonetik. Festschrift zum fünfzigjährigen Bestehen des Instituts für Kommunikationsforschung und Phonetik der Universität Bonn* (S. 1-16). Hamburg: Buske.

Wodarz, Hans Walter (1996). Menzerath, Paul. In: Stammerjohann, Harro (ed.), *Lexicon grammaticorum. Who's Who in the History of World Linguistics* (S. 627-628). Tübingen: Niemeyer.

Karl-Heinz Best