Glottometrics 16
2008

Dedicated to Viktor V. Levickij

on the Occasion of his 70th Birthday

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. 16 (2008), Lüdenscheid: RAM-Verlag, 2008. 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 16 (2008) **ISSN 2625-8226**

Contents

Glottometrics 16, 2008, 1-16

Runes: complexity and distinctivity

Ján Mačutek[1]*, Bratislava*

Abstract. Complexity of several types of runes is investigated using the composition method introduced by Altmann (2004). First models for distributions of components and connections are suggested. Letter distinctivity is studied.

Keywords: script complexity, components, connections, distinctivity, runes

1. Method

The first method for measuring script complexity, which will be followed also in this paper, was introduced by Altmann (2004). We briefly recall it here. Any letter, sign etc. in any writing system consists of points (assigned 1 point), straight lines (2 points) and arches (3 points). An arch does not exceed 180° (e.g., “U” has one arch, but “C” consists of two arches). These components can be connected continuously (1 point, e.g., two arches in “C”), crisply (2 points, e.g., straight lines in “V”) or there can be crossings (3 points, e.g., in “X”). The complexity of a letter/sign etc. is the sum of its components´ and connections´ complexities. In order to reduce ambiguities, we refine slightly the complexity measuring:

a) in the points where more than two components are connected, connections of all pairs are taken into account (e.g., "Y" consists of three lines which are connected in one point, hence it has three conections), mathematically speaking, if *n* components touch or cross each other in a point then there are $\binom{n}{2}$ connections in the point;
b) there is a continuous contact if a point is placed on a line or on an arch (e.g., the rune ᛂ, cf. Table 4, rune no. 7);
c) if there are more possibilities for assigning complexity to a letter, use the one which evaluates to the minimum value.

The research is in its beginning stage; so far only two writing systems were investigated, namely Latin (Arial and Courier New fonts in Altmann 2004) and Oriya (one of Indian scripts, cf. Mohanty 2007). We present a similar analysis of runic scripts.

2. Runes and their complexity

Runic scripts, named futharks after the first six characters, were used mainly among Germanic tribes in Scandinavia, northwestern Europe and the British Isles (but some Viking inscriptions were found also in other part of Europe). The earliest inscriptions are dated from the first century. The runes were replaced by the Latin alphabet in the tenth and eleventh century in Britain, even sooner in continental Europe. However, in some regions of Scandinavia they survived much longer (cf. Page 1987 and Elliott 1996). Due to a relatively long

[1] E-mail: jmacutek@yahoo.com

time period and a considerable geographic diversity of areas where runes were used several more or less similar futharks (with different inventory sizes) were developped.

As with any other non-printed script, there is no “officialy” unified version and many characters exhibit a variety of shapes. The present study neither has the ambition to cover all futharks, or all variants of particular runes within a futhark, nor it aims at particular character development (which is an interesting problem investigated by Hegenbarth-Reichardt and Altmann 2008 on the example of Egyptian hieroglyphs, but these aspects of runic scripts are left for future research). Alternative choices of rune shapes may lead to slightly different results. Still another possibility is to take the average complexity for all known variants of a character.

Table 1
Germanic Futhark

	letter	transliteration	components	connections	complexity
1.	ᚠ	f	3x2	2x2	10
2.	ᚢ	u	2+3	2	7
3.	ᚦ	th[2]	3x2	3x2	12
4.	ᚨ	a	3x2	2x2	10
5.	ᚱ	r	4x2	5x2	18
6.	ᚲ	k	2x2	2	6
7.	ᚷ	g	2x2	3	7
8.	ᚹ	w	3x2	3x2	12
9.	ᚺ	h	3x2	2x2	10
10.	ᚾ	n	2x2	3	7
11.	ᛁ	i	2	-	2
12.	ᛃ	j[3]	4x2	2x2	12
13.	ᛇ	ï[4]	3x2	2x2	10
14.	ᛈ	p	5x2	4x2	18
15.	ᛉ	R[5]	3x2	3x2	12
16.	ᛊ	s	4x2	3x2	14
17.	ᛏ	t	3x2	3x2	12
18.	ᛒ	b	5x2	7x2	24
19.	ᛖ	e	4x2	3x2	14
20.	ᛗ	m	4x2	4x2+3	19
21.	ᛚ	l	2x2	2	6
22.	ᛜ	ŋ	4x2	4x2	16
23.	ᛟ	o	4x2	3x2+3	17
24.	ᛞ	d	4x2	4x2+3	19

[2] as the consonant in the beginning of the English word “think”

[3] as the consonant in the beginning of the English word “year”

[4] an uncertain vowel in the region of i

[5] a palatalized r-sound

Table 2
Anglo-Saxon Futhork

	letter	**transliteration**	**components**	**connections**	**complexity**
1.	ᚠ	f	3x2	2x2	10
2.	ᚢ	u	2+3	2	7
3.	ᚦ	th	3x2	3x2	12
4.	ᚩ	o	5x2	4x2	18
5.	ᚱ	r	4x2	5x2	18
6.	ᚳ	c	2x1	2	6
7.	ᚷ	g	2x2	3	7
8.	ᚹ	w	3x2	3x2	12
9.	ᚻ	h	4x2	4x2	16
10.	ᚾ	n	2x2	3	7
11.	ᛁ	i	2	-	2
12.	ᛄ	j	3x2	3x3	15
13.	ᛇ	ï	3x2	2x2	10
14.	ᛈ	p	5x2	4x2	18
15.	ᛉ	x	3x2	3x2	12
16.	ᛋ	s	3x2	2x2	10
17.	ᛏ	t	3x2	3x2	12
18.	ᛒ	b	5x2	7x2	24
19.	ᛖ	e	4x2	3x2	14
20.	ᛗ	m	4x2	4x2+3	19
21.	ᛚ	l	2x2	2	6
22.	ᛝ	ŋ	4x2	2x2+2x3	18
23.	ᛟ	œ	4x2	3x2+3	17
24.	ᛞ	d	4x2	4x2+3	19
25.	ᚪ	a	4x2	3x2	14
26.	ᚫ	æ	3x2	2x2	10
27.	ᚣ	y	2x2+3	2	9
28.	ᛠ	ea	5x2	5x2	20
29.	ᚸ	G[6]	6x2	6x2+3	27
30.	ᛣ	k	3x2	3x2	12
31.	ᛤ	K[7]	7x2	6x2+3x3	35

[6] a variant pronunciation of g

[7] a variant pronunciation of k

Table 3
Long-branch Futhark

	letter	transliteration	components	connections	complexity
1.	ᚠ	f	3x2	2x2	10
2.	ᚢ	u	2+3	2	7
3.	ᚦ	th	2+3	2x2	9
4.	ᚬ	ą	3x2	2x2	10
5.	ᚱ	r	2x2+3	3x2	13
6.	ᚴ	k	2x1	2	6
7.	ᚼ	h	3x2	3x3	15
8.	ᚾ	n	2x2	3	7
9.	ᛁ	i	2	-	2
10.	ᛅ	a	2x2	3	7
11.	ᛋ	s	3x2	2x2	10
12.	ᛏ	t	3x2	3x2	12
13.	ᛒ	b	5x2	7x2	24
14.	ᛘ	m	3x2	3x2	12
15.	ᛚ	l	2x2	2	6
16.	ᛦ	R	3x2	3x2	12

Table 4
Short Twig Futhark

	letter	transliteration	components	connections	complexity
1.	ᚠ	f	3x2	2x2	10
2.	ᚢ	u	2+3	2	7
3.	ᚦ	th	2+3	2x2	9
4.	ᚬ	ą	3x2	2x2	10
5.	ᚱ	r	2x2+3	3x2	13
6.	ᚴ	k	2x1	2	6
7.	ᚽ	h	2+1	1	4
8.	ᚿ	n	2x2	2	6
9.	ᛁ	i	2	-	2
10.	ᛆ	a	2x2	2	6
11.	ᛌ	s	2+1	1	4
12.	ᛐ	t	2x2	2	6
13.	ᛓ	b	3x2	2x2	10
14.	ᛙ	m	2+1	1	4
15.	ᛚ	l	2x2	2	6
16.	ᛧ	R	2	-	2

Table 5
Staveless Runes

	letter	transliteration	components	connections	complexity
1.		f	2	1	3
2.		u	3	-	3
3.		th	2	-	2
4.		ą	2	-	2
5.		r	3	-	3
6.		k	2	1	3
7.		h	2	-	2
8.		n	2	-	2
9.		i	2	-	2
10.		a	2	-	2
11.		s	2	-	2
12.		t	2	-	2
13.		b	2	-	2
14.		m	2x1	-	2
15.		l	2	-	2
16.		R	2x1	-	2

Altmann (2008) and Mohanty (2007) present several hypotheses associated with script complexity and its distribution (cf. Tables 6–9 below for runes complexity distributions; SR are omitted from these considerations, as the associated distributions are defined on two points only, which is not enough for analysis), among them the following.

H1: The smaller the letter inventory, the smaller is the range of complexity.
H2: The smaller the letter inventory, the smaller is the variance of complexity.
H3: The distribution of complexities is uniform.
H4: The greater the mean complexity, the greater is its range and variance.

In fact, H4 is a property of the uniform distribution, so it is an obvious corollary of H3.

Table 6
Distribution of complexities for Germanic Futhark

C	f_C	***C***	f_C	***C***	f_C	***C***	f_C	***C***	f_C	***C***	f_C	***C***	f_C	***C***	f_C
2	1	**5**	0	**8**	0	**11**	0	**14**	2	**17**	1	**20**	0	**23**	0
3	0	**6**	2	**9**	0	**12**	5	**15**	0	**18**	2	**21**	0	**24**	1
4	0	**7**	3	**10**	4	**13**	0	**16**	1	**19**	2	**22**	0		

Table 7

Distribution of complexities for Anglo-Saxon Futhork

C	f_C	**C**	f_C	**C**	f_C	**C**	f_C	**C**	f_C	**C**	f_C	**C**	f_C	**C**	f_C	**C**	f_C
2	1	**6**	3	**10**	4	**14**	2	**18**	4	**22**	0	**26**	0	**30**	0	**34**	0
3	0	**7**	2	**11**	0	**15**	1	**19**	2	**23**	0	**27**	1	**31**	0	**35**	1
4	0	**8**	0	**12**	5	**16**	1	**20**	1	**24**	1	**28**	0	**32**	0		
5	0	**9**	1	**13**	0	**17**	1	**21**	0	**25**	0	**29**	0	**33**	0		

Table 8

Distribution of complexities for Long-branch Futhark

C	f_C	**C**	f_C	**C**	f_C	**C**	f_C	**C**	f_C	**C**	f_C	**C**	f_C	**C**	f_C
2	1	**5**	0	**8**	0	**11**	0	**14**	0	**17**	0	**20**	0	**23**	0
3	0	**6**	2	**9**	1	**12**	3	**15**	1	**18**	0	**21**	0	**24**	1
4	0	**7**	3	**10**	3	**13**	1	**16**	0	**19**	0	**22**	0		

Table 9

Distribution of complexities for Short Twig Futhark

C	f_C	**C**	f_C	**C**	f_C	**C**	f_C	**C**	f_C	**C**	f_C
2	2	**4**	3	**6**	5	**8**	0	**10**	3	**12**	0
3	0	**5**	0	**7**	1	**9**	1	**11**	0	**13**	1

As the Pearson χ^2 goodness-of-fit test is not reliable for such low frequencies, we follow the approach chosen by Mohanty (2007), i.e., we perform the run test about the mean. Denote *I* the inventory size, *R* the range of complexities, $\overline{C}$ the mean complexity and σ_C the standard deviation of complexities. If the data are uniformly distributed, all expected frequency values are E = I/(R+1). A run is a sequence of frequencies which are either all greater than *E* or all smaller than *E*, e.g., for STF we have E = 16/(11+1) = 1.33 and the runs $\left[\underline{2},\underline{0},\underline{3},\underline{0},\underline{5},\underline{1,0,1},\underline{3},\underline{0,0,1}\right]$, i.e., we have 8 runs. Next, denote $n = R + 1$, n_1 the number of frequencies smaller than *E* and n_2 the number of frequencies greater than *E* (in this case $n = 12$, $n_1 = 8$, $n_2 = 4$). The number of runs can be considered random (and, consequently, the distribution can be considered uniform) if

$$z = \frac{|r - E(r)| - 0.5}{\sigma_r} < 1.96,$$

where *r* is the number of runs, $E(r) = 1 + \frac{2n_1n_2}{n}$ and $\sigma_r = \sqrt{\frac{2n_1n_2(2n_1n_2 - n)}{n^2(n-1)}}$. We obtain $z_{GF} = 0.39$, $z_{ASF} = 0.85$, $z_{LBF} = 0.21$ and $z_{STF} = 0.80$, which means that in all four cases the uniform distribution is a good model for the distribution of complexities.

In the following table we summarize some other results. It is to be noted that because of the method refinement (cf. Section 1) our characteristics of Latin, font Arial and Latin, font Courier New are slightly different from the ones obtained by Altmann (2004) and used by Mohanty (2007). In our calculations the letter “Y” (Latin Arial) has the complexity 12, the letters “M”, “N”, “R” and “Y” (all Latin Courier New) have the complexities 34, 24, 21 and 24, respectively.

Table 10

Some script characteristics

script	***I***	***R***	$\overline{C}$	σ_C	**script**	***I***	***R***	$\overline{C}$	σ_C
Latin Arial	26	14	9.81	4.00	Germanic Futhark	24	22	12.25	5.21
Latin Courier New	26	27	18.38	6.83	Anglo-Saxon Futhork	31	33	14.03	6.87
Staveless Runes	16	1	2.25	0.45	Long-branch Futhark	16	22	10.13	4.94
Short Twig Futhark	16	11	6.56	3.12					

Our results corroborate the hypotheses H3 (cf. *z*-variables above) and H4 (the correlation between $\overline{C}$ and R is 0.8981, and between $\overline{C}$ and σ_C, 0.9569). On the other hand, the hypotheses H1 and H2 are questionable, e.g., more complex Latin fonts will have high mean complexities and according to H4 also high ranges and variances, while the inventory will remain unchanged.

As the complexities are uniformly distributed and the mean and the median of the uniform distribution are the same, we apply the median test to compare the mean complexities of different scripts. Consider two scripts with the inventory sizes I_1, I_2. Merge the two data sets (i.e., the sample size is $I_1 + I_2$ now) and find the common median m. Denote n_{B1}, n_{B2} the numbers of complexities below m in the first and in the second script respectively (for each complexity which is equal to the median add 0.5). Analogously n_{A1}, n_{A2} are the numbers of complexities above m (e.g., if we compare Germanic Futhark and Anglo-Saxon Futhork, we have $m = 12$, $I_1 = 24$, $I_2 = 31$, $n_{B1} = 12.5$, $n_{A1} = 11.5$, $n_{B2} = 13.5$ and $n_{A2} = 17.5$). The medians in the two samples (hence also the mean complexities) are significantly different if

$$v = \frac{4\left(\left|n_{B1}n_{A2} - n_{A1}n_{B2}\right| - \dfrac{I_1 + I_2}{2}\right)^2}{(I_1 + I_2)I_1 I_2} > \chi_1^2(0.95) = 3.841.$$

The results of testing are summarized in Table 11 below (significant differences are highlighted in bold).

Table 11

Test for differences between mean complexities

	Latin Arial	Latin Courier New	Germanic Futhark	Anglo-Saxon Futhork	Long-branch Futhark	Short Twig Futhark	Staveless Runes
Latin Arial	-	**9.31**	1.06	1.89	0.06	3.19	**105.08**
Latin Courier New	**9.31**	-	**5.08**	2.17	**12.22**	**17.06**	**100.18**
Germanic Futhark	1.06	**5.08**	-	0.13	0.94	**8.07**	**98.18**
Anglo-Saxon Futhork	1.89	2.17	0.13	-	2.35	**6.21**	**82.23**
Long-branch Futhark	0.06	**12.22**	0.94	2.35	-	3.13	**81.28**
Short Twig Futhark	3.19	**17.06**	**8.07**	**6.21**	3.13	-	**69.03**
Staveless Runes	**105.08**	**100.18**	**98.18**	**82.23**	**81.28**	**69.03**	-

An aspect which has not been studied so far is the number of components (i.e., the number of points, straight lines and arches together, regardless of the type) and connections (again, the number of all connections). Tentatively we suggest the Poisson distribution $(P_x = e^{-\lambda}\lambda^x/x!, \lambda > 0)$ as a model in both cases, i.e., the numbers of components and connections are controlled by the respective means. A character containing a number of components (or connections) which highly exceeds the mean has a very low probability. At the same time, the higher the mean, the higher variability in the number of components or connections is allowed. Of course it may be necessary to modify the very simple model or to replace it with a different one when data from more scripts are available. We run the usual χ^2 goodness-of-fit test (hence the parameter λ is estimated by the minimum χ^2 method), the value of $P \geq 0.05$ indicates a satisfactory fit.

Table 12

Number of components

	Latin Arial	**Latin Courier New**	**Germanic Futhark**	**Anglo-Saxon Futhork**	**Long-branch Futhark**	**Short Twig Futhark**
1	3	3	1	1	1	2
2	8	5	5	5	6	10
3	10	3	8	11	8	4
4	5	4	8	8	0	
5		7	2	4	1	
6		3		1		
7		1		1		
λ	1.7734	2.7815	2.3384	2.5068	N[8]	N
χ^2	2.2907	6.2147	4.9267	3.3103	N	N
P	0.3181	0.2859	0.1772	0.6523	N	N

[8] not enough degrees of freedom, the minimum χ^2 method cannot be used

Table 13
Number of connections

	Latin Arial	**Latin Courier New**	**Germanic Futhark**	**Anglo-Saxon Futhork**	**Long-branch Futhark**	**Short Twig Futhark**
0	3		1	1	1	2
1	5	1	5	6	5	9
2	8	5	5	4	4	4
3	9	4	6	8	5	1
4	1	3	3	5	0	
5		7	3	4	0	
6		5	0	0	0	
7		0	1	2	1	
8		0		0		
9		1		1		
λ	2.1975	3.2600	2.8034	3.3401	2.1034	N
χ^2	6.3043	6.4355	1.2630	5.7191	2.9436	N
P	0.0977	0.2661	0.9387	0.4554	0.4004	N

3. Distinctivity

A method for measuring distinctivity (characters are decomposed into their distinctivity components, then all permutation of the components are compared, finally the difference between the characters is the minimum over all permutations) was presented by Antić and Altmann (2005). The paper contains also a table with differences between distinctivity components; however, it is to be noted that minor mistakes were found in the table, hence values listed there cannot be used without checking their correctness. We follow the approach, again with some refinement, namely the following:

a) there are three possible positions for connecting points on lines or arches – on both ends and in the middle,
b) distinctivity components can differ in types (arch, line, point); if they are of the same type, they can differ in orientation (i.e., four orientations for the lines are —, |, /, and \; cf. Antić and Altmann 2005 for eight orientations of arches); components of the same type and orientation can have different types, numbers or positions of connection points (e.g., the vertical lines in the runes ᚠ,ᚨand ᛉ are different – the first one has two middle crisp connection points, the second one has one crisp connection point at its upper end and one in the middle, the third one has only one crisp connection point in the middle),
c) the difference between components of different types is the sum of their weights (arch 3, line 2, point 1) plus the sum of weights of all their connection points (continuous 1, crisp 2, crossing 3)
d) the difference between components of the same type but of different orientation is the weight of their orientation difference (cf. Antić and Altmann 2005) plus the sum of weights of all their connection points
e) if two components of the same type have the same orientation (e.g., two horizontal lines), their difference is the sum of the weights of the connection points which distinguish them (e.g., the difference between the vertical lines in the runes ᚠand ᚨ is 4 –

the first one has two crisp connection points in the middle, the second one has one crisp connection point in the middle and one at its upper end, hence the distinguishing points are one in the middle and one at the upper end, both of them crisp)

f) if two different characters have zero difference according to the previous rules (i.e., they differ only by lengths of their components or by their positions on the line – positions, not orientations), their difference is assigned the value 0.5 (e.g., the runes ᛁ and ᛁ or ⁝ and ⁝); in other cases (i.e., their difference according to the rules a) – e) is nonzero) lengths or positions on the line are not taken into account.

Then, the mean character distinctivity is the sum of its differences with respect to all other characters, divided by $I-1$. The mean distinctivity of a writing system $\bar{D}$ is the mean of all mean character distinctivities.

Distinctivities of the five above mentioned futharks (i.e., runic “alphabets”) are presented in Tables 14-23 below.

Table 14

Differences between Germanic Futhark characters

	ᚠ	ᚢ	ᚦ	ᚨ	ᚱ	ᚲ	ᚷ	ᚹ	ᚺ	ᚾ	ᛁ	ᛃ	ᛇ	ᛈ	ᛉ	ᛊ	ᛏ	ᛒ	ᛖ	ᛗ	ᛚ	ᛜ	ᛟ	ᛞ
ᚠ	0	19	10	16	18	11	17	14	15	18	12	15	12	20	8	15	16	22	19	25	16	19	23	29
ᚢ	19	0	23	15	23	14	16	19	19	15	7	22	15	27	17	26	13	33	21	31	9	30	28	31
ᚦ	10	23	0	14	8	10	20	4	9	20	16	15	16	16	6	11	10	18	10	16	14	15	15	20
ᚨ	16	15	14	0	8	11	17	10	9	16	12	15	16	16	12	15	8	18	15	21	6	20	20	25
ᚱ	18	23	8	8	0	14	24	4	13	24	20	11	16	8	10	7	10	10	7	13	14	12	13	17
ᚲ	11	14	10	11	14	0	10	10	11	11	7	8	11	18	8	12	8	24	12	22	5	16	14	22
ᚷ	17	16	20	17	24	10	0	20	17	7	9	18	17	28	14	22	14	34	22	20	11	26	16	20
ᚹ	14	19	4	10	4	10	20	0	9	20	16	15	12	12	6	11	6	14	6	12	10	15	15	16
ᚺ	15	19	9	9	13	11	17	9	0	16	12	16	17	21	7	16	11	23	14	16	10	20	20	24
ᚾ	18	15	20	16	24	11	7	20	16	0	8	19	18	28	14	23	14	34	22	26	10	27	21	26
ᛁ	12	7	16	12	20	7	9	16	12	8	0	15	12	24	10	19	10	30	18	28	6	23	21	28
ᛃ	15	22	15	15	11	8	18	15	16	19	15	0	11	10	9	4	9	16	14	24	13	8	14	24
ᛇ	12	15	16	16	16	11	17	12	17	18	12	11	0	12	12	15	8	18	14	25	12	19	23	21
ᛈ	20	27	16	16	8	18	28	12	21	28	24	10	12	0	18	6	14	6	11	21	18	10	16	17
ᛉ	8	17	6	12	10	8	14	6	7	14	10	9	12	18	0	13	12	20	12	18	12	17	17	22
ᛊ	15	26	11	15	7	12	22	11	16	23	19	4	15	6	13	0	13	12	10	20	17	4	10	20
ᛏ	16	13	10	8	10	8	14	6	11	14	10	9	8	14	12	13	0	20	8	18	4	17	17	18
ᛒ	22	33	18	18	10	24	34	14	23	34	30	16	18	6	20	12	20	0	17	23	24	8	18	23
ᛖ	19	21	10	15	7	12	22	6	14	22	18	14	14	11	12	10	8	17	0	12	10	14	16	10
ᛗ	25	31	16	21	13	22	20	12	16	26	28	24	25	21	18	20	18	23	10	0	22	25	23	8
ᛚ	16	9	14	6	14	5	11	10	10	10	6	13	12	18	12	17	4	24	12	22	0	21	19	22
ᛜ	19	30	15	20	12	16	26	15	20	27	23	8	19	10	17	4	17	8	14	25	21	0	10	24
ᛟ	23	28	15	20	13	14	16	15	20	21	21	14	23	14	17	10	17	18	16	23	19	10	0	22
ᛞ	29	31	20	25	17	22	20	16	24	26	28	24	21	17	22	20	18	23	10	8	22	24	22	0

Table 15

Mean distinctivities of Germanic Futhark characters

ᚠ	16.91	ᚨ	14.57	ᚷ	18.22	ᚾ	19.00	ᛇ	15.30	ᛊ	13.96	ᛖ	13.65	ᛜ	17.39
ᚢ	20.57	ᚱ	13.22	ᚹ	12.00	ᛁ	15.78	ᛈ	16.39	ᛏ	12.09	ᛗ	20.30	ᛟ	17.87
ᚦ	13.74	ᚲ	12.57	ᚺ	15.00	ᛃ	14.13	ᛉ	12.78	ᛒ	20.22	ᛚ	13.35	ᛞ	21.26

Table 16

Differences between Anglo-Saxon Futhork characters

	ᚠ	ᚢ	ᚦ	ᚩ	ᚱ	ᚳ	ᚷ	ᚹ	ᚻ	ᚾ	ᛁ	ᛄ	ᛇ	ᛈ	ᛉ	ᛋ	ᛏ	ᛒ	ᛖ	ᛗ	ᛚ	ᛝ	ᛟ	ᛞ	ᚪ	ᚫ	ᚣ	ᛠ	ᛥ	ᛣ	ᚸ
ᚠ	0	19	10	16	18	12	17	14	21	18	12	19	20	20	8	13	16	22	19	25	16	25	23	21	16	16	18	18	37	12	41
ᚢ	19	0	23	27	23	13	16	19	29	15	7	20	15	27	17	15	13	33	21	31	9	30	28	35	21	15	2	25	48	17	52
ᚦ	10	23	0	16	8	10	20	4	15	20	16	21	16	16	6	13	10	18	10	16	14	25	15	12	10	14	22	14	33	6	37
ᚩ	16	27	16	0	8	18	28	12	15	28	24	29	16	8	14	21	18	6	15	21	18	26	20	25	6	12	26	6	25	18	29
ᚱ	18	23	8	8	0	14	24	4	15	24	20	25	12	8	10	13	10	10	7	13	14	23	13	17	2	8	22	10	29	10	33
ᚳ	12	13	10	18	14	0	11	10	16	10	6	15	10	22	8	15	8	24	16	22	4	23	19	22	12	6	12	20	39	4	43
ᚷ	17	16	20	28	24	11	0	20	27	7	9	5	17	28	14	17	14	34	22	20	11	14	16	20	22	17	15	26	32	14	37
ᚹ	14	19	4	12	4	10	20	0	19	20	16	21	12	12	6	9	6	14	6	12	10	25	15	16	6	10	18	10	33	6	37
ᚻ	21	29	15	15	15	16	27	19	0	26	22	28	18	23	17	28	20	21	22	24	20	30	24	16	13	14	27	21	36	16	40
ᚾ	18	15	20	28	24	10	7	20	26	0	8	5	16	28	14	17	14	34	22	26	10	21	21	26	22	16	14	26	39	14	37
ᛁ	12	7	16	24	20	6	9	16	22	8	0	13	12	24	10	12	10	30	18	28	6	23	21	28	18	12	9	22	41	10	45
ᛄ	19	20	21	29	25	15	5	21	28	5	13	0	19	29	15	18	15	35	23	21	15	18	18	21	23	19	19	27	34	15	32
ᛇ	20	15	16	16	12	10	17	12	18	16	12	19	0	12	12	13	8	18	14	25	6	21	19	29	10	8	14	14	37	12	41
ᛈ	20	27	16	8	8	22	28	12	23	28	24	29	12	0	18	17	14	6	11	21	18	22	16	25	10	16	26	6	25	18	29
ᛉ	8	17	6	14	10	8	14	6	17	14	10	15	12	18	0	11	12	20	12	18	12	19	17	18	8	12	16	16	35	8	39
ᛋ	13	15	13	21	13	15	17	9	28	17	12	18	13	17	11	0	7	23	10	20	11	26	20	24	15	15	13	15	38	11	42
ᛏ	16	13	10	18	10	8	14	6	20	14	10	15	8	14	12	7	0	20	8	18	4	23	17	22	12	8	12	12	35	4	39
ᛒ	22	33	18	6	10	24	34	14	21	34	30	35	18	6	20	23	20	0	17	23	24	28	18	27	12	18	32	8	23	20	27
ᛖ	19	21	10	15	7	16	22	6	22	22	18	23	14	11	12	10	8	17	0	10	12	26	16	18	9	15	19	9	32	12	36
ᛗ	25	31	16	21	13	22	20	12	24	26	28	21	25	21	18	20	18	23	10	0	22	24	23	8	15	21	29	19	42	18	46
ᛚ	16	9	14	18	14	4	11	10	20	10	6	15	6	18	12	11	4	24	12	22	0	23	19	26	12	6	8	16	39	8	43
ᛝ	25	30	25	26	23	23	14	25	30	21	23	18	21	22	19	26	23	28	26	24	23	0	10	24	21	25	31	20	34	23	39
ᛟ	23	28	15	20	13	19	16	15	24	21	21	18	19	16	17	20	17	18	16	23	19	10	0	22	15	20	29	18	24	17	29
ᛞ	21	35	12	25	17	22	20	16	16	26	28	21	29	25	18	24	22	27	18	8	26	24	22	0	19	25	33	23	42	18	46
ᚪ	16	21	10	6	2	12	22	6	13	22	18	23	10	10	8	15	12	12	9	15	12	21	15	19	0	6	20	12	31	12	35
ᚫ	16	15	14	12	8	6	17	10	14	16	12	19	8	16	12	15	8	18	15	21	6	25	20	25	6	0	14	18	37	8	41
ᚣ	18	2	22	26	22	12	15	18	27	14	9	19	14	26	16	13	12	32	19	29	8	31	29	33	20	14	0	24	47	16	51
ᛠ	18	25	14	6	10	20	26	10	21	26	22	27	14	6	16	15	12	8	9	19	16	20	18	23	12	18	24	0	23	16	27
ᛥ	37	48	33	25	29	39	32	33	36	39	41	34	37	25	35	38	35	23	32	42	39	34	24	42	31	37	47	23	0	35	5
ᛣ	12	17	6	18	10	4	14	6	16	14	10	15	12	18	8	11	4	20	12	18	8	23	17	18	12	8	16	16	35	0	39
ᚸ	41	52	37	29	33	43	37	37	40	37	45	32	41	29	39	42	39	27	36	46	43	39	29	46	35	41	51	27	5	39	0

Table 17

Mean distinctivities of Anglo-Saxon Futhork characters

ᚠ	18.73	ᚱ	14.90	ᚻ	22.10	ᛇ	16.53	ᛏ	14.30	ᛚ	15.20	ᚪ	14.83	ᛡ	33.67
ᚢ	22.17	ᚳ	15.47	ᚾ	19.93	ᛈ	18.50	ᛒ	21.50	ᛝ	24.07	ᚫ	15.73	ᛣ	14.57
ᚦ	15.67	ᚷ	19.13	ᛁ	17.73	ᛉ	14.73	ᛖ	16.23	ᛟ	19.40	ᚣ	21.27	ᛤ	37.23
ᚩ	18.37	ᚹ	14.07	ᛄ	20.57	ᛋ	17.40	ᛗ	22.03	ᛞ	23.60	ᛠ	17.37		

Table 18

Differences between Long-branch Futhark characters

	ᚠ	ᚢ	ᚦ	ᚬ	ᚱ	ᚴ	ᚼ	ᚾ	ᛁ	ᛅ	ᛋ	ᛏ	ᛒ	ᛘ	ᛚ	ᛦ
ᚠ	0	19	15	12	21	6	19	18	12	16	13	16	22	8	16	12
ᚢ	19	0	13	19	13	13	20	15	7	15	15	13	33	17	9	17
ᚦ	15	13	0	15	8	13	24	19	11	19	23	21	37	17	17	17
ᚬ	12	19	15	0	15	12	19	16	12	18	19	12	22	12	10	8
ᚱ	21	13	8	15	0	15	24	19	15	21	20	13	29	17	9	13
ᚴ	6	13	13	12	15	0	15	12	6	10	10	12	24	4	10	8
ᚼ	19	20	24	19	24	15	0	5	13	5	18	15	35	15	15	15
ᚾ	18	15	19	16	19	12	5	0	8	8	17	14	34	14	10	14
ᛁ	12	7	11	12	15	6	13	8	0	8	12	10	30	10	6	10
ᛅ	16	15	19	18	21	10	5	8	8	0	16	14	34	14	12	14
ᛋ	13	15	23	19	20	10	18	17	12	16	0	7	23	11	11	11
ᛏ	16	13	21	12	13	12	15	14	10	14	7	0	20	12	4	4
ᛒ	22	33	37	22	29	24	35	34	30	34	23	20	0	20	24	20
ᛘ	8	17	17	12	17	4	15	14	10	14	1	12	20	0	12	8
ᛚ	16	9	17	10	9	10	15	10	6	12	11	4	24	12	0	8
ᛦ	12	17	17	8	13	8	15	14	10	14	11	4	20	8	8	0

Table 19

Mean distinctivities of Long-branch Futhark characters

ᚠ	15.00	ᚦ	17.93	ᚱ	16.80	ᚼ	17.13	ᛁ	11.33	ᛋ	15.07	ᛒ	27.13	ᛚ	11.53
ᚢ	15.87	ᚬ	14.73	ᚴ	11.33	ᚾ	14.87	ᛅ	14.93	ᛏ	12.47	ᛘ	12.73	ᛦ	11.93

Table 20

ifferences between Short Twig Futhark characters

	ᚠ	ᚢ	ᚦ	ᚬ	ᚱ	ᚴ	ᚽ	ᚿ	ᛁ	ᛆ	ᛌ	ᛐ	ᛓ	ᛙ	ᛚ	ᛧ
ᚠ	0	19	15	12	21	6	14	12	12	10	14	14	0.5	14	16	12
ᚢ	19	0	13	19	13	13	9	13	7	13	9	9	19	9	9	7
ᚦ	15	13	0	15	8	13	13	13	11	13	13	17	15	13	17	11
ᚬ	12	19	15	0	15	12	14	6	12	12	14	16	12	14	10	12
ᚱ	21	13	8	15	0	15	17	9	15	15	17	15	21	17	9	15
ᚴ	6	13	13	12	15	0	8	6	6	4	8	8	6	8	10	6
ᚽ	14	9	13	14	17	8	0	8	2	9	2	9	14	2	8	2
ᚿ	12	13	13	6	9	6	8	0	6	6	8	10	12	8	4	6

ᛁ	12	7	11	12	15	6	2	6	0	6	2	6	12	2	6	0.5
ᛆ	10	13	13	12	15	4	9	6	6	0	9	4	10	9	10	6
ᛌ	14	9	13	14	17	8	2	8	2	9	0	9	14	2	8	2
ᛐ	14	9	17	16	15	8	9	10	6	4	9	0	14	9	6	6
ᛓ	0.5	19	15	12	21	6	14	12	12	10	14	14	0	14	16	12
ᛙ	14	9	13	14	17	8	2	8	2	9	2	9	14	0	8	2
ᛚ	16	9	17	10	9	10	8	4	6	10	8	6	16	8	0	6
ᛧ	12	7	11	12	15	6	2	6	0.5	6	2	6	12	2	6	0

Table 21
Mean distinctivities of Short Twig Futhark characters

ᚠ	12.77	ᚦ	13.33	ᚱ	14.80	ᚽ	8.73	ᛁ	7.03	ᛌ	8.73	ᛓ	12.77	ᛚ	9.53
ᚢ	12.07	ᚭ	13.00	ᚴ	8.60	ᚿ	8.47	ᛆ	9.07	ᛐ	10.13	ᛙ	8.73	ᛧ	7.03

Table 22
Differences between Staveless Runes characters

	╿	⧹̅	ı̅	⸜	ʅ̅	╽̲̅	⊥	⸌̅	ǀ	⁄	╹	ˊ	⸝	⁚	ˋ	꞉
╿	0	7	2	3	7	2	2	3	2	3	2	3	3	2	3	2
⧹̅	7	0	5	5	4	7	5	5	5	5	5	5	5	5	5	5
ı̅	2	5	0	1	5	2	0.5	1	0.5	1	0.5	1	1	4	1	4
⸜	3	5	1	0	5	3	1	0.5	1	2	1	2	2	4	0.5	4
ʅ̅	7	4	5	5	0	7	5	5	5	5	5	5	5	5	5	5
╽̲̅	2	7	2	3	7	0	2	3	2	3	2	3	3	4	3	4
⊥	2	5	0.5	1	5	2	0	1	0.5	1	0.5	1	1	4	1	4
⸌̅	3	5	1	0.5	5	3	1	0	1	2	1	2	2	4	0.5	4
ǀ	2	5	0.5	1	5	2	0.5	1	0	1	0.5	1	1	4	1	4
⁄	3	5	1	2	5	3	1	2	1	0	1	0.5	0.5	4	2	4
╹	2	5	0.5	1	5	2	0.5	1	0.5	1	0	1	1	4	1	4
ˊ	3	5	1	2	5	3	1	2	1	0.5	1	0	0.5	4	2	4
⸝	3	5	1	2	5	3	1	2	1	0.5	1	0.5	0	4	2	4
⁚	2	5	4	4	5	4	4	4	4	4	4	4	4	0	4	0.5
ˋ	3	5	1	0.5	5	3	1	0.5	1	2	1	2	2	4	0	4
꞉	2	5	4	4	5	4	4	4	4	4	4	4	4	0.5	4	0

Table 23
Mean distinctivities of Staveless Runes characters

╿	3.07	ı̅	1.97	ʅ̅	5.20	⊥	1.97	ǀ	1.97	╹	1.97	⸝	2.33	ˋ	2.33
⧹̅	5.20	⸜	2.33	╽̲̅	3.33	⸌̅	2.33	⁄	2.33	ˊ	2.33	⁚	3.77	꞉	3.77

The runes can be ordered from the least to the most distinctive as follows (we note that a character distinctivity is relative to all other characters in a writing system, hence distinctivities of two characters from two different writing systems cannot be compared).

Table 24
Summary on runes distinctivity

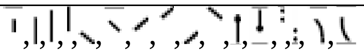
Futhark	ascending distinctivity	$\bar{D}$	*I*
Germanic Futhark	ᚹ,ᛏ,ᚲ,ᛉ,ᚱ,ᛚ,ᛗ,ᚦ,ᛊ,ᛃ,ᚨ,ᚺ,ᛇ,ᛁ,ᛈ,ᚠ,ᛜ,ᛟ,ᚷ,ᚾ,ᛒ,ᛖ,ᚢ,ᛞ	15.84	24
Anglo-Saxon Futhork	ᚹ,ᛏ,ᛣ,ᛉ,ᚩ,ᚱ,ᛚ,ᚳ,ᚦ,ᚫ,ᛖ,ᛇ,ᛠ,ᛋ,ᛁ,ᚪ,ᛈ,ᚠ,ᚷ,ᛟ,ᚾ,ᛡ,ᚣ, ᛒ,ᛗ,ᚻ,ᚢ,ᛞ,ᛝ,ᚸ,ᛤ	19.26	31
Long-branch Futhark	ᚴ,ᛁ,ᛚ,ᛦ,ᚴ,ᛘ,ᚬ,ᚾ,ᛅ,ᚠ,ᛋ,ᚢ,ᚱ,ᚼ,ᚦ,ᛒ	15.05	16
Short Twig Futhark	ᛁ,ᛧ,ᛆ,ᚴ,ᚿ,ᛌ,ᛐ,ᛓ,ᛚ,ᛙ,ᚢ,ᚠ,ᚭ,ᚬ,ᚦ,ᚱ	10.30	16
Staveless Runes		2.89	16

Table 24 is a good illustration of two facts. First, as was stated above, a character distinctivity depends on all other characters in the writing system. Hence, if the same character exists in two different writing systems, it can have (relatively) high distinctivity in one of them and (relatively) low in the other one (e.g., compare the distinctivity rank of the rune ᛁ in Germanic Futhark and Short Twig Futhark). Second, even if there (probably) is some relation between complexity and distinctivity, high (or low) complexity does not guarantee the respective distinctivity – the rune ᛁ (which has the lowest complexity in both Germanic Futhark and Short Twig Futhark) being an example again.

Altmann (2008) presented also some hypotheses associated with complexity and distinctivity:

H5: The greater the letter inventory, the smaller the mean distinctivity of letters.

H6: The greater the letter inventory, the greater the mean complexity.

H7: The greater the complexity of a letter, the more distinctive it is.

Our results contradict H5 (cf. Table 24, Antić and Altmann 2005 obtained $\bar{D}$=17.87 for Latin Arial with I =26). Even without the numbers, if H6 and H4 are true, they imply the corollary the greater the letter inventory, the greater the range of complexities, and letters with low complexities are supposed to be highly distinctive from very complex letters. Anyway, the hypothesis will be definitely rejected (or better corroborated) only when more results are available.

Finally we present two other aspects of distinctivity, namely, distribution of distinctivity components (e.g., in Germanic Futhark 7 components occur in one letter only, 7 of them in two letters, 1 in 3 letters, etc., cf. Table 25) and distribution of letter distinctivities. For distinctivity components, there are too many distributions which fit well the data and searching for a model will be postponed until more writing systems are investigated.

Table 25. *Distribution of distinctivity components*

	1	2	3	4	5	6	7	8	9	10	11	12
Latin Arial	15	11	5	1	1							
Germanic Futhark	7	7	1	2	0	0	2	0	1			
Anglo-Saxon Futhork	5	10	4	1	0	2	1	1	0	1	0	1
Long-branch Futhark	6	4	4	1	1							
Short Twig Futhark	5	2	5	1								
Staveless Runes	4	0	2	2								

Distinctivities were pooled automatically with the statistical software R (in Table 26 we present ranks of classes and the respective frequencies). Only the Hyperpoisson distribution ($P_x = \frac{a^x}{{}_1F_1(1;b;a)b^{(x)}}$, cf. Wimmer and Altmann 1999) yields a good fit with respect to all data. Again, we note that the model (i.e., the distribution of distinctivities) is tentative only.

Table 26
Distribution of distinctivities

	1	2	3	4	5	6	7	8	a	b	χ^2	P
Latin Arial	1	6	6	6	2	3	0	2	3.0339	1.3023	2.1805	0.7026
Germanic Futhark	9	5	4	2	4				37.0719	55.4985	0.2011	0.9043
Anglo-Saxon Futhork	6	15	8	0	1	1			0.5908	0.2363	0.1225	0.7264
Long-branch Futhark	10	5	0	1					N	N	N	N
Short Twig Futhark	2	7	1	5	1				1.2930	0.3694	5.9310	0.0515
Staveless Runes	4	6	0	2	2	0	0	2	7.4131	8.3167	4.7071	0.1945

Acknowledgement

The author is grateful to Michael Lerche Nielsen for his help and comments on the runic script. The runes were copied from Odd Einar Haugen’s webpage (http://helmer.aksis.uib.no/Runefonter/PDF-filar/AlfabetGullhornet.pdf, http://helmer.aksis.uib.no/Runefonter/PDF-filar/AlfabetGullskoen.pdf).
The research was supported by the grant VEGA 1/3016/06.

References

Altmann, G. (2004). Script complexity. *Glottometrics 8, 68-74.*

Altmann, G. (2008). Towards a theory of script. In: Altmann, G., Fan, F. (eds.), *Analysis of Script. Properties of Characters and Writing Systems: 145-160.* Berlin: de Gruyter (in press).

Antić, G., Altmann, G. (2005). On letter distinctivity. *Glottometrics 9, 46-53.*

Elliott, R.W.V. (1996). The Runic Script. In: Daniels, P.T., Bright, W. (eds.), *The World's Writing Systems: 333-339*. Oxford: Oxford University Press.

Hegenbarth-Reichardt, I., Altmann, G. (2008). On the decrease of complexity from hieroglyphs to demotic symbols. In: Altmann, G., Fan, F. (eds.), *Analysis of Script. Properties of Characters and Writing Systems: 101-110.* Berlin: de Gruyter (in press).

Mohanty, P. (2007). On the script complexity and the Oriya script. In: Grzybek, P., Köhler, R. (eds.), *Exact Methods in the Study of Language and Text: 473-484.* Berlin: de Gruyter.

Page, R.I. (1987). *Runes*. London: British Museum Press.

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

Glottometrics 16, 2008, 17-26

Writer's voice in the texts of “Peace and War” themes

Nadia Yesypenko, Chernivtsi[1]

Abstract. The article deals with the study of lexical semantic peculiarities of writer's voice in the novels of modern English and American literature. The statistical approach adopted in the present study is combined with the componential analysis of the word-stock in the novels and set in a linguistic framework that prioritizes lexis and semantics. The study investigates semantic verb classes that are very frequent in the analyzed novels. The lexical statistical approach to the data involves a comparative analysis of the usage of semantic verb classes in the novels of war versus peace theme.

Keywords: text, writer's voice, semantic class, statistical analysis, componential analysis, concept verbalization.

In linguistics the purpose of a scientific analysis is to identify and classify the elements of language and to find the (lawlike) dependencies between them. In literary studies the purpose is usually an adjunct to understanding, exegesis, and interpretation, and, in advanced state, the explanation of findings. In both cases, an extremely detailed and scrupulous attention is paid to the text.

Within the last decade, the availability of robust tools for text analysis has provided an opportunity for establishing the peculiarities of the writer's voice through the application of stylistic analysis of the text. In some forms of stylistic analysis, the numerical recurrence of certain stylistic features is used to make judgments about the nature and the quality of the writing. Stylistic analysis is a normal part of literary studies. It is practised as a part of understanding the possible meanings in the text. It is also generally assumed that the process of analysis will reveal the (good) qualities of the writing. However, it is important to recognize that the concept of style is much broader than just the "good style" of literary prose.

The present research focuses on the semantic layer of the author's writing style, author's selection of words for verbalization of different concepts in the text. It addresses the notions of

- text;
- writer's voice;
- semantic class.

Writing style means the way in which a particular writer manages his or her words and sentences. It is the creative part of writing, giving life to the writer's personality. An individual way in which an author expresses himself in writing may yield valuable information about his personality. It may show permanent peculiarities which recur and lend to the written expression a distinctive colour which is usually called the person's "style". A person's style corresponds to his characteristic manner, appearance, bearing, while the single utterance corresponds to his behavior in a specific situation. The situation may provoke the discharge of a specific emotion, and the written utterance is an expression of it (Schachtel 1977: 178).

Writer's voice is a literary term used to describe the individual writing style of an author. Voice is a combination of a writer's use of syntax, diction, punctuation, character development, dialogue within a given body of text or across several works (Webster

[1] Address correspondence to: jargar@ukr.net

Dictionary 1972:789). Voice can also be referred to as the specific fingerprint of an author, as every author has a different writing style.

All writers vary in their styles. Their sentences differ in length and complexity. Each individual tends to put the words together in particular patterns. Each writer's tone may be objective and distanced, or up-close and personal. Style involves all these choices and some more. Style is the intangible essence of what makes a person's writing unique. It is writer's voice that makes texts on the same theme written by different authors structured, verbalized and perceived in a different way.

As the study of writer's voice is "impossible without a text, which is a basis for a linguistic, philological and literary study" (Bachtin 1986:473), the need for text definition becomes more pressing. Lately, a new approach was formulated and implemented that addresses not only the description of text structure but text pragmatics, too. Text is viewed as a communication act closely connected to a speech act. Text analysis has a special focus on the survey of relations between a structure and function of the text, text and context, discourse and social communication, text and a speech act situation, text classification into genres and types (Nikolina 2006: 281-283).

But in modern linguistics there is no distinct definition of the term "text". Some scientists associate text with a literary work, stating that "text" is most likely to mean a piece of literary writing which is the subject of study. A.Grišunin (1998) and L. Babenko (2004) define "text" as a verbal discourse in which all language units (from phoneme to a sentence) are realized; text exists in the form of a written document. We consider the definition of text by I.Gal´perin the most precise and accurate: "Text is a product of creative discourse process characterized by completion; objectified in a form of a written document; consists of a title and a number of language units which are bound by different types of lexical, grammatical, logical, stylistic connections; it has purposefulness and pragmatics" (1981:18). In a more general setting, text is any sequence of spatially or temporally ordered entities (not only linguistic ones and not only written ones). However, definitions are neither true nor false; they are conventions, hence one may adhere even to an ad hoc definition.

The specific features of a literary text depend on the subject of the writer’s artistic mastering and presentation which can be a human life, surrounding reality, etc. They are reflected in the contents of a literary work when they fully acquire a verbal form and get individual and completed embodiment in the text (Giršman 1996:730). The function of the language of a literary work is to depict images of the world.

The expressiveness and deterministic properties of the language combined with its communicative properties reproduce various verbal fabrics of a literary work and its vivid system of images. It enables us to reveal a semantic palette of the word and the diversity of semantic nuances (Chrapčenko 1976:129f.).

The semantic layer of the individual writer's voice represents several general ways of the verbal presentation of objects and emotional colouring of the given world. Words are interlaced in the integrity, which forms the descriptive level of the realization of the language world view.

As was stated above, reality is reflected in a word; that is why it is necessary to analyze, classify and define the dominant semantic classes of words in a literary work which can be verbal representatives of the surrounding world.

The lexical semantic structure of language has been given a thorough investigation and analysis; however the partitioning of vocabulary into semantic classes presents serious difficulties. It is due both to the absence of sufficiently effective and objective methods of classification and to the fact that all entities of language have a fuzzy identity (Altmann 1996). Thus, Dixon (1991) says, that words of any language can be grouped into a certain number of lexical classes, which he calls semantic types, if a general semantic component and common grammatical qualities can be singled out. Every semantic type in language is associated with a certain class of words. He underlines that size (big, little, long), colour (red, blue), age (new, old) and estimation (good, bad), in most languages belong to adjectives, while words with concrete reference (woman, hand, water) belong to nouns. Semantic types expressing motion (go, throw), actions (cut, burn), perception (see, hear) and conversation (say, ask, tell) refer to the class of verbs (1991: 76-78). It must be remarked that this view is based on European languages.

A semantic class contains words that share a semantic property. Semantic classes may intersect. The member of each semantic class has some features in common, thus distinguishing them from the members of other semantic classes (Levin 1998).

The current methods of study of the semantic word structure – statistical, contextual, structural, psycholinguistic and componential – constitute a research paradigm.

In the study of the semantic layer of the individual writer's voice the method of componential analysis (O´Grady, Archibald 2000) and the statistical method (the chi-square criterion of independence) (Levickij 2004) have been applied. Componential analysis has been used to analyze the meaning of certain types of words (verbs) in terms of semantic features. An obvious advantage of this approach is that it allows us to group entities into semantic classes by means of meaning generalizations. Componential analysis has given its most impressive results in the study and classification of verb meaning. A typical component of verb meaning is the concept GO, which is associated with change of various sorts. The Go-concept is often manifested in the meaning of verbs other than just *go*. The verbs might have somewhat different senses like denoting the entity undergoing change or expressing the endpoint of that change. Componential analysis reveals subtle semantic nuances and helps to determine the semantic type that particular verbs can belong to (O´Grady, Archibald 2000:241).

The statistical approach claims that to study a semantic layer of a style is to analyze style mathematically by breaking down the word stock into semantic classes and noting what features are statistically more common in certain styles.

The analysis of the realization of semantic classes in the text with the help of statistical methods has resulted in a wide variety of algorithms that use the distributional hypothesis to discover many aspects of semantics, by applying statistical techniques to large corpora of verbs selected from literary texts. The chi-square criterion of independence makes it possible to define the similarity and the difference of the semantic classes' frequency.

In this article we aim: (1) to study the semantic classes of verbs in the novels of war and peace theme by modern English and American writers; (2) to establish differences or similarities in the use of semantic classes of verbs depending on the subject (genre, contents) of works and the individual writer's voice. In order to accurately perform the analysis of the semantic layer of the individual writing style on the basis of semantic verb classes, semantically annotated corpora are needed. The verbs have been selected from the following novels: Evelyn Waugh (“Vile Bodies”, “A Handful of Dust”), Ernest Hemingway (“For Whom The Bell Tolls”, “Fiesta: The Sun Also Rises”), Kurt Vonnegut (“Slaughterhouse 5”, “Cat’s Cradle”), Irvin Show (“Young Lions”, “Two Weeks in of Another Town”), Joseph Heller (“Catch-22”, “Something Happened”), Norman Mailer (“The Naked and of The Dead”, “An American Dream”) and Ford Madox Ford (“Parade’s End”).

We put forward the hypothesis, that lexical semantic composition of the text depends on the followings factors:

1) the subject of a literary work;
2) the writer's voice.

Consequently, in texts belonging to different authors, there will be different frequency of the semantic classes of verbs and every author will use different semantic classes.

To study the semantic peculiarities of the verbs we classified them into several semantic classes. This classification is based on the capacity of the words of one class to demonstrate belonging of all of the classified lexemes to the same semantic type. The general content is either obviously plugged into a lexeme, or is recognized there at the second step of semantic reduction. The lexemes of a definite class contain the same type of information in the meaning structure and the same grammatical forms. The classification is strictly built on the semantic basis. Thus, 27 semantic classes of verbs have been singled out: verbs of motion/ removing: *approach, run, pass, sail;* verbs of process, change, development: *increase, wax, vary, turn*; verbs of beginning/end of action: *start, commence, finish, quit*; verbs of physical action: *throw, lean, nod, beat*; engender verbs: *produce, arch, build, form*; destroy verbs: *shrink, collapse, wound, choke*; verbs of successful/unsuccessful action implementation: *fail, succeed, achieve, miss*; verbs of attempt: *try, endeavor;* verbs of sound emission: *groan, sound, whir*; verbs of light phenomena: *shine, light, glare*; verbs of temperature phenomena: *burn, heat*; verbs of nature phenomena: *snow, flood, blow, rain*; verbs of communication: *speak, answer, utter*; verbs of moral impact/effect: *warn, prepare, serve*; verbs of social activity: *lead, retire, organize*; position verbs: *stand, lie, hang*; verbs of existence: *be, exist, remain*; modality verbs: *could, need, must*; verbs of human relations: *marry, help, kiss, invite, entertain*; verbs of reference: *indicate, mean, matter, seem*; verbs of emotional psychological impact: *amuse, frighten, insult*; verbs of ownership/loss: *lose, buy, get, possess*; verbs of physiological state: *sweat, weigh, sign*; verbs of perception: *listen, hear, feel, look*; verbs of mental activity: *think, remember, know, suppose, consider*; verbs of subjective assessment: *like, love, hate, mistrust, judge*; verbs of emotional psychological state: *wonder, wish, hope, wait*.

The calculation of the actual realization of semantic verb classes in the texts by seven authors shows uneven frequency of their use in the probed works (see Table 1).

Table 1
Frequency of the lexical semantic classes of verbs

Lexical semantic classes	Evelyn Waugh	Ernest Hemingway	Kurt Vonnegut	Irvin Show	Joseph Heller	Norman Mailer	F.M.Ford	Together
1. Verbs of Motion/Removing	324	408	180	418	288	395	216	2229
2. Verbs of Process, Change, Development	54	30	27	83	104	156	85	539
3. Verbs of Beginning/End of Action:	66	121	45	63	142	114	98	649
4. Verbs of Physical Action	259	378	286	393	296	305	241	2158
5. Engender Verbs	71	83	107	113	105	101	94	674
6. Destroy Verbs	27	44	62	62	78	64	38	375
7. Verbs of Successful/Unsuccessful Action Implementation	10	6	9	13	47	46	15	146
8. Verbs of Attempt	17	35	40	55	57	34	17	255
9. Verbs of Sound Emission	4	10	19	28	18	22	11	112
10. Verbs of Light Phenomena	3	6	2	5	20	42	21	99
11. Verbs of Temperature Phenomena	1	11	1	2	9	7	2	33
12. Verbs of Nature	10	9	0	8	6	10	2	45

Phenomena								
13. Verbs of Communication	301	452	372	360	258	295	346	2384
14. Verbs of Moral Impact/Effect	73	98	108	78	104	62	66	589
15. Verbs of Social Activity	49	33	50	86	90	69	63	440
16. Position Verbs	24	95	24	51	45	42	43	324
17. Verbs of Existence	553	683	742	467	720	631	559	4355
18. Modality Verbs	212	297	223	178	285	307	313	1815
19. Verbs of Human Relations	41	37	51	72	74	50	40	365
20. Verbs of Reference	53	32	39	42	66	56	58	346
21. Verbs of Emotional Psychological Impact	15	14	12	16	35	13	19	124
22. Verbs of Ownership/Loss	172	184	192	142	249	278	220	1437
23. Verbs of Physiological State	13	18	28	38	43	58	23	221
24. Verbs of Perception	141	265	126	170	170	194	96	1162
25. Verbs of Mental Activity	187	257	186	265	298	233	227	1653
26. Verbs of Subjective Assessment	29	53	17	61	81	35	49	325
27. Verbs of Emotional Psychological State	66	110	64	140	246	142	117	885
Total								

Verbs belonging to the semantic class of *verbs of existence* (4355 words) are more frequently used. *Verbs of communication* comprise 2384 samples, *verbs of motion/removing* amount to 2229 units and the *verbs of physical action* include 2158 words.

However, the determination of the usage frequency does not constitute the complete statistical analysis of the given subject-matter. It remains unrevealed whether the usage frequency of the semantic verb classes in the novels of a certain writer substantially exceeds some theoretically expected quantity. Therefore for reliable quantitative analysis of data presented in Table 1, the criterion *z* has been applied. The most widespread formula for the calculation of *z* is as follows:

$$z = \frac{n_{ij} - \frac{n_{i\bullet} n_{\bullet j}}{n}}{\sqrt{\frac{n_{i\bullet} n_{\bullet j}\left(n - n_{i\bullet}\right)\left(n - n_{\bullet j}\right)}{n^2\left(n-1\right)}}} \qquad (1)$$

where n_{ij} is the frequency in the i, j cell;

$n_{i\bullet}$ is the sum of the rows;

$n_{\bullet j}$ is the sum of the columns;

n is the sum of all frequencies in Table 1.

The result of the calculation shows that the authors avoid some verb classes (if *z* is less than -1.96) and prefer another verbs in their texts (if *z* is more than 1.96). The other semantic classes are considered to be neutral. We present *z* data indicating the calculation by *P* (preferred), *A* (avoided), *N* (neutral) verb classes.

Table 2
The realization of verb semantic classes in the novels

Lexical semantic classes	Evelyn Waugh	Ernest Hemingway	Kurt Vonnegut	Irvin Show	Joseph Heller	Norman Mailer	F.M.Ford
1. Verbs of Motion/Removing	P	P	A	P	A	P	A
2. Verbs of Process, Change, Development	N	A	A	N	N	P	N
3. Verbs of Beginning/End of Action:	N	N	A	A	P	N	N
4. Verbs of Physical Action	N	N	N	P	A	A	A
5. Engender Verbs	N	A	P	N	N	N	N
6. Destroy Verbs	A	A	P	N	N	N	N
7. Verbs of Successful/Unsuccessful Action Implementation	N	A	A	N	P	P	N
8. Verbs of Attempt	A	N	N	P	P	N	A
9. Verbs of Sound Emission	A	N	N	P	N	N	N
10. Verbs of Light Phenomena	A	A	A	A	N	P	P
11. Verbs of Temperature Phenomena	N	P	N	N	N	N	N
12. Verbs of Nature Phenomena	P	N	A	N	N	N	N
13. Verbs of Communication	N	P	P	N	A	A	P
14. Verbs of Moral Impact/Effect	N	N	P	N	N	A	N
15. Verbs of Social Activity	N	A	N	P	P	N	N
16. Position Verbs	A	P	A	N	N	N	N
17. Verbs of Existence	P	N	P	A	N	A	N
18. Modality Verbs	N	N	N	A	N	N	P
19. Verbs of Human Relations	N	A	N	P	N	N	N
20. Verbs of Reference	P	A	N	N	N	N	P
21. Verbs of Emotional Psychological Impact	N	N	N	N	P	N	N
22. Verbs of Ownership/Loss	N	A	N	A	N	P	P
23. Verbs of Physiological State	A	A	N	N	N	P	N
24. Verbs of Perception	N	P	N	N	N	N	A
25. Verbs of Mental Activity	N	N	N	N	N	N	N
26. Verbs of Subjective Assessment	N	N	A	P	P	A	N
27. Verbs of Emotional Psychological State	P	P	A	P	A	P	A

Values of $z > 1.96$ (denoted as P) where empirical frequencies of the semantic verb class usage exceed the theoretically expected values, underline the author's promotion of the selected verb classes in the semantic text structure and provide material for further analyses.

Statistically relevant z of six semantic verb classes are characteristic for Ernest Hemingway's novels: *verbs of position*, *verbs of perception*, *verbs of communication, verbs of motion/removing, verbs of temperature phenomena, engender verbs*. Consequently, the dominant semantic verb classes in Ernest Hemingway's novels emphasize dynamic presentation of reality.

The greatest number of statistically relevant z indexes (eight cases) is discovered in works by Irvin Show: *verbs of motion/removing*, *verbs of physical action*, *verbs of attempt*, *verbs of sound emission*. The less relevant indexes are characteristic for *verbs of social activity*, *verbs of human relation*, *verbs of mental activity* and *verbs of subjective assessment*. The analysis of verbs in Irvin Show's novels divides the verb-stock into two main categories – “motion, action“ and “social life“.

Eight cases of statistically relevant indexes are found in the novels by Joseph Heller: *verbs of emotional psychological state, verbs of successful/unsuccessful action implementation*, *verbs of subjective assessment*, *verbs of beginning/end of action*, *verbs of emotional psychical impact, verbs of attempt*, *verbs of social activity*, *destroy verbs*. The dominant semantic verb classes underline the author's description of the heroes' internal world.

A half of the semantic verb classes in Norman Mailer's novels show statistically relevant results. Among them: v*erbs of process, change, development, verbs of light phenomena*, *verbs of successful/unsuccessful action implementation*, *verbs of physiological state, verbs of ownership/loss*, *verbs of motion/removing*. The semantic verb classes in works by Norman Mailer reflect the author's perception of reality as a dynamic process; the verbs of the light phenomena create an appropriate background of the narration.

The statistical analysis of the verb classes in the novels by F.M.Ford finds the advantage of empiric values in five cases. Most relevant is the value for the class of *verbs of modality,* then *verbs of ownership/loss, verbs of light phenomena, verbs of communication, verbs of reference*.

Five cases of statistically relevant z indexes are found for Kurt Vonnegut: *verbs of existence*, *verbs of communication, verbs of moral impact/effect, engender verbs* and *destroy verbs*.

The novels by Irvin Show are marked by the least number of statistically relevant z indexes: *verbs of motion/removing, verbs of existence, verbs of nature phenomena* and *verbs of reference*.

The diversity of semantic verb classes used by the authors for the depiction of common concepts – war and peace – can be accounted for individual creative approach in writing and the authors' individual priorities of the concept perception.

Despite the variety of semantic verb classes used in the analyzed texts, several common verb classes are exploited by the seven authors. Thus, a definite similarity in the writer's voice of the authors can be traced.

The next step of the research is to distinguish similarity in the semantic verb classes' realization. The test for the prominence of the diagonal, proposed by G.Altmann (1987, cf. also Schulz, Altmann 1988), is applied to compare pairs of authors. A comparison can be found in a sample table.

Table 3
Comparison of the verb semantic classes usage
in the novels of E. Waugh and E. Hemingway

		Evelyn Waugh			
		A	N	P	$n_{i\cdot}$
Ernest Hemingway	A	3	6	1	10
	N	2	7	2	11
	P	1	3	1	5
	$n_{\cdot j}$	6	16	4	26 =n

To measure the association between the usage frequencies of the semantic verb classes by all the authors, the value of χ^2 is found in the following formula:

$$\chi^2 = \frac{n\left(n\sum_i n_{ii} - \sum_i n_{i\cdot}n_{\cdot i}\right)^2}{\sum_i n_{i\cdot}n_{\cdot i}\left(n^2 - \sum_i n_{i\cdot}n_{\cdot i}\right)} \qquad (2)$$

where $\sum_i n_{i\cdot}n_{\cdot i}$ is the sum of the products of marginal frequencies having the same index; $\sum_i n_{ii}$ is the sum of the numbers on the diagonal.

Two authors are similar in the verb classes usage if χ^2 is more than 3.84. The results of the calculation are stated in Table 4.

Table 4
Measure of similarity in the semantic verb classes usage

	Evelyn Waugh	Ernest Hemingway	Kurt Vonnegut	Irvin Show	Joseph Heller	Norman Mailer	Ford Madox Ford
Evelyn Waugh		9,83	9,55	9,49	9,50	9,79	9,63
Ernest Hemingway				9,54		9,54	9,47
Kurt Vonnegut					9,41		9,58
Irvin Show							
Joseph Heller						9,60	9,56
Norman Mailer							9,58
Ford Madox Ford							

The data in Table 4 indicate a certain similarity in the usage of semantic verb classes by all the writers under study. The likeness or difference of writers' voice is established through a comparative assessment of the degree of similarity of the analyzed pairs of writers.

Thus, Evelyn Waugh shows similar tendencies in the verb-stock to the rest of the authors. F.M.Ford has common verb classes with five authors: [F.M.Ford + Evelyn Waugh, Ernest Hemingway, Kurt Vonnegut, Norman Mailer, Joseph Heller]. Four cases of likeness are characteristic for the styles of Ernest Hemingway and Joseph Heller: [Ernest Hemingway + Evelyn Waugh, Irvin Show, Norman Mailer, F.M.Ford]; [Joseph Heller + Evelyn Waugh, Kurt Vonnegut, Norman Mailer, F.M.Ford]. Three cases of similar verb usage are found for Kurt Vonnegut and Norman Mailer: [Kurt Vonnegut + Evelyn Waugh, Joseph Heller, F.M.Ford]; [Norman Mailer + Evelyn Waugh, Ernest Hemingway, Joseph Heller]. Only two cases of similar usage of semantic verb classes are found in the novels by Irvin Show and Evelyn Waugh, Irvin Show and Ernest Hemingway.

The presence of the statistically fixed similarity in the realization of semantic classes of verbs in the writer's voice does not testify the absence of individual features of every author. It only specifies common characteristics of styles. It can be explained by similar world perception by the writers.

The realization of semantic verb classes in the novels by Irvin Show, Kurt Vonnegut, and Norman Mailer proved to be the most individual and original in terms of verbs selection as the least number of similar cases of the verb usage in the styles of the above stated authors were found.

The writer's voice of the analyzed authors has common features in the verb realization in the text. The current research is meant to extend the analysis of the writer's voice in providing a focus for verbal presentation of war and peace concepts in the novels of the two themes.

The results of the given research show that writer's voice varies with the individual author. Writer's voice of the analyzed authors is in strong connection to the author's character, plot, setting, and theme of the novel. Style includes the multitude of choices fiction writers make, consciously or not, in the process of writing a story. Writer's voice of the seven authors encompasses strategic choices of concepts (war and peace) verbalized in text and it also includes the tactical choices of word usage (semantic verb classes). The research shows that in the process of creating a novel, these choices appear to become the writer’s voice.

The lexical semantic layer of the text depends on the individual writer's voice and the concept reflected in the theme of a novel. All analyzed writers vary in their verb-stock as appropriate to the theme. They select one verb class or another to suit their purpose.

References

Altmann, G. (1987). Tendenzielle Vokalharmonie. *Glottometrika 8, 104-112.*

Altmann, G. (1996). The nature of linguistic units. *Journal of Quantitative Linguistics 3, 1-7.*

Babenko L. (2004). *Philological analyses of text. Fundamentals of theory, principles, aspects of analyses.* Moscow: Academic Avenue.

Bachtin M. (1986). *Literary Critical Articles*. Moscow: Fiction.

Chrapčenko M. (1976). *Creative individuality of an author and development of literature*. Kiev: Dnipro

Dixon R. (1991). *A New Approach to English Grammar on Semantic Principles*. Oxford: Oxford University Press.

Gal´perin I. (1981). *Text as an object of linguistic study*. Moscow: Science.

Giršman M. (1996). *Selected articles. Artistry, literature, rhythm, style, dialogue, mentality.* Donetsk:Lebed'.

Grišunin A. (1998). *Research Aspects of Text Studies.* Moscow: Heritage.

Levickij V. (2004). *Quantitative methods in linguistics.* Chernivtsi: Ruta.

Levin B. (1998). *English verb classes and alternations*. Chicago: University of Chicago Press.

Nikolina N. (2006). *Modern Belles-Lettres Text // Text as a Dynamic System.* Moscow: Institutional Technology.

O´Grady, W., Archibald, J. (2000). *Contemprorary linguistic analysis. An introduction.* Fourth edition. Toronto: Addison Wesley Longman.

Schachtel, E.G. (1977). The Analysis of Style in Writing. *Contemporary. Psychoanalisis. 13:178-199.*

Schulz, K.-P., Altmann, G. (1988). Lautliche Strukturierung von Spracheinheiten. *Glottometrika 9, 1-48.*

Selivanova O. (2006). *Modern Linguistics: Encyclopedia*. Poltava: Dovkilia-K.

Webster Dictionary (1972). *The Living Webster Dictionary of the English Language.* Chicago: The English-Language Institute of America.

Glottometrics 16, 2008, 27-30

Word length in Persian

Karl-Heinz Best, Göttingen

Abstract. The aim of this paper is to show that word lengths in some Persian texts follow the 1-displaced Hyperpoisson distribution. The findings lend support to the theory of word length distributions (Wimmer et alii 1994, Wimmer & Altmann 1996) once more.

Keywords: Persian, word length, Hyperpoisson distribution

Preamble

After examinations of word length in about 50 languages in the Göttingen Project *Quantitative Linguistics* (Best 2001), Persian word length data can be presented for the first time. The data originate from Rypka´s (1936) study of metre in three New Persian texts. They are

- *Shāh-Nāme*[1] (Book of Kings) by Firdausí;
- *Yūsuf ó Zuleichā* (Romance; by Rypka ascribed to Firdausí; according to Wilpert (1997: 456) the text was written by Amānī in ca 1083);
- *Garshāsp-Nāmē* (Book of Garshāsp) by Asadí.

These works originate from the early epoch of New Persian, namely from the time between 1010 and 1083. For his study Rypka selected from each text 500 verses (1000 half verses). Beside many other results he presents also data obtained from the three texts on word length measured in terms of number of syllable in the three texts. The arbitrary delimitation of text parts is not optimal (Best 2006: 39) but it did not have any negative effect on the modelling of word length in these texts.

Theoretical background

The theoretical background of the present research – just as with all the respective studies in the Göttingen Project – is the theoretical approach by Wimmer et al (1994) and Wimmer & Altmann (1996). Their general hypothesis is that different lengths of linguistic units occur in texts in accordance with some theoretically substantiated distributions. These distributions can be used for modelling length of linguistic units, and have proved effective in a great number of examinations (Best 1998, 2001). The one-displaced Hyperpoisson distribution has been tested on Rypka´s data and turned out to be adequate; this is almost always the case with data from older Indo-European languages (Best 1998: 158).

[1] The manner of writing follows Wilpert (1988).

The model of word length distribution in Persian

We fitted the Hyperpoisson distribution in 1-displaced form to the data because they do not contain zero-syllabic words. The formula is

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

a and *b* being parameters, $_1F_1$ (1; b; a) is the confluent hypergeometric function, i.e.

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

and $b^{(x-1)}$ = b (b+1) (b+2)...(b+x-2). The results of fitting (1) to the Persian texts are presented in Table 1. The computations were performed using the Altmann-Fitter (1997).

Fitting the 1-displaced Hyperpoisson distribution to Persian texts

The following symbols and abbreviations are contained in Table 1: *x* is the syllabic length of words, n_x is the number of words having length x in the given text, NP_x is the theoretical number of these words computed by means of the model; *a* and *b* are the parameters of the distribution, X^2 is the value of the chi-square. Since the text passages evaluated by Rypka are relatively long, we used rather the discrepancy coefficient $C = X^2/N$ as goodness-of-fit criterion (*N* = total number of words). The fitting should fulfil the criterion *C* # 0.01; this condition is met in all cases as shown in the table.

Table 1
Fitting the 1-displaced Hyperpoisson distribution to Persian texts

	Shāh-Nāme		*Yūsuf ó Zuleichā*		*Garshāsp-Nāmē*	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	832	835.62	951	957.65	1043	1041.93
2	1910	1918.31	1988	2001.90	2128	2125.82
3	1210	1223.45	1217	1213.21	1147	1128.87
4	526	453.08	499	429.94	308	344.57
5	89	118.22	68	107.66	88	73.80
6	10	28.33	12	24.63	14	12.17
7					1	1.84
a =	0.8831		0.8535		0.7179	
b =	0.3847		0.4083		0.3519	
*X*² =	31.013		32.331		7.565	
C =	0.0068		0.0068		0.0016	

Figure 1 illustrates the result using the text *Shāh-Nāme*:

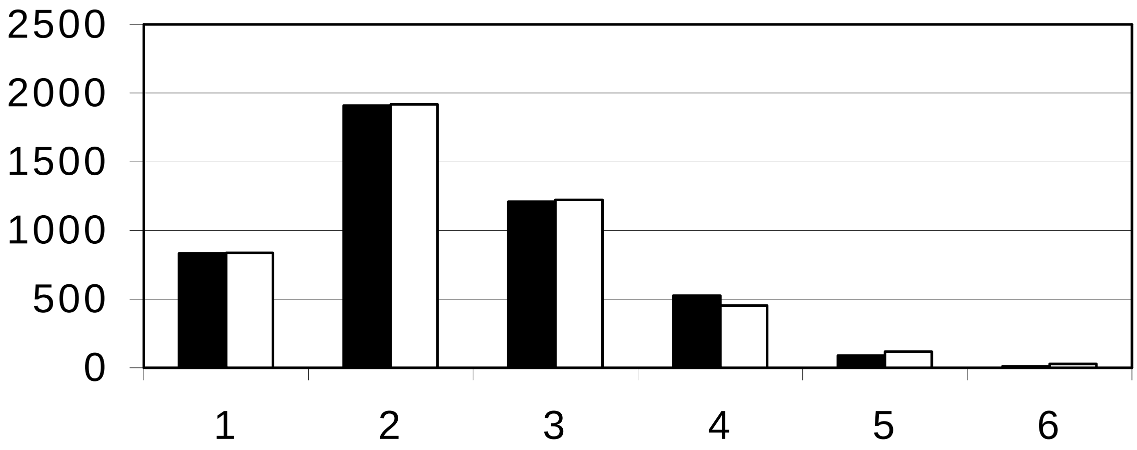


Figure 1. Fitting the 1-displaced Hyperpoisson distribution to *Shāh-Nāme*:

The black columns represent the empirical values, the white ones those computed with the aid of the Hyperpoisson distribution. Small differences can be seen in the classes of 4- and 5-syllable words, which are not sufficiently represented in the data.

Discussion and conclusion

The 1-displaced Hyperpoisson distribution can be successfully fitted to all the three texts. Again, it turns out to be an adequate model of word length distribution for texts of older Indo-European languages. Of course, three texts are not a very strong evidence. But there is some additional experience corroborating it.

1. It is in agreement with the results from other Indo-European languages: almost always the Hyperpoisson distribution is an adequate model for Old English, Old High German, Old Icelandic, Gothic or Latin texts. The exceptions are some Latin poems and an epos in verses; which is probably the influence of the text type (for bibliography see http://wwwuser.gwdg.de/~kbest/).

2. Further support of this result is a study by a student from Göttingen, Manaz Nolte, who evaluated 17 Persian reading-book texts (2003, not published). 13 of these texts follow the Hyperpoisson model. Still better result can be achieved using some other distributions (Cohen-Poisson d., Singh-Poisson d.), which are substantiated by the above-mentioned theory, too. That means, in the course of development of Indo-European languages, the patterns of word length seem to change, too, a phenomenon that can be observed in English.

If one compares the results from the three texts processed by Rypka with the studies just mentioned, one obtains a consistent picture: word lengths are distributed in texts in a law-like manner. However, their distribution is not always the same; it may depend on language, time, author, genre and possibly still other conditions. Even the difference between counting words forms or lemmas can influence the result (Grotjahn & Altmann 1993). In many cases, these factors are visible only in different parameter values of distributions, but in some other ones, one must consider another attractor, another distribution, as seems to be the case with New Persian texts. More data from Persian and other Iranian languages would be necessary in order to corroborate or to reject the adequacy of the 1-displaced Hyperpoisson distribution in word length fitting.

.

References

Best, Karl-Heinz (1998). Results and Perspectives of the Göttingen Project on Quantitative Linguistics. *Journal of Quantitative Linguistics 5, 155-162.*

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

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

Grotjahn, Rüdiger, Altmann, Gabriel (1993). Modelling the Distribution of Word Length: Some Methodological Problems. In: Köhler, Reinhard, & Rieger, Burghard B. (eds.), *Contributions to quantitative linguistics* (S. 141-153). Dordrecht u.a.: Kluwer.

Rypka, Jan (1936). La métrique du mutaqárib épique persan. *Travaux du Cercle Linguistique de Prague 6. Études dédiées au quatrième congrès de linguistes, 192-207.* (Reprint: Nendeln: Kraus Reprint 1968)

Wilpert, Gero von (Hrsg.) (1997). *Lexikon der Weltliteratur. Band 1: Biographisch-bibliographisches Handwörterbuch nach Autoren und anonymen Werken A – K. 3., neubearbeitete Auflage.* München: Deutscher Taschenbuch Verlag.

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

Glottometrics 16, 2008, 31-37

Zipf´s mean and language typology

Ioan-Iovitz Popescu, Bucharest
Gabriel Altmann, Lüdenscheid

Abstract. Zipf´s law is not only an expression of the rank-frequency relationship of words but it also enables us to make statements about some morphological features of language, too. In the present study, several indicators are proposed and their mutual relations are studied. The data are taken from 20 languages.

Keywords: Zipf´s law, analytism, synthetism, hapax legomena

In a previous article (Popescu, Altmann 2008) we have shown that in the rank-frequency distributions of word *forms*, hapax legomena (words occurring once) occupy a specific number of ranks, a matter of fact generally known. A function of this number is a characteristic of synthetism/analytism of a language. Zipf´s curve crosses the sequence of hapax legomena (or its prolongation) at a special place depending on the morphological complexity of language. In a strongly synthetic language like Hungarian the empirical hapax legomena are situated for the most part above the Zipfian function, and in a strongly analytic language like Hawaiian, they are situated mostly below it. Thus the fitting of Zipf´s function in the form of a non-linear regression to rank-frequency data reveals not only the validity of this law, but its (say, least square) deviation in the domain of hapax legomena characterizes a language morphologically.

A logical consequence of this finding is the fact that if Zipf´s curve (sequence) runs mostly below the hapax legomena, then its mean must be smaller that the empirical mean

$$(1) \qquad M_E = \frac{1}{N}\sum_{r=1}^{V} rf_r\,,$$

where N = text length (number of tokens), V = vocabulary (= number of word form types), r = rank, f_r = frequency at rank r. Similarly, if Zipf´s curve runs mostly above the hapax legomena, its mean must be greater than that of the empirical values in (1). In order to quantitatively express this distance we set up a new indicator B in the form

$$(2) \qquad B = \frac{M_E - M_F}{M_E}\,,$$

where M_F denotes the mean of the fitting curve

(3) $f(r) = c/r^a$.

The indicator B has the following properties:

if $B > 0$, then the language tends to contain synthetic phenomena
if $B < 0$, then the language tends to get analytic

if $B = 0$, the language is balanced, containing both types of phenomena.

The greater |B|, the more the language tends to a morphological extreme. As an example consider the frequency count of word forms in the Hawaiian text Hw 05: Moolelo Mokuna III (taken from Popescu et al. 2008, see also Table 1 below). The empirical mean yields M_E = 68.7388. Now, using iterative fitting of (3) we obtain the curve $f(r) = 592.6243r^{0.7267}$. Its mean yields M_F = 170.3493. Inserting these two values in (2) we obtain

$B(Hawaii\ 05)$ = (68.7388 – 170.3493)/68.7388 = -1.4782.

Since the value of B is a direct consequence of the index A denoting the course of Zipf´s curve in its positional relation to hapax legomena and expressed formally as

$$(4) \qquad A = \frac{c}{(V - HL/2)^a},$$

where c is the scaling constant of Zipf´s curve, V is the vocabulary of text, $HL/2$ is the half of the range of hapax legomena, and a is the Zipfian exponent; <A, B> must yield a very rigorous relation, especially if one takes the means of all texts written in one language.

Another indicator playing the same role as A is the Zipf curve end frequency, that is, the value of the theoretical Zipf curve in point V, i.e. at the highest rank = V, yielding

$$(5) \qquad C = \frac{c}{V^a}$$

which is low in strongly synthetic languages and high in strongly analytic languages.

In Table 1 the results from 100 texts in 20 languages are presented. It can be shown that text length N does not play any role. Since we do not fit a distribution but a curve, the size plays a role only in computing the mean (since $N = \sum f(r)$).

Table 1
Indicators A, B and C from 100 texts in 20 languages
(B = Bulgarian, Cz = Czech, E = English, G = German, H = Hungarian, Hw = Hawaiian, I = Italian, In = Indonesian, Kn = Kannada, Lk = Lakota, Lt = Latin, M = Maori, Mq = Marquesan, Mr = Marathi, R = Romanian, Rt = Rarotongan, Ru = Russian, Sl = Slovenian, Sm = Samoan, T = Tagalog)

ID	V	HL	Zipf a	Zipf c	M_E	M_F	B	A	C
B 01	400	298	0.6850	41.8602	116.4139	109.6275	0.0583	0.9507	0.6909
B 02	201	153	0.5704	17.6950	63.1108	65.6908	-0.0409	1.1292	0.8593
B 03	285	212	0.5550	20.9975	87.5379	93.6461	-0.0698	1.1798	0.9114
B 04	286	222	0.6169	23.6917	91.5569	87.2274	0.0473	0.9790	0.723
B 05	238	187	0.6202	22.0499	75.3153	72.848	0.0328	1.0090	0.7405
Cz 01	638	517	0.7473	54.2844	205.6006	154.7747	0.2472	0.6416	0.4352
Cz 02	543	412	0.7169	51.9648	162.8963	139.7022	0.1424	0.8013	0.5692
Cz 03	1274	964	0.8028	175.4805	311.3947	268.0846	0.1391	0.8261	0.564
Cz 04	323	241	0.6228	23.3822	108.8831	97.3214	0.1062	0.8562	0.6401
Cz 05	556	445	0.8722	77.1944	164.7137	107.4763	0.3475	0.4864	0.3114
E 01	939	662	0.7657	145.9980	216.7004	216.0852	0.0028	1.0783	0.773
E 02	1017	735	0.7434	180.1325	202.6156	242.9598	-0.1991	1.4610	1.0468

E 03	1001	620	0.8179	254.7482	192.996	207.047	-0.0728	1.2123	0.8953
E 04	1232	693	0.8712	385.9532	223.1696	223.4339	-0.0012	1.0449	0.7836
E 05	1495	971	0.8009	319.1386	286.4662	313.164	-0.0932	1.2529	0.9148
E 07	1597	1075	0.7568	300.1258	303.6303	364.9494	-0.2020	1.5416	1.1301
E 13	1659	736	0.8034	811.1689	219.5143	343.8041	-0.5662	2.5688	2.1
G 05	332	250	0.6935	32.8211	105.5599	90.6857	0.1409	0.8129	0.5858
G 09	379	302	0.6523	32.5565	117.9433	109.0793	0.0752	0.9431	0.677
G 10	301	237	0.6053	21.8114	100.7583	92.9696	0.0773	0.9331	0.6893
G 11	297	232	0.5895	19.9677	100.9872	93.5783	0.0734	0.9320	0.696
G 12	169	141	0.6062	14.3627	59.9203	53.4282	0.1083	0.8888	0.6408
G 14	129	107	0.5755	10.8110	47.5543	42.7453	0.1011	0.8977	0.6595
G 17	124	84	0.5515	13.1021	39.8311	42.2041	-0.0596	1.1531	0.9179
H 01	1079	844	1.2268	214.2708	304.7397	69.6929	0.7713	0.0749	0.0407
H 02	789	638	1.1865	122.0057	253.4014	63.2871	0.7502	0.0824	0.0446
H 03	291	259	1.2114	44.9653	107.2308	28.3793	0.7353	0.0950	0.0466
H 04	609	509	0.9549	74.8581	205.1592	97.1793	0.5263	0.2753	0.1642
H 05	290	250	0.8168	30.9795	104.7337	65.8429	0.3713	0.4784	0.3018
Hw 03	521	255	0.7932	329.6012	69.9367	117.9251	-0.6862	2.8821	2.3069
Hw 04	744	347	0.7633	678.1305	75.0495	174.0335	-1.3189	5.3384	4.359
Hw 05	680	302	0.7267	592.6243	68.7388	170.3493	-1.4782	6.2199	5.1825
Hw 06	1039	500	0.7816	1081.7823	91.914	230.7216	-1.5102	5.8855	4.7463
I 01	3667	2514	0.7266	509.5979	677.9826	865.2727	-0.2762	1.7784	1.3109
I 02	2203	1604	0.7488	305.6487	457.5523	505.2243	-0.1042	1.3468	0.9596
I 03	483	382	0.7895	56.8099	146.0597	110.6116	0.2427	0.6427	0.432
I 04	1237	848	0.7014	153.3448	275.2637	315.9784	-0.1479	1.3948	1.0391
I 05	512	355	0.6524	54.5840	134.0469	145.64	-0.0865	1.2306	0.9322
In 01	221	166	0.5809	18.2346	71.4973	71.1092	0.0054	1.0420	0.7926
In 02	209	147	0.5915	19.1717	66.3995	66.5723	-0.0026	1.0509	0.8132
In 03	194	130	0.5417	15.6229	62.7781	65.5138	-0.0436	1.1233	0.9005
In 04	213	145	0.4877	11.9156	74.8338	75.8346	-0.0134	1.0683	0.8721
In 05	188	121	0.5374	19.4218	53.3671	63.8473	-0.1964	1.4347	1.1645
Kn 003	1833	1373	0.6072	66.4545	576.1998	539.1967	0.0642	0.9223	0.6936
Kn 004	720	564	0.5237	22.1001	261.3076	240.2214	0.0807	0.9144	0.7048
Kn 005	2477	1784	0.6621	124.5588	705.5287	664.299	0.0584	0.9480	0.7054
Kn 006	2433	1655	0.5809	95.9573	657.818	740.4287	-0.1256	1.3181	1.0353
Kn 011	2516	1873	0.5786	77.0267	764.0881	767.8495	-0.0049	1.0862	0.8297
Lk 01	174	127	0.6416	23.4838	50.0667	52.6722	-0.0520	1.1474	0.8575
Lk 02	479	302	0.7731	139.2126	89.0171	112.9533	-0.2689	1.5798	1.1788
Lk 03	272	174	0.7512	71.8668	57.7355	68.9918	-0.1950	1.4240	1.066
Lk 04	116	80	0.6792	18.7509	35.3927	34.4326	0.0271	0.9901	0.7429
Lt 01	2211	1792	0.7935	109.3668	771.113	461.7444	0.4012	0.3666	0.2427
Lt 02	2334	1878	0.8047	160.3530	716.6397	474.286	0.3382	0.4729	0.3126
Lt 03	2703	2049	0.6366	109.5291	803.9286	754.7652	0.0612	0.9695	0.7158
Lt 04	1910	1359	0.6505	129.2023	484.4184	525.0506	-0.0839	1.2627	0.9486
Lt 05	909	737	0.5877	34.1056	319.8213	278.5167	0.1291	0.8449	0.6225
Lt 06	609	521	0.5293	19.3370	230.4608	202.4373	0.1216	0.8726	0.6494

M 01	398	202	0.7680	185.4091	63.9248	95.7958	-0.4986	2.3386	1.8677
M 02	277	146	0.8197	123.4636	50.5234	62.835	-0.2437	1.5787	1.2285
M 03	277	133	0.7902	147.8281	46.2162	65.9788	-0.4276	2.1571	1.7364
M 04	326	192	0.8353	137.7184	58.6804	70.9494	-0.2091	1.4664	1.0958
M 05	514	239	0.7484	297.2460	69.4287	125.8978	-0.8133	3.3897	2.7807
Mq 01	289	91	0.8030	240.0615	44.6326	67.1753	-0.5051	2.9102	2.5361
Mq 02	150	86	0.7440	46.4870	33.6324	40.1976	-0.1952	1.4370	1.1177
Mq 03	301	138	0.9795	225.2046	50.6561	50.0045	0.0129	1.0853	0.841
Mr 001	1555	1128	0.6293	78.3965	450.1638	443.8837	0.0140	1.0210	0.769
Mr 018	1788	1249	0.6685	128.5531	454.4077	477.8562	-0.0516	1.1470	0.8606
Mr 026	2038	1486	0.6224	101.6971	559.1975	584.868	-0.0459	1.1758	0.8867
Mr 027	1400	846	0.6166	120.0829	312.5678	408.1721	-0.3059	1.7214	1.3789
Mr 288	2079	1534	0.6304	100.2890	588.117	589.042	-0.0016	1.0857	0.8122
R 01	843	606	0.6720	73.6423	228.9908	228.8815	0.0005	1.0739	0.7961
R 02	1179	908	0.7567	115.8007	328.5853	272.9949	0.1692	0.7930	0.5489
R 03	719	567	0.7175	60.8094	218.4913	182.4494	0.1650	0.7771	0.5423
R 04	729	573	0.6673	52.4236	222.1083	200.3455	0.0980	0.8993	0.6445
R 05	567	424	0.6746	48.1009	169.812	155.514	0.0842	0.9157	0.6677
R 06	432	353	0.6349	30.3691	141.4417	126.7049	0.1042	0.8995	0.6444
Rt 01	223	127	0.8575	123.9533	38.7252	48.4559	-0.2513	1.6008	1.2009
Rt 02	214	128	0.7469	83.2271	39.0686	55.5682	-0.4223	1.9726	1.5128
Rt 03	207	98	0.7208	78.6409	40.7635	55.916	-0.3717	2.0454	1.6835
Rt 04	181	102	0.7359	60.2092	37.5232	48.3329	-0.2881	1.6749	1.3128
Rt 05	197	73	0.6917	87.0541	37.9226	55.5516	-0.4649	2.5959	2.2528
Ru 01	422	316	0.6538	36.1404	129.4329	120.6856	0.0676	0.9437	0.6945
Ru 02	1240	946	0.7713	138.5450	323.625	278.493	0.1395	0.8251	0.5696
Ru 03	1792	1365	0.7106	158.2659	454.9782	445.0264	0.0219	1.0851	0.7719
Ru 04	2536	1850	0.7181	234.3457	598.9348	614.624	-0.0262	1.1661	0.8419
Ru 05	6073	4395	0.7826	775.3826	1215.696	1249.8376	-0.0281	1.2063	0.8488
Sl 01	457	364	0.7467	44.1840	146.7963	113.0045	0.2302	0.6665	0.4561
Sl 02	603	423	0.6846	68.9001	153.3246	162.5056	-0.0599	1.1571	0.8609
Sl 03	907	651	0.7685	115.2402	235.1974	207.9875	0.1157	0.8651	0.6147
Sl 04	1102	701	0.9187	334.8100	213.7368	179.9701	0.1580	0.7633	0.537
Sl 05	2223	1593	0.7232	240.2785	502.7643	535.6631	-0.0654	1.2572	0.9122
Sm 01	267	119	0.8285	177.1858	41.4405	59.8669	-0.4446	2.1315	1.7297
Sm 02	222	96	0.7752	123.5355	38.3578	55.1037	-0.4366	2.2641	1.8745
Sm 03	140	75	0.6858	58.1896	26.4554	40.6778	-0.5376	2.4320	1.9639
Sm 04	153	76	0.7925	89.0771	27.3927	38.2418	-0.3961	2.0738	1.6539
Sm 05	124	66	0.7161	46.3093	25.915	34.9991	-0.3505	1.8312	1.4673
T 01	611	465	0.7624	120.0367	133.617	144.6973	-0.0829	1.2995	0.902
T 02	720	540	0.7803	144.5780	157.1779	163.5462	-0.0405	1.2297	0.8522
T 03	645	447	0.7652	167.7334	119.4537	151.5339	-0.2686	1.6447	1.1877

For the sake of an easier survey we present in Table 2 the means of the above indicators for individual languages. It can easily be seen that the individual languages occupy mostly the same rank with all three indicators, i.e. the indicators are only different expressions of the same property. In order to display the relationships graphically, we use all texts and present the relation <*A*, *B*> in Figure 1 and <*A, C*> in Figure 2. Since the indicators *A* and *C* are both some functions of *V*, they are linked linearly: $C = 0.8408A – 0.0985$. However, *B* and *A* express synthetism/analytism from different points of view, hence their relationship is not quite linear. Nevertheless, we suppose a power curve which must, however, attain also negative values, hence we combine two functions and obtain

$$B = k(A^{-r} – A^{-s}),$$

in our case

$$B = 0.5331(A^{-0.1963} – A^{0.6861})$$

yielding $R^2 = 0.9859$. This curve can be used for typological purposes, too.

Table 2
Means of indicators *A*, *B* and *C* in 20 languages

Language	mean A		Language	mean B		Language	mean C
Hungarian	0.2012		Hungarian	0.6309		Hungarian	0.1196
Czech	0.7223		Czech	0.1965		Czech	0.5040
Latin	0.7982		Latin	0.1612		Latin	0.5819
Romanian	0.8931		Romanian	0.1035		Romanian	0.6407
German	0.9372		Slovenian	0.0757		Slovenian	0.6762
Slovenian	0.9418		German	0.0738		German	0.6952
Kannada	1.0378		Russian	0.0349		Russian	0.7453
Russian	1.0453		Kannada	0.0146		Bulgarian	0.7850
Bulgarian	1.0495		Bulgarian	0.0055		Kannada	0.7938
Indonesian	1.1438		Indonesian	-0.0501		Indonesian	0.9086
Marathi	1.2302		Italian	-0.0744		Italian	0.9348
Italian	1.2787		Marathi	-0.0782		Marathi	0.9415
Lakota	1.2853		Lakota	-0.1222		Lakota	0.9613
Tagalog	1.3913		Tagalog	-0.1307		Tagalog	0.9806
English	1.4514		English	-0.1617		English	1.0919
Marquesan	1.8108		Marquesan	-0.2291		Marquesan	1.4983
Rarotongan	1.9779		Rarotongan	-0.3597		Rarotongan	1.5926
Samoan	2.1465		Samoan	-0.4331		Samoan	1.7379
Maori	2.1861		Maori	-0.4385		Maori	1.7418
Hawaiian	5.0815		Hawaiian	-1.2484		Hawaiian	4.1487

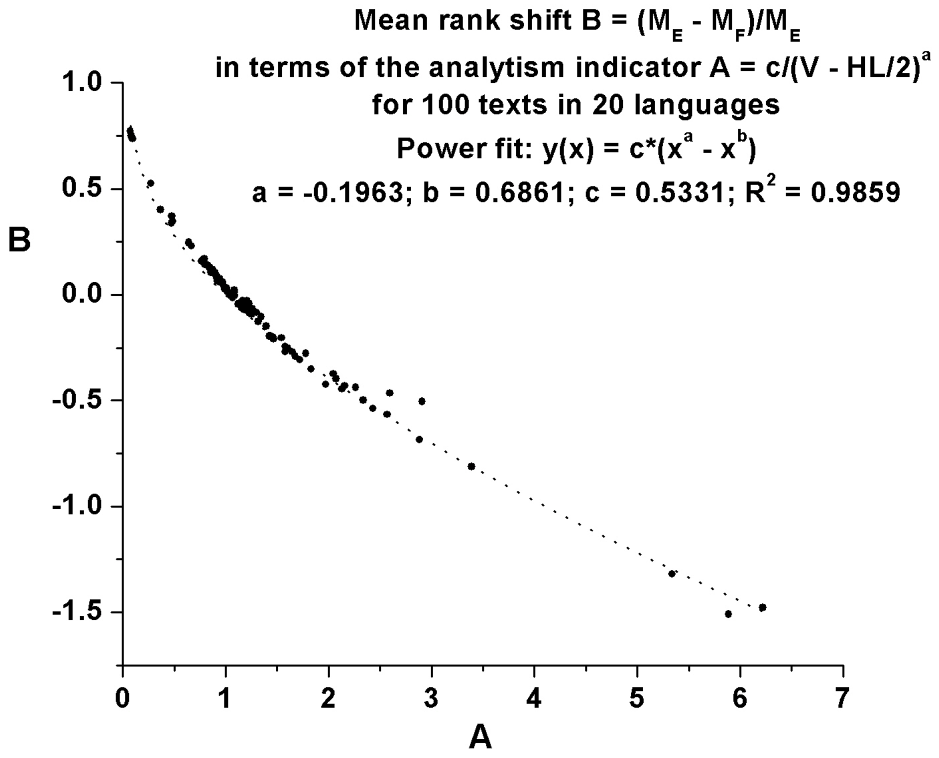


Figure 1. The relationship between indicators A and B

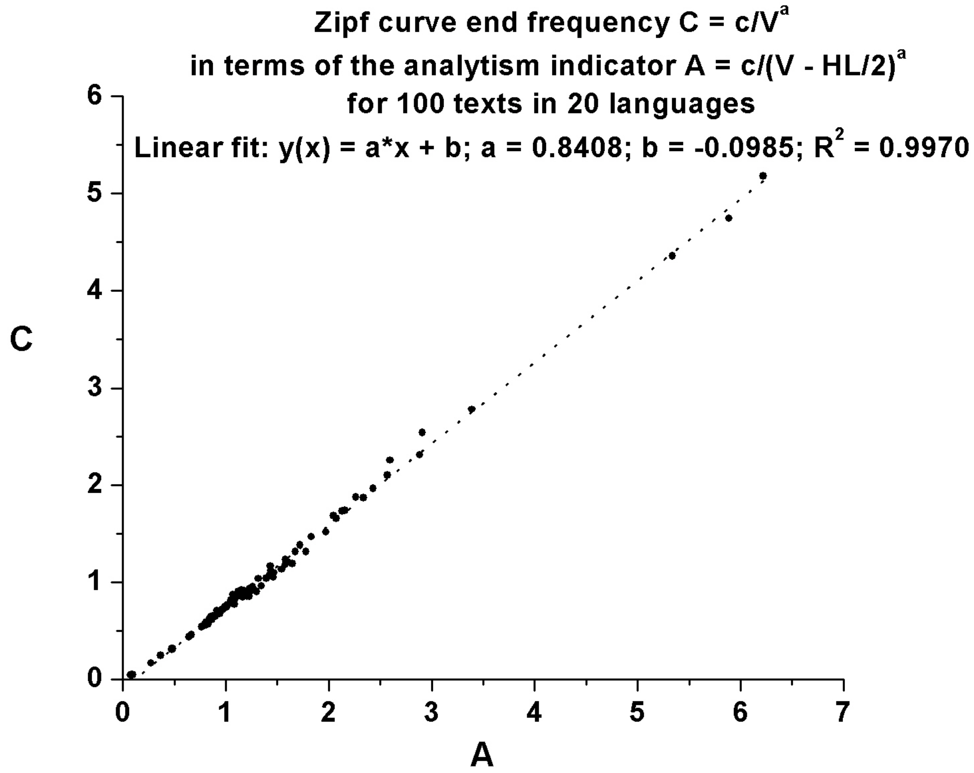


Figure 2. The relationship between indicators A and C

The fact that Zipf´s curve signalizes typological features means that in some cases it may display deviant behaviour when applied to rank-frequency data. Though in the overwhelming majority of fittings of Zipf´s (zeta) distribution to data one obtains very satisfactory results (cf. Popescu et al. 2008), the "best fit" or a fit crossing the hapax legomena exactly in their mean would, perhaps, bring some hint at the modification of Zipf´s curve in this domain. There are the following possibilities: (a) One varies the parameter "a" in order to obtain $M_E = M_F$ or $c/(V\text{-}HL/2)^a = 1$; (b) For B < 0 one uses a modification (e.g. Zipf-Mandelbrot, Lerch, Zipf-Alekseev) and for B > 0 another one. (c) One uses the same modification for both cases but with different parameters. (d) One uses a quite different way of reasoning. Using these possibilities one probably obtains a better fit, but the typological properties of the text (language) must be then inferred from different indicators. In any case we see that Zipf´s law yields deeper insights in language beyond the modelling of rank-frequency distributions.

References

Popescu, I.-I., Altmann, G. (2008) Hapax legomena and language typology (to appear)

Popescu, I.-I., Vidya, M.N., Uhlířová, L., Pustet, R., Mehler, A., Mačutek, J., Krupa, V., Köhler, R., Jayaram, B.D., Grzybek, P., Altmann, G. (2008). *Word frequency studies* (to appear).

Glottometrics 16, 2008, 38-45

Wortlängenverteilungen in französischen Briefen eines Autors

Nora Heinicke,[1] *Universität Göttingen*

Abstract. Earlier studies examining regularities in texts of modern French came to the result that the frequency with which words of different lengths are used in texts can be described by the Hirata-Poisson distribution. This study aims to clarify whether this rule also applies to letters written by the French author Charles Baudelaire in the 19th century. Furthermore this paper analyses whether the letters are homogeneous regarding the distribution of their word lengths.

Résumé. Jusqu'ici des études ayant analysé les lois dans les textes du français moderne ont conclu que c´est la distribution de Hirata-Poisson qui règle la fréquence avec laquelle des mots de longueurs différentes sont utilisés dans les textes. Le but de l´étude consiste à découvrir - au travers de l'analyse de lettres françaises d´un auteur du 19ième siècle - si c´est la même loi de la langue qui détermine la fréquence des longueurs du mot. Par ailleurs, ce travail comprend une recherche dans laquelle il s´agit de déterminer si ces lettres sont homogènes quant à la distribution de ses longueurs du mot.

Keywords: French, word length, Hirata-Poisson distribution, Ord´s criterion

0. Ausgangspunkt dieser Untersuchung ist die Frage, ob die Häufigkeit, mit der Wörter verschiedener Länge in französischen Briefen eines Autors verwendet werden, Gesetzmäßigkeiten folgt, wie vorhergehende Untersuchungen zu französischem Sprachmaterial (vgl. auch Feldt, Janssen & Kuleisa, 1997; Dieckmann & Judt, 1996; Wimmer & Altmann, 1996) erwarten lassen, und wenn dies zutrifft, ob es wiederum die Hirata-Poisson-Verteilung ist, die diese Gesetzmäßigkeiten regelt.

Da bislang primär Texte des modernen Französisch untersucht wurden, befasst sich diese Arbeit mit Daten des etwas älteren Französisch. Ausgewählt wurden Briefe Charles Baudelaires, die im Zeitraum zwischen 1832 und 1866 entstanden. Bei den 21 untersuchten Texten handelt es sich um Briefe Baudelaires an seine Verwandten und Freunde; der durchschnittliche Umfang der Briefe beträgt 350 Worte (längster Text: 628 Worte, kürzester Text: 131 Worte).

1. Briefe sind ein geeignetes Untersuchungsobjekt, da sie meist spontan, ohne längere Unterbrechungen und Neubearbeitungen und in einem natürlichen Wortlängenrhythmus verfasst werden, d.h., sie sind in sich relativ homogen. Sie sind im Funktionalstil der Alltagsrede geschrieben, Fachvokabular wird größtenteils ausgespart (vgl. Ammermann, 1997: 63; 2001: 62; Bartels, Strehlow, 1997: 71). Doch durch die große Zeitspanne, in der die Briefe verfasst wurden, und die wechselnde Adressierung können sich Unterschiede bezüglich der Wortlängenhäufigkeitsverteilung der Texte untereinander ergeben, da sich der Stil des Autors entwickelt und somit verändert haben kann (vgl. Ammermann, 1997: 66). Deshalb wird im

[1] Address correspondence to: Nora Heinicke, Roter Berg 4a, 38162 Cremlingen

Folgenden des Weiteren untersucht, ob die Texte hinsichtlich der Wortlängenverteilung untereinander ähnlich sind.

2. Um die Häufigkeit von Wörtern unterschiedlicher Länge, gemessen an der Anzahl der Silben pro Wort, untersuchen zu können, muss man „Wort“ und „Silbe“ im Voraus definieren. Als „Wort“ wird hier das „orthographische Wort“ verstanden, welches eine ununterbrochene Kette an Graphemen ist, die von Interpunktionszeichen und Leerstellen eingegrenzt wird (Trennungs- und Bindestriche, Apostrophe und Querstriche gelten hierbei nicht als Interpunktionszeichen). Die Anzahl der Silben pro Wort bemisst sich an der Zahl der Vokale bzw. Diphthonge/ Triphthonge im Wort (vgl. Best, 2006: 24). Somit werden die gleichen Kriterien für „Wort“ und „Silbe“ verwendet wie bei Feldt, Janssen und Kuleisa (1997). Bei Unklarheiten und in Zweifelsfällen, bedingt durch die orthographischen und phonetischen Besonderheiten des Französischen, werden Wörterbücher und Aussprachewörterbücher zur Hilfe genommen.

3. Die Überschriften und Ortsangaben in den Texten werden im Gegensatz zu Anrede und Schlussformel (hier inklusive des Autorennamens) nicht mitgezählt. Ausführungen des Autors, die der Schlussformel folgen, werden als Textbestandteil gewertet und ausgezählt.

Wie bei Feldt, Janssen, Kuleisa (1997) und Dieckmann, Judt (1996) werden die vokallosen, apostrophierten und dementsprechend nullsilbigen Worte wie *l`*, *d`* nicht als eigene Wortlängenklasse in die Tabellen aufgenommen, sondern als phonetischer Bestandteil ihrer Nachbarwörter angesehen. In verschiedenen vorhergehenden Modellierungsversuchen zeigte sich nämlich, dass die entsprechende Verteilung sowohl mit als auch ohne die nullsilbigen Worte an die Texte angepasst werden kann.

Abkürzungen wie *M* für *Monsieur*, *Mme* für *Madame*, *1er* für *premier* oder *etc.* für *et cetera* werden in ihrer gesprochenen Form gewertet. Zahlwörter werden wie folgt aufgeführt: 1983 = *mille-neuf-cent-quatre-vingt-trois*: 5 Wörter, 22 = *vingt-deux*: 2 Wörter.

Feldt, Janssen und Kuleisa (1997) folgend werden die mit Bindestrich verbundenen Worte unterschiedlich bewertet, je nachdem, ob sie als lexikalisierte Form angesehen werden können oder nicht: *voulez-vous* = 2 Wörter, *peut-être* = 1 Wort, auch Eigennamen wie *Saint-Victor* = 1 Wort.

Halbvokale wie in *lui*, *moi*, *serai* werden nicht mitgezählt, da sie nicht als Silbenträger fungieren, das „e-muet“ wird je nach Umgebung gesprochen oder nicht und dementsprechend gewertet (vgl. Bollée, 2002: 30f; Klein, 1963: 91f; Dieckmann und Judt, 1996: 159).

Zweisilbige Worte, deren zweiter Teil durch eckige Klammern von dem Ersten getrennt ist und deren erster Teil allein ebenfalls eine lexikalisierte Form darstellt, werden in diesem Kontext als zwei Wörter mit unterschiedlicher Silbenzahl bewertet: *vis[ite] = vis + visite* = 2 Wörter, *vol[ume] = vol + volume* = 2 Wörter.

4. Folgende Texte wurden für die Untersuchung ausgewählt:

T1 - Charles Baudelaire: Brief an Alphonse Baudelaire, Ende August oder Anfang September 1835. In: Baudelaire, Charles; *Correspondance 1, Janvier 1832 – Février 1860*. Hrsg. von Claude Pichois. Paris: Gallimard 1973. S. 34f.
T2 - Charles Baudelaire: Brief an Madame Aupick, 22. März 1837. In: Ebd. S. 38.
T3 - Charles Baudelaire: Brief an Madame Aupick, 5. Dezember 1837. In: Ebd. S. 48.
T4 – Charles Baudelaire: Brief an Colonel Aupick, etwa 18. Juni 1839. In: Ebd. S. 72f.
T5 – Charles Baudelaire: Brief an Madame Sabatier, 25. September 1857. In: Ebd. S. 429.
T6 – Charles Baudelaire: Brief an Madame Aupick, 17. November 1858. In: Ebd. S. 525.
T7 – Charles Baudelaire: Brief an Madame Aupick, 6. Juni 1856. In: Ebd. S. 349f.

T8 – Charles Baudelaire: Brief an Colonel Aupick, 26, Februar 1839. In: Ebd. S. 66ff.
T9 – Charles Baudelaire: Brief an Ernest Feydeau, 14. Juni 1858. In: Ebd. S. 506ff.
T10 – Charles Baudelaire: Brief an Madame Aupick, etwa 20. Januar 1860. In: Ebd. S. 661f.
T11 – Charles Baudelaire: Brief an Auguste Poulet-Malassis, 11. März 1860. In: Baudelaire, Charles; *Correspondance 2, Mars 1860 – Mars 1866*. Hrsg. von Claude Pichois. Paris: Gallimard 1973. S. 8f.
T12- Charles Baudelaire: Brief an Madame Aupick, 5. März 1866. In: Ebd. S. 625f.
T13 - Charles Baudelaire: Brief an Jules Troubat, 5. März 1866. In: Ebd. S. 626f.
T14 - Charles Baudelaire: Brief an Sainte-Beuve, 2. Januar 1866. In: Ebd. S. 562f.
T15 - Charles Baudelaire: Brief an Madame Aupick, 12 Januar 1866. In: Ebd. S.567.
T16 - Charles Baudelaire: Brief an Madame Aupick, etwa 15. März 1860. In: Ebd. S. 16.
T17 - Charles Baudelaire: Brief an Eugène Crépet, etwa 10. April 1860. In: Ebd. S. 21f.
T18 - Charles Baudelaire: Brief an Paul de Molènes, 12. Mai 1860. In: Ebd. S. 42.
T19 - Charles Baudelaire: Brief an Madame Aupick, 5. August 1860. In: Ebd. S. 71.
T20 - Charles Baudelaire: Brief an Auguste Poulet-Malassis, 8. September 1860. In: Ebd. S.90f.
T21 - Charles Baudelaire: Brief an Eugène Crépet, 8. November 1860. In: Ebd. S. 104.

5. An die Daten wurde die Hirata-Poisson-Verteilung angepasst, weil alle bisher untersuchten französischen Texte dieser Verteilung folgen mit Hilfe des Altmann-Fitters (1997). Da die nullsilbigen Worte in den Texten nicht gesondert gezählt werden, muss die Formel in 1-verschobener Form verwendet werden.

Die Formel lautet folgendermaßen:

$$P_x = \sum_{i=0}^{\left[\frac{x-1}{2}\right]} \binom{x-1-i}{i} \frac{e^{-a} a^{x-1-i}}{(x-1-i)!} b^i (1-b)^{x-1-2i}, \qquad x = 1, 2, \ldots$$

und lässt sich darstellen z.B. als Randomisierung des Poisson-Parameters durch die Normalverteilung. Dabei sind *a* und *b* die Parameter der Funktion. Außerdem werden in den Tabellen folgende Werte angegeben:

x	Wortlänge, gemessen an der Zahl der Silben
n_x	Anzahl der Wörter im Text mit Länge *x*
NP_x	nach der Hirata-Poisson-Verteilung berechnete theoretische Werte
X_k^2	Chiquadrat mit *k* Freiheitsgraden
P	Überschreitungswahrscheinlichkeit des Chiquadrats

Die Anpassung wird als zufrieden stellend betrachtet, wenn $P \geq 0.05$.

Die Anpassung der 1-verschobenen Hirata-Poisson-Verteilung an die Daten der 21 Briefe erbrachte folgende Resultate, die in Tabelle 1 angegeben sind:

Tabelle 1
Wortlängenverteilung in Baudelaires Texten

x	T1		T2		T3	
	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	214	216.1458	136	136.0000	156	155.6904
2	40	38.2540	40	40.2809	61	62.0491
3	37	33.0887	12	10.4409	15	13.7518
4	5	8.5114	1	2.2782	2	2.5087
	$a = 0.3144$; $b = 0.4370$ $X_1^2 = 2.0120$; P = 0.15		$a = 0.3290$; $b = 0.1000$ $X_1^2 = 0.9519$; P = 0.32		$a = 0.4074$; $b = 0.0218$ $X_1^2 = 0.2348$; P = 0.62	

x	T4		T5		T6	
	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	333	332.6915	82	82.0000	185	183.7998
2	107	108.4126	33	33.0000	58	58.0481
3	30	26.3188	15	12.0554	20	22.4826
4	3	5.5772	1	3.9446	8	6.6695
	$a = 0.3518$; $b = 0.0739$ $X_1^2 = 1.7245$; P = 0.18		$a = 0.4684$; $b = 0.1409$ $X_1^2 = 2.9173$; P = 0.08		$a = 0.3882$; $b = 0.1865$ $X_1^2 = 0.5474$; P = 0.45	

x	T7		T8		T9	
	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	267	268.1548	445	438.1890	340	340.9639
2	90	90.2123	120	121.4876	114	113.7135
3	40	35.5285	42	53.0534	47	44.5320
4	7	8.5491	16	11.5962	9	10.6357
5	1	2.5552	5	3.6739	3	3.1548
	$a = 0.4123$; $b = 0.1840$ $X_2^2 = 1.7955$; P = 0.40		$a = 0.3598$; $b = 0.2296$ $X_2^2 = 4.5781$; P = 0.10		$a = 0.4084$¸ $b = 0.1835$ $X_2^2 = 0.3994$; P = 0.81	

x	T10		T11		T12	
	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	261	260.5411	240	237.5717	388	388.7892
2	87	86.8051	82	80.9143	120	119.8268
3	28	29.3922	20	26.2691	42	39.8129
4	9	8.2616	10	7.2450	7	8.4764
5	-	-	-	-	2	2 0947
	$a = 0.3904$; $b = 0.1467$ $X_1^2 = 0.1332$; P = 0.71		$a = 0.3931$; $b = 0.1337$ $X_1^2 = 2.5832$; P = 0.10		$a = 0.3631$; $b = 0.1512$ $X_2^2 = 0.3834$; P = 0.82	

x	T13		T14		T15	
	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	250	249.7935	333	333.6143	202	200.5393
2	90	89.7457	119	119.3678	81	79.5015
3	35	35.7333	44	38.9708	19	24.2504
4	9	8.9767	5	8.8499	8	5.4489
5	3	2.7507	2	2.1972	1	1.2599
	a = 0.4377; b = 0.1793 X_2^2 = 0.0386; P = 0.98		a = 0.4106; b = 0.1285 X_2^2 = 2.3438; P = 0.30		a = 0.4387; b = 0.0965 X_2^2 = 2.4237; P = 0.29	

x	T16		T17		T18	
	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	120	119.5736	178	177.8521	186	186.4861
2	49	49.1651	78	78.0491	67	67.3134
3	15	15.6851	27	27.2559	25	22.6167
4	4	3.6786|	9	8.8429	5	6.5838
5	1	0.8976|	-	-	-	-
	a = 0.4578; b = 0.1018 X_1^2 = 0.0712; P = 0.78		a = 0.4958; b = 0.1148 X_1^2 = 0.0053; P = 0.94		a = 0.4170; b = 0.1345 X_1^2 = 0.6349; P = 0.42	

x	T19		T20		T21	
	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	108	108.2220	374	374.7368	152	151.8887
2	35	34.8881	131	131.4627	52	51.9730
3	10	8.9127	55	52.6900	17	17.2776
4	1	1.9772	13	13.0913	5	4.8607
5	-	-	2	3.2313|	-	-
6	-	-	1	0.7879|	-	-
	a = 0.3527; b = 0.0861 X_1^2 = 0.6164; P = 0.43		a = 0.4298; b = 0.1839 X_2^2 = 0.3634; P = 0.83		a = 0.3973; b = 0.1389 X_1^2 = 0.0085; P = 0.92	

Die Ergebnisse zeigen, dass die 1-verschobene Hirata-Poisson-Verteilung in allen Fällen an die Daten der 21 Briefe zufrieden stellend angepasst werden kann.

6. Des Weiteren wird untersucht, ob es sich bei den hier untersuchten Briefen um eine homogene Textklasse handelt. Die Anwendung des Ordschen Kriteriums, welches die Momente m_1, m_2 und m_3 der verwendeten Verteilung nutzt, ermöglicht eine graphische Veranschaulichung der Verhältnisse in den Texten. Hierbei stellt man zuerst die Größen I und S auf, die sich aus $I = m_2/m_1$ und $S = m_3/m_2$ berechnen.
Dabei sind:

$$m_1 = \frac{1}{N}\sum x f_x \; ; \qquad m_r = \frac{1}{N}\sum (x - m_1)^r f_x , \quad r \geq 2 ,$$

mit m_1 als Mittelwert der Verteilung, m_2 ist die Varianz und m_3 die Schiefe oder Asymmetrie der Verteilung (Best 2006: 68).

Die Ergebnisse sind in Tabelle 2 und Abbildung 1 dargestellt.

Tabelle 2
Das Ordsche Kriterium in französischen Briefen

Text	m_1	m_2	m_3	I	S
1	1.4358	0.5972	0.7276	0.4159	1.2184
2	1.3545	0.3876	0.4057	0.2862	1.0467
3	1.4145	0.4222	0.4080	0.2985	0.9664
4	1.3721	0.3985	0.4085	0.2904	1.0251
5	1.5038	0.5248	0.4530	0.3490	0.8632
6	1.4502	0.5722	0.7374	0.3946	1.2887
7	1.4815	0.5805	0.6869	0.3918	1.1833
8	1.4331	0.6277	1.0267	0.4380	1.6357
9	1.4815	0.6083	0.8128	0.4106	1.3362
10	1.4416	0.5323	0.6477	0.3692	1.2168
11	1.4318	0.5294	0.6882	0.3697	1.3000
12	1.4168	0.5114	0.6709	0.3610	1.3119
13	1.5142	0.6632	0.9210	0.4380	1.3887
14	1.4573	0.5305	0.6359	0.3640	1.1987
15	1.4727	0.5644	0.7436	0.3832	1.3175
16	1.5026	0.5992	0.7737	0.3988	1.2912
17	1.5445	0.6179	0.6682	0.4001	1.0814
18	1.4664	0.5316	0.5752	0.3625	1.0820
19	1.3766	0.4036	0.4126	0.2932	1.0223
20	1.5087	0.6527	0.9123	0.4326	1.3977
21	1.4469	0.5304	0.6289	0.3666	1.1857

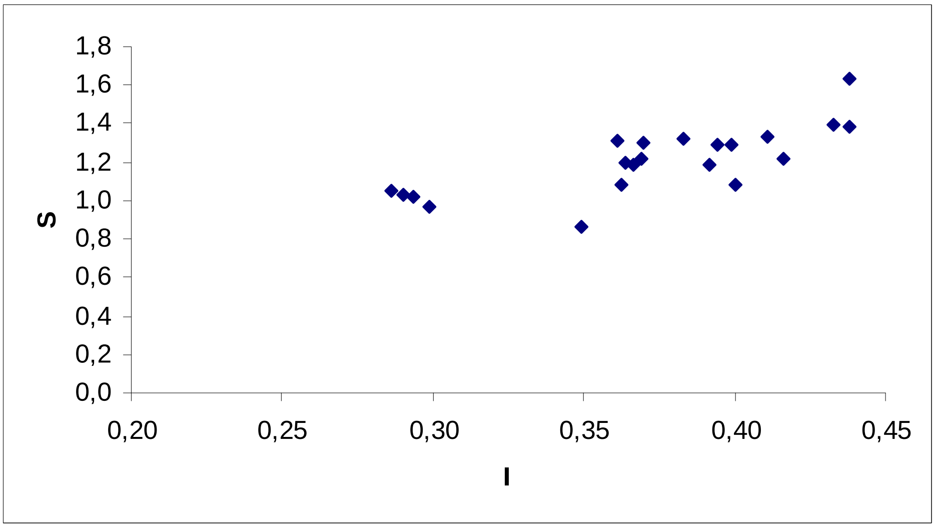


Abb. 1 Das Ordsche Kriterium in französischen Briefen

Wie die Graphik zeigt, ist bei dem untersuchten Sprachmaterial unter Berücksichtigung des Ordschen Kriteriums eine erhebliche horizontale Streuung der Wortlängen zwischen den einzelnen Texten zu beobachten. Jedoch liegen alle Texte im schmalen Bereich der Beta-Pascal-Verteilung. Von diesem Standpunkt her ist ihre Homogenität akzeptabel.

Die Homogenität aller Daten kann mit einem Chiquadrat-Test überprüft werden. Der übliche Chiquadrat-Test für Homogenität ergibt nach Einsetzung von Nullen bis zur Klasse $x = 6$ dort, wo es keine Werte gibt, $X^2 = 149.83$, was mit 100 Freiheitsgraden und $P = 0.001$ eine Heterogenität signalisiert. Testet man aber mit der Informationsstatistik *2I*, so bekommt man $2I = 136.37$, was abzüglich aller 32 Nullen $2I = 104.37$ ergibt. Mit 100 *FG* ergibt sich damit $P = 0.36$, was eine Homogenität signalisiert. Offenbar lässt sich das Problem nur für kleinere Textgruppen, z.B. Briefe an die gleiche Person, lösen.

7. Als Resultat der Untersuchung kann festgehalten werden, dass die Anpassung der 1-verschobenen Hirata-Poisson-Verteilung an die Textdateien in allen Fällen zufrieden stellende Ergebnisse erbrachte. Alle Texte erfüllen das Kriterium $P \geq 0.05$. Somit hat sich gezeigt, dass die Häufigkeitsverteilungen der Wortlängen in den französischen Briefen eines Autors, trotz der unterschiedlichen Länge und Adressierung, einem einzigen Modell unterliegen. Die Wortlängenverteilung des älteren Französisch folgt somit denselben Gesetzmäßigkeiten wie das schon zuvor untersuchte moderne Französisch.

Doch obwohl es sich bei den 21 Briefen Charles Baudelaires um Texte ein und derselben Klasse handelt, ist die Homogenität der Verteilungen etwas fraglich. Dies kann daran liegen, dass Charles Baudelaire im Verlauf der 34 Jahre, in denen er die 21 Briefe verfasste, seinen Stil änderte und/oder dass er seinen Stil dem jeweiligen Adressaten entsprechend anpasste und abwandelte. Ob dies der Fall ist bzw. welche Gründe hierfür noch vorliegen können, bedarf weitergehender Forschung zu den Briefen Baudelaires bzw. anderer Autoren.

Mit der Feststellung, dass die Hirata-Poisson-Verteilung ein gutes Modell auch für etwas ältere französische Briefe darstellt, kommt diese Untersuchung zu dem gleichen Ergebnis wie Knopp (1998), die in einer unveröffentlichten Arbeit je 20 Briefe von Voltaire und außerdem französische Briefe von Leibniz und Friedrich dem Großen auswertete. Bleibt nur die Frage offen, ob dieses Modell auch für andere Textsorten und noch ältere Texte gleich gut geeignet ist. In einer Seminararbeit (Schultz 2001) erwies sich, dass an 10 Gedichte Baudelaires die geometrische Verteilung angepasst werden konnte. Es ist auch durchaus denkbar, dass im Französischen ebenso wie im Englischen mit der Sprachentwicklung ein Modellwechsel von der Hyperpoisson-Verteilung, die selbst eine Verallgemeinerung der Poisson-Verteilung ist, hin zu anderen Verteilungen stattgefunden hat. Die Hirata-Poisson-Verteilung ist eine Zusammensetzung aus Poisson-Verteilung und Normalverteilung; gleichzeitig ist sie eine durch die Null-Eins-Verteilung verallgemeinerte Poisson-Verteilung und eine Faltung der Poisson-Verteilung mit der Doublet-Poisson-Verteilung (vgl. Wimmer, Altmann 1999: 25). Es ist anzunehmen, dass im Laufe der Entwicklung einer dieser Prozesse stattgefunden hat.

Literatur

Primärliteratur

Baudelaire, C. (1973). *Correspondance 1, Janvier 1832 – Février 1860. C.* Pichois (Hg.), Paris: Gallimard.

Baudelaire, C. (1973). *Correspondance 2, Mars 1860 – Mars 1866. C.* Pichois (Hg.), Paris: Gallimard.

Sekundärliteratur

Ammermann, S. (1997). Untersuchungen zur Wortlängenhäufigkeit in Briefen Kurt Tucholskys. In: K.-H. Best (Hg.), *Glottometrika* 16 (S. 63-70), Trier: Wissenschaftlicher Verlag Trier.

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

Bartels, O., & Strehlow, M. (1997). Zur Häufigkeit von Wortlängen in deutschen Briefen im 19. Jahrhundert und in der ersten Hälfte des 20. Jahrhunderts (Bismarck, Brecht, Kafka, Th. Mann, Tucholsky). In: K.-H. Best (Hg.), *Glottometrika 16* (S. 71-76), Trier: Wissenschaftlicher Verlag Trier.

Best, K.-H. (2006). *Quantitative Linguistik: Eine Annäherung*. 3., überarb. und erw. Aufl.. Göttingen: Peust & Gutschmidt.

Bollée, A. (2002). *Französische Phonologie und Orthographie*. http://web.uni-bamberg .de/split/sprachlabor/skripten/franzoesische_phonologie_und_orthographie. pdf (10.10.2007).

Dieckmann, S., & Judt, B. (1996). Untersuchung zur Wortlängenverteilung in französischen Pressetexten und Erzählungen. In: P. Schmidt (Hg.), *Glottometrika 15* (S. 158-163), Trier: Wissenschaftlicher Verlag Trier.

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

Knopp, A. (1998). *Wortlängen in französischen Briefen deutscher und französischer Verfasser.* Staatsexamensarbeit, Göttingen.

Schultz, M. (2001). *Wortlängen in deutscher und französischer Fassung von Gedichten Baudelaires.* Seminararbeit, Göttingen.

Wimmer, G., & Altmann, G. (1996). The theory of word length: some results and generalizations. In: P. Schmidt (Hg.), *Glottometrika 15* (S. 112-133), Trier: Wissenschaftlicher Verlag Trier.

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

Nachschlagewerke

Klein, H.-W. (1963). *Phonetik und Phonologie des heutigen Französisch.* München: Max Hueber.

Martinet, A., & Walter, H. (1973). *Dictionnaire de la prononciation française dans son usage réel.* Paris: France Expansion.

Robert, P. (2002). *Le Nouveau Petit Robert. Dictionnaire alphabétique et analogique de la langue française.* J. Rey-Debove u. A. Rey (Hgg.), Paris: Dictionnaires Le Robert 2002.

Software

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

Glottometrics 16, 2008, 46-56

Modelling polysemy in different languages: a continuous approach

Emmerich Kelih[1]*, Graz*

Abstract. This paper investigates the frequency of polysemy in six genetically unrelated languages. It can be shown that these distributions can be described by a power model, developed by the Estonian linguist Juhan Tuldava. Furthermore, interrelations between descriptive parameters of the analyzed empirical distributions have been obtained. Special attention has been paid to the behaviour of the parameters of the theoretical models, taking into account different influence factors (language analyzed, sample size, parts of speech).

Keywords: frequency of polysemy, parameter behaviour, interrelations

0. Introduction

Since Zipf (1935, 1949) it is a well known fact that the number of meanings follows certain regularities. These regularities can be explained by two opposing forces: (1) the forces of unification and (2) the forces of diversification; e.g. the most efficient way for the speaker would be one word with many different meanings, and for the hearer, one word with only one meaning. The first “force” (i.e. the speaker's economy) tends to reduce the effort of encoding, while the second “force” (i.e. the hearer’s economy) leads to a minimization of decoding effort. In other words, we are concerned with the principle of least effort, which competes with the necessity to communicate efficiently. The interaction of the forces of unification and diversification results – as on every other level in language – in a compromise in the form of self organization, e.g. in a specific shape of probability distributions of the number of meanings of words.

The focus in our contribution is not on a discrete model[2] for this distribution, but rather on an empirical re-analysis of a continuous model, developed by Tuldava (1979, 1998). According to Wimmer/Altmann (2005:792) it is basically irrelevant, whether linguistic regularities are being described by discrete or by continuous models. Both approaches are approximations to linguistic reality and they are – as shown in Mačutek/Altmann (2007) – transformable into one another. However, neither such theoretical problems nor a survey of the state of the art needs to be presented here. Cf. Levickij (2005) and Hoffmann (2001) for a comprehensive overview of quantitative studies of lexical polysemy.

The focus of this paper will be on the following problems:

1. empirical verification of the continuous model, developed by Tuldava (1979, 1998),
2. the integration of his model in a synergetic approach,
3. the interrelations between descriptive parameters of the analyzed empirical distributions,
4. the behaviour of the parameters of the theoretical models, taking into account the following factors of influence, such as the language analysed, the sample size of the examined dictionaries and parts of speech.

[1] Address correspondence to: Inst. für Slawistik, Merangasse 70/1; 8010 Graz. Austria. E-mail: emmerich.kelih@uni-graz.at

[2] Cf. the theoretical deduction of adequate discrete models in Wimmer/Altmann (1999) and the empirical corroboration of these models in Kelih (2007).

1.1 Continuous model for polysemy

One of the well known models[3] for the frequency of polysemy was developed by Tuldava (1979) and Tuldava (1998: 120). He postulates this modified exponential function to be adequate for modelling the number of meanings:

(1) $$.y = ae^{-b\sqrt{x}}$$

In formula (1) *y* denotes the relative frequency of words with a given number of meanings, *x* the number of meanings; *a* and *b* are parameters and *e* is the basis of the natural logarithm. According to Tuldava (1998: 120) the root of the number of meanings is a new measurement unit of the "semantic extent". The sequence of natural numbers 1,2,3 … (i.e. the root of 1,2,3…) marks different degrees of polysemy in languages.

Tuldava (1998:120) calculated, using the above mentioned formula, the theoretically expected relative number of meanings for three languages (Russian, Hungarian, English)[4]. All data are – as is the practice in quantitative analysis of polysemy – based on monolingual explanatory dictionaries. The starting point for the modelling is the relative frequency of words with 1,2,3 … meanings. Further details are given in Table 1. However, Tuldava (1979, 1998:120) did not perform a test or give an indicator which would give deeper information on the goodness of fit of the tested models. Therefore, we re-analyzed the data from Tuldava (1998, 1979), using an iterative approximation of the parameters *a* and *b* and calculating the determination coefficient *D* (cf. Table 1).

Table 1
Re-analysis: Data by Tuldava (1979, 1998)

x	**English**		**Hungarian**		**Russian (verbs)**	
	y		y		y	
	obs.	exp.	obs.	exp.	obs.	exp.
1	0.427	0.4263	0.504	0.5155	0.615	0.622
2	0.203	0.2049	0.265	0.2251	0.254	0.2159
3	0.117	0.1168	0.118	0.1192	0.071	0.0959
4	0.072	0.0727	0.052	0.0697	0.03	0.0483
5	0.048	0.0479	0.024	0.0435	0.013	0.0265
6	0.035	0.0328	0.013	0.0284	0.007	0.0153
7	0.023	0.0232	0.008	0.0192	0.003	0.0093
8	0.016	0.0168	0.005	0.0133	0.002	0.0058
9	0.013	0.0124	0.003	0.0094	0.002	0.0038
10	0.009	0.0093	0.002	0.0068	0.002	0.0025
11	0.0073	0.0071	0.0014	0.005	0	0.0017
12	0.006	0.0055	0.0012	0.0037	0	0.0011
13	0.0053	0.0042	0.009	0.0028	0	0.0008
14	0.0034	0.0033	0.007	0.0021	0.01	0.0006
15	0.0032	0.0026	0.007	0.0016	0	0.0004
> 15	0.014	0.0021	0.002	0.0013	0	0.0003
parameter	a	b	a	b	a	b
	2.4997	1.7688	3.8117	2.0007	8.0033	2.5546
D	0.9992		0.9891		0.9924	

[3] Further continuous models were developed by Krylov/Jakubovskaja (1977) and Polikarpov (1987).

[4] Tuldava (1979, 1998) took the data for Russian (verbs) from Krylov/Jakubovskaja (1977), for English from Višnjakova (1976) and for Hungarian from Papp (1967).

For all the three languages a determination coefficient *D* > 0.98 is obtained (cf. Table 1). This result must be interpreted as a convincing empirical verification of the model proposed by Tuldava (1979, 1998).

This first positive result is our starting point for a further empirical analysis on a larger data basis: we have based our study on 45 polysemy frequency distributions from Russian, English, German, Maori, Hungarian and Polish (cf. for details see Table 2).

Table 2
Analyzed languages and used resources

No.	Language	Specification[5]	Sample size[6] (N)	Source
1		Dic.-comp.; Ve;	2765	Levickij et al. (1999)
2		Dic.-comp., No.,	3278	
3		Dic.-comp., Adj.;	490	
4		Dic.-comp.; (no. 1-3)	6533	
5	Maori	Dic.-comp.;	7689	Wimmer/Altmann (1999)
6	Russian	Dic.-comp.; 11-14. century	2394	Andreevskaja (1990)
7		Dic.-comp.; 15.-17. century	2953	
8		Dic.-comp.; 18. century	3420	
9		Dic.-comp.; 19. century	4110	
10		Dic.-comp; 20. century	4185	
11		Dic.-comp., (no. 6-10)	17062	
12	English	Dic.-comp.; Adj.;	7191	Višnjakova (1976)
13		Dic.-comp.; Adv.;	287	
14		Dic.-comp.; No.;	15673	
15		Dic.-comp.; Ve.;	2796	
16		Dic.-comp.; (no. 12-16)	25947	
17	Russian	Dic.-comp.; SO; Ve.;	9502	Krylov/Jakubovskaja (1977)
18		Dic-sa.; SO; Ve. (I,K,S);	1329	
19		Dic-sa.; SSRLJA; Ve. (I,K,S);	2711	
20		Dic-comp; SO; Ve.;	10570	Krylov (1982)
21		Dic-comp; SO; No.;	16748	
22		Dic-comp; (no. 20-21)	32559	
23		Dic-comp; MAS;	82159	Polikarpov (1987)
24		Dic-comp; SO (9. edition)	57003	
25		Dic-comp; SSRLJA;	120481	
26	English	Dic-comp; HO;	44372	
27		Dic-comp; Sho;	79801	
28	Russian	Dic-sa.; MAS;	3931	Polikarpov/Krjukova (1989)
29		Dic-sa.; MAS; Adj.	431	
30		Dic-sa.; MAS; Adv.;	138	
31		Dic-sa.; MAS; No.;	1716	
32		Dic-sa.; MAS; Ve.;	1613	
33		Dic-sa.; MAS	3203	

[5] The abbreviations are as follows: Dic.-comp.: complete dictionary; Dic-sa.: sample from dictionary; No.: nouns; Ve.: verbs; Adj.: adjectives; Adv.: Adverb; I,K,S: sample of lexemes with initial letters I, K and S.; SO: Slovar Ožegova, SSRLJA: Slovar' sovremennogo russkogo literaturnogo jazyka; MAS Slovar' russkogo jazyka pod. red. A.P. Evgen'evoj; HO: Hornby: Oxford Advanced Learner's Dictionary of Current English; Sho: Shorter Oxford English Dictionary. See the bibliographical references for further details on used issues, edition etc.

[6] The sample size N is the number of analyzed words.

34		Dic-sa.; SO;	3971	
35		Dic-sa., SO; Adj.	446	
36		Dic-sa.; SO; Adv.	136	
37		Dic-sa.; SO; No.;	1731	
38		Dic-sa.; SO; Ve.;	1630	
39	German	Dic-sa.; No.;	5919	Schierholz (1991)
40	Hungarian	Dic.-who.;	59574	Papp (1967)
41	Polish	Dic-sa.; No.;	13356	Hammerl (1991)
42		Dic-sa.; Ve.;	6053	
43		Dic-sa.; Adj.;	8777	
44		Dic-sa.; Adv.	1391	
45		Dic.-who.; (no. 41-45)	29577	

Our specific choice of data allows us to analyze whether the discussed Tuldava model is suitable for all the different languages used in this study. Analyzing the 45 data samples by calculating the parameters *a* and *b* (iterative approximation) and the determination coefficient, we get a very clear result: the average determination coefficient $\bar{D}$ = 0.9950, with a minimum of *D* = 0.9704 and a maximum of *D* = 0.9999. In other words, the model proposed by Tuldava (1998, 1979) seems to be suitable and adequate for all six languages.

The calculated determination coefficient *D* and the parameter *a* and *b* for every analyzed sample are in Table 3; furthermore, we have included two additional descriptive parameters, the average polysemy $\bar{x}$ and the relative frequency of words with only one meaning p_1. These two parameters will be used in a further analysis in chapter 2.

Table 3
Descriptive, theoretical parameters and *D*

No.	$\bar{x}$	$\mathbf{p_1}$	**D**	**a**	**b**
1	2.0886	0.5009	0.9904	3.77	-2.00
2	2.0799	0.4793	0.982	3.31	-1.90
3	2.2959	0.4347	0.9876	2.5	-1.72
4	2.0998	0.4851	0.988	3.42	-1.93
5	1.5763	0.6647	0.9997	12.47	-2.93
6	1.6817	0.7026	0.9978	24.47	-3.55
7	1.3356	0.786	0.9999	45.42	-4.06
8	1.2535	0.8228	0.9999	69.53	-4.44
9	1.2545	0.8236	0.9999	72.38	-4.48
10	1.2645	0.8117	0.9999	58.05	-4.27
11	1.3307	0.797	0.9999	54.1	-4.22
12	2.5574	0.4148	0.9918	2.24	-1.66
13	1.4286	0.7108	0.9954	17.64	-3.21
14	2.1437	0.5552	0.9999	5.98	-2.38
15	3.5293	0.2711	0.9712	0.91	-1.13
16	2.3997	0.4821	0.9997	3.55	-2.00
17	1.642	0.6151	0.9913	7.98	-2.55
18	1.6561	0.6185	0.9922	8.27	-2.58
19	2.1498	0.5242	0.9916	4.69	-2.18
20	1.5553	0.6662	0.9981	12.24	-2.91
21	1.372	0.7477	0.9994	26.05	-3.55

22	1.434	0.7204	0.9993	19.98	-3.32
23	1.503	0.7293	0.9998	26.1	-3.58
24	1.3764	0.7748	0.9999	41.24	-3.97
25	1.6973	0.634	0.9992	9.89	-2.74
26	1.3596	0.8161	0.9992	94.41	-4.75
27	2.0114	0.576	0.9999	6.81	-2.47
28	1.6566	0.6115	0.995	7.81	-2.54
29	1.5128	0.6473	0.9907	9.72	-2.70
30	1.3841	0.7101	0.995	16.07	-3.12
31	1.4953	0.6824	0.9977	13.73	-3.00
32	1.8574	0.5226	0.9868	4.23	-2.07
33	1.3562	0.6912	0.9704	12.61	-2.89
34	1.4077	0.7318	0.9993	22.38	-3.42
35	1.287	0.7848	0.9983	37.4	-3.86
36	1.2059	0.8162	0.9973	47.27	-4.06
37	1.3114	0.777	0.9994	35.38	-3.82
38	1.5595	0.6644	0.9988	12.22	-2.91
39	2.7363	0.4396	0.9998	2.68	-1.81
40	1.9455	0.5073	0.9862	3.85	-2.00
41	1.5419	0.6664	0.9986	12.19	-2.90
42	1.9091	0.5174	0.9888	4.10	-2.05
43	1.2706	0.8212	0.9999	73.88	-4.50
44	1.2919	0.8009	0.9999	52.49	-4.19
45	1.5248	0.6882	0.9998	15.37	-3.11

1.2 Integration of Tuldava’s model into the Wimmer/Altmann approach

In addition to the first empirical findings it will be shown that Tuldava’s model (1979, 1998: 120) can easily be integrated into the theoretical framework of Wimmer/Altmann (2005). According to Wimmer/Altmann (2005: 795) the model of Tuldava (1979, 1998) is a special case of a more common formula (“unified theory”). Hence this law of polysemy is a special case, which can be deduced from formula:

$$\frac{dy}{y} = \frac{-bc}{x^{1-c}}dx \quad .$$

It results in

(2) $y = Ce^{-bx^c}$, whereas Tuldava (1998: 120) fixes c = ½. So the model gets this final form:

(3) $y = Ce^{-b\sqrt{x}}$.

As shown above, this model describes the behaviour of the polysemy distribution in all six languages and thus the Wimmer/Altmann (2005) approach has been indirectly confirmed.

In the next chapter the attention will be drawn to the interrelation between parameters from the empirical frequency distributions and the statistical behaviour of the parameters *C* and *b*, that is in our case, the parameters *a* and *b*.

2. Empirical findings: Interrelations

2.1. Interrelation between relative frequency of words with one meaning and average polysemy

The frequency distribution of polysemy in the analyzed languages does not only follow a general, theoretically integrated model, but also shows an interesting and systematic picture with respect to the behaviour of the empirical parameters.

In dealing with frequency data of polysemy we are concerned with a *natural rank frequency*, e.g. the occurrence of meanings is a monotone decreasing curve from the first frequency class on. It is very likely that this monotony is responsible for a direct interrelation between the average polysemy $\bar{x}$ and the relative frequency of words with one meaning p_1. A priori we postulate that with a decreasing mean value $\bar{x}$ the frequency of p_1 increases. Interestingly enough we have not obtained a linear interrelation, but a monotonous decreasing power function. Cf. the visualization of this relation in Figure 1.

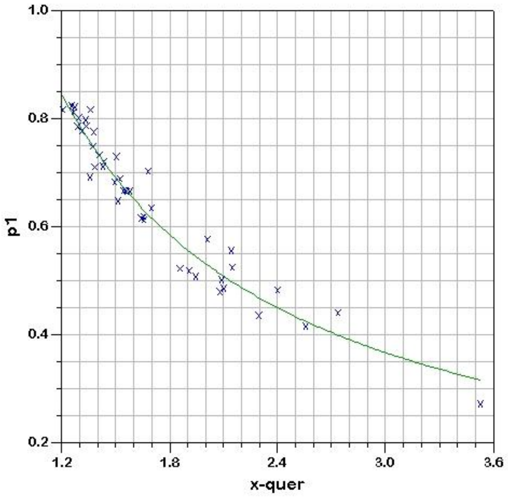


Figure 1: Dependency of $\bar{x}$ on p_1

A simple power function in the form $p_1 = c^d$ suffices to describe this interrelation. With d = -0.9138 a satisfying D = 0.95 can be obtained. This result is a strong empirical evidence for a harmonious relation between $\bar{x}$ and p_1.

Of course, it is certain that adding more data to our analysis the parameter will shift, but nevertheless we propose that the curve will definitely have a similar shape as the above one. So the forces of self organization are observable on the descriptive level already. The relative frequency of words with one meaning is predictable with only the mean value of the distribution.

2.2 Interrelations between empirical parameters and parameter *a*

In addition to the described relations above on the empirical level some more dependencies between the parameter *a*, the mean value $\bar{x}$ and the relative frequency of words with one meaning p_1 have been noticed. A priori we postulate a direct dependency between the parameter *a* and p_1, because the parameter *a* controls the “shift” of the curve on the *y*-axis. Hence the frequency of words with *one* meaning is controlled by *a*. Therefore it should hold true that with a decrease of p_1 the parameter *a* also decreases and because of the known dependency of p_1 on $\bar{x}$ (cf. Figure 1) the parameter *a* increases with a decreasing mean value $\bar{x}$. These two assumptions are already confirmed in a visualization of the mentioned dependencies (cf. Figure 2a and 2b).

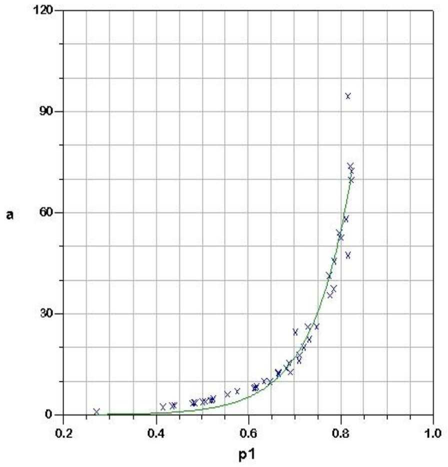


Figure 2a
Interrelation between p_1 and parameter *a*

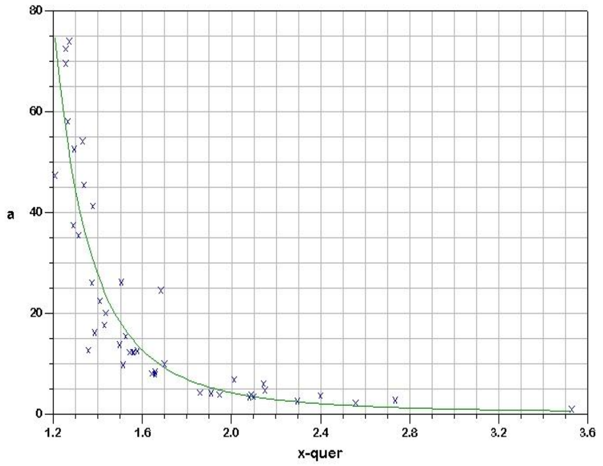


Figure 2b
Interrelation between $\bar{x}$ and parameter *a*

The first interrelation between p_1 and the parameter *a* can be captured by the simple formula: $a = c \exp(dp_1)$ with $D = 0.94$ (parameter $c = 0.0048$ and $d = 11.68$). For the interrelation between $\bar{x}$ and the parameter *a* (Table 2) the model $a = g\exp(-h/\bar{x})$ is suitable: With the parameters $g = 0.0553$ and $h = -8.6909$ a reliable[7] $D = 0.82$ is obtainable (cf. Figure 2b).

From Figure 2a it can be seen that from approximately $p_1 > 0.80$ the parameter *a* rises sharply. This observation is explainable by the fact that at this point a minimum of polysemy is reached. Above this point a “normal” and efficient communication is probably no longer possible. A similar behaviour is shown by the mean polysemy $\bar{x}$, which may never equal 1, since in this case a language would have no polysemy at all, e.g. this would lead to a severe complication and inefficiency of the communication act. Therefore the self-regulated behaviour of *a*, p_1 and $\bar{x}$ is a necessary precondition of the language system.

3.3. Parameter *a* and *b:* language specificity

The specific behaviour of the parameter *a* is the starting point for further analysis of this parameter. In chapter 1 a general cross linguistic valid model, based on Tuldava’s approach, has been found. Even if the existence of polysemy is supposed to be a “linguistic universal” (cf. Levickij 2006: 161f.; Croft 2003), the question of the language specificity of polysemy-distributions must be raised. In other words, are the parameters of our model specific for a certain language or not? In case of such specificity the conceptual power of our approach would be confirmed, since the general and the specific behaviour of the frequency distribution can be described at the same time.

To get an impression about this behaviour, we have calculated the mean values of the parameters $\bar{a}, \bar{b}$ from Tuldava’s model and the mean value of the average polysemy $\bar{x}(1)$ for Russian, German, English and Polish. For Maori and Hungarian the single values were taken as the basis for our interpretation (cf. Table 4). Because of an unbalanced number of sources per language the following comments and interpretations are preliminary and should be understood as a first attempt to a parameter interpretation of polysemy distributions.

[7] Dataset no. 26 has been excluded from the analysis, because of its unusual behaviour of the parameter *a* (outlier). Qualitatively (i.e. concerning the homogeneity of the data, type of the dictionary etc.) this decision cannot be justified for the time being.

Table 4
Number of sources, parameter *a* und *b* and mean values

Language	Number of sources (*n*)	Parameter $\bar{a}$	Parameter $\bar{b}$	$\bar{x}(1)$
Polish	5	31.6094	-3.3480	1.5077
Russian	26	26.8922	-3.3372	1.4823
English	7	18.7896	-2.5132	2.2042
Maori	1	12.4700	-2.9300	1.58
Hungarian	1	3.8460	-2.0000	1.95
German	5	3.1360	-1.8699	2.2601

For the time being only a simple qualitative interpretation of the parameters can be offered: The parameter $\bar{a}$ shows clear language specificity, because the values from all languages differ widely. We get the following "order" of languages: Polish, Russian, English, Maori, Hungarian and German (cf. Table 4). Due to the unbalanced number of sources (n) no deeper statistical analyzes are possible. Nevertheless, it is noticeable that the range of the parameter $\bar{b}$ is shorter than the range of parameter *a*. Furthermore, a direct, but statistically not significant, dependency between the parameter $\bar{a}$ and $\bar{b}$ is observable. Thus we postulate that both parameters contain some information about the languages examined. This assumption is supported by the fact that due to the different morphological structures of the languages (and presumably in dependency of the word length) polysemy is adopted in different ways. So it is very likely that morphology does have a significant influence on the specific shape of the distribution of polysemy. See also the considerations by Polikarpov (1979) on polysemy in dependency on the language type (analytic vs. synthetic).

3.4. Parameter *a* and *b*: sample size

The next step deals with the question: to what extent does the sample size of the analyzed dictionaries influence the parameters. We hypothesize that polysemy increases with an increasing lexicon size, since a larger dictionary should contain more meanings than a smaller one. To analyze this assumed relation only data from complete dictionaries will be used (data no. 5-10, 16, 23-27, 40, 45).

In fact empirically neither between the sample size N and the parameter *a*, nor between N and parameter *b* a dependency has been observed (cf. Figure 3a and 3b). One reason for the missing dependency is the high variation of the parameters. Another factor could be the small number of analyzed dictionaries.

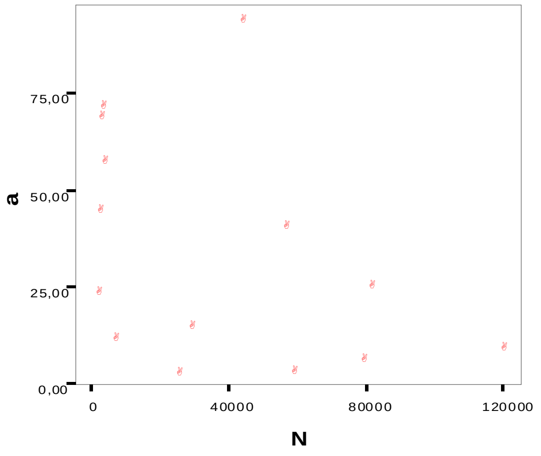


Figure 3a. Interrelation between N and a

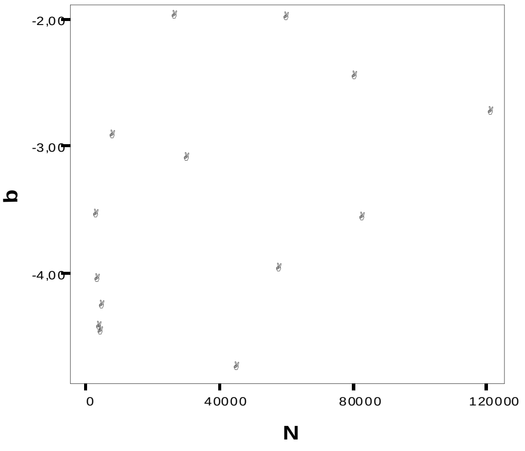


Figure 3b. Interrelation between N and b

So – at least based on the analyzed data – we propose that there is no clear interrelation between the sample size and the parameters. This observation is based on the fact that between $\bar{x}$ **,** p_1 and N no correlations are observable. Thus the analyzed languages and parts of speech have more influence on the frequency distribution of polysemy. The latter factor will be analyzed in the next chapter.

3.5. Parameter *a* and *b*: parts of speech as an influence factor

To end our contribution, the impact of parts of speech on polysemy will be analyzed. Here our main focus will be only on verbs, nouns and adjectives, due to the lack of reliable data for other parts of speech. For a first impression, without taking into consideration the individual languages, the average polysemy $\bar{x}(1)$ and the mean values of the parameters *a* and *b* ($\bar{a}$ **,** $\bar{b}$) were calculated.

It turns out that the verbs are very active with regard to their polysemy ($\bar{x}$**(**1) = 1.99), followed by nouns and adjectives. The parameter $\bar{a}$ of the analyzed verbs shows a rather "independent" behaviour, e.g. the values are very low (cf. Table 5), whereas the values for nouns and adjectives are much higher.

The average polysemy $\bar{x}$ (1) is directly responsible for the different values of the parameter $\bar{a}$ within the different parts of speech. Having only three sets of data at our disposal (cf. Table 5), it can be shown that with a decreasing average polysemy $\bar{x}$ an increase of the parameter $\bar{a}$ is observable. Because of the inadequacy of data no final interpretation can be offered.

Table 5
Parts of speech and parameter $\bar{a}$ and $\bar{b}$

Part of speech	**No. of sources (n)**	***$\bar{x}$ (1)***	***Parameter $\bar{a}$***	**Parameter $\bar{b}$**
Verbs	9	1.9941	6.4906	-2.2636
Nouns	7	1.8115	14.1885	-2.7649
Adjectives	5	1.7847	25.1468	-2.8870

Finally, we conclude our paper with an interpretation of the dependency of the parameter *a* with respect to language and parts of speech: All verbs in the languages (except for German) have the highest average of polysemy, which influences directly the values of the parameters $\bar{a}$ and $\bar{b}$. In regard to other parts of speech no clear results have been obtained (cf. Table 6).

Table 6
Language specific data for parts of speech

Language	**No. of sources (n)**	**Parts of speech**	**$\bar{x}$ (1)**	**Parameter $\bar{a}$**	**Parameter $\bar{b}$**
Russian	6	Verb	1.73	8.27	-2.53
	3	Noun	1.39	25.05	-3.45
	2	Adjective	1.39	23.56	-3.28
German	1	Verb	2.09	3.77	-1.85
	2	Noun	2.41	2.99	-1.90
	1	Adjective	2.3	2.30	-1.72
English	1	Verb	3.53	0.91	-1.13
	1	Noun	2.14	5.98	-2.38
	1	Adjective	2.56	2.24	-1.66
Polish	1	Verb	1.9091	4.11	-2.05
	1	Noun	1.5419	12.19	-2.91
	1	Adjective	1.2706	73.88	-4.50

3. Conclusions

The following conclusions are of interest for the further quantitative studies on polysemy:

1.3 The power model, developed by Tuldava (1979, 1998), can be easily integrated into the Altmann/Wimmer (2005) approach.

1.4 The discussed model is suitable for six different languages, e.g. cross linguistic evidence for modelling the polysemy is given.

1.5 The average polysemy $\bar{x}$ and relative frequency of words with one meaning p_1 are related systematically.

1.6 The parameters of the theoretical model give further information about (i) the language and (ii) the parts of speech. Interestingly enough no dependencies of the parameters on the sample size have been found.

Nevertheless these findings are preliminary and only further systematic analyzes will give deeper insights into the statistical characteristics of the frequency of polysemy.

References

Andreevskaja, A.V. (1990): Kvantitativnoe issledovanie polisemii korenych slov russkogo jazyka XI-XX vekov. In: *Kvantitativnaja lingvistika i avtomatičeskij analiz tekstov 6,* 3-11. [= Učenye zapiski tartuskogo gosudarstvennogo universiteta, 912]

Croft, W. (2003): *Typology and Universals.* Cambridge: University Press.

Hammerl, R. (1991): *Untersuchungen zur Struktur der Lexik: Aufbau eines lexikalischen Basismodells.* Trier: Wissenschaftlicher Verlag.

Hoffmann, Ch. (2001): Polylexie lexikalischer Einheiten in Texten. In: Uhlířová, L., Wimmer, G., Altmann, G., Köhler, R. (eds.), *Text as a linguistic paradigm: levels, constituents, constructs. Festschrift in honour of Luděk Hřebíček:* Trier: WVT, *76-97.*

Kelih, E. (2007): Diskretes Modell für die Polysemie: Neue empirische Evidenz, in: *Glottotheory, 1.* [submittted]

Köhler, R.; Altmann, G.; Piotrowski, R.G. (eds.) (2005): *Quantitative Linguistik. Quantitative Linguistics. Ein internationales Handbuch. An International Handbook.* Berlin u.a.: Walter de Gruyter. [= Handbücher zur Sprach- und Kommunikationswissenschaft, 27]

Krylov, Ju.K. (1982): Ob odnoj paradigme lingvističeskich raspredelenij. In: *Trudy po lingvostatistike 8: Lingvostatistika i vyčislitel'naja lingvistika: 80-102.* [= Učenye zapiski Tartuskogo Gosudarstvennogo Universiteta, 628]

Krylov, Ju.K.; Jakubovskaja, M.D. (1977): Statističeskij analiz polisemii kak jazykovoj universalii i problema semantičeskogo toždestva slova. *Naučno-techničeskaja informacija, Serija 2, 3, 1-6.*

Levickij, V. (2005): Polysemie. In: Köhler, R.; Altmann, G.; Piotrowski, R.G. (eds.) (2005), *458-464.*

Levickij, V.V. (2006): *Semasiologija.* Vinica: Nova Knyga.

Levickij, V.V.; Kijko, J.J.; Spolnicka, S.V. (1996): Quantitative Analysis of Verb Polysemy in Modern German. *Journal of Quantitative Linguistics. 3(2), 132-135.*

Levickij, V.V.; Drebet, V.V.; Kiiko, S.V. (1999): Some Quantitative Characteristics of Polysemy of Verbs, Nouns and Adjectives in the German Language. *Journal of Quantitative Linguistics 6,* 2, *172-187.*

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

Papp, F. (1967): O nektorych količestvennych charakteristikach slovarnogo sostava jazyka. *Slavica 7, 51-58.*

Polikarpov, A.A. (1979): *Ėlementy teoretičeskoj sociolingvistiki: nekotorye predposylki, rezul'taty i perspektivy pričinnogo podchoda v obščej semiotike i jazykozanii*. Moskva: Izdatel'stvo MGU.

Polikarpov, A.A. (1987): Polisemija: sistemno-kvantitativnye aspekty. *Kvantitativnaja lingvistika i avtomatičeskij analiz tekstov 3,* 135-154. [= Učenye zapiski tartuskogo gosudarstvennogo universiteta, 774]

Polikarpov, A.A.; Krjukova, O.S. (1989): O sistemnom sootnešenii kratkogo i srednogo tolkovych slovarej russkogo jazyka. *Kvantitativnaja lingvistika i avtomatičeskij analiz tekstov 5, 111-125*. [= Učenye zapiski tartuskogo gosudarstvennogo universiteta, 872]

Schierholz, Stefan (1991): *Abstraktheit, Häufigkeit und Polysemie deutscher Substantive*. Tübingen: Niemeyer. [= Linguistische Arbeiten, 269]

Tuldava, Ju.A. (1979): O nekotorych kvantitativno-sistemnych charakteristikach polisemii. *Linguistica XI, 107-141*. [= Učenye zapiski Tartuskogo gosudarstvennogo universiteta, 502]

Tuldava, Ju. (1998): Probleme und Methoden der quantitativ-systemischen Lexikologie. Trier: Wissenschaftlicher Verlag. [= Quantitative Linguistics, vol. 59] (= German translation of Tuldava 1987).

Višnjakova, S.M. (1976): Opyt statističeskogo issledovanija mnogoznačnosti slov v anglijskom jazyke. In: Guseva, E.K.; Andrjuščenko, V.M.; Revzin, I.I. (eds.) (1976): *Vyčislitel'naja lingvistika: 168-178.*. Moskva: Nauka.

Wimmer, G.; Altmann, G. (1999): Rozdelenie polysémie v maorčine. In: Genzor, J.; Ondrejovič, S. (eds.) (1999): *Pange lingua. Zbornik na počest' Viktora Krupu: 17-25*. Bratislava: Veda.

Wimmer, G., Altmann, G. (2005): Unified derivation of some linguistic laws. In: Köhler, R.; Altmann, G.; Piotrowski, R.G. (eds.) (2005), *791-807*.

Zipf, G.K. (1935): *The Psycho-Biology of Language. An Introduction to Dynamic Philology.* Boston: Hougthon Mifflin Company. [Neuauflage in Zipf, G.K. (1965), Cambridge/Massachusetts: M.I.T. Press]

Zipf, G.K. (1949): *Human Behavior and the Principle of Least Effort. An Introduction to Human Ecology.* Cambridge/Massachusetts. [Reprint in Zipf, G.K. (1972), New York: Hafner Publishing Company]

Glottometrics 16, 2008, 57-62

Zur Verteilung rhythmischer Einheiten in russischer Prosa

Marina Knaus, Göttingen[1]

Abstract. The purpose of this paper is to present some further evidence for the validity of the well-known law concerning the distribution of language entities of different lengths in texts. To this end it is shown that in 20 Russian texts of Tolstoj the length of rhythmic units abides by the 1-displaced binomial distribution.

Keywords: rhythm, Russian, binomial distribution

1. Ziel

Die Idee, deutsche Prosatexte auf die Verteilung rhythmischer Einheiten hin zu untersuchen, stammt von dem deutschen Psychologen Karl Marbe (1904), als ihm beim Lesen von Goethes „Sankt Rochusfest zu Bingen“ und Heines „Harzreise“ ein deutlicher Unterschied in der Rhythmik der Werke aufgefallen war. Die Untersuchungen von Marbe und einigen seiner Schüler und Kollegen gewannen neue Aktualität, als in der Quantitativen Linguistik die Hypothese aufkam, dass sprachliche Einheiten unterschiedlicher Länge sich in Texten gemäß bestimmten Sprachgesetzen verteilen sollten (Altmann 1988; Wimmer u.a. 1994). Tests ergaben, dass die 1-verschobene Hyperpoisson-Verteilung gut mit den Textabschnitten von Goethe und Heine übereinstimmte (Best 2001). Dieses Ergebnis ließ sich mit einigen weiteren Untersuchungen von Marbes Kollegen und Schülern wiederholen, gelang aber nicht immer (Best 2006a, b). Der Grund für gelegentlich schlechte Ergebnisse wird darin zu suchen sein, dass die Texte auf ungeeignete Weise bearbeitet wurden, indem man willkürlich Textabschnitte bildete. Derzeit ist festzustellen, dass bei besserer Datenaufnahme sehr gute Ergebnisse erzielt werden; die Datenbasis ist aber noch recht gering und soll mit diesem Beitrag etwas erweitert werden.

Diese Arbeit befasst sich mit der Verteilung rhythmischer Einheiten in russischen Prosatexten. Es geht also darum, zu prüfen, ob die rhythmischen Einheiten in russischer Prosa einem der Gesetzesvorschläge genügen, die Wimmer u.a. (1994) entwickelt haben.

2. Definition: Rhythmische Einheit

Als *rhythmische Einheit* bezeichnet man die Sprecheinheit im Text von einer betonten Silbe zur nächsten. Die Länge rhythmischer Einheiten ergibt sich aus der Anzahl der unbetonten Silben einschließlich der voranstehenden betonten. Gibt es keine unbetonte Silbe zwischen zwei betonten, so hat die rhythmische Einheit die Länge 1, bei nur einer unbetonten Silbe zwischen zwei betonten handelt es sich um eine rhythmische Einheit der Länge 2. Wenn zwei unbetonte zwischen zwei betonten Silben auftreten, ist es eine rhythmische Einheit der Länge 3, usw. (vgl. Best 2005).

[1] Adress correspondence to: Marina.Knaus@web.de

Die Ermittlung der Längen rhythmischer Einheiten gestaltet sich jedoch um einiges schwieriger als die des Wortes oder des Satzes, da „manche Wörter unterschiedliche Betonungen zulassen (*wéshalb und weshálb)*“ (Best 2005: 210) und die Auffassung des Textes an den Rezipienten gebunden ist.

Best schlägt aus diesen Gründen vor, die „Akzentuierungen eher als Momentaufnahmen des Textverständnisses durch den jeweiligen Bearbeiter auf[zu]fassen“ (ebd.).

3. Textauswahl

Der Untersuchung liegen 20 vollständige russische Prosatexte von Tolstoj aus „Izbranije sotčinenija“. Tom tretij. Moskva 1989. – „Ausgewählte Aufsätze”. 3. Band. Moskau 1989.) zugrunde. Die Texte sind zwischen 645 und 1348 Wörtern lang. Da es möglichst homogene Texte desselben Autors und aus einem Werk sein sollten, mussten zwei etwas kürzere Texte (unter 700 Wörtern) verwendet werden. Die Überschriften der Texte wurden nicht berücksichtigt. Sechs Texte enthielten vereinzelt französische und deutsche Wörter und sehr kurze Sätze. Diese wurden bei der Akzentuierung und Zählung ebenfalls berücksichtigt, da es unwahrscheinlich ist, dass diese geringe Menge an Wörtern das Ergebnis ernsthaft beeinflussen würde.

4. Modellanpassung

Vor der Akzentuierung wurden zunächst die Wörter jedes einzelnen Textes gezählt, um sicherzustellen, dass der Text nicht zu kurz bzw. auch zu lang ist. Die Akzentsetzung erfolgte beim langsamen lauten Lesen. Aus den akzentuierten Texten wurden dann die Tabellen erstellt.

Mit dem *Altmann – Fitter* (1997) konnte an die gewonnenen Daten in allen Fällen die Binomialverteilung angepasst werden. Nur bei Text 17 ist das Ergebnis nicht zufriedenstellend; es ist in diesem Fall aber auch nicht so schwach, dass man die Anpassung verwerfen müsste. Eine Anpassung der modifizierten Binomialverteilung gelingt mit *P* = 0.10, was zeigt, dass auch dieser Text mit der Verteilung rhythmischer Einheiten unterschiedlicher Länge nicht ganz chaotisch ist.

Die Formel für die 1-verschobene Binomialverteilung lautet:

$$P_x = \binom{n}{x-1} p^{x-1} q^{n-x+1}, \qquad x = 1, 2, ..., n+1$$

5. Ergebnisse

Die Tabellen zeigen das Ergebnis der Anpassung der 1-verschobenen Binomialverteilung an die 20 Texte. Ergibt die Anpassung, dass $P \geq 0.05$ ist, so gilt das Ergebnis als zufriedenstellend. Anpassungen mit $0.01 \leq P < 0.05$ werden nicht mehr als zufriedenstellend angesehen, werden aber noch toleriert (s. Text 17).

Legende zu den Tabellen:

x - Klasse der rhythmischen Einheit (s.o.)
n_x - Anzahl der rhythmischen Einheiten der jeweiligen Klasse im Text
NP_x - Anzahl der rhythmischen Einheiten der jeweiligen Klasse aufgrund der Anpassung der 1-verschobenen Binomialverteilung
n, p - Parameter der Binomialverteilung
X^2 - Chiquadrat
FG - Freiheitsgrade
P - Überschreitungswahrscheinlichkeit des Chiquadrats
\| - zusammengefasste Klassen

Die Ergebnisse:

	Text 1		Text 2		Text 3		Text 4	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	46	54.21	98	96.13	90	88.02	51	55.74
2	148	143.70	252	253.63	210	221.14	152	147.59
3	168	158.72	294	286.79	267	246.92	178	170.99
4	93	93.50	164	180.16	160	160.83	116	113.19
5	28	30.98	78	67.91	52	67.34	33	46.84
6	2	5.48|	15	15.36	22	18.80	16	12.40
7	2	0.40|	1	2.03	5	3.50|	2	2.05|
8					1	0.45|	1	0.20|
∑	487		902		807		549	
n =	6		7		9		8	
p =	0.3064		0.2737		0.2182		0.2487	
X^2 =	2.804		3.712		7.351		6.268	
FG =	3		4		4		4	
P =	0.42		0.45		0.12		0.18	

	Text 5		Text 6		Text 7		Text 8	
X	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	72	76.40	104	107.43	85	80.75	100	96.05
2	206	197.86	291	289.20	172	180.90	204	218.02
3	225	227.73	339	324.40	177	177.31	223	212.09
4	160	152.90	184	194.07	112	99.31	118	114.62
5	59	65.99	57	65.31	27	34.76	34	37.17
6	15	18.99	15	11.72|	8	7.79	5	7.23|
7	7	4.13	3	0.88|	1	1.18	1	0.78|
8							1	0.04|
∑	744		993		582		686	
n =	9		6		8		7	
p =	0.2234		0.3097		0.2188		0.2449	
X^2 =	4.534		4.674		4.052		2.132	
FG =	4		3		4		3	
P =	0.34		0.20		0.40		0.55	

	Text 9		Text 10		Text 11		Text 12	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	89	89.49	101	96.84	102	112.20	135	128.01
2	249	238.55	192	203.35	289	292.06	248	261.92
3	235	254.35	197	186.83	359	325.80	245	238.19
4	116	135.60|	94	98.08	189	201.91	126	126.36
5	62	36.15|	35	32.18	66	75.08	47	43.09
6	7	3.85|	5	6.76|	15	16.75|	6	9.80
7			1	0.96|	5	2.08|	2	1.64
8					1	0.11|		
∑	758		625		1026		809	
n =	5		8		7		9	
p =	0.3477		0.2079		0.2711		0.1852	
X^2 =	2.436		2.164		6.492		3.224	
FG =	1		3		3		4	
P =	0.12		0.54		0.09		0.52	

	Text 13		Text 14		Text 15		Text 16	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	90	87.82	93	92.59	107	103.64	161	175.80
2	198	205.23	212	208.79	185	192.66	409	365.99|
3	205	199.83	197	196.18	174	165.30	284	317.49|
4	110	103.78	95	98.31	79	86.67	147	146.88
5	23	30.31	23	27.71	35	30.99	41	38.23|
6	5	4.72|	7	4.17|	8	7.98	6	5.31|
7	1	0.31|	1	0.26|	1	1.76	2	0.31|
∑	632		628		589		1050	
n =	6		6		13		6	
p =	0.2803		0.2732		0.1251		0.2576	
X^2 =	2.768		3.851		2.401		1.99	
FG =	3		3		4		1	
P =	0.43		0.28		0.66		0.16	

	Text 17		Text 18		Text 19		Text 20	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	132	115.12	74	74.01	102	99.07	99	93.85
2	215	248.98|	166	180.52	198	202.58	203	214.73
3	241	230.79|	212	188.72	188	181.23	213	204.72
4	131	118.85	106	109.60	83	92.65	102	104.10
5	37	36.72	29	38.19|	37	29.60	33	29.77
6	1	6.81|	8	7.99|	4	6.88	2	4.83
7	0	0.70|	3	0.93|				
8	1	0.03|	2	0.05|				
∑	758		600		612		652	
n =	7		7		8		6	
p =	0.2360		0.2584		0.2036		0.2761	
X^2 =	8.97		4.722		4.506		3.309	
FG =	2		2		3		3	
P =	0.0113		0.09		0.21		0.35	

Zur Veranschaulichung der guten Ergebnisse diene die folgende Graphik zu Text 20 (Abb. 1):

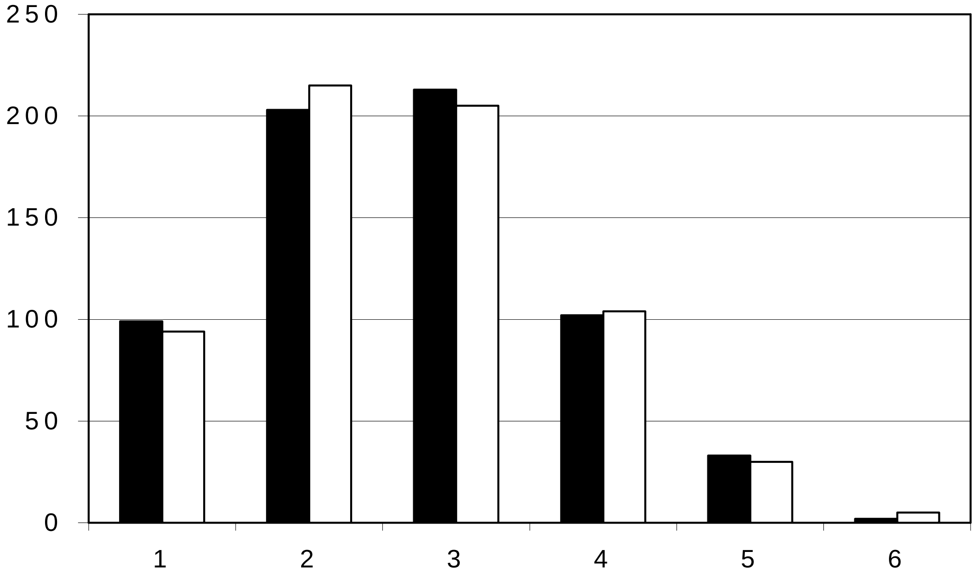


Abbildung 1. Anpassung der 1-verschobenen Binomialverteilung an Text Nr. 20

6. Zusammenfassung und Perspektive

Die Anpassung der 1-verschobenen Binomialverteilung ergab zufriedenstellende Ergebnisse; nur bei Text 17 ist das Ergebnis schwach, aber noch im Toleranzbereich. Die 20 untersuchten Texte folgen im Hinblick auf den Untersuchungsgegenstand also tatsächlich einem bestimmten Gesetz. Die rhythmischen Einheiten verhalten sich damit genau so wie die anderen untersuchten Entitäten, z.B. die Satz- und Wortlängen.

Um dieses Ergebnis abzusichern, müssen noch weitere Untersuchungen von Texten weiterer russischer Autoren und auch zu anderen Textsorten durchgeführt werden. Die bisher einzige Arbeit zum Russischen, die das gleiche Ziel verfolgt, Lehfeldt (2003), überprüft an vier russischen Prosatexten die Hypothese, die erweiterte positive Binomialverteilung entspreche einem Steuerungsmechanismus, welcher die Verteilung rhythmischer Einheiten in einem Text hervorbringe (vgl. Lehfeldt 2003: 171) und erhält sehr gute Ergebnisse für zwei Texte von Puschkin (P = 0.64 und P = 0.81). An die beiden südrussischen Dialekttexte kann diese Verteilung mit P = 0.06 und P = 0.08) ebenfalls noch erfolgreich angepasst werden. Das von Lehfeldt begründete Modell wurde auch auf die Texte von Tolstoj angewendet; die Ergebnisse sind etwas schlechter als die hier vorgestellten; der Vorteil der Binomialverteilung ist außerdem, dass sie einen Parameter weniger benötigt.

Zusätzlich sei auf Kagarov (1928) verwiesen, der vier Texte (Auszüge aus zwei Romanen und die ersten 1000 bzw. 10000 Silben eines Vortrags von Lenin) bearbeitet und unter anderen Gesichtspunkten betrachtet. Die Binomialverteilung kann an alle vier Textausschnitte angepasst werden; das Testergebnis bei der Anpassung der erweiterten positiven Binomialverteilung ist geringfügig schlechter, aber auch akzeptabel.

Vergleicht man die Ergebnisse von Marbe mit den in dieser Arbeit beschriebenen, so treten zwar Unterschiede in der Häufigkeitsverteilung zwischen den rhythmischen Einheiten in deutscher und russischer Prosa auf: sie folgen anscheinend verschiedenen Formen des Längenverteilungsgesetzes. Das von Wimmer u.a. (1994) vorgeschlagene zugrundeliegende Sprachgesetz ist jedoch dasselbe. Diese Ergebnisse müssen in Anbetracht der Tatsache, dass bisher nur recht wenige Texte ausgewertet wurden, noch als vorläufig betrachtet werden.

8. Literatur

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

Best, Karl-Heinz (2001). Zur Verteilung rhythmischer Einheiten in deutscher Prosa. In: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten: 162–166*. Göttingen: Peust & Gutschmidt.

Best, Karl-Heinz (2005). Längen rhythmischer Einheiten. In: Köhler, R., Altmann, G., Piotrowski, R. G. (Hrsg.): *Quantitative Linguistik /Quantitative Linguistics. Ein internationales Handbuch /An International Handbook: 208–214*. Berlin/ New York: de Gruyter.

Best, Karl-Heinz (2006a). Rhythmische Einheiten im Altgriechischen. *Göttinger Beiträge zur Sprachwissenschaft 13, 73-76.*

Best, Karl-Heinz (2006b). Lorenzo Bianchi (1889-1960). *Glottometrics 14, 72-74.*

Kagarov, E. G. (1928). O ritme prozaičeskoj reči. In: *Doklady Akademii Nauk SSSR*, (Serija B)*, 44-51.*

Lehfeldt, Werner (2003). *Akzent und Betonung im Russischen*. München: Verlag Otto Sagner.

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

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

9. Texte

Tolstoj, L. N. (1989). *Izbranije sotčinenija*. Tom tretij. Moskva.

(Text 1: Glava VI, 22-24; Text 2: Glava VII, 24-27; Text 3: Glava III, 13-16; Text 4: Glava XII, 34-36; Text 5: Glava XIII, 36-38; Text 6: Glava XIV, 38-42; Text 7: Glava XV, 42-44; Text 8: Glava XI, 31-34; Text 9: Glava XXIII, 68-71; Text 10: Glava XXVI, 76-78; Text 11: Glava II, 93-97; Text 12: Glava III, 97-99; Text 13: Glava V, 100-102; Text 14: Glava VII, 104-106; Text 15: Glava XIV, 121-122; Text 16: Glava XVIII, 130-134; Text 17: Glava XIX, 134-136; Text 18: Glava XX, 136-138; Text 19: Glava XXII, 139-141; Text 20: Glava XXVII, 149-151.

10. Software

Altmann-Fitter (1994). Lüdenscheid: RAM-Verlag.

Glottometrics 16, 2008, 63-79

Some properties of the Ukrainian writing system

Solomija Buk[1]*, Lviv*
Ján Mačutek[2]*, Bratislava*
Andrij Rovenchak[3]*, Lviv*

Abstract. We investigate the grapheme–phoneme relation in Ukrainian and some properties of the Ukrainian version of the Cyrillic alphabet.

Keywords: Ukrainian, phoneme-grapheme relation, script analysis.

1. Introductory remarks

Ukrainian is an East Slavic language spoken by about 40 million people in Ukraine and Ukrainian communities in neighboring states (Belarus, Moldova, Poland, Slovakia, Russia — especially in the so-called *Zelenyj Klyn* ‘Green wedge’ in the Far East Siberia from the Amur and Ussuri rivers eastwards to the Pacific), also in Argentina, Australia, Brazil, Canada, USA, and some others.

The features typical for modern Ukrainian are found already in the texts from 11th-12th cent. AD, they have been appearing systematically since 14th-15th cent. (Rusanivsjkyj 2004). Ukrainian uses the Cyrillic script. The Cyrillic alphabet, also known as *azbuka* (from old names of its first two letters **А** (**аꙁъ**) and **Б** (**боукы**)), has been traditionally used to write East and South Slavic languages (with the exception of modern Croatian and Slovenian), and also Romanian until 1860 (Jensen 1969: 491). As a result of political decisions it spread over a much larger area covering most (but not all) of languages in the former USSR, many of them using Latin or Arabic script before (cf. Comrie 1996b for a more detailed historical overview). Obviously, being applied in so different languages like Russian, Abkhaz, Tatar, Tajik or Chukchi (to give just a few examples) it had to represent much more phonemes than those occurring in Slavic languages, hence there are/were many language specific modifications of the alphabet (modified particular letters, diacritic marks or completely new letters, cf. Comrie 1996a). The Ukrainian version of the Cyrillic alphabet is called also *abetka* in vernacular from the names of the first two letters *a* and *be*. It consists of 33 letters:

< А а, Б б, В в, Г г, Ґ ґ, Д д, Е е, Є є, Ж ж, З з, И и, І і, Ї ї, Й й, К к, Л л, М м, Н н, О о, П п, Р р, С с, Т т, У у, Ф ф, Х х, Ц ц, Ч ч, Ш ш, Щ щ, ь, Ю ю, Я я >

[1] Department for General Linguistics, Ivan Franko National University of Lviv, 1 Universytetska St., Lviv, UA-79000, Ukraine, e-mail: solomija@gmail.com.
[2] Department of Applied Mathematics and Statistics, Comenius University, Mlynská dolina, 842 48 Bratislava, Slovakia, e-mail: jmacutek@yahoo.com.
[3] Department for Theoretical Physics, Ivan Franko National University of Lviv, 12 Drahomanov St., Lviv, UA-79005, Ukraine, e-mail: andrij@ktf.franko.lviv.ua, andrij.rovenchak@gmail.com.

When italicized, the following lowercase letters differ more or less significantly from the roman type: г — *г,* д — *д,* и — *и,* й — *й,* п — *п,* т — *т,* ш — *ш.*

Two letters are usually considered unique in the Ukrainian alphabet: < Ґ > and < Ї >. The first one denotes velar plosive [ɡ] and is used mainly in loan-words. This letter was first attested in 16th cent. and included in the alphabet by Meletius Smotrytsky in 1619 (Pivtorak 2004a). The use of < Ґ > was abolished by Stalin's regime in 1933 and reintroduced into the Ukrainian alphabet only in 1990. Regarding the uniqueness of this letter one must note however that < Ґ > was in use in the Belarusian orthography before 1933 but never officially revived until today except in some dissident editions (Katkouski and Rrapo n.d.), cf. also Barry 1997. The letter < Ї > in its modern phonetic value ([ji], see details below) was first attested in 1875 (Pivtorak 2004b) becoming thus the last standardized letter of the Ukrainian alphabet.

The apostrophe < ' > is not considered as a part of the alphabet but it plays an important role in the orthography (similar to that of the hard sign < ъ > in Russian), as described below. Ukrainian orthography is largely phonemic and thus can be referred to as a 'shallow' one (Coulmas 2004: 380). The deviations from the 'one letter to one sound' correspondence are few and the sound changes due to the assimilation are quite predictable and justified on the morphological level.

The letter < щ > always represents a two-phoneme combination /ʃʧ/ = /ʃ/ + /ʧ/.

The letter < ь > ('soft sign') is not given in a capital form, as it never stands in a word-initial position. This letter does not represent any sound but indicates the palatalization of a preceding consonant. The letters < є, ю, я > can represent one or two phonemes, depending on their position. When immediately following a consonant, they indicate the palatalization of the consonant and correspond to /ɛ, u, ɑ/, respectively. In a word-initial position, after < а, е, и, і, о, у, є, ї, ю, я >, < ь > and the apostrophe < ' > these letters represent two-phoneme combinations with /j/: /jɛ, ju, jɑ/. In modern Ukrainian the letter < ї > always corresponds to /ji/, nevertheless, it must be separated by the apostrophe from the preceding consonant. Historically, this letter originated from older < ѣ > and had a two-fold correspondence similar to < є, ю, я >, both /i/ with palatalizing a preceding consonant and /ji/. However, in modern literary Ukrainian the letter < і > replaced < ї > in the first case.

Originally, the palatalization of a preceding (mainly dental) consonant by < і > did not occur if this letter originated from older < о > (this fact is seen in < о > preserved in some word-forms: *стіл / стола, дім / дому, ніс / носа* versus *ніс / нести*). At present, such pronunciation, being influenced by the orthography, gradually becomes marginal though it is not considered incorrect.

Note that there is no special letter to indicate the palatalization before /ɔ/, unlike Russian or Belarusian < ё >. In Ukrainian, the combination < ьо > is used in this case.

We would like to note that in Ukrainian the letter < і > is used to represent the phoneme /i/. Within Slavic languages using the Cyrillic script only Belarusian has the same practice, in all the other orthographies the letter < и > is used to represent this phoneme. In Ukrainian, however, the grapheme < и > corresponds to the phoneme /ɪ/ (not to be mixed with close-central /ɨ/ typical for, e.g., Polish and Russian). Another difference with East Slavic orthographies is that the use of < е > is consistent with Slavic Latin and South Slavic Cyrillic practice. A special grapheme < є >, which inherited its outer form from old Cyrillic alphabet, has a two-fold nature serving to denote both the palatalization of a preceding consonant + /ɛ/ and /jɛ/, but not < е > — as in Russian and Belarusian.

2. Ukrainian phonetics

2.1. General

There is no universally accepted definition for the notion of phoneme in scientific literature. In this work, we consider a phoneme as a group of phonetically similar speech sounds. It is the smallest structural unit of language that can distinguish meaning of the words. This definition is close to, e.g., the Saint-Petersburg (Leningrad) School of phonology or American descriptivism versus, e.g., Moscow School of phonology. In particular, we consider the assimilatory changes within one morpheme as different phonemes but not as allophonic modifications. Our approach is consistent with the similar existing studies on the Slavic languages: Slovak (Nemcová and Altmann 2008) and Slovene (Kelih 2008).

It is commonly accepted that Ukrainian has 38 phonemes: 6 vowels and 32 consonants (Bilodid 1969; Ponomariv 2001; Žovtobrjukh and Khomenko 2004). The deviations from this number are linked with different approaches to the phonemic status of semi-palatalized and geminate consonants (Žovtobrjukh and Kulyk 1965: 109–110). The details of pronunciation further are given mainly according to Pohribnyj 1984. The IPA transcription is based on the tables given by Bilous (2005). We would also like to note an English-language source for Ukrainian phonetics (Zilynśkyj 1979).

2.2. Vowel phonemes

The vowel phonemes are /ɑ, ɛ, ɪ, i, ɔ, u/. For Ukrainian vowels, the difference between stressed and unstressed positions is not crucial. When unstressed, /ɑ/ has an allophone [ɐ], /ɔ/ has an allophone [o], this sound also slightly approaches /u/ if followed by a syllable containing /u/ or /i/, /u/ has an allophone [ʊ], the variations in the pronunciation of /i/ are very slight. Most problems concern the difference between unstressed /ɛ/ and /ɪ/. Depending on the phonetic environment, several variations of these sounds can be identified. In Ukrainian phonetic transcription based on the Cyrillic script they are denoted as [е$^{\text{и}}$] (closer to [ɛ]) and [и$^{\text{е}}$] (closer to [ɪ]). We will join them in one allophone [e] belonging to the phoneme /ɛ/ as it seems incorrect — within our approach — to relate one allophone with different phonemes.

2.3. Consonant phonemes

Consonants in Ukrainian appear, along with ordinary (‘hard’) forms, in palatalized (‘soft’) or semi-palatalized (‘semi-soft’) variants. The first group consists of 22 phonemes: /b, ʋ, ɦ, ɡ, d̪, ʒ, z̪, k, l̪, m, n̪, p, r, s̪, t̪, f, x, ʦ̪, ʧ, ʃ, ʣ̪, ʤ/. In the following text, we will not mark the dental character of the phonemes for simplicity.

The group of palatalized consonants consists of 10 phonemes: /j, dʲ, zʲ, lʲ, nʲ, rʲ, sʲ, tʲ, ʦʲ, ʣʲ/. There is no complete agreement about the nature of the palatalization of /rʲ/, sometimes it is considered as a semi-palatalized consonant (Ponomariv 2001: 16, 20). As there is no special IPA mark for semi-palatalization, we will use a superscript dotless ‘j’, e. g., /rʲ/. The palatalization of the consonants /bʲ, ʋʲ, ɦʲ, ɡʲ, ʒʲ, kʲ, mʲ, pʲ, fʲ, xʲ, ʧʲ, ʃʲ, ʤʲ/ is even weaker;

they are usually treated rather as the allophones of the respective ‘hard’ consonants, not as separate phonemes.

Ukrainian has the following sonorants: /ʋ, l, lʲ, m, nʲ, r, rʲ, j/. The labio-dental approximant /ʋ/ represented by < в > must not be mixed with, e. g., Polish or Russian fricative /v/, which falls into a pair with voiceless /f/. In Ukrainian, fricative /f/ is quite rare phoneme appearing only in loans and in onomatopoetic words. The Ukrainian phoneme /ʋ/ can appear in several allophonic modifications:

- non-syllabic [u] — [u̯] starts a syllable coda (*мав, був, мавпа, шовк*), in continuous speech this sound can be found in a word-initial position after a vowel of the preceding word (*а вперше* [ɐu̯pɛrʃe]).
- voiced labialized velar approximant [w] before /ɔ, u/ and voiced consonants (not after a vowel): *вниз, вона, вухо*;
- voiceless labialized velar approximant [ʍ] before voiceless consonants (not after a vowel): *вперше* [ʍpɛrʃe];
- semipalatalized labio-dental approximant [ʋʲ]: *він* [ʋʲin], *свято*[sʲʋʲɑto].

In a syllable-final position (being more precise, the first position of a syllable coda) the phoneme /j/ represented by < й > appears as a non-syllabic sound [i̯]: *хай, знайте*.

Sometimes the combinations of a vowel plus non-syllabic [u̯] or [i̯] are considered as diphthongs ([ɑu̯], [uu̯], [ɔi̯], etc.) but they are not phonemic in Ukrainian.

Most obstruents can be grouped into the “voiced–voiceless” pairs: /b/–/p/, /g/–/k/, /d/–/t/, /ʒ/–/ʃ/, /z/–/s/, /ʣ/–/ʦ/, /ʤ/–/ʧ/, /dʲ/–/tʲ/, /zʲ/–/sʲ/, /ʣʲ/–/ʦʲ/.

The articulation of the sound represented by < г > as voiced velar fricative [ɣ] instead of [ɦ] is incorrect. That is, the opposition between < г > and < х > (phonetically /ɦ/ and /x/) is not exact. Voiceless /f/ has no voiced counterpart.

No separate letters exist for the phonemes /ʣ/ and /ʤ/. They are represented by digraphs <дз> and <дж>, respectively: *дзвоник* /ʣʋɔnɪk/, *бджола* /bʤɔlɑ/. On the prefix-root boundary, however, these digraphs represent two phonemes: *надзвичайно* /nɑdzʋɪʧɑjnɔ/ (assimilates to /nɑʣzʋɪʧɑjnɔ/).

Also, no separate graphemes exist for the palatalized phonemes. To represent this feature, several techniques are used, see Table 2.

2.4. On the phonemic status of semi-palatalized consonants

As it was mentioned above, the following consonants have a semi-palatalized form: labials /bʲ, ʋʲ, mʲ, pʲ, fʲ/, velars /gʲ, kʲ, xʲ/, glottal /ɦʲ/, and postalveolar /ʒʲ, ʧʲ, ʃʲ, ʤʲ/. In Ukrainian, this phenomenon occurs mainly before < і >, and thus semi-palatalized sounds are the combinatorial allophones of the respective ‘hard’ consonants. However, in a few Ukrainian words labial < в > and < м > can appear before < я > or < ьо > (*свято, духмяний, тьмяний, цвьохнути*), having a sense-distinguishing role in, e.g., *свят* /sʲʋʲɑt/ (‘holiday’, Gen. Pl.) versus *сват* /sʋɑt/ (‘matchmaker’; ‘father of the son- or daughter-in-law’, Nom. Sing.) (Šerech 1951: 377). In the pronunciation of many speakers, there is a tendency to substitute semi-soft labials with a ‘labial + /j/’ combination: /ʋʲ/ → /ʋj/, /bʲ/ → /bj/, etc. (Bilodid 1969: 240), cf. also similar tendency in Polish (Swan 2002: 12). In loan-words, semi-palatal consonants, except postalveolar, can be found more frequently (*бюро, кюре, мюон, фюзеляж, ґяур*).

Semi-palatalized postalveolar /ʒʲ, ʧʲ, ʃʲ/ appear in most cases as geminate consonants in a stem-final position: *збіжжя*, *затишшя*, *ніччю*.

Semi-palatalized consonants are not found in the opposition of the respective hard consonants, except a very limited number of cases. Therefore, they are not treated as separate phonemes but as the allophones. However, it is possible that the phonological system of the Ukrainian language can change when the number of commonly used loans with semi-palatalized consonants becomes substantial.

2.5. On the phonemic status of geminates

In Ukrainian, geminate consonants appear mainly within morpheme boundaries. As a result of word formation, the gemination is produced by prefixation (*беззвучно*: *без* + *звучно*), suffixation (*законний: закон* + *н* + *ий*), or stem concatenation (*юннат*: *юн(ий)* + *нат(ураліст)*). In some Ukrainian words, geminates are preserved historically (*панна* ‘young lady; miss’, *манна* ‘manna’) and have a sense-distinguishing role (cf. *пана* ‘gentleman; sir’ Gen. Sing., *мана* ‘delusion’). Another source of geminates is connected with the loss of jers in the suffix < *ьj > (Bethin 1992): *знання*, *зілля*, *життя*, *сіллю*, *ніччю*, *затишшя*, *збіжжя*, *відповіддю*, *маззю*. This produces geminate dentals /dʲ, zʲ, lʲ, nʲ, sʲ, tʲ, ʦʲ, ʣʲ/ and postalveolar /ʒʲ, ʧʲ, ʃʲ/ (the phoneme/ʤ/ is too rare to occur in this position). It is interesting that labials /b, ν, m, p, f/, as well as /r/, are not geminated in such situations but appear as a ‘consonant + /j/’ combination: *любов’ю*, *верф’ю*, *пір’я*. In all the described cases, the geminate consonants are generally treated as a sequence of two identical phonemes, not a separate phoneme (Bilodid 1969; Ponomariv 2001; Žovtobrjukh and Kulyk 1965).

2.6. Assimilation

In modern Ukrainian, the regressive assimilation occurs in some consonant clusters. The following types of the assimilation are known (Pohribnyj 1984; Ponomariv 2001; Žovtobrjukh and Kulyk 1965; see also Wetzels and Mascaró 2001 for comparison with some other languages):

1) ***Regressive voicing and devoicing***

- A voiceless consonant followed by a voiced obstruent undergoes the voicing: *боротьба* /bɔrɔdʲbɑ/, *просьба* /prɔzʲbɑ/, *якби* /jɑgbɪ/, *вокзал* /νɔgzɑl/, *хоч би* /xɔʤbɪ/.
- A voiced /ɦ/ represented by < г > is devoiced when followed by a voiceless consonant: *нігті* /nʲixtʲi/, *легко* /lɛxkɔ/, *дьогтю* /dʲɔxtʲu/.
- The prefix and the preposition given by < з > is devoiced before voiceless consonants: *зсипати* /ssɪpɑtɪ/, *зсунути* /ssunutɪ/, *зцідити* /sʦʲidɪtɪ/, *з хати* /sxɑtɪ/. However, this effect is not universal and even denied by some authors (Ponomariv 2001: 18). Note, in particular, that such assimilation is reflected in orthography before < к, п, т, ф, х >: *скласти*, *спитати*. As a rule, < з > in the prefixes < роз- > and < без- > is not devoiced.
- It must be noted that voiced consonants are not devoiced when followed by voiceless: *ложка* /lɔʒkɑ/, *казка* /kɑzkɑ/, *кладка* /klɑdkɑ/. Additionally, there is no final devoicing in Ukrainian: *хліб* /xlʲib/, *сад* /sɑd/, *низ* /nɪz/.

2) ***Assimilation by place and manner of articulation***

- Dentals before (hushing) sibilants become (hushing) sibilants: *зшити* /ʃʃɪtɪ/.
- Hushing sibilants before dentals become dentals: *дощі* /dɔsʲʦʲi/.
- The stop represented by < т > before < ч, ш > becomes /ʧ/: *коротший* /kɔrɔʧʧɪj/.
- The stop represented by < т > before < ц > becomes /ʦ/: *коритце* /kɔrɪʦʦɛ/.

3) ***Regressive palatalization***

- Dentals followed by a soft consonant are themselves palatalized: *кінський* /kʲinʲsʲkɪj/, *пісня* /pʲisʲnʲɑ/, *дні* /dʲnʲi/.
- Dentals represented by < с, з, ц, дз > followed by a semisoft labial are palatalized: *свято* /sʲʋʲɑto/, *сміх* /sʲmʲix/, *цвіт* /ʦʲʋʲit/, *звір* /zʲʋʲir/.

3. Phoneme-grapheme relation

In this section we present and analyze graphemic representations of Ukrainian phonemes.

Table 1
Vowels

Phoneme	Graphemes	Comments and examples
/ɑ/	< а > < я >	*сам* *яр, м'яз, зняв*
/ɛ/	< е > < є > < и >*	*тер* *твоє, мене,* *мине*
/i/	< і > < ї >	*ліс* *з'їм*
/ɪ/	< и >	*сир*
/ɔ/	< о >	*гора*
/u/	< у > < ю >	*вулик* *знаю, ллю*

* In unstressed positions only, see Sec. 2.2.

In the table below, the following abbreviations are used for certain sets of graphemes:

- < і, я, ю, є > = < IOT > ('iotated', softening a preceding consonant);
- < з, с, дз, ц, н, л, д, т > = < DEN > (dentals);
- < б, п, в, м, ф > = < LAB > (labials);
- < б, г, ґ, д, ж, з, дж, дз > = < VOB > (voiced obstruents, to distinguish from sonorants).

Table 2
Consonants

Phoneme	Graphemes	Comments and examples
/b/	< б > < п >	*брат* before < VOB >: *крепдешин*

Phoneme	Graphemes	Comments and examples
/ʋ/	< в >	*вага, вона*
	< ф >*	before < VOB >: the root *афган…*
/ɦ/	< г >	*гора, луг*
	< хг >	in loan-words: *бухгалтер, цейхгауз*
	< х >**	before < VOB >: *їх друг*
/ɡ/	< ґ >	*ґрунт*
	< к >	before < VOB >: *якби, вокзал*
/d/	< д >	*дар, рід*
	< т >	before < VOB >: *п’ятдесят*
/d^j/	< д >	followed by < IOT >: *дяк*
		followed by soft < DEN >: *дня*
	< дь >	*відповідь*
	< т >	before soft < VOB >: *кіт дівся*
	< ть >	before < VOB >: *боротьба*
/ʒ/	< ж >	*жир*
	< з >	followed by < ж, ш, ч, дж >: *зжати*
	< ш >	before < VOB >: *наш друг*
/z/	< з >	*за, віз*
	< с >	before < VOB >: *юрисдикція*
	< ст >	before < VOB >: *шістдесят*
/z^j/	< з >	followed by < IOT >: *зілля*
		followed by soft < DEN >: *лазня*
		followed by semi-soft < LAB >: *звір*
	< зь >	*лізь*
	< ж >	followed by soft < с, ц >: *мажся*
	< с >	before soft < VOB >: *мус дійти*
	< сь >	before < VOB >: *просьба*
/j/	< й >	*його, мільйон, гай*
	< ї >	In modern Ukrainian, always = /ji/: *їжак, з’їв, країна*
	< я >	If preceded by the apostrophe < ’ >, < ь >,
	< ю >	a vowel or in a word-initial position: *я, моя, мільярд,*
	< є >	*п’ю, б’є, знаю*
/k/	< к >	*кава*
/l/	< л >	*ласка*
/l^j/	< л >	followed by < IOT >: *люба*
		followed by soft < DEN >: *ллє*
	< ль >	*сіль*
/m/	< м >	*мама*
/n/	< н >	*наш*
	< нт >	followed by < ст >: *студентство*
/n^j/	< н >	followed by < IOT >: *няв*
		followed by soft < DEN >: *кінський*
	< нь >	*кінь*
	< нт >	followed by the suffix < ськ >: *студентський*

Phoneme	Graphemes	Comments and examples
/p/	< п >	*пара*
/r/	< р >	*рот*
/r^j/	< р >	followed by < IOT >: *ряд*
	< рь >	only before < о >: *трьох*
/s/	< с >	*сон*
	< з >	followed by a voiceless consonant in some cases: *зсип*
	< ст >	followed by < с, н >: *шістнадцять*
/s^j/	< с >	followed by < IOT >: *сім*
		followed by soft < DEN >: *слід*
		followed by semi-soft < LAB >: *сміх, світ*
	< сь >	*колись*
	< ш >	followed by soft < с, ц >: *смієшся*
	< ст >	followed by < ськ > or soft < ц >: *роялістський, кістці*
/t/	< т >	*тихо, кіт*
/t^j/	< т >	followed by < IOT >: *тіло, тягти*
		followed by soft < DEN >: *новітній*
	< ть >	*ходить*
/f/	< ф >	*фонтан*
/x/	< х >	*хата*
	< г >	followed by < к, т > in some words: *нігті, легко*
/ʦ/	< ц >	*цей, цнота*
	< т >	followed by < ц >: *коритце*
	< тс >	*тсуга, спортсмен, братство*
/ʦ^j/	< ц >	followed by < IOT >: *цілувати*
		followed by soft < DEN >: *міцні*
		followed by semi-soft < LAB >: *цвіт*
	< ць >	*цього, кінець*
	< ч >	followed by soft < с, ц >: *сорочці*
	< т >	followed by soft < ц >: *винуватця*
	< ть >	the verbal cluster <ться> corresponds to /ʦ^{j}ʦ^{j}ɑ/: *сміється*
	< с >	in the verbal cluster < ться >
/ʧ/	< ч >	*чай*
	< щ >	= /ʃʧ/: *ще, дощ*
	< т >	followed by < ш, ч >: *коротший, тітчин*
/ʃ/	< ш >	*шум, ваш*
	< щ >	= /ʃʧ/: *щока*
	< с >	followed by < ш >: *вирісши*
	< з >	followed by < ш, ч > in a word-initial position: *зшити*
	< ст >	followed by < ч >: *невістчин*
	< ч >	followed by < н >, in some words only: *ячний*
/ʣ/	< дз >	*дзвонити*
	< ц >	followed by < VOB >: *плацдарм*
	< д >	followed by < с, ц, з >: *звідси*

Phoneme	Graphemes	Comments and examples
/ʣʲ/	< дз >	followed by < IOT >: *дзінь*
		followed by soft < DEN > or semi-soft < LAB >: *дзвякнути*
	< дзь >	*ґедзь*
	< ц >	before soft < VOB >: *буц діда*
	< ць >	before soft < VOB >: *лиць багато*
	< д >	followed by soft < с, ц >: *одинадцять*
	< дь >	followed by soft < DEN >: *підводься*
/ʤ/	< дж >	*бджола*
	< ч >	followed by < VOB >: *хоч би*
	< д >	followed by < ж, ш, ч >: *швидше*

* Grapheme < ф > before < VOB > appears as [v] being the voiced counterpart of [f]. This sound is not typical for Ukrainian. Thus, it can be treated as a combinatorial allophone of /f/.

** Grapheme < х > before < VOB > appears as [ɣ] being the voiced counterpart of [x]. Such situation occurs in native Ukrainian on the word boundaries: *тих днів*. In other situations the use of [ɣ] means incorrect pronunciation and thus [ɣ] can be treated as a voiced allophone of /x/. Cf. also similar voicing in Polish (Swan 2002: 16).

These two ambiguous possibilities will not be considered as separate graphemic representations in the following analysis of the grapheme-phoneme relation.

Bernhard and Altmann (2008) proposed the Shenton–Skees-geometric distribution

$$P_x = p(1-p)^{x-1}\left[1+a\left(x-\frac{1}{p}\right)\right], \quad x = 1,2,\dots,$$

with parameters $0 < p \le 1$ and $0 \le a \le \frac{1}{1-p} - 1$ (cf. Mačutek 2008a) as a model. In the table below one finds the distribution of graphemic representations, where x is the number of possibilities how a phoneme can be represented in writing, $f(x)$ is the number of phonemes with x graphemic representations (i.e., 10 phonemes are represented by 1 grapheme, 12 phonemes by 2 graphemes, etc) and $NP(x)$ are expected frequencies. Our results provide another corroboration of their hypothesis.

Table 3

Fitting the Shenton–Skees-geometric distribution to graphemic representations

phonemes	***x***	***f*(*x*)**	**NP(*x*)**
/ɪ, ɔ, ν, k, l, m, p, r, t, f/	1	10	10.29
/ɑ, i, u, b, ɦ, g, d, lʲ, n, rʲ, tʲ, x/	2	12	10.99
/ɛ, ʒ, z, nʲ, s, ʦ, ʧ, ʣ, ʤ/	3	9	7.50
/ dʲ, sʲ/	4	2	4.40
/zʲ, j/	5	2	2.39
/ʦʲ, ʃ, ʣʲ/	6	3	2.44
	a = 0.7105 χ^2 = 1.90 p = 0.5737 P = 0.59 DF = 5		

It is to be noted that the Shenton–Skees-geometric distribution (see above) yields a satisfactory fit ($P = 0.52$) also in the case that the two consonant allophones discussed under Table 2 are taken into consideration.

In the following we present a study of orthographic uncertainty in Ukrainian. We only note that some other properties (economy of script system, graphemic size, graphemic load of letters, letter utility) were investigated by Bernhard and Altmann (2008), Best and Altmann (2005), Kelih (2008) and Nemcová and Altmann (2008). As the number of analyzed languages is too small to allow constructions of models, we do not examine the properties in this paper. When more languages are investigated, Ukrainian data relevant for this direction of research can be easily mined from Tables 1, 2 and 3.

The mean orthographic uncertainty $\bar{U}$ of Ukrainian phonemes defined as follows:

$$\bar{U} = \frac{1}{N}\sum_{n} f_n \log_2 n ,$$

where f_n is the number of phonemes represented by n graphemes (cf. Bernhard and Altmann 2008), yields the value $\bar{U} = 1.1227$. Mean orthographic uncertainties $\overline{U_1}, \overline{U_2}$ in two languages are significantly different if

$$z = \frac{\overline{U_1} - \overline{U_2}}{\sqrt{V(\bar{U}_1) - V(\bar{U}_2)}} > 1.96,$$

$V(\bar{U}_1), V(\bar{U}_2)$ being estimation of the uncertainties variances (it holds $V(\bar{U}) = \dfrac{s^2}{0.48N\bar{x}^2}$, where $\bar{x}^2$ and s^2 are the sample mean and variance of the distribution of graphemic representations, cf. Table 3). The test was derived by Bernhard and Altmann (2008). For Ukrainian we obtain $\bar{x} = 2.5526$, $s^2 = 2.1420$ and $V(\bar{U}) = 0.018022$.

Table 4 below contains the comparison of mean uncertainties for the orthographies of six languages (the z-values are the values of the test statistics for Ukrainian compared with the language in the respective column, significant differences are highlighted in bold). The data are taken from Bernhard and Altmann (2008) for Italian, Best and Altmann (2005) for German and Swedish, Kelih (2008) for Slovene and Nemcová and Altmann (2008) for Slovak. Note that all these orthographies are based on the Latin script. Interestingly, the orthographic uncertainty of Ukrainian is significantly higher than the ones of the other two Slavic languages (Slovak and Slovene). The relatively high value for Ukrainian can be justified either by many assimilation possibilities or by a different writing system, namely Cyrillic; of course other factors cannot be excluded at this stage of research. Comparisons with other Cyrillic-based orthographies can help find an answer.

Table 4
Mean uncertainty in various writing systems

Language	German	Italian	Slovak	Slovene	Swedish	Ukrainian
$\bar{U}$	0.965	0.5641	0.7586	0.7841	0.797	1.1227
z-value	1.00	**3.59**	**2.10**	**2.09**	1.75	-

4. Ukrainian version of Cyrillic: complexity and distinctivity

When talking about the script complexity, distinctivity, etc., it is to be noted that the properties of a writing system depend also on a chosen font. We apply the composition method proposed by Altmann (2004), later slightly improved by Mačutek (2008b). In this method, a point is given a measure 1, a straight line corresponds to 2, an arch not exceeding 180° corresponds to 3; a continuous connection gets the weight 1, a crisp one 2 and a crossing evaluates to 3. Evaluation can be seen in Table 5 below.

Table 5
Complexity of Cyrillic letters (font Arial)

letter	Transliteration	components	connections	complexity
А	a	3×2	3×2	12
Б	b	2×2+3	3×2	13
В	v	2+2×3	4×2	16
Г	h	2×2	2	6
Ґ	g	3×2	2×2	10
Д	d	5×2+3	6×2	25
Е	e	4×2	3×2	14
Є	je	2+2×3	1+2	11
Ж	ž	3×2+3×3	2×1+2×2+3	26
З	z	4×3	2×1+2	16
И	y	3×2	2×2	10
І	i	2	—	2
Ї	ji	2+2×1	—	4
Й	j	3×2+3	2×2	13
К	k	2×3+2×2	2×2+1	18
Л	l	2×2+3	2×2	11
М	m	4×2	3×2	14
Н	n	3×2	2×2	10
О	o	2×3	2×1	8
П	p	3×2	2×2	10
Р	r	2+3	2×2	9
С	s	2×3	1	7
Т	t	2×2	2	6
У	u	2+3	2	7
Ф	f	2×3+2	2×1+2×3	16
Х	kh	2×2	3	7
Ц	c	4×2	3×2	14
Ч	č	2+3	2	7
Ш	š	4×2	3×2	14
Щ	šč	5×2	4×2	18
Ь	*soft sign**	2+3	2×2	9

Ю	ju	2×2+2×3	2×2+2×1	16
Я	ja	2×2+3	3×2	13

* A non-phonemic character, often transliterated as < ’ > or < j >, cf., e.g., Buk and Rovenchak (2004).

Mohanty (2007) supposed that the distribution of complexities was uniform. The hypothesis was successfully tested by him for the Oriya script and by Mačutek (2008b) for the Latin and Runic scripts.

Table 6
Distribution of complexities

C	f_C	**C**	f_C	**C**	f_C	**C**	f_C	**C**	f_C	**C**	f_C	**C**	f_C	**C**	f_C	**C**	f_C
2	1	**5**	0	**8**	1	**11**	2	**14**	4	**17**	0	**20**	0	**23**	0	**26**	1
3	0	**6**	2	**9**	2	**12**	1	**15**	1	**18**	1	**21**	0	**24**	0		
4	1	**7**	4	**10**	4	**13**	3	**16**	4	**19**	0	**22**	0	**25**	1		

We perform the run test about the mean to test the uniformity of the distribution. Denote *I* the inventory size, *R* the range of complexities, $\overline{C}$ the mean complexity and σ_C the standard deviation of complexities (we only note that for Cyrillic we have $\overline{C}$ = 11.79 and σ_C = 5.24). If the data are uniformly distributed, all expected frequency values are $E = \frac{I}{R+1}$. A run is a sequence of frequencies which are either all greater than *E* or all smaller than *E*. Hence we have $E = \frac{33}{24+1} = 1.32$ and the runs [$\underline{1,0,1,0}$, $\underline{2,4}$, $\underline{1}$, $\underline{2,4,2}$, $\underline{1}$, $\underline{3,4}$, $\underline{1}$, $\underline{4}$, $\underline{0,1,0,0,0,0,0,0,1,1}$], i.e., 9 runs. Next, denote $n = R + 1$, n_1 the number of frequencies smaller than *E* and n_2 the number of frequencies greater than *E* (in this case $n = 25$, $n_1 = 17$, $n_2 = 8$). The number of runs can be considered random (and, consequently, the distribution can be considered uniform) if

$$z = \frac{|r - E(r)| - 0.5}{\sigma_r} < 1.96,$$

where *r* is the number of runs, $E(r) = 1 + \frac{2n_1n_2}{n}$ and $\sigma_r = \sqrt{\frac{2n_1n_2(2n_1n_2 - n)}{n^2(n-1)}}$. We obtain $z = 1.13$, which means that the uniform distribution is a good model for the distribution of complexities also in this case.

Mačutek (2008b) suggested the Poisson distribution ($P_x = \frac{e^{-\lambda}\lambda^x}{x!}, \lambda > 0$) as a model for both the number of components and the number of connections. As can be seen in the following Table 7, Cyrillic is no exception, with an excellent fit for connections. For the number of components there are not enough degrees of freedom and the usual χ^2 goodness of fit test cannot be used (but at least intuitively the shape of the histogram is very similar to Poisson frequencies).

Table 7
Fitting the Poisson distribution to the numbers
of components and connection

	components	**connections**
0		2
1	1	6
2	9	10
3	12	8
4	8	5
5	1	1
6	2	1
λ = 2.49, χ^2 = 1.52, *P* = 0.91, *DF = 5*		

A method for measuring distinctivity of letters was introduced and described in details by Antić and Altmann (2005). In short, letters are decomposed into components (i.e., points, straight lines and arches), with orientations and connection points having differentiating functions. Differences between components are assigned weights, a difference between two letters is the minimum of sums of all components differences over all possible components permutations. Some minor refinements were added by Mačutek (2008b). Differences between letters of the Cyrillic alphabet are given below.

Table 8
Differences between Cyrillic letters

	А	Б	В	Г	Ґ	Д	Е	Є	Ж	З	И	І	Ї	Й	К	Л	М
А	0	26	39	15	14	33	24	24	50	38	17	16	18	20	34	18	18
Б	26	0	13	11	17	27	15	22	45	29	22	17	19	21	25	13	28
В	39	13	0	24	30	31	28	29	45	29	30	22	24	28	20	21	36
Г	15	11	24	0	6	25	12	14	39	28	11	6	8	14	23	7	17
Ґ	14	17	30	6	0	19	16	20	45	34	9	12	14	12	29	9	19
Д	33	27	31	25	19	0	23	35	59	48	28	31	33	26	38	18	31
Е	24	15	28	12	16	23	0	26	51	40	21	18	20	24	31	19	24
Є	24	22	29	14	20	35	26	0	34	22	25	14	16	23	19	17	31
Ж	50	45	45	39	45	59	51	34	0	29	45	33	35	42	27	41	51
З	38	29	29	28	34	48	40	22	29	0	34	22	24	32	23	30	40
И	17	22	30	11	9	28	21	25	45	34	0	12	14	3	29	17	10
І	16	17	22	6	12	31	18	14	33	22	12	0	2	15	17	13	18
Ї	18	19	24	14	8	33	20	16	35	24	14	2	0	17	19	15	20
Й	20	21	28	14	12	26	24	23	42	32	3	15	17	0	27	15	13
К	34	25	20	23	29	38	31	19	27	23	29	17	19	27	0	24	35
Л	18	13	21	7	9	18	19	17	41	30	17	13	15	15	24	0	23
М	18	28	36	17	19	31	24	31	51	51	10	18	20	13	35	23	0
Н	14	17	30	10	8	27	16	20	45	34	17	12	14	20	25	13	23
О	28	25	25	18	24	24	30	14	30	20	24	12	14	21	19	20	30

П	14	17	30	6	4	19	16	20	45	34	13	12	14	16	29	9	15
Р	28	6	11	13	19	33	21	22	39	23	19	11	13	18	19	15	25
С	26	24	25	16	22	37	28	6	30	16	22	10	12	20	15	19	28
Т	19	15	24	4	10	29	16	18	39	28	11	6	8	14	23	11	17
У	23	20	24	14	20	33	26	16	36	24	18	8	10	16	19	16	21
Ф	39	35	35	28	34	48	40	26	28	34	34	22	24	31	29	30	40
Х	20	23	29	12	18	37	24	21	42	30	17	9	12	20	26	19	22
Ц	20	23	36	12	6	21	20	26	51	40	15	18	20	18	35	15	22
Ч	24	17	21	13	19	34	21	15	35	23	19	7	9	17	16	16	25
Ш	20	23	36	16	10	25	20	26	51	40	19	18	20	22	35	19	26
Щ	26	29	29	18	12	23	26	32	57	46	21	24	26	24	41	21	28
Ь	28	6	11	17	19	33	21	22	39	23	19	11	13	18	19	19	29
Ю	25	19	29	29	20	34	26	20	43	29	28	20	22	27	23	16	34
Я	24	14	22	14	20	33	18	22	38	31	15	15	17	15	22	19	21

	Н	О	П	Р	С	Т	У	Ф	Х	Ц	Ч	Ш	Щ	Ь	Ю	Я
А	14	28	14	28	26	19	23	39	20	20	24	20	26	28	25	24
Б	17	25	17	6	24	15	20	35	23	23	17	23	29	6	19	14
В	30	25	30	11	25	24	24	35	29	36	21	36	42	11	29	22
Г	10	18	6	13	16	4	14	28	12	12	13	16	18	17	18	14
Ґ	8	24	4	19	22	10	20	34	18	6	19	10	12	19	20	20
Д	27	38	19	33	37	29	33	48	37	21	34	25	23	33	34	33
Е	16	30	16	21	28	16	26	40	24	20	21	20	26	21	26	18
Є	20	14	20	22	6	18	16	26	21	26	15	26	32	22	20	22
Ж	45	30	45	39	30	39	36	28	42	51	35	51	57	39	43	38
З	34	20	34	23	16	28	24	34	30	40	23	40	46	23	29	31
И	17	24	13	19	22	11	18	34	17	15	19	19	21	19	28	15
І	12	12	12	11	10	6	8	22	9	18	7	18	24	11	20	15
Ї	14	14	14	13	12	8	10	24	12	20	9	20	26	13	22	17
Й	20	21	16	18	20	14	16	31	20	18	17	22	24	18	27	15
К	25	19	29	19	15	23	19	29	26	35	16	35	41	19	23	22
Л	13	20	9	15	19	11	16	30	19	15	16	19	21	19	16	19
М	23	30	15	25	28	17	21	40	22	22	25	26	28	29	34	21
Н	0	24	8	19	22	14	20	34	18	14	15	14	20	19	16	20
О	24	0	24	19	8	18	14	14	20	30	14	30	36	19	24	22
П	8	24	0	19	22	10	20	34	18	10	19	14	16	23	20	20
Р	19	19	19	0	18	13	14	29	20	25	11	27	31	4	21	15
С	22	8	22	18	0	16	13	22	18	28	11	28	34	18	22	19
Т	14	18	10	13	16	0	14	28	12	12	13	16	18	17	22	14
У	20	14	20	14	13	14	0	25	15	26	11	26	32	14	23	21
Ф	34	14	34	29	22	28	25	0	31	40	24	40	46	29	34	32
Х	18	20	18	20	18	12	15	31	0	24	16	24	30	20	26	20
Ц	14	30	10	25	28	12	26	40	24	0	25	4	6	25	26	26
Ч	15	14	19	11	11	13	11	24	16	25	0	25	31	11	18	14
Ш	14	30	14	27	28	16	26	40	24	4	25	0	6	25	26	26

Щ	20	36	16	31	34	18	32	46	30	6	31	6	0	31	32	32
Ь	19	19	23	4	18	17	14	29	20	25	11	25	31	0	21	15
Ю	16	24	20	21	22	22	23	34	26	26	18	26	32	21	0	23
Я	20	22	20	15	19	14	21	32	20	26	14	26	32	15	23	0

The mean distinctivity of a letter is the sum of its differences with respect to all other letters in the alphabet divided by $I-1$.

Table 9
Mean distinctivities of Cyrillic letters

А	24.84	Д	31.75	И	20.34	Л	18.25	Р	19.53	Х	21.94	Ь	20.09
Б	20.97	Е	24.03	І	15.25	М	26.03	С	20.47	Ц	22.72	Ю	24.62
В	27.53	Є	21.81	Ї	17.22	Н	19.69	Т	16.78	Ч	18.53	Я	27.16
Г	15.53	Ж	41.22	Й	20.66	О	22.16	У	19.88	Ш	23.91		
Ґ	18.06	З	30.53	К	25.53	П	18.69	Ф	31.88	Щ	28.28		

The mean distinctivity is the mean of all mean distinctivities, for Cyrillic we obtain $\overline{D} = 22.55$, cf. Antić and Altmann (2005) and Mačutek (2008b) for distinctivity analysis of Latin and Runic scripts.

Several other hypotheses, their partial corroboration or criticism, some tests and tentative models were presented in Mohanty (2007), Altmann (2008) and Mačutek (2008b). However, we postpone the analysis until more data are available.

Acknowledgement

J. Mačutek was supported by the research grant VEGA 1/3016/06.

References

Altmann, G. (2004). Script complexity. *Glottometrics 8, 68-74.*

Altmann, G. (2008). Towards a theory of script. In: Altmann, G., Fan, F. (eds.), *Analysis of Script. Properties of Characters and Writing Systems: 145-160.* Berlin: de Gruyter (in press).

Antić, G., Altmann, G. (2005). On letter distinctivity. *Glottometrics 9, 46-53.*

Barry, R.K. (ed.) (1997). *ALA-LC Romanization Tables: Transliteration Schemes for Non-Roman Scripts.* Washington: Library of Congress.

Bernhard, G., Altmann, G. (2008). The phoneme–grapheme relationship in Italian. In: Altmann, G., Fan, F. (eds.), *Analysis of Script. Properties of Characters and Writing Systems: 11-21.* Berlin: de Gruyter.

Best, K.-H., Altmann, G. (2005). Some properties of graphemic systems. *Glottometrics 9, 29-39.*

Bethin, Ch.Y. (1992). Iotation and gemination in Ukrainian. *The Slavic and East European Journal 36, 275-301.*

Bilodid, I. K. (ed.) (1969). *Sučasna ukrajinsjka literaturna mova. Vstup. Fonetyka [Modern Ukrainian literary language. Introduction. Phonetics]*. Kyiv: Naukova dumka.

Bilous, T. (2005). *IPA for Ukrainian.* Available from: http://www.vesna.org.ua/txt/ bilous/UkrIPA.pdf. Accessed 21 February 2008.

Buk, S., Rovenchak, A. (2004). Rank–frequency analysis for functional style corpora of Ukrainian. *Journal of Quantitative Linguistics 11, 161-171.*

Comrie, B. (1996a). Adaptations of the Cyrillic Alphabet. In: Daniels, P.T., Bright, W. (eds.), *The World's Writing Systems: 700-726*. Oxford: Oxford University Press.

Comrie, B. (1996b). Script Reform in and after the Soviet Union. In: Daniels, P.T., Bright, W. (eds.), *The World's Writing Systems: 781-784*. Oxford: Oxford University Press.

Coulmas, F. (2004). *The Blackwell Encyclopedia of Writing Systems.* Oxford: Blackwell.

Jensen, H. (1969). *Die Schrift in Vergangenheit und Gegenwart.* 3rd ed. Berlin: VEB Deutscher Verlag der Wissenschaften.

Katkouski, U., Rrapo J. (n.d.) *Introduction to Belarusian Alphabet.* http://www. pravapis.org/art_belarusian_alphabet.asp. Accessed 19 February 2008.

Kelih, E. (2008). The phoneme–grapheme relationship in Slovene. In: Altmann, G., Fan, F. (eds.), *Analysis of Script. Properties of Characters and Writing Systems: 59-71.* Berlin: de Gruyter.

Mačutek, J. (2008a). On the distribution of graphemic representations. In: Altmann, G., Fan, F. (eds.), *Analysis of Script. Properties of Characters and Writing Systems: 73-76.* Berlin: de Gruyter.

Mačutek, J. (2008b). Runes: complexity and distinctivity. *Glottometrics 16,1-16.*

Mohanty, P. (2007). On the script complexity and the Oriya script. In: Grzybek, P., Köhler, R. (eds.), *Exact Methods in the Study of Language and Text: 473-484.* Berlin: de Gruyter.

Nemcová, E., Altmann, G. (2008). The phoneme–grapheme relation in Slovak. In: Altmann, G., Fan, F. (eds.), *Analysis of Script. Properties of Characters and Writing Systems: 77-85.* Berlin: de Gruyter.

Pivtorak, H. P. (2004a). G. In: Rusanivsjkyj, V. M., Taranenko, O. O., et al. (eds.), *Ukrajinsjka mova: Encyklopedija [Ukrainian Language: Encyclopedia]. 2nd ed.: 127.* Kyiv: Ukrajinska encyklopedija.

Pivtorak, H. P. (2004b). Ji. In: Rusanivsjkyj, V. M., Taranenko, O. O., et al. (eds.), *Ukrajinsjka mova: Encyklopedija [Ukrainian Language: Encyclopedia]. 2nd ed.: 241.* Kyiv: Ukrajinska encyklopedija.

Pohribnyj, M. I. (ed.) (1984). *Orfoepičnyj slovnyk [Orthoepic dictionary].* Kyiv: Radjansjka škola.

Ponomariv, O. D. (ed.) (2001). *Sučasna ukrajinsjka mova [Modern Ukrainian language].* Kyjiv: Lybidj.

Rusanivsjkyj, V. M. (2004). Ukrajinsjka mova. In: Rusanivsjkyj, V. M., Taranenko, O. O., et al. (eds.), *Ukrajinsjka mova: Encyklopedija [Ukrainian Language: Encyclopedia]. 2nd ed.: 716-718.* Kyiv: Ukrajinska encyklopedija.

Swan, O. E. (2002). *A grammar of contemporary Polish*. Bloomington, IN: Slavica Publishers.

Šerech, Ju. (1951). *Narys sučasnoji ukrajinsjkoji literaturnoji movy [An outline of modern literary Ukrainian]*. Munich: Molode zyttia publishing Co.

Wetzels, W. L.; Mascaró J. (2001). The Typology of Voicing and Devoicing. *Language 77, 207-244.*

Zilynśkyj, I. (1979). *A phonetic description of the Ukrainian language.* Cambridge: Harvard University Press.

Žovtobrjukh, M. A., Kulyk, B. M. (1965). *Kurs sučasnoji ukrajinsjkoji literaturnoji movy [Course of modern literary Ukrainian]*. Kyiv: Radjansjka škola.

Žovtobrjukh, M. A., Khomenko, L. M. (2004). Fonema. In: Rusanivsjkyj, V. M., Taranenko, O. O., et al. (eds.), *Ukrajinsjka mova: Encyklopedija [Ukrainian Language: Encyclopedia]. 2nd ed.: 760-761*. Kyiv: Ukrajinska encyklopedija.

Glottometrics 16, 2008, 80-110

Some problems of musical texts

Zuzana Martináková, Banská Bystrica[1]
Ján Mačutek, Bratislava
Ioan-Iovitz Popescu, Bucharest
Gabriel Altmann, Lüdenscheid

Abstract. The aim of this article is to find fixed points and regularities in musical texts, set up statistical tests for their comparison and observe their development. The analysis is based on rank-frequency distributions of pitches. The following indicators are described: the *h*-point and its angle, the *a*-indicator, the *H*-point and the *H*-coverage having an affinity to the golden section, and the *A*-ratio. Different curves capturing the trends are proposed. The analysis has been performed on 266 compositions of 12 European composers from Palestrina to Ligeti.

Key words: h-point, a-indicator, H-coverage, A-ratio, rank-frequency distribution, musical texts

1. Introduction

From the general point of view a musical composition is an organized sequence of the musical sounds[2], musical shapes (motives)[3], and musical sections, sentences, parts, movements, etc., just as linguistic texts are sequences of phonemes, syllables, words, phrases, clauses and sentences. However, both the matter of which they are made and the aim of their production, as well as the inventories of units, are different. Any comparison of their inventory sizes is, nevertheless, futile. But whatever the material or functional background of musical sequences, up to a certain level they display repetitions. Sentences in linguistic texts repeat seldom (except for very colloquial ones), and texts[4], linguistic or musical, never.

The units of musical or linguistic texts are not given a priori; they are constructed by us conceptually. In speech, there is only a stream of sounds with tones, stress and intonation, but without blanks, diacritics, or clear sentence ends. But even this stream can be seen differently by a physicist and a linguist. The physicist constructs waves; the linguist constructs linguistic units and segments the text in many ways. In music a staccato sequence differs musically from a legato sequence, but they are equal as sequence. The segmentation of music (metric segmentation in bars, or ametric segmentation) is not given; it results from a certain rhythm

[1] Address correspondence to: zuzanamartinakova@yahoo.com

[2] We have to distinguish: 1. a simple sound = tone, defined by a combination of pitch, duration, timbre (type of instrumental sound or human voice), intensity and articulation (such as legato, arco, pizzicato, staccato, etc.); 2. a complex sound = vertical set of several simple sounds (intervals, accords, clusters), defined by its composition (the set of participating simple sounds) and its configuration (the set of the order relations on this set, i.e. a 5-tuple consisting of the configurations of pitch, duration, intensity, timbre and articulation); 3. a simple sound group = horizontal sequence of simple sounds (tones), defined by melody, rhythm, meter, colour etc., 4. a complex sound group = sequence of (simple and) complex sounds.

[3] A musical shape = motif representing the smallest meaningful semiotic unit of a musical text; while a simple sound is the smallest structural unit of a musical text.

[4] A musical text is an actual configuration of musical elements, such as a composition, folk, popular, jazz etc.; a song; an improvisation etc. There are various representations of musical texts (scores, instructions, graphical representation, oral tradition etc.).

and meter a posteriori. Hence there is no “natural” unit in musical texts produced by humans. In spite of this, in musical sequences one can observe certain regularities which may but need not be conscious. Those which are conscious are used purposefully by the author; just as a text is partitioned into sentences and chapters, a musical text has sections, parts and movements, etc. But some regularities, local or global, are concealed and must be brought to light by formal methods. In general one says that a special segmentation is prolific if it allows us to discover regularities some of which may be laws. Laws cannot be learnt but they are abode by. If a special order decays – as can be seen in the contemporary music – other order replaces it. The task of science is to capture this order, its decay and the emergence of new order. Needless to say, the transition from one order to another is accompanied by deviations, outliers, extremes and a surface chaos which leads to new equilibria.

A sequence of musical events (sounds) has as many properties as we are able to construct conceptually. Some of them are “more objective”, e.g. pitch, duration, intensity, timbre, articulation, density (complex sounds); others are latent and can be interpreted emotionally, e.g. sad, uneasy, magnificent etc. Some of the properties can be measured quite easily; some necessitate personal judgements which are not always unique. Here we shall restrict ourselves to a surface property, namely the frequency of individual tones identified by their pitches. This can be performed either with pencil and paper or using a program which does it automatically. For this purpose we have used Reinhard Köhler’s computer program QUAMS (= Quantitative Analysis of Musical Structures) created in 1995/1996, providing distributions from MIDI data, which can also order all used tone pitches in the musical text according to type and frequency, i.e. the program is able to establish rank-frequency distributions of pitch values.[5]

The simplest problem is the computation of the rank-frequency distribution of tone pitches and finding the appropriate theoretical distribution. As has been shown (cf. Köhler, Martináková-Rendeková 1995, 1998; Martináková 1997, 1998; Wimmer, Wimmerová 1997, Martináková-Rendeková 2002, 2003, 2004, 2007) the negative hypergeometric distribution is an adequate model, in most cases also in linguistic texts (cf. Popescu et al. 2007). However, it is not known as yet how to interpret the individual parameters even if their motion is known (cf. Martináková 2007)

Here we shall study some other properties of the rank-frequency distribution.

2. The *h*-point and the *a*-indicator

The *h*-point of a rank-frequency distribution, $f = f(r)$, is a fixed point that can be computed in various ways (cf. Popescu 2007; Popescu et al. 2007; Popescu, Altmann 2007). These ways result from its definition proper, that is to find the point $(r, f(r))$ at which $r = f(r)$, i.e. the rank is equal to frequency. As illustrated in Table 1, in Beethoven´s Sonata No. 6 the *h*-point is located at $r = 46 = f(r)$.

[5] We are deeply grateful to Prof. Reinhard Köhler who kindly placed this program at our disposal.

Table 1
Rank-frequency distribution of tone pitches in Beethoven´s Sonata No. 6

r	f(r)	r	f(r)	r	f(r)	r	f(r)
1	404	16	185	31	85	46	46
2	316	17	181	32	83	47	42
3	303	18	167	33	79	48	36
4	302	19	156	34	78	49	31
5	281	20	150	35	77	50	31
6	278	21	146	36	72	51	29
7	275	22	138	37	70	52	15
8	265	23	129	38	69	53	13
9	247	24	127	39	64	54	11
10	227	25	122	40	59	55	6
11	214	26	113	41	57	56	6
12	208	27	110	42	54	57	5
13	200	28	94	43	53	58	3
14	192	29	89	44	53	59	3
15	187	30	87	45	48		

In some cases for all *r* there is no equal *f*(*r*) and one computes it either exactly (by fitting and interpolation) or one takes that *r* whose *absolute* difference to *f*(*r*) is minimum. For example in Beethoven´s Sonata No. 28 we have

rank r	frequency f(r)	r - f(r)
45	63	-18
46	59	-13
47	56	-9
48	52	-4
49	46	3
50	41	9
51	40	11
52	39	13

where the minimal absolute difference is 3 corresponding to *r* = *h* = 49.

It has been shown in linguistics that the *h*-point depends on the length of the texts according to the relationship $N = ah^2$, as originally proposed by Hirsch (2005) in scientometrics for the citations count. The indicator

$$a = \frac{N}{h^2} \tag{1}$$

has successfully been used in linguistic text analysis (cf. Popescu et al. 2007; Mačutek, Popescu, Altmann 2007) and brought relevant typological results. The same simple power trend can be seen now in musical texts, as illustrated in Figure 1. Thus, if we compute the *a*-indicators for Beethoven´s Sonatas, as shown in Table 2, we obtain a zero trend, as expected. The almost constant values of *a* can also be found in the last column of Table 2 and in Figure 2. The mean of all sonatas is $\bar{a}$ = 4.35. A comparison with Skrjabin having $\bar{a}$ = 2.84 shows that the differences are considerable and can have their causes. However, tests for differences must be performed (see below).

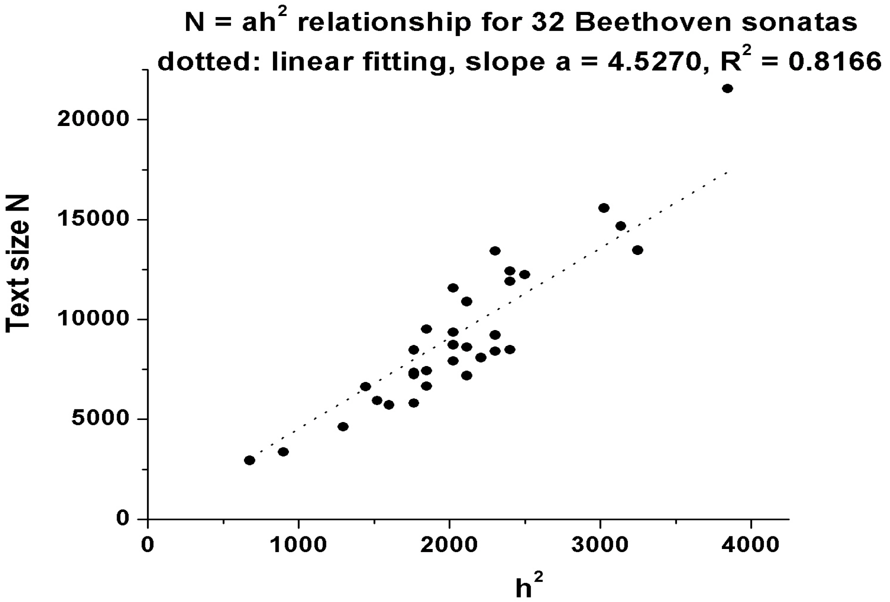


Figure 1. The dependence of *h* on *N* for 32 Beethoven sonatas. Roughly we have $N = ah^2$.

Table 2
The *a*-indicator for Beethoven´s Sonatas

ID	Text	N	h	$a = N/h^2$	ID	Text	N	h	$a = N/h^2$
LvB01	Sonata 1	7332	42	4,16	LvB17	Sonata 17	7905	45	3,9
LvB02	Sonata 2	9340	45	4,61	LvB18	Sonata 18	12428	49	5,18
LvB03	Sonata 3	11915	49	4,96	LvB19	Sonata 19	3362	30	3,74
LvB04	Sonata 4	12248	50	4,9	LvB20	Sonata 20	2937	26	4,34
LvB05	Sonata 5	7229	42	4,1	LvB21	Sonata 21	14682	56	4,68
LvB06	Sonata 6	7171	46	3,39	LvB22	Sonata 22	5802	42	3,29
LvB07	Sonata 7	9201	48	3,99	LvB23	Sonata 23	15575	55	5,15
LvB08	Sonata 8	8396	48	3,64	LvB24	Sonata 24	4619	36	3,56
LvB09	Sonata 9	5706	40	3,57	LvB25	Sonata 25	5930	39	3,9
LvB10	Sonata 10	6623	38	4,59	LvB26	Sonata 26	7416	43	4,01
LvB11	Sonata 11	10898	46	5,15	LvB27	Sonata 27	6643	43	3,59
LvB12	Sonata 12	9497	43	5,14	LvB28	Sonata 28	8467	49	3,53
LvB13	Sonata 13	8461	42	4,8	LvB29	Sonata 29	21559	62	5,61
LvB14	Sonata 14	8597	46	4,06	LvB30	Sonata 30	8713	45	4,3
LvB15	Sonata 15	11581	45	5,72	LvB31	Sonata 31	8075	47	3,66
LvB16	Sonata 16	13439	48	5,83	LvB32	Sonata 32	13468	57	4,15

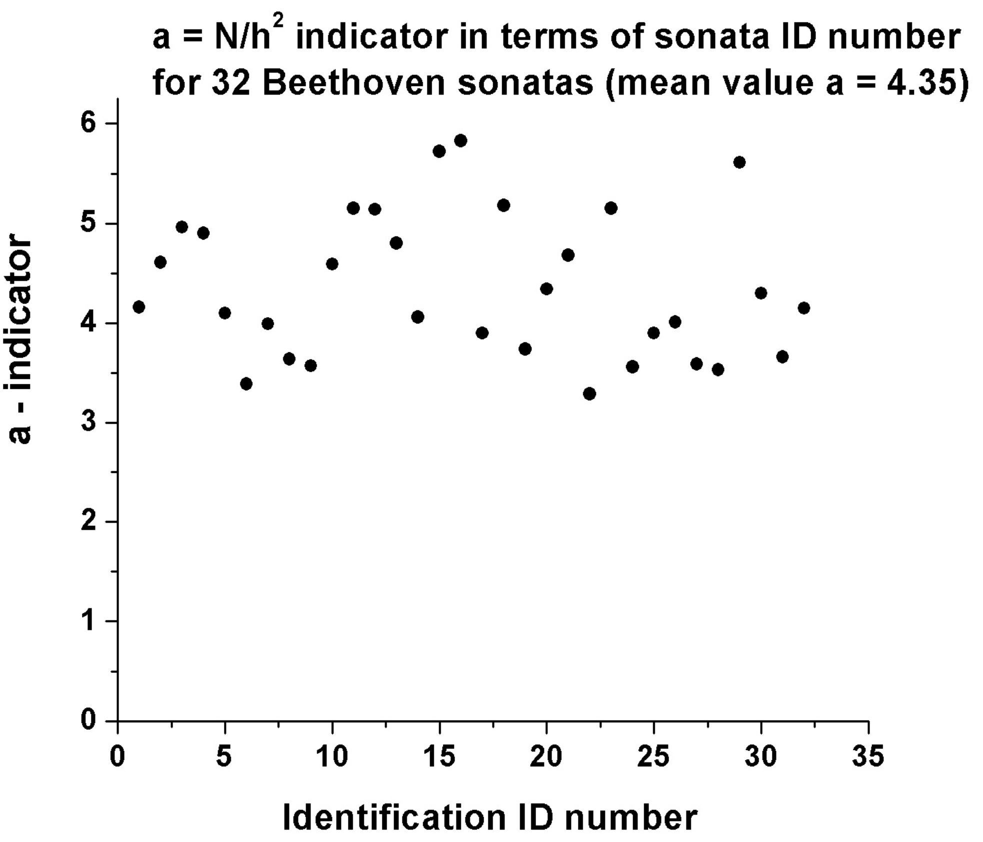


Figure 2. The *a*-values of Beethoven Sonatas

The following hypotheses can be set up in connection with the *a*-indicator: (1) The (mean) indicator *a* is significantly different with different composers either in its mean value or its dispersion. (2) It is significantly different for genres. (3) It may display a certain development tendency in the history of music and it is different for historical musical styles. (4) It is different for compositional language created in different national cultures. Since tests for the *a*-indicators were made possible (cf. Mačutek et al. 2007; Popescu et al. 2008) all hypotheses could be tested. Here we shall restrict ourselves to the comparison of Beethoven and Skrjabin. The basic data of Skrjabin are shown in Table 3, the *a*-indicator is shown in Figure 3.

Table 3
The *a*-indicator for Skrjabin´s compositions

ID	Text	N	h	$a = N/h^2$
Skr01	Prelude op. 27 – No 1	355	12	2.47
Skr02	Prelude op. 27 – No 2	222	10	2.22
Skr03	Prelude op. 31 – 1	651	16	2.54
Skr04	Prelude op. 31 – 4	155	7	3.16
Skr05	Prelude op. 33 – 2	195	8	3.05
Skr06	Prelude op. 33 – 3	212	9	2.62
Skr07	Prelude op. 35 – 2	362	11	2.99
Skr08	Prelude op. 37 – No 1	212	9	2.62
Skr09	Prelude op. 37 – No 2	91	5	3.64
Skr14	Piece op. 2, No 1	1150	20	2.88
Skr15	Etude op. 8, No 4	747	17	2.58
Skr16	Etude op. 8, No 5	1541	21	3.49
Skr17	Etude op. 8, No 12	2301	27	3.16
Skr18	Poem op. 32 – No 1	981	16	3.83
Skr19	Poème tragique op.34	1001	16	3.91
Skr20	Etude op. 42, No 4	787	18	2.43
Skr21	Etude op. 42, No 5	3088	32	3.02
Skr22	Sonate No 5, op. 53	7761	50	3.10

Skr10	Prelude op. 48 – 2	224	9	2.77	Skr23	Sonate No 9, op. 68	4014	40	2.51
Skr11	Prelude op. 59	709	17	2.45	Skr24	Poem op. 69 – No 2	539	14	2.75
Skr12	Prelude op. 67 – 1	338	12	2.35	Skr25	Dance op. 73 – No 1: Guirlandes	694	16	2.71
Skr13	Prelude op. 74 – 3	228	11	1.88	Skr26	Dance op. 73 – No 2: Flammes sombres	1051	20	2.63

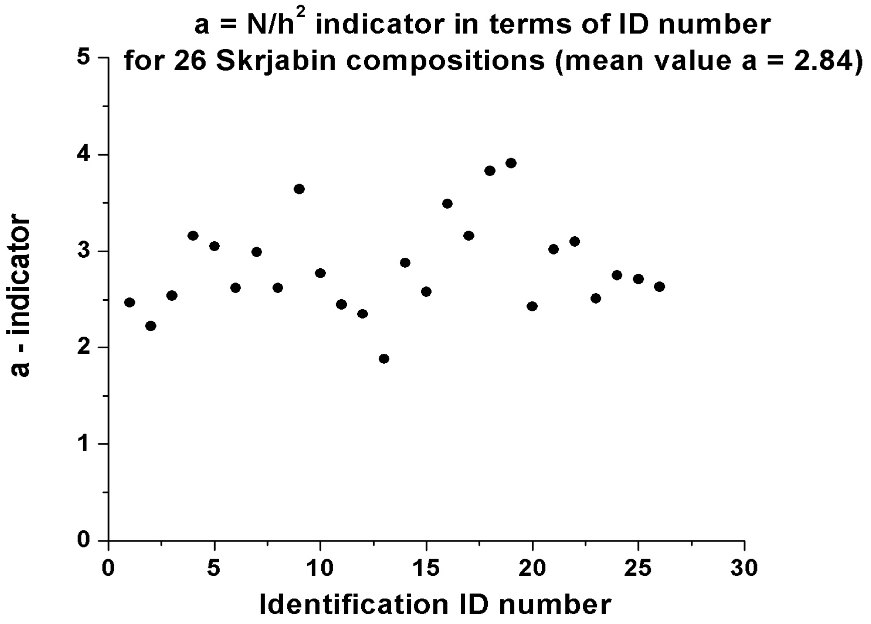


Figure 3. The *a*-indicator in compositions by Skrjabin

The optical difference to Beethoven is evident (Skrjabin´s *a*-indicators are placed deeper than those of Beethoven) but we perform a usual test for averages starting from empirical data. We set up the (simplified) criterion

$$(2) \qquad z = \frac{\bar{a}_1 - \bar{a}_2}{\sqrt{Var(\bar{a}_1) + Var(\bar{a}_2)}}$$

which is a standard normal variable (as a matter of fact, with small sample sizes it is a *t*-variable). The individual values can be computed from the above tables mechanically (e.g. by Excel). The variance of *a*-values of Beethoven is $Var(a_1)$ = 0.50 and $Var(\bar{a}_1)$ = 0.50/32 = 0.015625, that of Skrjabin is $Var(\bar{a}_2)$ = 0.00889. Inserting all these values in (2) we obtain

$$z = \frac{4.35 - 2.84}{\sqrt{0.015625 + 0.00889}} = 9.64\ ,$$

telling us that concerning the *a*-indicator and the given compositions, the two composers are very different. The test can be made finer if one estimates a common variance (in that case we would obtain *z* = 9.10).

Nevertheless, the individual composers themselves need not be as homogeneous as they seem when compared with other composers. However, the test between two individual *a*-indicators must be performed in a different way (cf. Mačutek, Popescu, Altmann 2007). The statistics (2) can be used again, but in this case a new problem arises, namely, we do not know the variances of the *a*-indicators. As their theoretical properties are not known, they were estimated from a simulation study. The simulations follow the idea described in Mačutek, Popescu and Altmann (2007), which we recall here in short (Beethoven's Sonata 1 will serve as an example).

Table 4
The *a*-indicator for Palestrina´s Masses

ID	Text	N	h	a = N/h^2
Pls01	Ascendo 1, Motetto	1856	19	5.14
Pls02	Ascendo 2, Kyrie	898	15	3.99
Pls03	Ascendo 3, Gloria	1348	17	4.66
Pls04	Ascendo 4, Credo	2120	19	5.87
Pls05	Ascendo 5, Sanctus	595	14	3.04
Pls06	Ascendo 5, Benedictus	563	14	2.87
Pls07	Ascendo 7, Agnus Dei I	431	13	2.55
Pls08	Ascendo 8, Agnus Dei II	487	14	2.48
Pls09	Ave Regina Chant	137	6	3.81
Pls10	Ave Regina Kyrie	687	15	3.05
Pls11	Ave Regina Gloria	1357	17	4.7
Pls12	Ave Regina Credo	2355	19	6.52
Pls13	Ave Regina Sanctus	436	13	2.58
Pls14	Ave Regina Benedictus	505	13	2.99
Pls15	Ave Regina Agnus Dei I	396	13	2.34
Pls16	Ave Regina Agnus Dei II	402	13	2.38
Pls17	Missa Papae Kyrie	995	16	3.89
Pls18	Missa Papae Gloria	1437	17	4.97
Pls19	Missa Papae Credo	2385	19	6.61
Pls20	Missa Papae Sanctus	1060	16	4.14
Pls21	Missa Papae Benedictus	644	13	3.81
Pls22	Missa Papae Agnus Dei I	711	15	3.16
Pls23	Missa Papae Agnus Dei II	793	14	4.05
Pls24	Missa Veni Kyrie	669	14	3.41
Pls25	Missa Veni Gloria	1013	15	4.5
Pls26	Missa Veni Credo	1596	19	4.42
Pls27	Missa Veni Sanctus	722	14	3.68
Pls28	Missa Veni Benedictus	576	14	2.94
Pls29	Missa Veni Agnus Dei I	343	13	2.03
Pls30	Missa Veni Agnus Dei II	415	14	2.12

We generated 7332 (there are 7332 tones in the sonata) random numbers from the negative hypergeometric distribution[6] with the parameters K = 3.4690, M = 0.8257, n = 59 (parameter values for which the best fit is obtained), and we found the *h*-point and *a*-indicator in this sample. The random number generation is repeated 100 times, resulting in 100 *a*-indicators from samples with the same size and distribution as tone pitches frequencies in Beethoven's Sonata 1. Next, we compute the variance of the 100 *a*-indicators. The procedure is repeated 10 times, i.e., we have 10 variance values, each of them being a variance of 100 *a*-indicators. Their mean is an estimation of the *a*-indicator variance.

We recommend larger number of random samples for a historical or comparative study; here we mainly aim at the method introduction.[7]

Nine compositions were chosen for testing differences between *a*-indicators. Recall that the difference is significant if the *z*-statistics value is less than –1.96 or more than 1.96. The results are shown in Table 5.

[6] The adequateness of the negative hypergeometric distribution for the rank-frequency distribution of tone pitches has been shown in Martináková-Rendeková (2005)

[7] Short simulation programs (cf. also Section 4) written in *R* can be sent upon request (jmacutek@yahoo.com).

Table 5
Tests for differences between *a*-indicators (Beethoven, Palestrina, Skrjabin)

	LvB01	LvB18	LvB31	Pls01	Pls19	Pls29	Skr01	Skr13	Skr19
LvB01	0	**-2.24**	1.18	-1.63	**-3.65**	**2.64**	**2.05**	1.27	0.28
LvB18	**2.24**	0	**3.53**	0.07	**-2.12**	**3.90**	**3.28**	1.84	1.40
LvB31	-1.18	**-3.53**	0	**-2.54**	**-4.50**	**2.06**	1.47	1.00	-0.28
Pls01	1.63	-0.07	**2.54**	0	-1.88	**3.46**	**2.92**	1.77	1.24
Pls19	**3.65**	**2.12**	**4.50**	1.88	0	**4.83**	**4.31**	**2.54**	**2.61**
Pls29	**-2.64**	**-3.90**	**-2.06**	**-3.46**	**-4.83**	0	-0.42	0.08	-1.67
Skr01	**-2.05**	**-3.28**	-1.47	**-2.92**	**-4.31**	0.42	0	0.31	-1.27
Skr13	-1.27	-1.84	-1.00	-1.77	**-2.54**	-0.08	-0.31	0	-1.04
Skr19	-0.28	-1.40	0.28	-1.24	**-2.61**	1.67	1.27	1.04	0

Consider now the dispersion of the *a*-values. Using the unbiased estimators of the variance, we obtain for Skrjabin *Var*(*a*) = 0.2403, for Beethoven *Var*(*a*) = 0.5197, but for Palestrina *Var*(*a*) = 1.5249 though his mean is *a* = 3.76, i.e. it is positioned between Skrjabin and Beethoven, as can be seen in Table 6. Automatically the hypothesis arises whether the dispersion of the *a*-indicators displays a historical development.

To this end we compare the work of some other composers as shown in Table 6.

Table 6
Mean and unbiased variance of *a* of all composers

Name	Mean year	Mean *a*	Variance of *a*
Palestrina (1525-1594)	1560	3.76	1.5249
Gesualdo (1560?-1613)	1587	2.73	0.1810
Monteverdi (1567-1643)	1605	4.60	1.1942
Bach (1685-1750)	1718	3.35	0.2013
Mozart (1756-1791)	1774	5.74	1.0534
Beethoven (1770-1827)	1799	4.35	0.5197
Liszt (1811-1886)	1849	3.01	0.2173
Skrjabin (1872-1915)	1894	2.84	0.2403
Schoenberg (1874-1951)	1913	2.97	0.9905
Stravinsky (1882-1971)	1927	3.56	1.5824
Shostakovich (1906-1975)	1940	2.97	0.7273
Ligeti (1923-2006)	1965	2.20	0.1583

Observing the values of *a* as shown in Figure 4, we can see that the existing trend is clearly divided in two parts: the first from Palestrina up to Mozart, the second from Mozart down to Ligeti. The first part cannot be captured by any simple curve but the second part displays a monotone linear decreasing trend (R^2 = 0.73) as can be seen in Table 7, yielding *a* = 29.4730 – 0.0138t, where *t* is the given mean year.

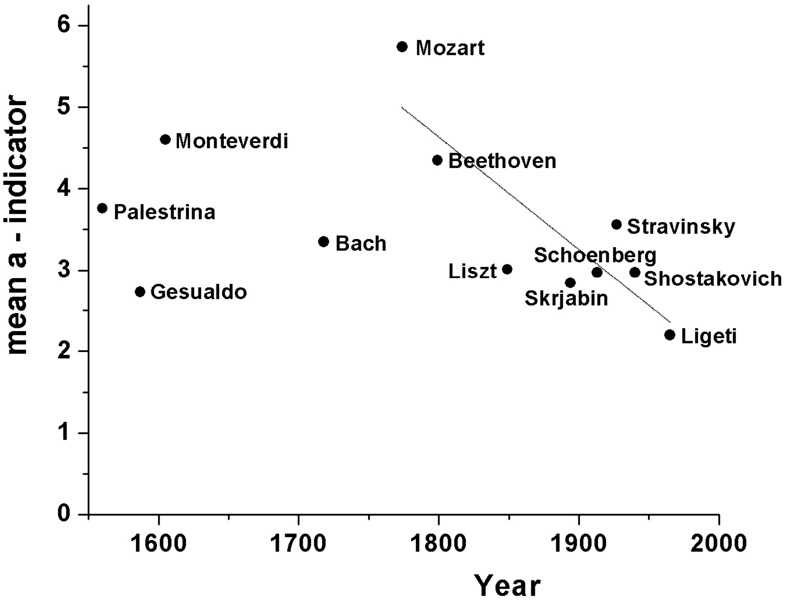


Figure 4. The trend of *a*-values

Table 7
The *a*-trend beginning with Mozart

Year	*a*-observed	*a*-computed
1774	5.74	4.99
1799	4.35	4.65
1849	3.01	3.96
1894	2.84	3.34
1913	2.97	3.07
1927	3.56	2.88
1940	2.97	2.70
1965	2.20	2.36

We conjecture that the complete trend is curvilinear and concave but the *a*-indicator should be computed from the complete work of each composer. This is unfortunately a very tiresome task that can be performed only partially in the future.

3. The view angles

In linguistic texts the *h*-point is considered a control position: the writer subconsciously looks at the top and the end of the distribution (the top is represented by f_1 – the greatest frequency, the end by the text vocabulary *V*) and controls their development. The angle of the *h*-point is metaphorically called "writer´s view". But the situation is quite different in music. The tone pitches are not parallels of words but rather of phonemes or letters. The composer cannot develop any other tones than those given by the instruments, but a speaker develops words continuously. Hence the LNRE (large number of rare events) theory does not hold for this aspect of music. Nevertheless, it can be shown that the rank-frequency of pitches abides by the negative hypergeometric distribution, which is used also in modelling the rank-frequency of letters or phonemes. A further difference is the fact that the angle of "writer´s view" in linguistic texts converges to the golden section 1.618. (cf. Popescu, Altmann 2007) but phonemes/letters or tone pitches do not. Nevertheless, the angle can be characteristic of composition, author, style, genre, historical epoch, etc., just as it is with other properties of rank-frequency distributions (cf. Martináková 2007).

Consider the *h*-point and the cosine of its angle as presented in Figure 5. The cosine can be computed as

$$\cos\alpha = \frac{-[h(f_1 - h) + h(n - h)]}{[h^2 + (f_1 - h)^2]^{1/2}[h^2 + (n - h)^2]^{1/2}}$$

where f_1 is the greatest frequency and *n* is the inventory of pitches. For example Sonata 1 by Beethoven in which $h = 42$, $f_1 = 537$, $n = 59$ yields

$$\cos\alpha = -[42(537-42) + 42(59-42)]/\{[42^2 + (537-42)^2]^{1/2}[42^2+(59-42)^2]^{1/2}\} = -0.9553$$

from which arccos(-0.9553) = 2.8416 radians. Evidently, these values drastically differ from those in linguistic texts concerning words which converge to the golden section.

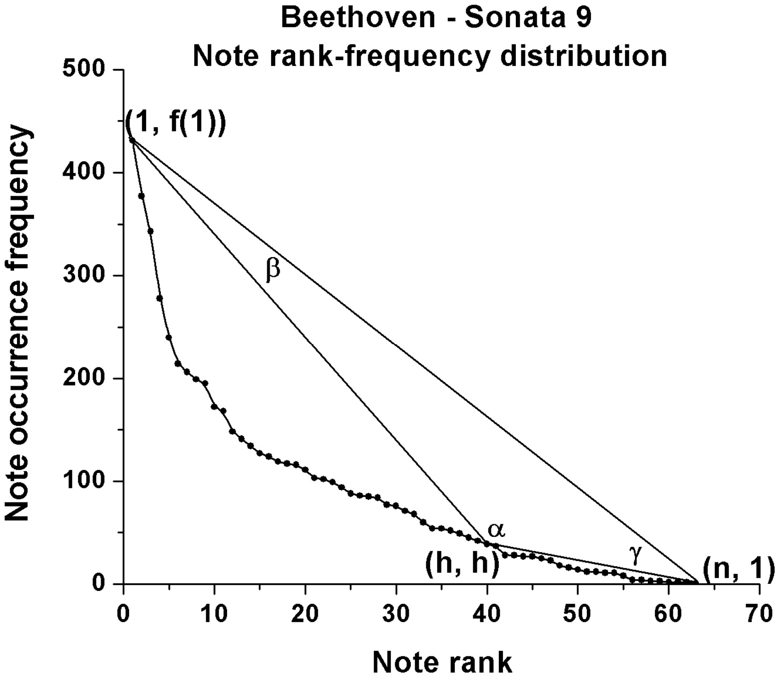


Figure 5a. The *h*-point and the angle α for a Beethoven composition

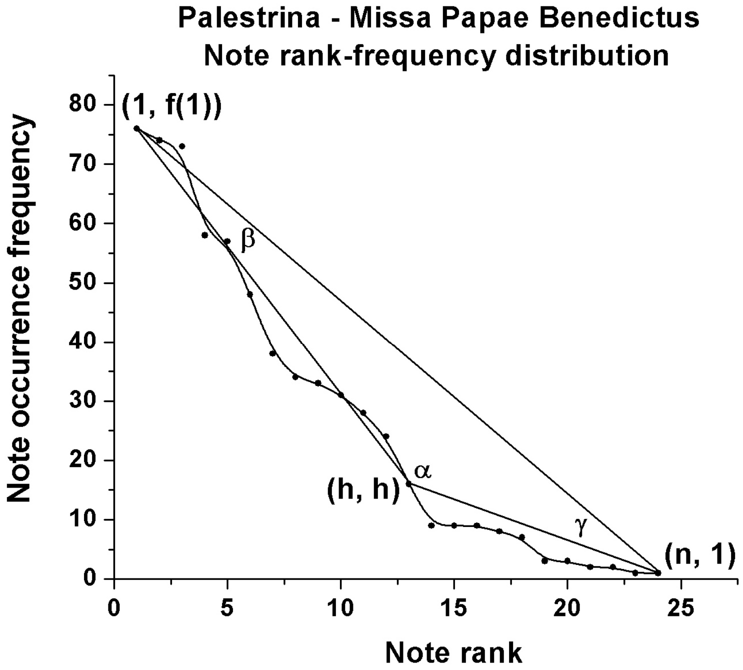


Figure 5b. The *h*-point and the angle *α* for a Palestrina composition

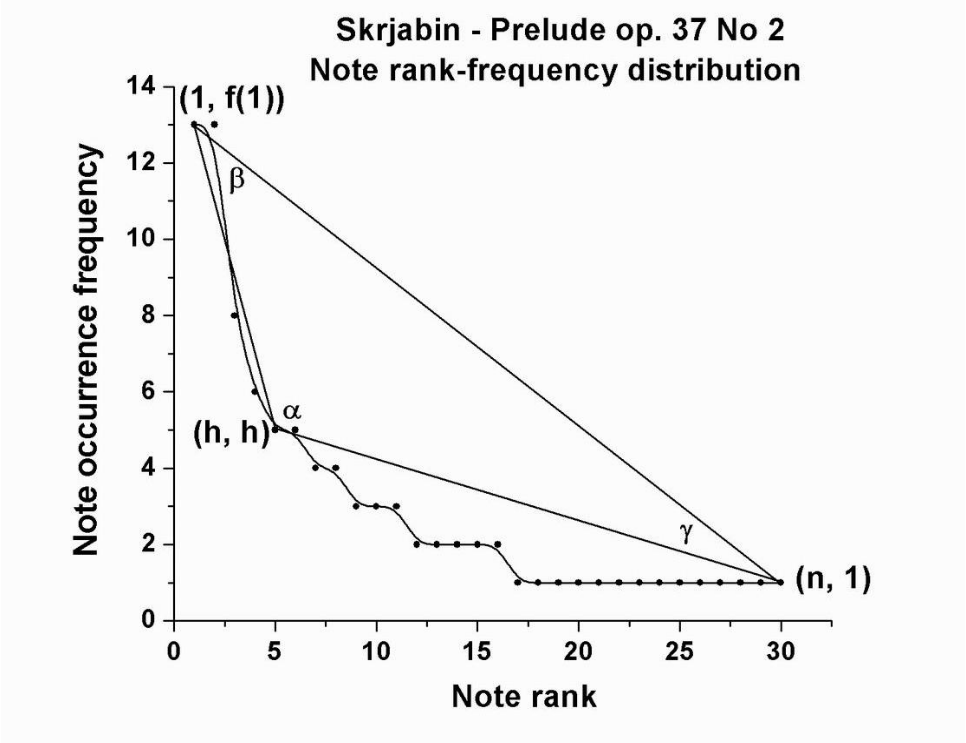


Figure 5c. The *h*-point and the angle α for a Skrjabin composition

As can be seen in Figure 6, the angles with Palestrina do not depend on composition length, and the angles *β* and *γ* are so acute that they cannot be used for characterization.

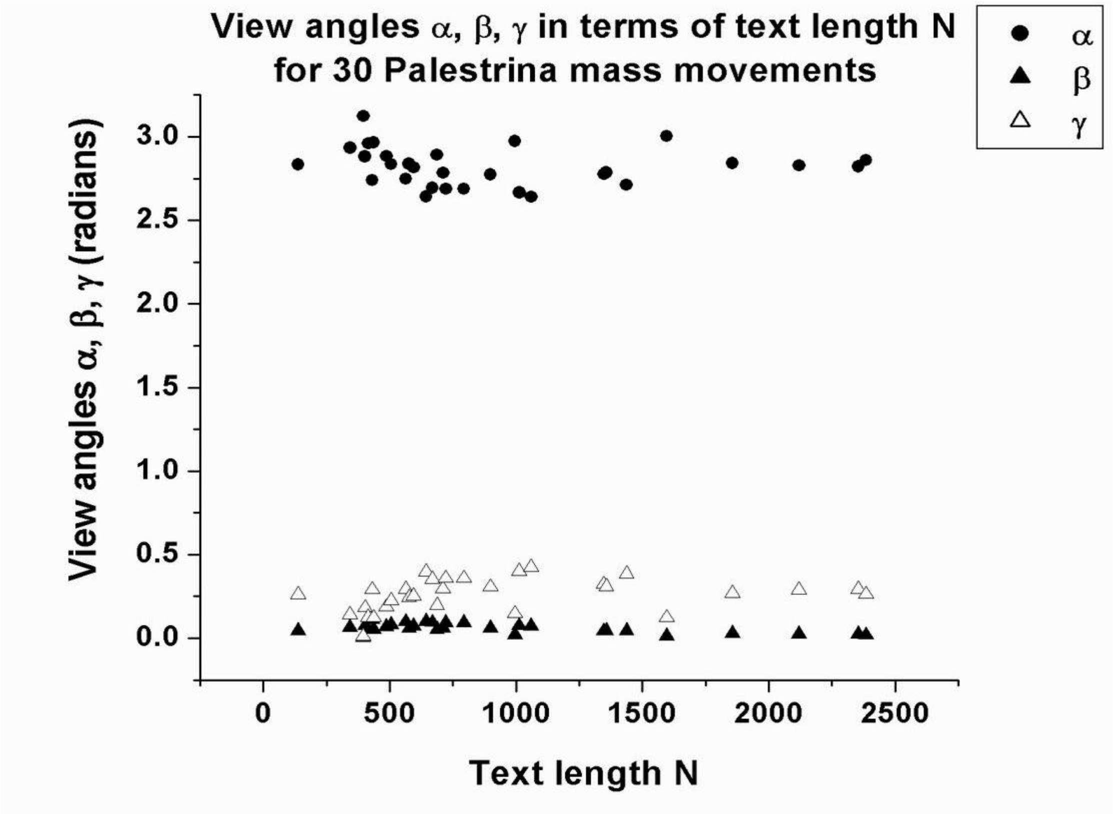


Figure 6. The angles of the triangle of the pitch distribution with Palestrina

Again, we can look whether there is a development of this angle in time. In Table 8 one can see the mean α radians with different composers

Table 8
The alpha radians with different composers

Name	Mean α radians
Palestrina	2.8212
Gesualdo	2.6053
Monteverdi	2.6945
Bach	2.6510
Mozart	2.6929
Beethoven	2.8340
Liszt	2.5526
Skrjabin	2.5582
Schoenberg	2.5449
Stravinsky	2.5615
Shostakovich	2.5515
Ligeti	2.8005

The mean α radians seem to represent a constant which does not change in the course of time and displays only a random oscillation. Hence this indicator is evidently a musical constant having a value α = 2.6557 ± 0.1071, almost coincident with the mathematical (Euler's or Napier's) number e = 2.71828...

3. Searching for the golden section

In natural language texts the golden section has been found as the limit of the α radians of the *h*-point of the rank-frequency distribution of words. However, as mentioned above, in music, simple notes do not correspond to words in language but rather to phonemes or letters. Hence if we believe in the existence of the golden section in the distribution of pitches, we must search for it differently. Let us begin with presenting the ranks and the frequencies in logarithmic form as can be seen in Table 9 for Sonata 5 by Beethoven. The natural logarithm of the rank is in the third column, the logarithm of the frequency in the fourth. If we draw a diagram, the logarithmic presentation has approximately the form of a concave monotone decreasing function, as illustrated in Figure 7.

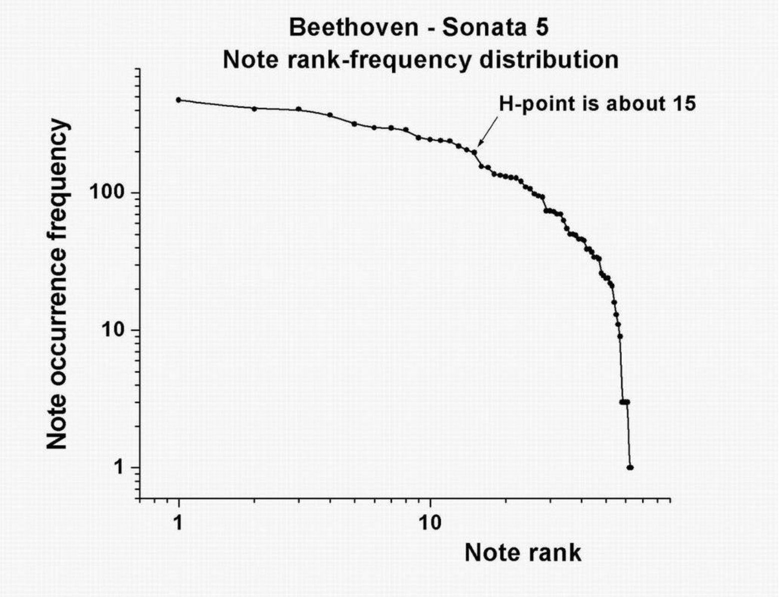


Figure 7. *h*-point definition

However, one can see that the first part of this curve has a rather linear form. Let us seek the end of the straight line. To this end we first take the first three values (of *log(r)* and *log(f(r))*) and compute the straight line. We obtain *log(f(r))* = 6.1457 – 0.1454 *log*® and the determination coefficient is R^2 = 0.8668. We add the next value and compute the straight line again. In this way we continue up to *r* = 18. The straight line exists if the determination coefficient R^2 oscillates or even increases, as can be seen in the sixth column of Table 9. Beginning with point *r* = 15 the determination coefficient begins to decrease because the points change the direction. Hence point *r* = 15 is the last point of the straight line.

Now let us compute the cumulative relative frequencies of the first part of the rank-frequency distribution as shown in the seventh column of Table 9. As can be seen, *F*(15) = 0.6159 represents that value which is the nearest to the golden proportion 0.618. This *r*-point will be called *H* and the cumulative frequency *F*(*H*) is called *H*-coverage.

Table 9
Computation of the *H*-point (Beethoven Sonata No 5)

Rank r	Frequency f(r)	ln(r)	ln(f(r))	ln(f(r)) = a-b ln(r)	R^2	F(r)
1	473	0.0000	6.1591			0.0654
2	407	0.6931	6.0088			0.1217
3	407	1.0986	6.0088	6.1457-0.1454x	0.8668	0.1780
4	369	1.3863	5.9108	6.1519-0.1636x	0.9213	0.2291
5	317	1.6094	5.7589	6.1760-0.2159x	0.8675	0.2729
6	298	1.7918	5.6971	6.1927-0.2451x	0.8876	0.3142
7	296	1.9459	5.6904	6.1970-0.2517x	0.9123	0.3551
8	288	2.0794	5.6630	6.1988-0.2540x	0.9285	0.3949
9	252	2.1972	5.5294	6.2161-0.2748x	0.9221	0.4298
10	244	2.3026	5.4972	6.2288-0.2889x	0.9263	0.4635
11	240	2.3979	5.4806	6.2365-0.2970x	0.9341	0.4967
12	239	2.4849	5.4765	6.2395-0.2999x	0.9418	0.5298
13	219	2.5649	5.3891	6.2499-0.3095x	0.9434	0.5601
14	206	2.6391	5.3279	6.2628-0.3208x	0.9416	0.5886
15	197	2.7081	5.2832	6.2758-0.3318x	0.9400	0.6159
16	155	2.7726	5.0434	6.3111-0.3605x	0.8939	0.6373
17	153	2.8332	5.0304	6.3394-0.3825x	0.8798	0.6585
18	137	2.8904	4.9200	6.3724-0.4074x	0.8627	0.6774
……	….	….	….	….	….	….

Of course, this lengthy computation is not always necessary because *H* can be determined visually or using a very quick method by means of Excel. The *H*-point is given by the rank at which *r***f*(*r*) becomes a maximum, as shown in Table 10 for the same data and in Figure 8.

Table 10
Computing the *H*-point (Beethoven Sonata No. 5)

Rank *r*	Frequency *f*(*r*)	*r***f*(*r*)
1	473	473
2	407	814
3	407	1221
4	369	1476
5	317	1585
6	298	1788
7	296	2072
8	288	2304
9	252	2268
10	244	2440
11	240	2640
12	239	2868

13	219	2847
14	206	2884
15	197	2955
16	155	2480
17	153	2601
18	137	2466
19	134	2546
20	131	2620
…………	…………………	………………

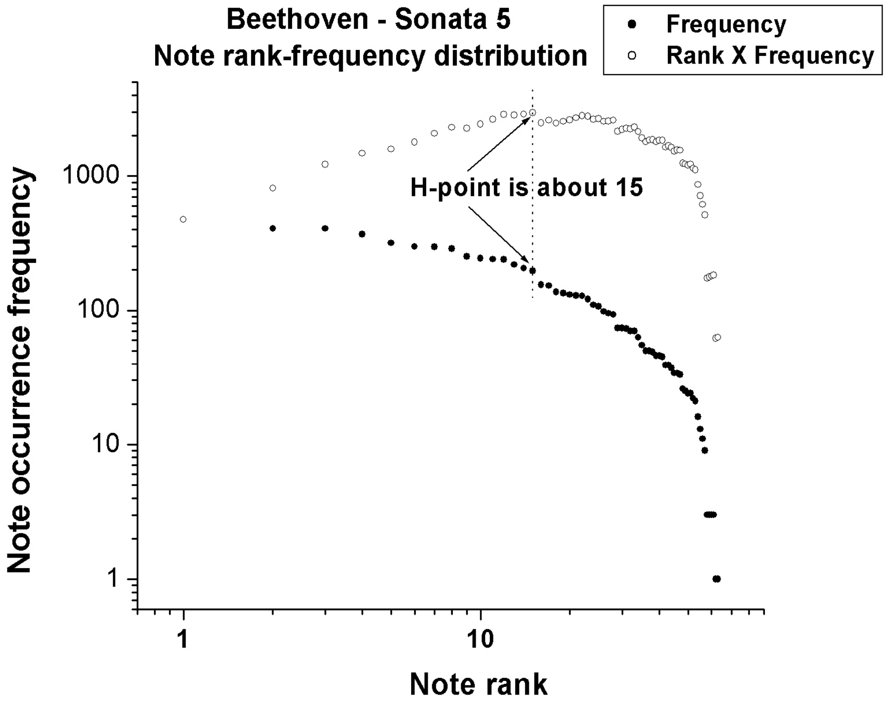


Figure 8. Determination of the *H*-point

F(*H*) is not always exactly 0.618 but it tends to this number. We must take into account that in written compositions the composers can make changes in the score a posteriori and cause thereby deviations, while in improvisations the agreement could be almost exact. To this end examinations in this direction should be made.

In order to show that this point displays a certain stability and is part of the composition we show in Figure 9 the *F*(*H*)-coverage for all Sonatas of Beethoven. The coverage does not change either with the length of the composition or with Beethoven´s age, and its mean for all Sonatas is 0.617 ± 0.057 where 0.057 is the standard deviation σ (see Table 10). Possibly the partitioning of the Sonatas in their parts would bring still better agreement.

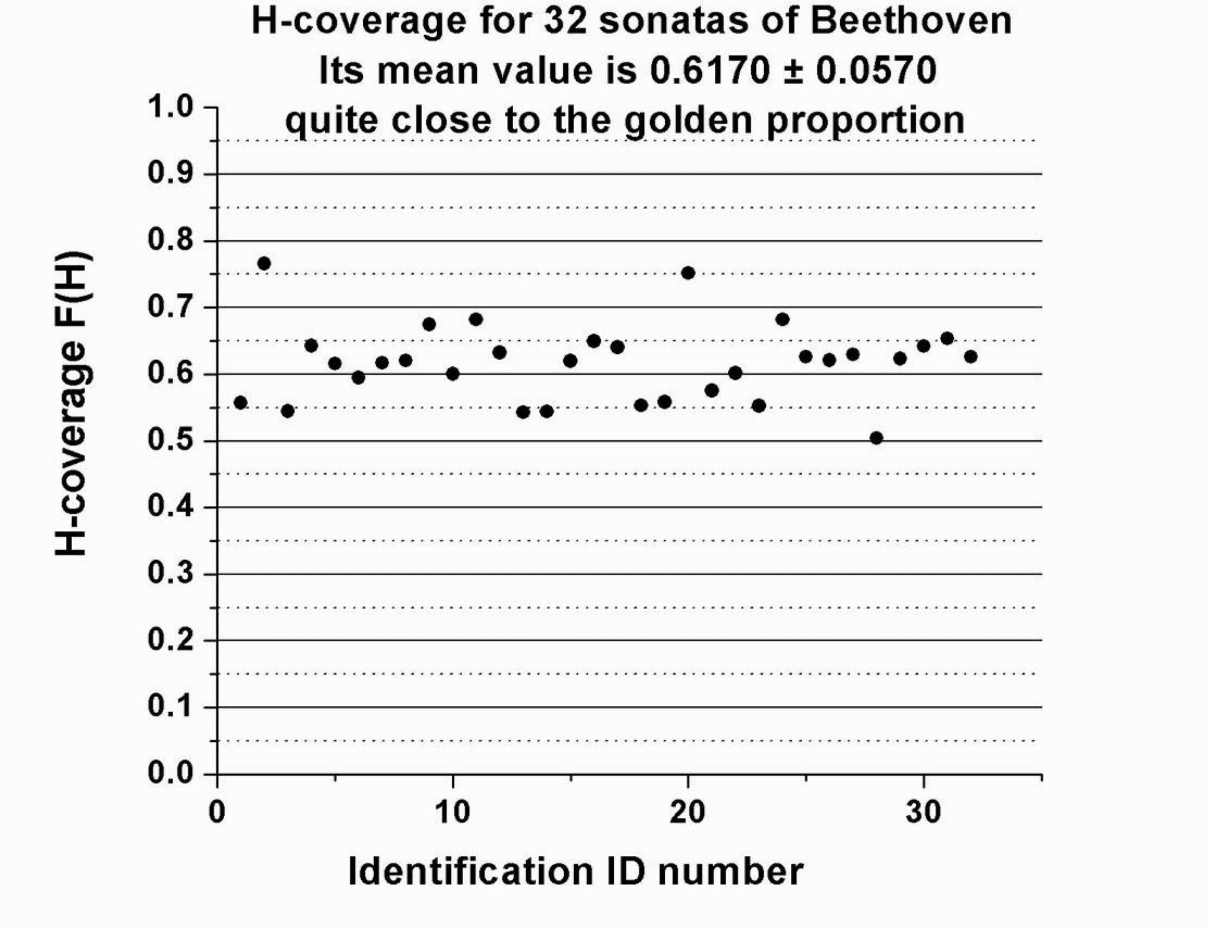
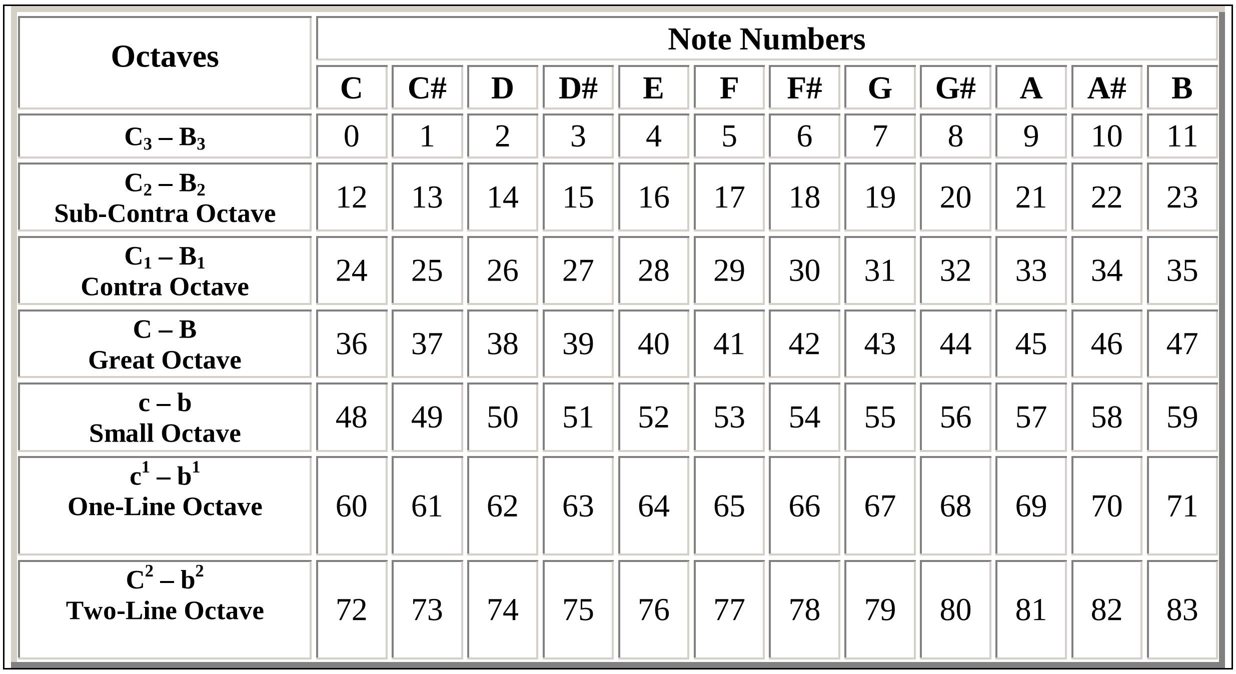
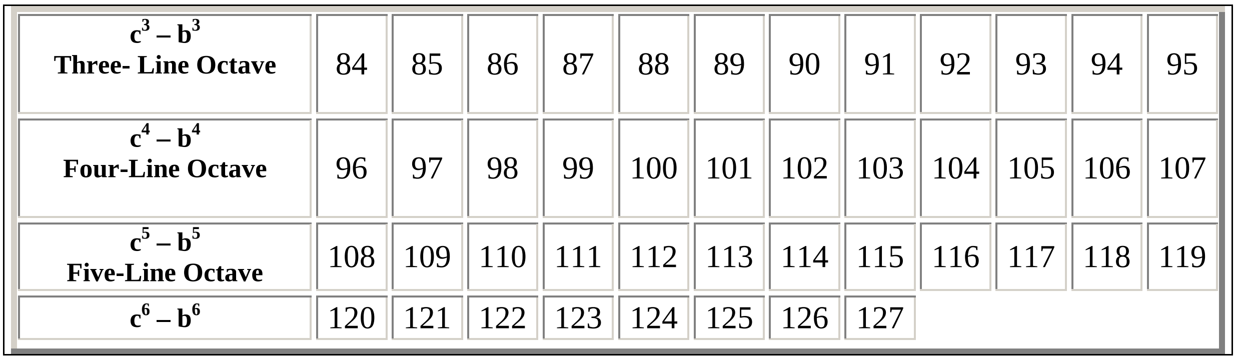


Figure 9. The *F*(*H*) for Beethoven´s Sonatas

In (linguistic) text analysis one knows that the most frequent words are synsemantics but in music we must look for the function of these pitches. Let us start from the usual marking of tones as shown in Figure 10, where the middle c is at piano keyboard (c^1 = 60).



Octaves	**Note Numbers**											
	C	**C#**	**D**	**D#**	**E**	**F**	**F#**	**G**	**G#**	**A**	**A#**	**B**
$C_3 – B_3$	0	1	2	3	4	5	6	7	8	9	10	11
$C_2 – B_2$ **Sub-Contra Octave**	12	13	14	15	16	17	18	19	20	21	22	23
$C_1 – B_1$ **Contra Octave**	24	25	26	27	28	29	30	31	32	33	34	35
C – B **Great Octave**	36	37	38	39	40	41	42	43	44	45	46	47
c – b **Small Octave**	48	49	50	51	52	53	54	55	56	57	58	59
$c^1 – b^1$ **One-Line Octave**	60	61	62	63	64	65	66	67	68	69	70	71
$C^2 – b^2$ **Two-Line Octave**	72	73	74	75	76	77	78	79	80	81	82	83



c^3 – b^3 **Three- Line Octave**	84	85	86	87	88	89	90	91	92	93	94	95
c^4 – b^4 **Four-Line Octave**	96	97	98	99	100	101	102	103	104	105	106	107
c^5 – b^5 **Five-Line Octave**	108	109	110	111	112	113	114	115	116	117	118	119
c^6 – b^6	120	121	122	123	124	125	126	127				

Figure 10. MIDI note numbers for the tone pitch and octave designation according to the Helmholtz System used in this article

The musicological interpretation of the points *H* and *h* could be, for example, the *Sonata No. 5* by Beethoven shown in Table 11, as follows: the Sonata is composed in tonal system in *C minor key* (1. movement – Allegro molto e con brio), in *A-flat major key* (2. movement – Adagio molto), *C minor key* (3. movement – Finale, last chord is in *C– major*, similar as in modal system where the last chord is mostly in major version: last cadence: minor sub-dominant triad: *f-ab-c*; diminished seventh (vii7 in minor keys): *h-d-f-ab* and tonic in major version: *c-e-g*.

In the first most frequent 15 tones (to the point *H*) we can see only the basic tones of the C minor key: *c-d-eb-f-g-ab-bb* in natural version (cf. Aeolian modus).

The tones from the point *H* to *h* (15-42) are:

1. the same tones but placed also in other octaves;
2. one most frequent new tone: *b* – it is very important as major seventh which is the basic tone (mediant) in the dominant (*g-b-d*);
3. one less frequent tone: *d-flat* – it is the basic tone in *A-flat major key* in the second movement;
4. two diesis: *e*, *a* – depend on the leading tones in melody and chromatization (*e* is also the mediant in major version of tonic triad *c-e-g* and *a* is the mediant for subdominant triad *f-a-c*);

After the point **h** we can find except for the mentioned tones (but also in more extreme octaves) the last 12th tone *f#/g-flat*.

Table 11
Pitches corresponding to ranks and frequencies in Beethoven´s Sonata 5

Rank	Freq	Pitch	Name	Rank	Freq	Pitch	Name	Rank	Freq	Pitch	Name
1	473	6300	e-flat^1	22	128	7100	b^1	43	39	3400	B-flat/A#$_1$
2	407	6000	c	23	121	5000	d	44	37	5400	g-flat/f#
3	407	5500	g	24	110	4600	B-flat/A#	45	34	4500	B-flat/A#
4	369	5100	e-flat	25	107	6100	d-flat/c#^{1}	46	34	3100	G$_1$
5	317	6700	g^1	26	98	8000	a-flat^2	47	33	6600	g-flat/f#^{1}
6	298	5600	a-flat	27	95	4400	A-flat	48	26	7800	g-flat/f#^{2}
7	296	7200	c^2	28	93	8200	b-flat/a#^{2}	49	25	3200	A-flat$_1$
8	288	5800	b-flat/a#	29	74	7300	d-flat/c#^{2}	50	24	8100	a^2
9	252	5300	f	30	74	6400	e^1	51	24	4200	G-flat/F#
10	244	6500	f^1	31	73	3900	E-flat/D#	52	22	8500	d-flat/c#^{3}
11	240	7500	e-flat^2	32	70	8600	d^3	53	21	4900	d-flat/c#

12	239	6800	a-flat^1
13	219	4800	c
14	206	6200	d^1
15	197	7000	b-flat/a#^{1}
16	155	7400	d^2
17	153	7700	f^2
18	137	7900	g^2
19	134	8400	c^3
20	131	5900	b
21	129	4300	G
33	70	4700	B
34	63	8700	e-flat^3
35	55	8300	b^2
36	50	6900	a^1
37	50	5200	e
38	49	8900	f^3
39	46	3600	C
40	46	4100	F
41	45	7600	e^2
42	39	5700	a
54	16	3500	B$_1$
55	13	3800	D
56	11	8800	e^3
57	9	3300	A$_1$
58	3	4000	E
59	3	2900	F$_1$
60	3	3000	G-flat/F#$_1$
61	3	3700	D-flat/C#
62	1	9000	g-flat/f#^{3}
63	1	2700	E-flat/D#$_1$

The computation of *H* and *F*(*H*) is shown in Tables 1A to 12A in the Appendix. As can be seen in Tables 1A to 12A and presented collectively in Table 12, the mean *F*(*H*) seems to develop. With Palestrina it does not acquire its ideal form; with Bach it acquires its purest form, thereafter an oscillation begins. This statement is very preliminary because we studied only some works by several composers. A more extensive investigation is necessary in order to attain better founded statements. In any case we have shown that something like the golden proportion exists directly in the frequencies of pitches.

Table 12
Survey of *H*-coverages

Composer	**mean F(H)**	**σ**
Palestrina	0.7530	0.0893
Gesualdo	0.6160	0.0249
Monteverdi	0.6183	0.0972
Bach	0.6180	0.0703
Mozart	0.6076	0.0530
Beethoven	0.6170	0.0570
Liszt	0.6231	0.0692
Skrjabin	0.5766	0.0813
Schoenberg	0.6268	0.0208
Stravinsky	0.7556	0.1079
Shostakovich	0.6746	0.0886
Ligeti	0.6986	0.0491

Since the computation of *H* is not always unequivocal but we are aware of its existence, the following algorithm can be proposed a posteriori: (a) Plot the ranks and frequencies of pitches in double-logarithmic scale. (b) Determine the *H*-point optically as the last point on the straight line beginning with ln(f_1). (c) Compute stepwise the linear regression starting from the point <0, ln(f_1)> down to the point yielding the last maximum determination coefficient. (d) If the optical and the computed *H*-point coincide, accept it. (e) If they do not coincide, choose that of the two points whose *F*(*H*) is nearer to 0.618. (f) Check the computation by the rank *H* corresponding to max[r*$f(r)$]. (g) Generally, a major downwards bend of the actual distribution defines the *H*-point, as illustrated in Figure 11. This implies that it is located at the maximum of the difference $\Delta f = f_{\text{actual}} - f_{\text{fitting}}$, as shown in Figure 12. It is to be noticed, however, that the parasite maxima at lower ranks should be discarded. Moreover, this last method should be applied cautiously, inasmuch as irregular actual distributions may produce a few Δ*f* maxima before the occurrence of the major distribution bend.

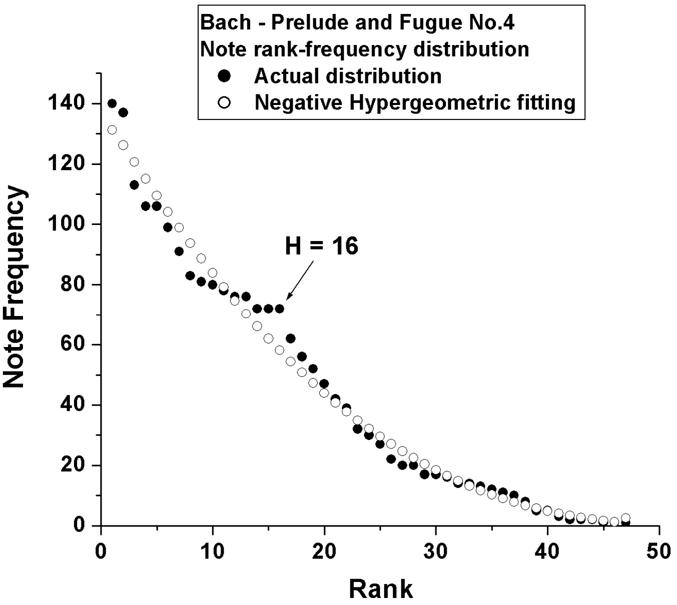


Figure 11. The *H*-point as a distribution break up

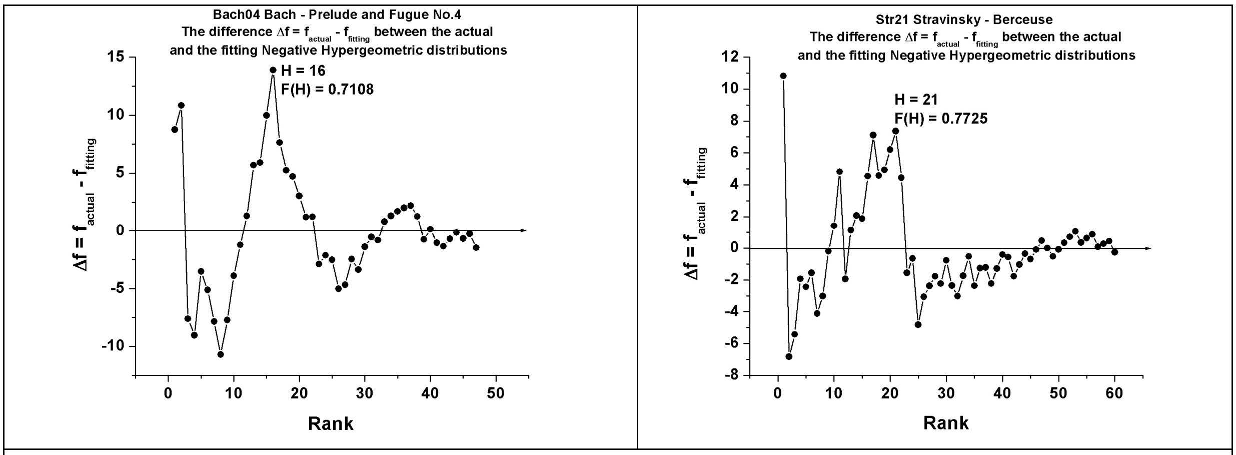


Figure 12. The *H*-point as the maximum of the $\Delta f = f_{actual} - f_{fitting}$ difference

The differences between *F*(*H*) coverages can again be tested using formula (2). Variances were estimated from simulations (cf. Section 2). Again, nine compositions (by Beethoven, Palestrina and Skrjabin) were chosen.

Table 13
Tests for differences between some *F*(*H*) coverages

	LvB01	LvB02	LvB28	Pls01	Pls15	Pls23	Skr01	Skr07	Skr14
LvB01	0	**-2.28**	0.56	**-2.74**	0.09	**-3.02**	0.69	0.80	**-1.78**
LvB02	**2.28**	0	**2.88**	-0.79	1.78	-1.24	**2.40**	**2.54**	0.08
LvB28	-0.56	**-2.88**	0	**-3.26**	-0.33	**-3.49**	0.27	0.37	**-2.26**
Pls01	**2.74**	0.79	**3.26**	0	**2.24**	-0.47	**2.81**	**2.95**	0.74
Pls15	-0.09	-1.78	0.33	**-2.24**	0	**-2.54**	0.50	0.58	-1.51

Pls23	**3.02**	1.24	**3.49**	0.47	**2.54**	0	**3.07**	**3.20**	1.14
Skr01	-0.69	**-2.40**	-0.27	**-2.81**	-0.50	**-3.07**	0	0.08	**-2.06**
Skr07	-0.80	**-2.54**	-0.37	**-2.95**	-0.58	**-3.20**	-0.08	0	**-2.17**
Skr14	1.78	-0.08	**2.26**	-0.74	1.51	-1.14	**2.06**	**2.17**	0

As can be seen, significant differences can arise even within the work of one composer and about half of the differences are significant. Hence *F*(*H*) seems to be a very sensitive characteristic of the composition.

Consequently, the question arises whether *F*(*H*) is a historically changing phenomenon or simply a text characteristic. Its “ideal value” attained by Bach displays a motion beginning with Palestrina and ending (preliminarily) with Ligeti, but this motion is not very smooth. In any case one can see a concave course. A special representation of this trend is shown in Figure 13, where we plotted the dependence $\langle$time, Log $A\rangle$ with $A = 1/\left|[F(H) - 0.618034]\right|$ as a merit indicator. Clearly we have to deal with the time development of a couple of concurring processes, firstly a fast rising one and secondly a slowly decaying one. Most intuitive appears the comparison of this compound motion in terms of the difference of two exponential functions as follows

$$y(t) = c\left[\exp\left(-\frac{t-t_0}{T_{fall}}\right) - \exp\left(-\frac{t-t_0}{T_{rise}}\right)\right]$$

where *y* is the considered musical merit indicator (here Log *A*), *t* is the time, t_0 is the time origin, *c* is a scaling factor, T_{rise} is the rise time of the "musical phenomenon", and T_{fall} is its decay time. This is a slightly modified 4 parameter Box-Lucas2 fitting exponential function built in the Origin 6.1 program (see more in Box, Lucas 1959). As it is illustrated in Figure 13, the musical golden proportion impetus has a maximum located in the mid of the 17th century, a rise time $T_{rise} \approx 75$ years, and a decay time $T_{fall} \approx 150$ years, hence a width of about $W = T_{rise} + T_{fall} = (75 + 150)$ years = 225 years, heralding and covering the brilliant epoch of Bach, Mozart, and Beethoven. On the other hand, the oldest composers considered in the present paper and belonging to the beginning of this motion are Palestrina, Gesualdo and Monteverdi after Leonardo da Vinci (1452–1519), Michelangelo (1475–1564), and Luca Pacioli (1445–1514) with his *Divina Proportione* (1509). Consequently, it appears that the whole musical golden proportion inspiration appears as a late echo of the Renaissance that spans roughly the 14th through the 17th century.

This development can be seen in Table 14 and Figure 13.

Table 14
Fitting A =1/|meanF(H)-0.618034| by Box-Lucas and impulse functions

Composer	Year	mean F(H)	A	Log A	$(\text{Log A})_{\text{Box-Lucas}}$	$(\text{Log A})_{\text{impulse}}$
Palestrina	1560	0.7530	7.409	0.870	0.802	0.797
Gesualdo	1587	0.6160	491.642	2.692	2.817	2.821
Monteverdi	1605	0.6183	3759.398	3.575	3.594	3.599
Bach	1718	0.6180	29411.765	4.469	3.848	3.842
Mozart	1774	0.6076	95.841	1.982	3.062	3.057
Beethoven	1799	0.6170	967.118	2.985	2.709	2.706

Liszt	1849	0.6231	197.394	2.295	2.072	2.072
Skrjabin	1894	0.5766	24.135	1.383	1.596	1.600
Schoenberg	1913	0.6268	114.077	2.057	1.425	1.429
Stravinsky	1927	0.7556	7.269	0.861	1.308	1.313
Shostakovich	1940	0.6746	17.678	1.247	1.208	1.213
Ligeti	1965	0.6986	12.412	1.094	1.034	1.040

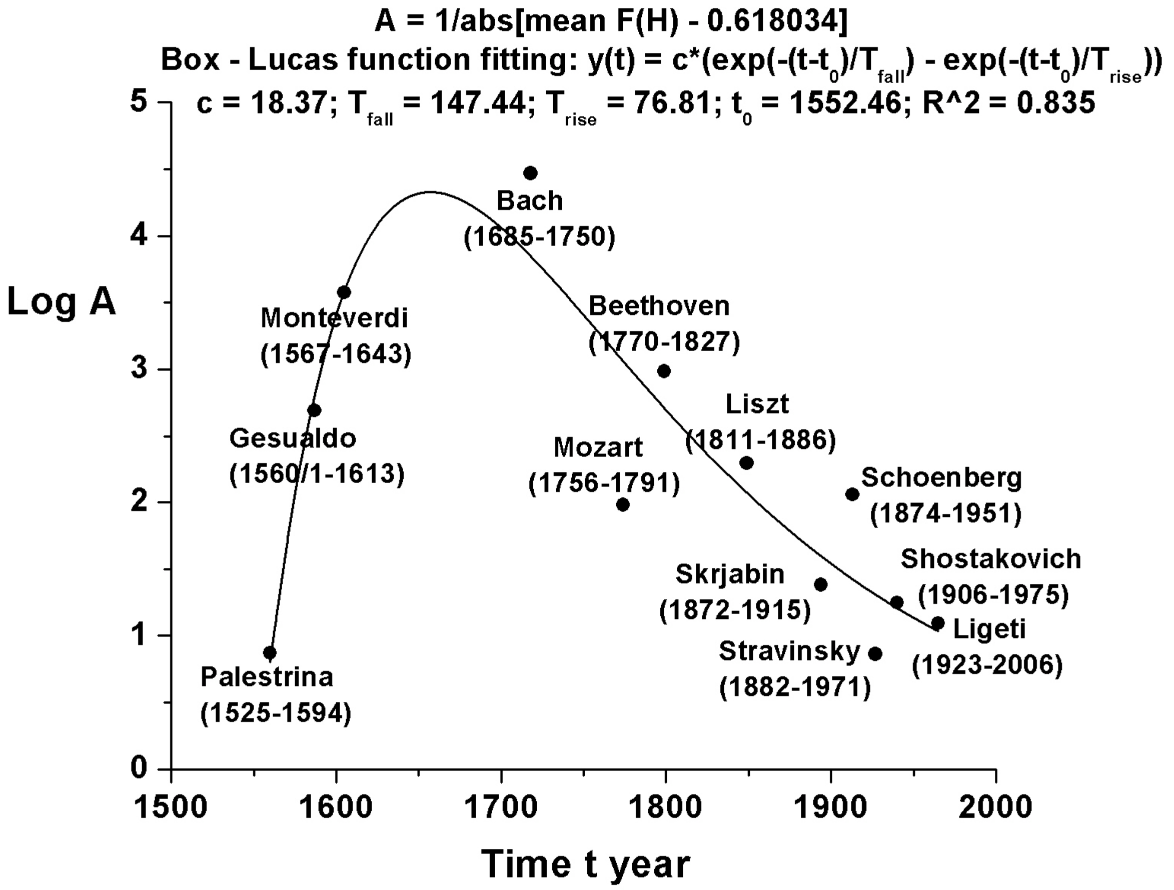


Figure 13. The musical echo of the Renaissance golden proportion
as revealed by the evolution of Log A (4 parameter Box-Lucas function fitting)

Another possibility is the use of the impulse function having three parameters and defined as

$$y(t) = c\exp\left(-\frac{t-t_0}{T}\right)\left[1-\exp\left(-\frac{t-t_0}{T}\right)\right]$$

yielding the results in Table 14 and Figure 14. The coincidence of both Box-Lucas and impulse function fitting is remarkable.

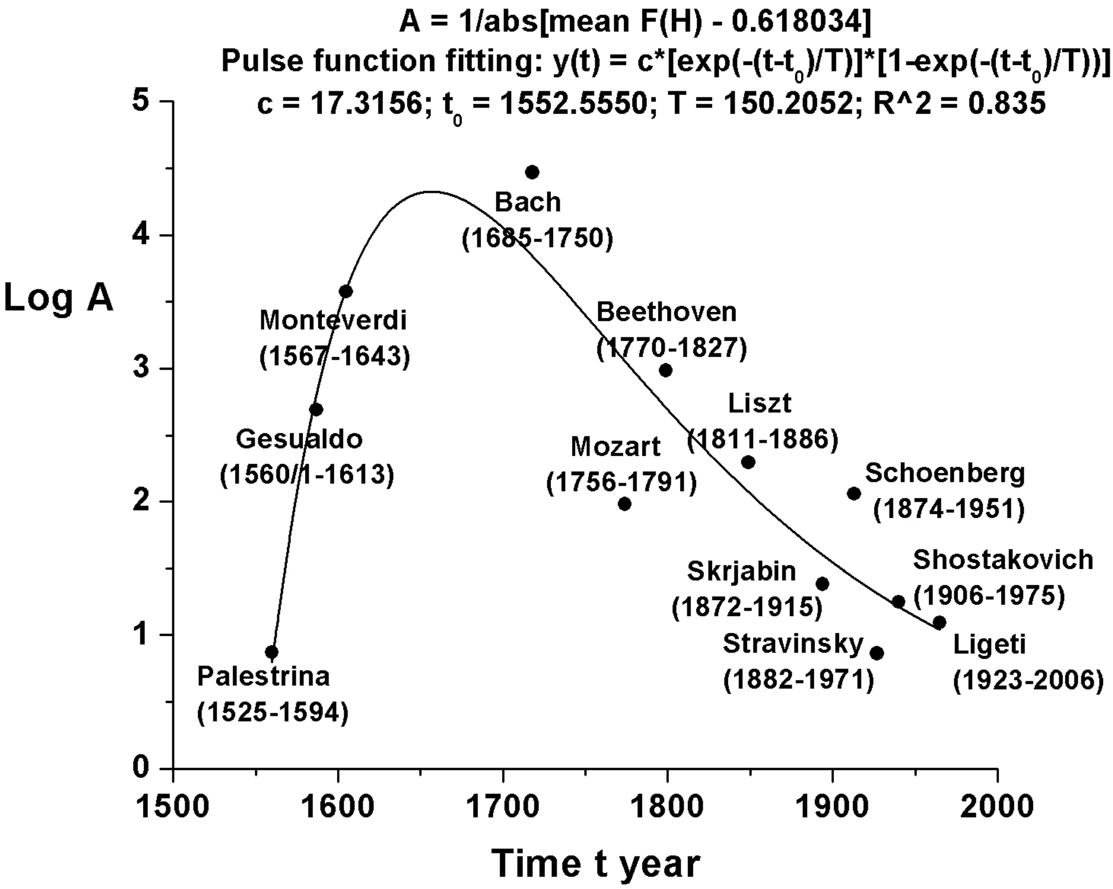


Figure 14. The musical echo of the Renaissance golden proportion as revealed by the evolution of Log A (3 parameter impulse function fitting)

Acknowledgement

J. Mačutek was supported by the research grant VEGA 1/3016/06.

References

Box, G. E. P., Lucas, H. L. (1959). Design of experiments in non-linear situations. *Biometrika XLVI, 77-90.*

Hirsch, J. E. (2005). An index to quantify an individual´s scientific research output. *Proceedings of the National Academy of Sciences of the USA 102, 16569-16572.* Cf. http://arxiv.org/PS_cache/physics/pdf/0508/0508025.pdf

Köhler, R., Martináková-Rendeková, Z. (1995). Niekoľko poznámok k systémovo-teoretickej analýze hudby. *Hudobno-pedagogické interpretácie (Nitra) 3, 51-58.*

Köhler, R., Martináková-Rendeková, Z. (1998). A systems theoretical approach to language and music. In: Altmann, G., Koch, W.A. (eds.), *Systems. New Paradigms for Human Sciences: 514-546.* Berlin – New York: Walter de Gruyter.

Mačutek, J., Popescu, I.-I., Altmann, G. (2007). Confidence intervals and tests for the h-point and related text characteristics. *Glottometrics 15, 45-52.*

Martináková, Z. (1997). Nové metódy kvantitatívnej analýzy hudby. Aplikácia niektorých počítačových programov na skladby v MIDI dátach. In: Martináková, Z., (ed), *Zborník Metódy analýzy a interpretácie hudby z historického a systematického aspektu I: 67-74.* Bratislava: Vysoká škola múzických umení.

Martináková, Z. (1998). Niektoré aspekty systémovej teórie v hudbe. In: Martináková, Z. (ed.), *Zborník Metódy analýzy a interpretácie hudby z historického a systematického aspektu II, 84-93*. Bratislava: Vysoká škola múzických umení.

Martináková-Rendeková, Z. (2000). Systems theoretical modelling in musicology. In: Mastorakis, N. (ed.), *Mathematics and Computers in Modern Science. Acoustics and Music, Biology and Chemistry, Business and Economics: 122-127.* Athens: World Scientific and Engineering Society Press 2000.

Martináková-Rendeková, Z. (2002). Synergetische und systemtheoretische Aspekte der Musikanalyse. *Semiotische Berichte 1-4/02, 191-216.*

Martináková-Rendeková, Z. (2003). Systems theoretical modelling of the 20th century music (An endeavour to categorization). *WSEAS Transactions on Acoustics and Music 1, 1-6.*

Martináková-Rendeková, Z. (2004). Rank-frequency distribution of pitch in musical texts using Altmann-Fitter 2.0 and Reinhard Köhler's QUAMS computer programs. *WSEAS 2004*. In: www.wseas.org

Martináková-Rendeková, Z. (2007). Different Parameters of the negative hypergeometric distribution as a discriminating feature for musical or composer's style. In: *Sytems theory and scientific computation. Proceedings of the 7th WSEAS International Conference (ISTASC'07): 217-222.* WSEAS Press.

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

Popescu, I.-I.; Altmann, G. (2006). Some geometric properties of word frequency distributions. *Göttinger Beiträge zur Sprachwissenschaft 13, 87-98.*

Popescu, I.-I. (2007). Text ranking by the weight of highly frequent words. In: Grzybek, P., Köhler, R. (eds.), *Exact methods in the study of language and text: 555-565*. Berlin: de Gruyter.

Popescu, I.-I.; Best, K-H.; Altmann, G. (2007). On the dynamics of word classes in text. *Glottometrics 14, 61-74.*

Popescu, I.-I., Altmann, G. (2007). Writer´s view of text generation. *Glottometrics 15, 71-81.*

Popescu, I.I. et al. (2008*). Word frequency studies*. Berlin: Mouton de Gruyter (in print)

Wimmer, G., Wimmerová, S. (1997). Exaktnejšie formulácie zákonitostí v hudbe. In: Martináková, Z. (ed.), *Zborník Metódy analýzy a interpretácie hudby z historického a systematického aspektu I, 75-84*. Bratislava: Vysoká škola múzických umení.

Appendix

Table A1
H and *F*(*H*) for Palestrina

ID	Text	N	H	F(H)
Pls01	Ascendo 1. Motetto	1856	12	0.8475
Pls02	Ascendo 2. Kyrie	898	10	0.7728
Pls03	Ascendo 3. Gloria	1348	12	0.8435
Pls04	Ascendo 4. Credo	2120	9	0.7193
Pls05	Ascendo 5. Sanctus	595	9	0.7445
Pls06	Ascendo 5. Benedictus	563	8	0.7194
Pls07	Ascendo 7. Agnus Dei I	431	10	0.7610
Pls08	Ascendo 8. Agnus Dei II	487	12	0.8480
Pls09	Ave Regina Chant	137	3	0.7445

Pls10	Ave Regina Kyrie	687	11	0.8122
Pls11	Ave Regina Gloria	1357	8	0.6743
Pls12	Ave Regina Credo	2355	11	0.8191
Pls13	Ave Regina Sanctus	436	10	0.7729
Pls14	Ave Regina Benedictus	505	9	0.7525
Pls15	Ave Regina Agnus Dei I	396	7	0.5455
Pls16	Ave Regina Agnus Dei II	402	10	0.7886
Pls17	Missa Papae Kyrie	995	8	0.7035
Pls18	Missa Papae Gloria	1437	13	0.8984
Pls19	Missa Papae Credo	2385	9	0.7338
Pls20	Missa Papae Sanctus	1060	9	0.7481
Pls21	Missa Papae Benedictus	644	6	0.5994
Pls22	Missa Papae Agnus Dei I	711	10	0.7792
Pls23	Missa Papae Agnus Dei II	793	13	0.9067
Pls24	Missa Veni Kyrie	669	7	0.6099
Pls25	Missa Veni Gloria	1013	8	0.6614
Pls26	Missa Veni Credo	1596	10	0.7531
Pls27	Missa Veni Sanctus	722	11	0.8324
Pls28	Missa Veni Benedictus	576	9	0.7622
Pls29	Missa Veni Agnus Dei I	343	12	0.8630
Pls30	Missa Veni Agnus Dei II	415	7	0.5735
		$\overline{F(H)} = 0.7530 \pm 0.0893$		

Table A2
H and *F*(*H*) for Gesualdo

ID	Text	N	H	F(H)
Ges01	Belta, poi che te accendi	688	10	0.6221
Ges02	Deh, coprite il bel seno	591	9	0.6024
Ges03	Dolcissima mia vita	581	10	0.6145
Ges04	Itene, o miei sospiri	761	10	0.6491
Ges05	Moro, lasso, al mio duolo	671	11	0.6528
Ges06	O vos omnes	432	12	0.5833
Ges07	Merce grido piangendo	681	8	0.5918
		$\overline{F(H)} = 0.6166 \pm 0.0249$		

Table A3
H and *F*(*H*) for Monteverdi

ID	Text	N	H	F(H)
Mon01	Monteverdi - Dixit Dominus (Psalm 109)	3002	12	0,8028
Mon02	Monteverdi - Laudate pueri (Psalm 112)	1927	10	0,7286
Mon03	Monteverdi - Laetatus sum (Psalm 121)	2719	6	0,4777

Mon04	Monteverdi - Nisi Dominus (Psalm 126)	3138	6	0,5118
Mon05	Monteverdi - Lauda Jerusalem (Psalm 147)	2161	9	0,6858
Mon06	Monteverdi - Hymn: Ave maris stella	1411	7	0,5464
Mon07	Monteverdi - Magnificat	1240	7	0,5355
Mon08	Monteverdi - A un giro sol de’belli occhi	813	8	0,6335
Mon09	Monteverdi - Si, ch’io vorrei morire	886	9	0,6377
Mon10	Monteverdi - Vorrei baciarti, o Filli	2217	6	0,6229
		$\overline{F(H)} = 0.6183 \pm 0.0972$		

Table A4
H and *F*(*H*) for Bach

ID	Text	N	H	F(H)
Bach01	1. Prelude and Fugue No 1	1318	10	0,5948
Bach02	1. Prelude and Fugue No 2	1877	10	0,5685
Bach03	1. Prelude and Fugue No 3	2266	14	0,6827
Bach04	1. Prelude and Fugue No 4	2085	16	0,7108
Bach05	1. Prelude and Fugue No 5	1553	13	0,6542
Bach06	1. Prelude and Fugue No 6	1602	10	0,5449
Bach07	1. Prelude and Fugue No 7	2345	12	0,5970
Bach08	1. Prelude and Fugue No 8	2129	12	0,5867
Bach09	1. Prelude and Fugue No 9	1221	14	0,7322
Bach10	1. Prelude and Fugue No 10	2069	12	0,5988
Bach11	1. Prelude and Fugue No 11	1562	11	0,5583
Bach12	1. Prelude and Fugue No 12	1897	11	0,5651
Bach13	1. Prelude and Fugue No 13	1378	12	0,6277
Bach14	1. Prelude and Fugue No 14	1477	10	0,5423
Bach15	1. Prelude and Fugue No 15	2392	12	0,5560
Bach16	1. Prelude and Fugue No 16	1491	10	0,5265
Bach17	1. Prelude and Fugue No 17	1575	13	0,6832
Bach18	1. Prelude and Fugue No 18	1371	13	0,6207
Bach19	1. Prelude and Fugue No 19	1794	14	0,6711
Bach20	1. Prelude and Fugue No 20	3043	15	0,7026
Bach21	1. Prelude and Fugue No 21	1603	11	0,5958
Bach22	1. Prelude and Fugue No 22	1514	14	0,6955
Bach23	1. Prelude and Fugue No 23	1315	11	0,5932
Bach24	1. Prelude and Fugue No 24	2551	10	0,5076
Bach25	2. Prelude and Fugue No 1	1973	14	0,6984
Bach26	2. Prelude and Fugue No 2	1361	10	0,5871
Bach27	2. Prelude and Fugue No 3	1624	16	0,7956
Bach28	2. Prelude and Fugue No 4	2663	17	0,7570
Bach29	2. Prelude and Fugue No 5	2423	11	0,5761
Bach30	2. Prelude and Fugue No 6	1897	9	0,5071

Bach31	2. Prelude and Fugue No 7	1616	13	0,6714
Bach32	2. Prelude and Fugue No 8	1994	13	0,6153
Bach33	2. Prelude and Fugue No 9	1645	11	0,6170
Bach34	2. Prelude and Fugue No 10	2637	13	0,6435
Bach35	2. Prelude and Fugue No 11	2206	10	0,5254
Bach36	2. Prelude and Fugue No 12	1849	9	0,5203
Bach37	2. Prelude and Fugue No 13	2618	13	0,6429
Bach38	2. Prelude and Fugue No 14	2279	13	0,6441
Bach39	2. Prelude and Fugue No 15	2436	13	0,6831
Bach40	2. Prelude and Fugue No 16	2144	11	0,5896
Bach41	2. Prelude and Fugue No 17	2876	11	0,5741
Bach42	2. Prelude and Fugue No 18	4090	12	0,5689
Bach43	2. Prelude and Fugue No 19	1439	10	0,5587
Bach44	2. Prelude and Fugue No 20	2271	16	0,6319
Bach45	2. Prelude and Fugue No 21	4421	16	0,7356
Bach46	2. Prelude and Fugue No 22	2933	16	0,7092
Bach47	2. Prelude and Fugue No 23	2355	10	0,5176
Bach48	2. Prelude and Fugue No 24	1852	11	0,5767
		$\overline{F(H)} = 0.6180 \pm 0.0703$		

Table A5

H and *F*(*H*) for Mozart

ID	Text	N	H	F(H)
Moz01	Mozart D major K.284	10585	13	0,6357
Moz02	Mozart C major K.309	7577	10	0,5125
Moz03	Mozart A minor K.310	8117	15	0,653
Moz04	Mozart Bb major K.333	7496	12	0,6107
Moz05	Mozart A major K.331	9470	9	0,5583
Moz06	Mozart C minor K.457	6400	15	0,6570
Moz07	Mozart C major K.545	3628	12	0,6563
Moz08	Mozart D major K.311	7157	10	0,5391
Moz09	Mozart F major K.332	6868	14	0,6457
		$\overline{F(H)} = 0.6076 \pm 0.0530$		

Table A6

H and *F*(*H*) for Beethoven

ID	Text	N	H	F(H)
LvB01	LvB Sonata 1	7332	13	0,5573
LvB02	LvB Sonata 2	9340	24	0,7661
LvB03	LvB Sonata 3	11915	14	0,5446
LvB04	LvB Sonata 4	12248	18	0,6424

LvB05	LvB Sonata 5	7229	15	0,6159
LvB06	LvB Sonata 6	7171	17	0,5948
LvB07	LvB Sonata 7	9201	19	0,6172
LvB08	LvB Sonata 8	8396	18	0,6205
LvB09	LvB Sonata 9	5706	19	0,6746
LvB10	LvB Sonata 10	6623	14	0,6005
LvB11	LvB Sonata 11	10898	18	0,6822
LvB12	LvB Sonata 12	9497	16	0,6324
LvB13	LvB Sonata 13	8461	13	0,5426
LvB14	LvB Sonata 14	8597	12	0,5437
LvB15	LvB Sonata 15	11581	16	0,6198
LvB16	LvB Sonata 16	13439	19	0,6497
LvB17	LvB Sonata 17	7905	19	0,6405
LvB18	LvB Sonata 18	12428	13	0,5533
LvB19	LvB Sonata 19	3362	10	0,5580
LvB20	LvB Sonata 20	2937	15	0,7518
LvB21	LvB Sonata 21	14682	18	0,5752
LvB22	LvB Sonata 22	5802	18	0,6013
LvB23	LvB Sonata 23	15575	17	0,5526
LvB24	LvB Sonata 24	4619	18	0,6820
LvB25	LvB Sonata 25	5930	15	0,6260
LvB26	LvB Sonata 26	7416	17	0,6207
LvB27	LvB Sonata 27	6643	18	0,6294
LvB28	LvB Sonata 28	8467	15	0,5040
LvB29	LvB Sonata 29	21559	26	0,6232
LvB30	LvB Sonata 30	8713	19	0,6423
LvB31	LvB Sonata 31	8075	21	0,6537
LvB32	LvB Sonata 32	13468	23	0,6259
		$\overline{F(H)} = 0.6170 \pm 0.0570$		

Table A7
H and *F*(*H*) for Liszt

ID	Text	N	H	F(H)
Liszt01	Liszt - Concert Etude No.3 Un Sospiro	1495	19	0,6863
Liszt02	Liszt - Paganini Etude No.3 La Campanella	4278	17	0,6173
Liszt03	Liszt - Transzendental Etudes Eroica	3003	24	0,5744
Liszt04	Liszt - Transzendental Etudes Feux Follets	4420	23	0,6860
Liszt05	Liszt - Venezia e Napoli: 1. Gondoliera	2899	14	0,6609
Liszt06	Liszt - Venezia e Napoli: 2. Canzone	2211	13	0,6260
Liszt07	Liszt - Venezia e Napoli: 3. Tarantella	7731	14	0,4315
Liszt08	Liszt - Sonata h mol	15921	27	0,5892
Liszt09	Liszt - Hungarian Dance 1	2790	18	0,6441
Liszt10	Liszt - Hungarian Dance 5	1785	11	0,5322

Liszt11	Liszt - Hungarian Dance 6	3065	18	0,6803
Liszt12	Liszt - Hungarian Rhapsody	941	14	0,6865
Liszt13	Liszt - Liebestraume No. 3	1891	23	0,7002
Liszt14	Liszt - Valse Oubliee No.1	1861	16	0,6083
Liszt15	Liszt - Valse Oubliee No.2	4147	18	0,6294
		$\overline{F(H)} = 0.6231 \pm 0.0692$		

Table A8
H and *F*(*H*) for Skrjabin

ID	Text	N	H	F(H)
Skr01	Skrjabin Prelude op. 27 – No 1	355	10	0,4704
Skr02	Skrjabin Prelude op. 27 – No 2	222	9	0,6081
Skr03	Skrjabin Prelude op. 31 – 1	651	13	0,5453
Skr04	Skrjabin Prelude op. 31 – 4	155	9	0,5032
Skr05	Skrjabin Prelude op. 33 – 2	195	12	0,6308
Skr06	Skrjabin Prelude op. 33 – 3	212	9	0,5896
Skr07	Skrjabin Prelude op. 35 – 2	362	9	0,4586
Skr08	Skrjabin Prelude op. 37 – No 1	212	8	0,5189
Skr09	Skrjabin Prelude op. 37 – No 2	91	11	0,7363
Skr10	Skrjabin Prelude op. 48 – 2	224	10	0,4598
Skr11	Skrjabin Prelude op. 59	709	20	0,6897
Skr12	Skrjabin Prelude op. 67 – 1	338	9	0,5769
Skr13	Skrjabin Prelude op. 74 – 3	228	9	0,5921
Skr14	Skrjabin Piece op. 2, No 1	1150	16	0,7574
Skr15	Skrjabin Etude op. 8, No 4	747	9	0,5114
Skr16	Skrjabin Etude op. 8, No 5	1541	10	0,5120
Skr17	Skrjabin Etude op. 8, No 12	2301	11	0,5067
Skr18	Skrjabin Poem op. 32 – No 1	981	10	0,6575
Skr19	Skrjabin Počme tragique op.34	1001	11	0,6284
Skr20	Skrjabin Etude op. 42, No 4	787	10	0,5756
Skr21	Skrjabin Etude op. 42, No 5	3088	10	0,4828
Skr22	Skrjabin Sonate No 5, op. 53	7761	19	0,5588
Skr23	Skrjabin Sonate No 9, op. 68	4014	25	0,6682
Skr24	Skrjabin Poem op. 69 – No 2	539	11	0,6178
Skr25	Skrjabin Dance op. 73 – No 1 - Guirlandes	694	14	0,5130
Skr26	Skrjabin Dance op. 73 – No 2 – Flammes sombres	1051	13	0,6232
		$\overline{F(H)} = 0.5766 \pm 0.0813$		

Table A9

H and *F*(*H*) for Schoenberg

ID	Text	N	H	F(H)
Sch01	Verklaerte Nacht	15477	18	0.6144
Sch02	Mondestrunken	1197	16	0.6266
Sch03	Valse de Chopin	1146	16	0.6353
Sch04	Nacht (Passacaglia)	1108	23	0.6724
Sch05	Raub	661	14	0.6157
Sch06	Galgenlied	244	14	0.6116
Sch07	Die Kreuze	2042	15	0.6166
Sch08	Parodie	1329	20	0.6253
Sch09	O alter Duft	537	14	0.6089
Sch10	Piece for piano Op.33a	763	27	0.6619
Sch11	Six Little Piano Pieces Op.19	627	17	0.6061
		$\overline{F(H)} = 0.6268 \pm 0.0208$		

Table A10

H and *F*(*H*) for Stravinsky

ID	Text	N	H	F(H)
Str01	Adoration of the Earth	2490	19	0.7574
Str02	The Augurs of Spring	5139	12	0.6550
Str03	Ritual of Abduction	2794	16	0.6442
Str04	Spring Rounds	2805	34	0.8781
Str05	Ritual of the Rival Tribes	3267	36	0.8445
Str06	Procession of the Sage	738	23	0.6965
Str07	Dance of the Earth	1806	29	0.9147
Str08	The Sacrifice - Introduction	1994	23	0.7161
Str09	Mystic Circles	3085	15	0.6707
Str10	Glorification of the Chosen	1715	29	0.7767
Str11	Evocation of the Ancestors	1301	14	0.9101
Str12	Ritual Action of the Ancestors	2588	30	0.8876
Str13	Sacrificial Dance	5800	34	0.7445
Str14	The Firebird Suite (complete)	37659	28	0.7088
Str15	The Firebird Suite - Introduction	2919	36	0.9394
Str16	The Firebird's Dance	1015	19	0.9202
Str17	The Firebird Suite - Variations	3735	13	0.5971
Str18	The Princesses' Round Dance	1481	12	0.5692
Str19	The Infernal Dance	18912	22	0.6367
Str20	Berceuse	1877	21	0.7725
Str21	Finale	7733	23	0.7886
Str22	Symphony of Psalms 1	1878	24	0.7545

Str23	Symphony of Psalms 2	1494	20	0.6365
Str24	Symphony of Psalms 3	4214	27	0.714
		$\overline{F(H)} = 0.7556 \pm 0.1079$		

Table A11
H and *F*(*H*) for Shostakovich

ID	Text	N	H	F(H)
Sho01	Op.87 Prelude No.1 in C major	440	6	0.5545
Sho02	Op.87 Fugue No.1 in C major	172	6	0.7209
Sho03	Op.87 Prelude No.2 in A minor	323	8	0.6347
Sho04	Op.87 Fugue No.2 in A minor	247	10	0.6032
Sho05	Op.87 Prelude No.3 in G major	330	10	0.5606
Sho06	Op.87 Fugue No.3 in G major	407	9	0.7309
Sho07	Op.87 Prelude No.4 in E minor	429	7	0.5874
Sho08	Op.87 Fugue No.4 in E minor	453	6	0.6468
Sho09	Op.87 Prelude No.5 in D major	516	8	0.7267
Sho10	Op.87 Fugue No.5 in D major	312	7	0.6346
Sho11	Op.87 Prelude No.6 in B minor	321	14	0.6729
Sho12	Op.87 Fugue No.6 in B minor	367	13	0.6807
Sho13	Op.87 Prelude No.7 in A major	304	12	0.7928
Sho14	Op.87 Fugue No.7 in A major	483	15	0.8551
Sho16	Op.87 Fugue No.8 in F-sharp minor	390	13	0.7795
Sho17	Op.87 Prelude No.9 in E major	195	10	0.6821
Sho18	Op.87 Fugue No.9 in E major	573	8	0.6422
Sho19	Op.87 Prelude No.10 in C-sharp minor	430	20	0.6442
Sho20	Op.87 Fugue No.10 in C-sharp minor	404	7	0.5990
Sho21	Op.87 Prelude No.11 in B major	306	11	0.6830
Sho22	Op.87 Fugue No.11 in B major	611	8	0.6268
Sho23	Op.87 Prelude No.12 in G-sharp minor	476	11	0.7836
Sho24	Op.87 Fugue No.12 in G-sharp minor	480	7	0.5354
Sho25	Op.87 Prelude No.13 in F-sharp major	401	7	0.6509
Sho26	Op.87 Fugue No.13 in F-sharp major	250	8	0.7600
Sho27	Op.87 Prelude No.14 in E-flat minor	791	6	0.7155
Sho28	Op.87 Fugue No.14 in E-flat minor	394	6	0.5660
Sho29	Op.87 Prelude No.15 in D-flat major	1070	8	0.6654
Sho30	Op.87 Fugue No.15 in D-flat major	407	11	0.7101
Sho31	Op.87 Prelude No.16 in B-flat minor	354	7	0.6328
Sho32	Op.87 Fugue No.16 in B-flat minor	634	6	0.7319
Sho33	Op.87 Prelude No.17 in A-flat major	588	12	0.7823
Sho34	Op.87 Fugue No.17 in A-flat major	607	4	0.5634
Sho35	Op.87 Prelude No.18 in F minor	250	8	0.5840
Sho36	Op.87 Fugue No.18 in F minor	332	7	0.6145
Sho37	Op.87 Prelude No.19 in E-flat major	338	12	0.5740

Sho38	Op.87 Fugue No.19 in E-flat major	256	7	0.6367
Sho39	Op.87 Prelude No.20 in C minor	306	7	0.5523
Sho40	Op.87 Fugue No.20 in C minor	335	8	0.5910
Sho41	Op.87 Prelude No.21 in B-flat major	867	10	0.5686
Sho42	Op.87 Fugue No.21 in B-flat major	542	8	0.5923
Sho43	Op.87 Prelude No.22 in G minor	503	16	0.7435
Sho44	Op.87 Fugue No.22 in G minor	371	11	0.8032
Sho45	Op.87 Prelude No.23 in F major	378	10	0.7090
Sho46	Op.87 Fugue No.23 in F major	519	17	0.8266
Sho47	Op.87 Prelude No.24 in D minor	355	9	0.7042
Sho48	Op.87 Fugue No.24 in D minor	1015	10	0.5724
Sho49	Op.93 Symphony Nr.10 e-Moll - 1st Mov.	1056	11	0.8570
Sho50	Op.93 Symphony Nr.10 e-Moll - 2nd Mov.	790	10	0.7722
Sho51	Op.93 Symphony Nr.10 e-Moll - 3rd Mov.	259	9	0.8610
Sho52	Op.93 Symphony Nr.10 e-Moll - 4th Mov.	1194	11	0.6843
		$\overline{F(H)} = 0.6764 \pm 0.0886$		

Table A12
H and *F*(*H*) for Ligeti

ID	Text	N	H	F(H)
Lig01	Études pour piano 1 Désordre	3017	30	0,7676
Lig02	Étude 4: Fanfares	3142	26	0,6706
Lig03	Étude 5: Arc-en-ciel	3015	24	0,6577
		$\overline{F(H)} = 0.6986 \pm 0.0491$		

Glottometrics 16, 2008, 111-121

The relation between word length and sentence length: an intra-systemic perspective in the core data structure

Peter Grzybek[1], *Emmerich Kelih*[1], *Ernst Stadlober*[2]

Abstract. Word length and sentence length are systematically organized in texts and corpora. In recent attempts at the synergetic modeling of the relation between sentence length and word length, the importance of distinguishing intra-textual from inter-textual approaches has been emphasized. The present study focuses on the intra-textual level: with a particular emphasis on different text types, it is shown, under which conditions processes of inter-level self-regulation are operative, and when they fail to be efficient.

Keywords: Menzerath-Altmann law, word length, sentence length, interrelation, intra-systemic structure

1 Theoretical ruminations

The impact of word length (WL) and sentence length (SL) for purposes of text classification has been repeatedly documented (cf. Grzybek et al. 2005; Kelih et al. 2006; Antić at al. 2006). Extending these studies, Grzybek and Stadlober (2007) and Grzybek et al. (2007) have focused on the relationship between SL and WL, rather than on these two linguistic categories as separate phenomena in their own right.

In this context, the relevance of Arens' Law has been emphasized and submitted to some critical re-investigation. Arens' Law is an extension of the well-known Menzerath Law which, subsequent to its generalization and mathematical formulation by Altmann (1980) has also become known as Menzerath-Altmann Law. The latter aims at a theoretical description of the relation of linguistic units of different levels. Basically, it claims that the complexity or length of a particular (linguistic) component is a function of the length or complexity of the (linguistic) construct which it constitutes; it has been successfully applied in systems theoretical analyses other than linguistic as well (Altmann and Schwibbe 1989). The most general form of what is known today as the Menzerath-Altmann Law, has been suggested by Altmann (1980) in his seminal "Prolegomena to Menzerath's Law":

(1a) $$y = ax^{-b}e^{cx} \quad (a,b,c > 0),$$

with two special cases for c = 0, or b = 0, respectively, namely

(1b) $$y = ax^{-b}, \text{ and}$$

(1c) $$y = ae^{cx}$$

Only recently, Wimmer and Altmann (2005, 2006) have extended this approach in their "General derivation of some linguistic laws". It is based on the differential equation

[1] Institut für Slawistik, Universität Graz, Merangasse 70, A-8010 Graz, Austria; correspondence address: peter.grzybek@uni-graz.at

[2] Institut für Statistik, Technische Universität Graz, Steyrergasse 17/IV, A-8010 Graz, Austria.

(2) $$\frac{dy}{y} = \left(a_0 + \frac{a_1}{x} + \frac{a_2}{x^2} + \frac{a_3}{x^3} + \ldots \right) dx$$

resulting in the solution

(3) $$y = Ce^{a_0 x} x^{a_1} e^{-a_2/x - a_3/\left(2x^2\right) - \ldots}$$

With $a_3 = 0$ and $-a_2 = d$ in equation (3), they arrive at the addition of an optional factor $e^{d/x}$, thus obtaining six options with $d = 0$ for equations (1a-c), where $C = a$, $a_0 = c$, $a_1 = -b$:

(1d) $$y = ae^{d/x}$$

(1e) $$y = ax^{-b}e^{d/x}$$

(1f) $$y = ax^{-b}e^{cx}e^{d/x}$$

Anyway, equation (1a) is generally considered the most basic and commonly used “standard form” for linguistic purposes. With $b > 0$, it predicts a ***decrease*** in length or complexity of the linguistic components with an increase in length or complexity of the construct they constitute – in longer words, e.g., the syllables forming these words are predicted to be shorter than those forming shorter words.

These ruminations are of course of central importance for the relation between sentence length and word length. However, in his analyses of German literary prose texts, Arens (1965) observed an ***increase*** in sentence length going along with an increase of word length, thus obtaining a result seemingly contradictory to the expectations.

By way of a solution, Altmann (1983: 31), in his attempt to interpret these results in Menzerathian terms, pointed out that the Menzerath-Altmann Law as described above is likely to hold true only when one is concerned with the direct constituents of a given construct. In case of the SL-WL relation, however, an intermediate level may be assumed to come into play – such as, e.g., phrases or clauses as the direct constituents of the sentence. As a consequence, words might be seen as direct constituents of clauses or phrases, but only as indirect constituents of sentences. Therefore, in its direct form, the Menzerath-Altmann Law might fail to grasp the SL-WL relation. In this case, an increase in SL should indeed result in an increase of WL, and it should be expected to be of the Menzerathian non-linear form: with y symbolizing word length, z symbolizing phrase (or clause) length, and x symbolizing sentence length, we were thus concerned with two simultaneous relations, $y = az^{-b}e^{cz}$ and $z = a'x^{-b'}e^{c'x}$. Inserting the latter equation into the first, one obtains y as a function

(4) $$y = a''x^{b''} exp\left(-c''x + a'''x^{-b'}e^{c'x}\right)$$

.

However, in studies of direct relations between linguistic units of different levels, the “standard case” of the Menzerath-Altmann Law, i.e. $z = a'x^{-b'}$ and $y = az^{-b}$, has been sufficient. Following this line, one thus obtains $y = a''x^{b''}$, for the indirect relation between sentence length and word length, corresponding to equation (1b). From this perspective, Arens Law is a special case of the Menzerath-Altmann Law: the only difference between direct and indirect relations thus is that, in case of directly neighboring units, the exponents *-b* and *–b’* are negative (due to the predicted decline), whereas in case of indirectly related units, with intermediate levels, $b'' = (-b) \cdot (-b')$ will become positive. However, this would hold true only in case of deterministic relations, and in no case for averages.

2. Empirical findings

Despite the importance of Arens' Law for linguistic and non-linguistic analyses in the field of general systems theory, only few studies have explicitly referred to it. A possible reason for this might be that there seems to be only poor evidence in support of the theoretical assumptions, as recently pointed out by Grzybek and Stadlober (2007). Thus, Arens conducted no statistics at all to test his assumptions, and Altmann (1983) tested the goodness of the non-linear Menzerathian model with F-tests which are very likely to result in misleading interpretations in case of large sample sizes, typical for linguistic data. In fact, as a re-analysis of Arens' data shows, fitting equation (1b) results in a rather poor fit ($R^2 = 0.70$), which is far from being convincing, and consequently sheds doubt on the adequacy of the Menzerathian interpretation.

In an attempt to find some explanation for this poor result by way of a systematic re-analysis of the sentence length – word length problem, Grzybek and Stadlober (2007) and Grzybek et al. (2007) have pointed out a number of possible problems coming into play:

1. *Data Sparsity*. Both the Menzerath-Altmann Law and Arens Law as a special case of it are what one might term "laws of averages", consequently demanding for a sufficient amount of data points for averages to be reliable. However, due to the large variance of *SL*, an insufficient amount of observations may be available for quite a number of data points of the independent variable. As a consequence, the frequency of observations for each data point has to be guaranteed to prevent random results. In fact, by pooling data into specific classes (as is usual in *SL* analyses), Grzybek, Kelih & Stadlober (2007) arrived at values of $0.93 \leq R^2 \leq 0.97$, differences depending on the pooling procedure chosen.
2. *Data homogeneity and text typology*. Given the fact that Arens' original data were based on German literary texts only, the question arises in how far the conclusions made can be generalized and transferred to other text types, as well. Thus, enlarging Arens' text data base by adding literary and scientific prose texts, previously analyzed by Fucks (1955), Grzybek and Stadlober (2007) found the R^2 value to become significantly worse.
3. *Intra-textual vs. Inter-textual approach.* The initial idea of the Menzerath-Altmann Law has been to describe the relation between the constituting components of a given construct and this construct; consequently, the Menzerath-Altmann Law originally was designed in terms of an intra-textual law, relevant for the internal structure of a given text sample. Arens' data, however, are of a different kind, implying inter-textual relations, based on the calculation of sentence length and word length means (m_{SL}, m_{WL}) for each individual text sample, thus resulting in a vector of arithmetic means. Therefore, in their systematic analysis of 199 Russian texts, Grzybek et al. (2007) obeyed the need to clearly keep the intra-textual and inter-textual perspectives apart. Concentrating on the inter-textual level only, they conducted separate analyses for six different text types, on the one hand, and corpus analyses for the combined data. As a result, they found only very weak evidence on support of Arens Law on an inter-textual level: for the individual text types, the results were between $0.02 \leq R^2 \leq 0.26$, for the complete corpus they obtained $R^2 = 0.49$. This result coincides with previous observations that obviously, average word length is relatively stable within a given text type – and it is a matter of fact that there can be no variation of word length depending on varying sentence length, if the dependent variable word length displays only poor variation.

3. The intra-textual perspective

The present study concentrates on an analysis of the sentence length – word length relation from an intra-textual perspective. Table 1 represents the text data with relevant characteristics.

Table 1
Text corpus and sub-corpora

Text type	Author	Number of texts	Words		Sentences	
			abs.	rel.	abs.	rel.
Drama	A.P. Čechov	44	67 430	0.28	11125	0.47
Private letters	(various)	120	56 751	0.23	4178	0.18
Literary prose	L.N. Tolstoj	69	74 708	0.31	5 680	0.24
Comments	(various)	60	43 263	0.18	2 556	0.11
Corpus		293	242 152	1.00	23 539	1.00

As can easily be seen, the proportions of sentences and words clearly differ for the different text types; consequently, m_{SL} and m_{WL} significantly differ across text types, as has well been documented elsewhere. With this in mind, it will be interesting to analyze the sentence length – word length relation separately for each text type; yet, by way of a first approximation, Fig. 1 offers an overview for the whole corpus.

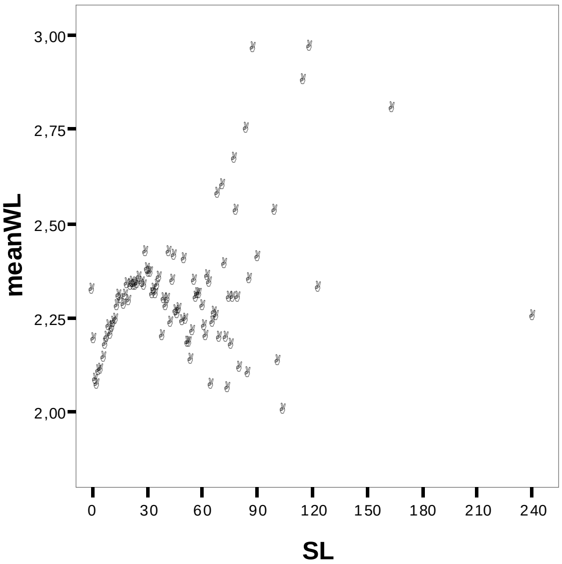
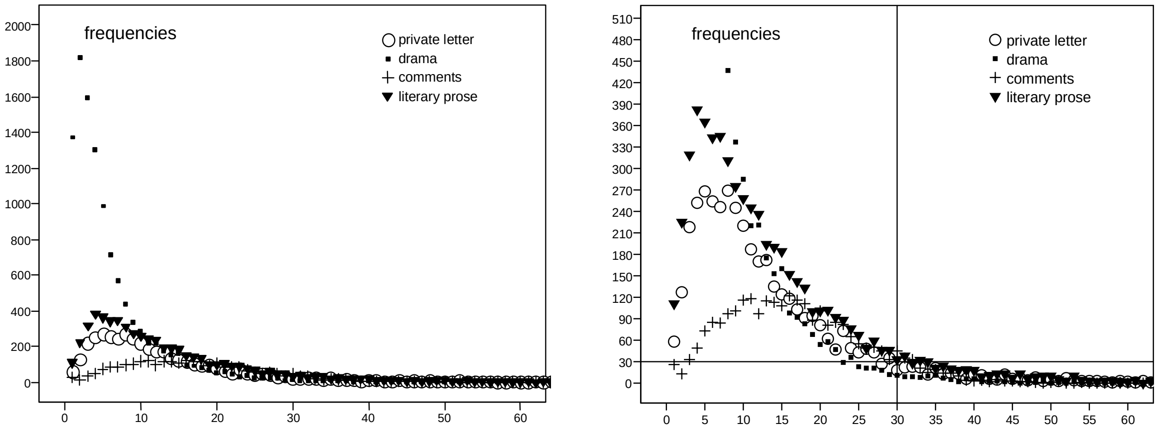


Fig. 1. Word length vs. sentence length: Total Corpus

An inspection of Figure 1 immediately shows the extreme variance of m_{WL} for long sentences with $SL \gtrsim 30$. It is well possible that we are concerned here with linguistic reasons, possibly coming into play; this possibility will be discussed in more detail below. Yet, another possibility must be checked first, which is of statistical rather than linguistic nature. In principle, this reason would concern short sentences as well as long sentences, but particularly long sentences, with $SL \gtrsim 30$, are likely to occur relatively rarely. So for a given SL, m_{WL} may be based on a few observations, only, causing a greater variation of m_{WL}. The increase of word length variation for sentences (and the resulting “loss” of a possibly existing systematic tendency in the WL-SL relation) might therefore be motivated by merely statistical reasons.

Figures 2 display the frequencies of particular SL occurrences; indeed, it can easily be seen, that for all four text types, it is just around $SL \approx 30$ that the frequency of sentences with the given length decreases to less than 30 observations per class.



(a) Complete sentence length distribution (b) Detailed insight ($f_{SL} < 450$);

Fig. 2. Sentence length distributions for four text-types

As a consequence, we exclude all occurrences with rare data points for m_{WL}, by way of an empirical rule of thumb, thus including only data where m_{WL} is based on 30 observations or more ($f_{SL} \geq 30$); we apply no pooling procedures for the remaining data with less observations, since the type of pooling may be an additional factor influencing the overall result.

Under these circumstances, guaranteeing the postulated minimum of 30 occurrences, a closer look at Figure 3 allows for a more detailed analysis of the overall trend of the core data structure.

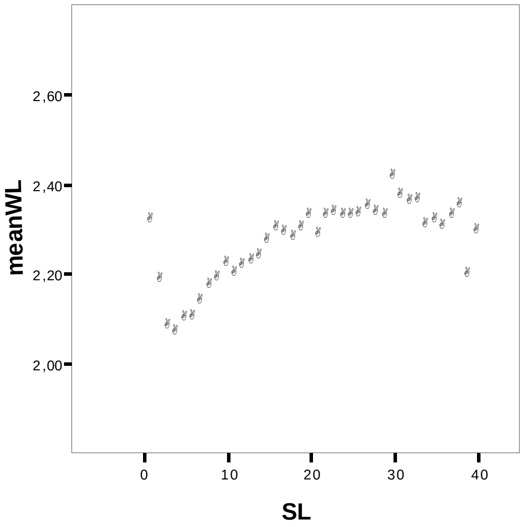


Fig. 3. Word length vs. sentence length: Restricted conditions

Generally speaking, one can now indeed observe a major tendency for longer sentences to be composed of longer words, as predicted by the hypothesis. Yet, there are two important deviations from this overall trend, characterized by two critical points:

1. In very short sentences, the *SL*–*WL* length relation seems to be differently organized as

compared to the bulk of data points: short sentences show a clear decline to a local minimum (in case of the complete corpus, at $SL = 4$), which shall be termed *lower critical point* (*LCP*), here. It goes without saying that, for other data material (particularly from other languages), this initial decreasing trend need not be obligatory, and the *LCP* may well be $LCP \neq 4$. Anyway, it seems reasonable to assume that we are concerned here with linguistic reasons for this tendency: obviously, very short sentences have no hyposyntactic sub-division and, as a consequence, do not ask for any inter-level Menzerathian control. A detailed analysis of these short sentences must be left for a separate analysis, particularly including *WL* frequency distributions for each of the *SL* classes. In future, it would be desirable to have a common model for all (short and long) sentences; yet, by way of a first approach, we exclude these short sentences from the present study, in order to better concentrate on the bulk of the material, hoping to grasp the general tendency by this procedure.

2. Whereas for sentences with $4 < SL < 30$, there seems indeed to be a general tendency for longer sentences to be composed of longer words (as predicted by the hypothesis), there seems to be an *upper critical point* (*UCP*) for longer sentences with $SL \gtrsim 30$. This point is clearly marked by the definite increase of word length variation for these sentences (cf. Figure 3), even after exclusion of occurrences with $f_{\text{SL}} < 30$. A detailed analysis of this phenomenon goes beyond the scope of this paper; yet, two alternatives lend themselves to interpretation:
   a. it is possible, that a minimum of $f_{\text{SL}} = 30$ is not sufficient for an average to become stable enough; in this case, we are still concerned with a statistical interpretation of the observed phenomenon,
   b. it does not seem unlikely that we are concerned her with a (psycho)-linguistically, rather than statistically motivated upper critical point (*UCP*): taking into account human processing limits, Miller’s magical rule of 7 ± 2 (and Yngve’s linguistic interpretation of it) might well hold true for clause length, and serve as a limitation of the length of clauses or phrases, and, as a consequence, of sentences. Thus, given an average clause length of 5-6 words per clause, the upper limit of information processing on this level might be reached, as a result “de-activating” the Menzerathian control.

In any case, in order to concentrate on the bulk of the material, thus hoping to obtain reliable information on the core of the data structure and grasp its overall tendency, we introduce three empirically motivated restrictions in this study :

(a) $f_{SL} > 30$,
(b) $m_{WL} > LCP$, and
(c) $SL < 30$.

With these empirical restrictions, it will now be interesting to look not only at the total corpus, but also at the specifics of each of the four different text types. Some basic characteristics of the relevant core data structures are represented in Table 2:

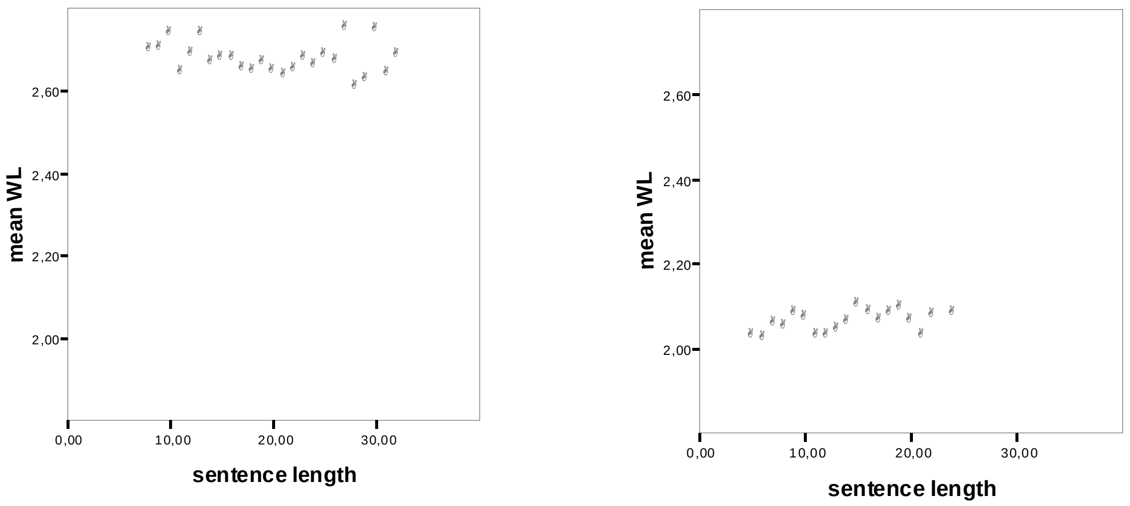
1. The Lower Critical Point (*LCP*) is defined as the minimal m_{WL} point subsequent to which there is a monotonous increase;
2. the Upper Critical Point (*UCP*) is determined by the empirical restriction of $f_{\text{SL}} > 30$;
3. the proportion (in %) of sentences is the percentage of data material representing the core data structure in the interval [*LCP*, *UCP*].

Table 2
Text corpus and sub-corpora

Text type	*LCP*	*UCP*	%
Drama	4	22	95.64
Private letters	3	27	90.45
Comments	7	32	94.20
Literary prose	2	31	93.30
Total	4	40	97.90

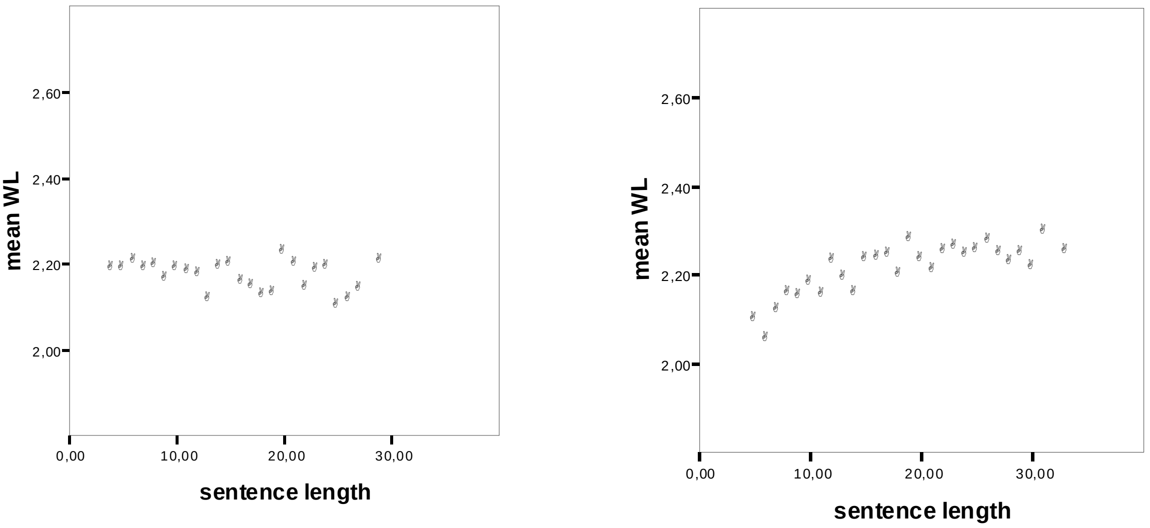
As can be seen, both *LCP* and *UPC* differ for the individual text types: Whereas the *LCP* ranges from $2 \leq LCP \leq 4$, the *UCP* ranges from $22 \leq UCP \leq 32$ (in case of the total corpus even reaching $UCP = 40$).

The core data structures for the four text types are represented in Figures 4a-d. With regard to the *SL*–*WL* relation, the results are extremely surprising: quite opposite to expectation, there is almost no increase in m_{WL} for three of the four text types: rather, in case of the comments, private letters, and dramatic texts, m_{WL} is almost stable across different *SL* classes. Only for the literary texts, we obtain a convincing fit of $R^2 = 0.88$ for the non-linear Menzerathian model, with parameter values $a = 1.93$ and $b = 0.05$.



(a) Comments

(b) Drama



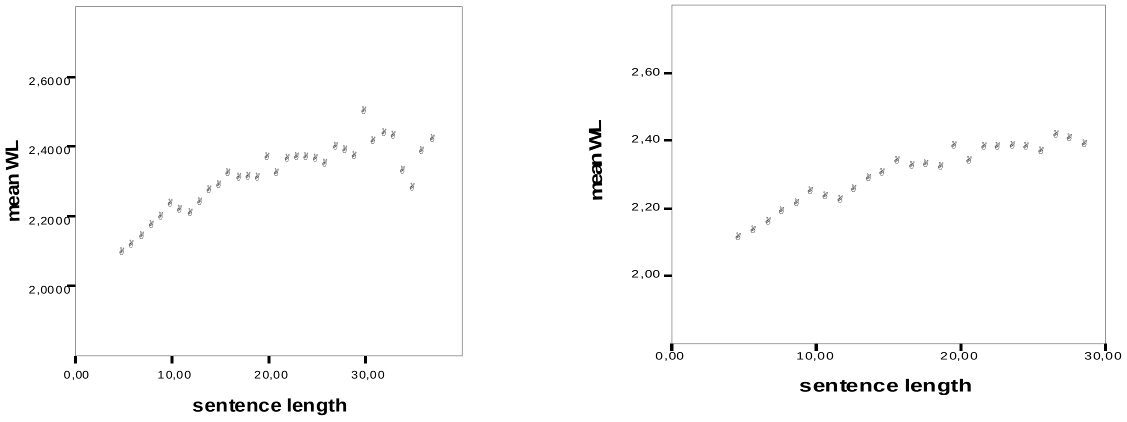
(c) Private letters

(d) Literary prose

Fig. 4. Word length vs. sentence length

In an attempt to find an interpretation of these findings, it seems reasonable to exclude any possible influence of the literary prose texts on the overall corpus. The easiest way to do this, is an additional analysis of a corpus consisting of all comments, private letters, and drama texts, but without the literary texts. This corpus of 167,444 words and 17,859 sentences contains 69.15% of the words and 75.87% of the sentences of the total corpus; its critical points are *LCP* = 4 (with m_{WL} = 2.07 at this point), and *UCP* = 37 (m_{WL} = 2.42).

Figure 5 (a) shows the *SL*-*WL* tendency for this particular corpus; again, like in the total corpus, there is a fluctuation of m_{WL} for *SL* > 30. Again discarding all sentences with *SL* > 30, however, the corpus of comments, drama texts and private letters, with R^2 = 0.87 (*a* = 1.88, *b* = 0.07), shows an almost identical tendency as the literary texts.



(a) Total corpus without literary texts

(b) Total corpus: Core data structure

Fig. 5. Word length vs. sentence length

Given these results for the partial corpus (without literary texts), let us now compare them to those for the total corpus. Again, concentrating on the core data structure of the total corpus, excluding short sentences, and cutting off the data at *SL* = 30, yields a convincing fit of the Menzerathian non-linear curve: with a determination coefficient of $R^2 = 0.96$. Interestingly enough, the parameter values $a = 1.88$ and $b = 0.07$ are almost identical with the one obtained for the corpus without the literary prose texts. Figure 5(b) illustrates the overall result.

We thus obtain a number of interesting results:

1. For three of the four analyzed text types (drama, comment, letters), no Menzerathian tendency can be confirmed; only for literary texts, a Menzerathian tendency (Arens Law) can be confirmed;
2. For a partial corpus consisting of these three text types, a Menzerathian (Arens Law) tendency can be confirmed; the same holds true for the total corpus of all four text types.

In attempting to find an answer to the alleged contradictions, it seems reasonable to pay attention to the obviously important factor of data heterogeneity: in case of the partial and total corpora, we are concerned with different text types, each characterized by specific WL and SL characteristics: thus, for the drama texts, we have $m_{\mathrm{WL}} = 2.04$ and $m_{\mathrm{SL}} = 6.06$, for the letters $m_{\mathrm{WL}} = 2.19$ and $m_{\mathrm{SL}} = 3.58$, and for the comments $m_{\mathrm{WL}} = 2.67$ and $m_{\mathrm{SL}} = 16.93$. Only taken together, merged into one common corpus of heterogeneous data, the Menzerathian tendency (Arens Law) appears to be at work. Let us term this phenomenon, which must be subjected to more empirical testing in future, *external textual heterogeneity*.

If this interpretation holds true, a similar hypothesis might be brought forth with regard to the literary texts, as well: in this case, it might well be possible that we are concerned with some kind of *internal textual heterogeneity*, literary texts characteristically being composed of dialogues, descriptive passages, narrative sequences, etc., all of which may well be shaped by different *WL* and *SL* characteristics.

Seen from this point, the emergence of the Menzerathian tendency (Arens Law) would have to be interpreted in terms of an index heterogeneity, at least as far as the external perspective is concerned – as to the internal perspective, only some rudimentary insights could be gained in this paper, and more systematic study is necessary in future.

4. Conclusion

The present study offers some important conclusions as to an interpretation of the *SL-WL* relation along the Altmann-Menzerathian line. Obviously, it seems to work, in case a number of pre-conditions are fulfilled:

- *Minimal sentence length.* For very short sentences *(SL* < 4), the Menzerathian tendency does not seem to play a crucial role; it seems reasonable that this circumstance is motivated by linguistic reasons only, sentences of this length not being subdivided into linguistic sub-units; it goes without saying that the resulting LCP may well be different (or even non-existing) for other languages.
- *Maximal sentence length.* For very long sentences *(SL* > 30), the Menzerathian tendency does not seem to play a crucial role; (psycho)linguistic reasons might be responsible for this circumstance, sentence regulation being at work only as long as a sub-division into sub-units of sentences can be cognitively controlled.
- *Minimal frequency.* Here, we are concerned with a predominantly statistical constraint: if there are not enough *(SL)* data points as a basis of m_{WL}, variance is too large to result in some kind of general tendency; accidentally, the *UCP* of *SL* around 30 coincides in

most of the data analyzed in this paper with the one explained by maximal SL.

- *Textual heterogeneity*. The Menzerathian principle seems to be of relevance for the *SL-WL* relation only in case sufficient linguistic heterogeneity is guaranteed: as long as the data material to be analyzed consists of homogenous texts (i.e., from a specific text type), *WL* seems to be regulated and, in fact, dominated, by this text type's specific *WL* organization. Only in case data from different text types are combined, the necessary textual heterogeneity is provided for the Menzerathian principle to come into play. It may well be that a literary text as a whole is characterized by this intrinsic heterogeneity, being composed of (homogeneous) text elements such as dialogues, descriptive and narrative sequences, auctorial comments, etc. This might be an explanation why the Menzerathian tendency can be observed in literary texts. It would be particularly interesting to see whether within literary texts, such homogeneous text elements can be isolated which, taken in isolation, do not display any Menzerathian tendencies, yet would, combined into a (heterogeneous) whole. A systematic test of this hypothesis must be left for future research, however.

In addition to these detailed problems, another open question is, if and how very short sentences on the one hand, and long sentences, on the other, can be integrated into one complex model. In other words: It will be an important future task to study (a) in how far the extreme ranges of word and sentence length are characterized by a diverging tendency as compared to the core data structure, and (b) if, both possibly heterogeneous tendencies can yet be incorporated into one overall model. Furthermore, the question of intrinsic heterogeneity, obviously characterizing literary texts, must be subjected to detailed analyses.

References

Altmann, G. (1980). Prolegomena to Menzerath's Law. *Glottometrika 2, 1–10.* Bochum: Brockmeyer,

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

Altmann, G., Schwibbe, M.H. (1989). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen.* Hildesheim: Olms.

Antić, G., Stadlober, E., Grzybek, P., and Kelih, E. (2006). Word length and frequency distributions. In: M. Spiliopoulou et al. (Eds.), *From data and information analysis to knowledge engineering: 310-318.* Berlin: Springer.

Arens, H. (1965). *Verborgene Ordnung. Die Beziehungen zwischen Satzlänge und Wortlänge in deutscher Erzählprosa vom Barock bis heute.* Düsseldorf: Pädagogischer Verlag Schwann.

Fucks, W. (1955): Unterschied des Prosastils von Dichtern und Schriftstellern. Ein Beispiel mathematischer Stilanalyse. *Sprachforum 1, 234–241.*

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

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

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

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

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

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

Wimmer, G.; Altmann, G. (2005). Towards a unified derivation of some linguistic laws. In: Grzybek, P. (ed.), *Contributions to the science of text and language. Word length studies and related issues: 329-337*. Dordrecht, NL: Springer.

Glottometrics 16, 2008, 122-131

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.

XXXII. Helmut Meier (1897-1973)

Vollständiger Name: Wilhelm Erich Helmut Meier. Geb. 20.12.1897 (Broitzem; der Ort wurde am 1.3.1974 nach Braunschweig eingemeindet), gest. 30.7.1973 (Braunschweig). 1912-1919 Lehrerseminar in Braunschweig, 1917 - Anfang 1919 Soldat. Ab 1919 Hilfslehrer (Braunschweig, Linnenkamp, Helmstedt, unterbrochen von Beurlaubungen), ab 1925 Lehrer in Braunschweig. 1939-1945 Militärdienst; danach wieder Lehrer in Braunschweig, zwischendurch 1946-1948 Dozent an der Kant-Hochschule für Lehrerbildung in Braunschweig (Didaktik, Mathematik); 1949 im Entnazifizierungsverfahren als „entlastet“ beurteilt. Tätigkeit als Lehrer bis zur Pensionierung 1963; auf eigenen Wunsch weitere Arbeit als Lehrer (im Angestelltenverhältnis). Für seine wissenschaftlichen Leistungen wurde ihm am 19.12.1964 der Ehrendoktor der Universität Hamburg (Dr. phil. h.c.) verliehen.

Meiers Bedeutung für die Quantitative Linguistik und die Sprachstatistik beruht darauf, dass er neben seiner Berufstätigkeit als Lehrer und Dozent jahrzehntelang in Anknüpfung an Kaeding (1897) sprachstatistische Erhebungen zum Deutschen durchgeführt hat (Aichele 2005, 18), die vor allem in seinem Hauptwerk (Meier 1964, ²1967) publiziert sind. Es handelt sich dabei um die bisher materialreichste und vielseitigste Zusammenstellung von Daten zum Deutschen. Seine Arbeit wurde nach dem 2. Weltkrieg von der DFG gefördert und kam auch der internationalen Hochschule für Pädagogik in Wiesbaden zugute. Außerdem führte Meier nach eigener Auskunft sprachstatistische Arbeiten im Auftrag der Universitätskliniken für Hals-, Nasen-, Ohrenkrankheiten in Freiburg und Marburg durch (Meier 1967: VIII, 301, 310) und war an der Entwicklung von Sprachtests für Zwecke der Audiometrie beteiligt.

Man findet in Meiers Buch (1964/67) u.a. Statistiken über die Häufigkeit von Satz- und Wortlängen, über die Häufigkeit, mit der Buchstaben und Laute im Deutschen verwendet werden, über die Häufigkeit grammatischer Erscheinungen (z.B.: wie oft erscheinen Substantive mit oder ohne bestimmte Begleitwörter wie Adjektive, Artikel oder Pronomen oder: wie häufig werden die verschiedenen Kasus verwendet?) oder auch zu der Frage, welche Themenbereiche in einem Wörterbuch wie stark vertreten sind. Diese Andeutungen mögen genügen.

Viele statistische Daten hat Meier neu erarbeitet; andere beruhen aber auch "nur" auf Umarbeitungen bereits vorhandenen Materials, darunter vor allem das von Kaeding (1897) (Meier 1967: 1). So hat Meier in der zweiten Auflage seines Hauptwerkes (Meier 1967) eine alphabetische Liste der Wörter aufgeführt, die bei Kaeding mindestens mit der Häufigkeit 10 aufgeführt sind, gefolgt von einer Rangliste der 7994 Wörter, die mindestens eine Häufigkeit von 51 aufweisen, sowie Listen der 2240 häufigsten Begriffswörter, geordnet nach Wortarten, die mindestens die Häufigkeit 500 bei Kaeding erreichen. Diese Daten geben also lediglich den Stand des Deutschen gegen Ende des 19. Jahrhunderts wieder. (Bleibt zu erwähnen, dass Meier wesentlich umfangreichere Ranglisten der Wörter bzw. Begriffswörter erarbeitet hat, aber nur deren Spitze im angegebenen Werk veröffentlichte.)

Etliche der von Meier dargebotenen Daten ließen sich für Zwecke der Quantitativen Linguistik verwenden, wobei sich erwies, dass seine Ergebnisse sich entsprechend bekannten Gesetzeshypothesen verhalten. Seine 100000-Laute-Zählung (Meier 1967: 250f.) bot Anlass, die Rangordnung der Laute und Phoneme daraufhin zu untersuchen, welchen Gesetzen sie unterliegen. Es konnte gezeigt werden, dass Laute und Phoneme sowohl in Poesie als auch in Prosa ebenso wie die aus beiden Bereichen zusammengefassten Daten Altmanns Modell (Altmann 1993) für beliebige Rangordnungen folgen (Best 2004/05). An die 20000 Sätze eines Mischtextes (Meier 1967: 186) konnte die Hyperpascal-Verteilung mit sehr gutem Ergebnis angepasst werden (Best 2002: 15).

Meiers sprachstatistische Arbeit wurde nicht nur zustimmend aufgenommen: So kritisiert Müller (1971: 123) ebenso wie Herdan (1966) an Meiers Hauptwerk, dass „die statistische Methodenlehre dem Autor gänzlich fremd ist.“ Herdan wirft ihm vor, dass er neue Entwicklungen ab 1955 nicht mehr zur Kenntnis genommen hat; manche neuere Arbeit habe er zwar genannt, aber offensichtlich sich nicht angeeignet.

Literatur

(Anmerkung: die heimatkundlichen und pädagogischen Publikationen Meiers werden hier nicht angeführt.)

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

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

Best, Karl-Heinz (2002). Satzlängen im Deutschen: Verteilungen, Mittelwerte, Sprachwandel. *GBS Göttinger Beiträge zur Sprachwissenschaft 7, 7-31.*

Best, Karl-Heinz (2004/05). Laut- und Phonemhäufigkeiten im Deutschen. *Göttinger Beiträge zur Sprachwissenschaft 10/ 11, 21-32.*

Gremminger, Günther (1951). Zu den Zählforschungen am deutschen Sprachschatz. *Muttersprache Jg. 1951, 173-174.*

Kaeding, Friedrich Wilhelm [Hrsg.] (1897). *Häufigkeitswörterbuch der deutschen Sprache. Festgestellt durch einen Arbeitsausschuß der deutschen Stenographie-Systeme. Erster Teil: Wort- und Silbenzählungen. Zweiter Teil: Buchstabenzählungen.* Steglitz bei Berlin: Selbstverlag des Herausgebers. Teilabdruck: *Grundlagenstudien aus Kybernetik und Geisteswissenschaften. Bd. 4/ 1963.*

Meier, Helmut (1935). Die Sprachstatistik im Dienste der Rechtschreibreform. *Nachrichtenblatt des Volksbundes für vereinfachte Rechtschreibung, Jg. 1935, 34f.*

Meier, Helmut (1951). Dreißig Jahre Zählforschungen am deutschen Sprachschatz. *Muttersprache Jg. 1951, 6-14.*

Meier, Helmut (1952). Erkenntnis und Verpflichtung. Zum künftigen Ausbau der Häufigkeitszählungen. *Muttersprache Jg. 1952, 250-252.*

Meier, Helmut (1952). Die tausend häufigsten Wortformen der deutschen Sprache. Sprachstatistik, Aufgabe und Verpflichtung. *Muttersprache Jg. 1952, 88-94.*

Meier, Helmut (1964). *Deutsche Sprachstatistik.* Hildesheim: Olms.

Rezensionen: Brock, Bernhard (1966), *Wirkendes Wort XVI, 209-211*; Daniels, Karlheinz (1965), *Muttersprache Jg. 1965, 273-280*; Eggers, Hans (1965), *Germanistik VI, 562*; Frank, Helmar (1964), *Grundlagenstudien aus Kybernetik und Geisteswissenschaft Heft 5, 126-127*; Hammerberg, Björn (1966), *Moderna Språk LX, 440-441*; Herdan, Gustav (1966), *Phonetica XIV, 111-114*; Marchl, Herbert (1965), *Beiträge zur Sprachkunde und*

Informations-verarbeitung Heft 7, 73-75; Moskovič, V.A. (1966), *Voprosy Jazykosnanija No. 6, 133-137.*

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

Müller, Werner (1971). Gedanken zu H. Meiers „Deutscher Sprachstatistik". *Muttersprache 81, 121-125.*

Die biographischen Informationen beruhen auf Auskünften und Dokumenten des Stadtarchivs der Stadt Braunschweig sowie des Niedersächsischen Landesarchivs – Staatsarchivs Wolfenbüttel, für deren Unterstützung hier gedankt sei.

Karl-Heinz Best

XXXIII. Adolf Busemann (1887-1967)

Vollständig: Adolf Hermann Heinrich Busemann, Dr. phil. (Göttingen), Dr. med. h. c. (Marburg), korrespondierendes Mitglied der Deutschen Vereinigung für Jugendpsychiatrie. Geb. 15.5.1887 (Emden), gest. 5.6.1967 (Marburg). Gymnasium Northeim 1897-1906, Studium der Psychologie in Göttingen 1906-1910 (Religion, Hebräisch, phil. Propädeutik). 1910 Prüfung für das höhere Lehramt. Lehrtätigkeit in Essen, Frankenberg und Bederkesa. Dazwischen 1917/18 Kriegsteilnahme im Landsturm. 1922-1925 zunächst als Oberlehrer, dann als Seminarstudienrat in Einbeck, 1925 wegen Auflösung des Lehrerseminars in den einstweiligen Ruhestand versetzt. 1924 Promotion in Göttingen, ab 1925 Greifswald, 1926 Habilitation in Greifswald. Bis 1928 Privatdozent (Medizinische Fakultät), danach beurlaubt, um an anderen Institutionen zu unterrichten (Prof. an den Pädagogischen Akademien Rostock, Breslau und Kiel). Ab 1932 wieder Privatdozent in Greifswald; danach „auf Grund des Gesetzes zur Wiederherstellung des Berufsbeamtentums 1934 in das Amt eines Volksschullehrers versetzt" (Mail v. Barbara Peters, Archiv der Universität Greifswald, 18.6.2007). WS 1934/35 und SS 1935 beurlaubt. 1937 aus gesundheitlichen Gründen in den dauernden Ruhestand versetzt. Übersiedelung nach Marburg; Personalgutachter beim Heer, 1942 aus dem aktiven Wehrdienst entlassen. 1943-1945 Psychologe am Hirnverletztenlazarett in Marburg. WS 1946/46 – SS 1948 Lehrveranstaltungen in Psychologie an der Universität Marburg. Bis 1954 Unterricht im Rahmen der „Lehrgänge zur Ausbildung von Sonderschullehrern" in Marburg. (Quellen: s. „Über Busemann".)

Das in der Quantitativen Linguistik am meisten beachtete Thema Busemanns ist der Aktionsquotient (Busemann 1925; 1948: 116, 139), der die Zahl der Verben und der Adjektive eines Textes zueinander in Relation setzt; dabei gilt ein Text, bei dem die Verben überwiegen, als aktiv und ein Text mit mehr Adjektiven als Verben als deskriptiv. Busemanns Daten beruhen hauptsächlich auf Niederschriften, das sind „provozierte schriftliche Selbstdarstellungen von rund 4000 Kindern und Jugendlichen" (Busemann 1926: 28); hinzu kommen einige spontansprachliche Quellen. Eine Diskussion der Probleme des Aktionsquotienten und Vorschläge für eine Verbesserung findet sich in Altmann (1978; 1988: 18ff.), eine weitere Behandlung in Altmann & Altmann (2005: 86-88). Tuldava (2005: 371, 376f.) reiht Busemanns Arbeit in die Forschungsgeschichte ein und geht auf die Arbeiten der Nachfolger ein.

Man findet bei Busemann aber noch weitere Themen, die für die Quantitative Linguistik von Bedeutung sind. So betrachtet er in (Busemann 1925: 90ff.) die Entwicklung der Wortlänge, indem er die relativen Anteile der Ein-, Zwei-, Drei- und Mehrsilber an der Sprachproduktion von Kindern bis zum Alter von 20 Jahren untersucht. Meist bleiben die Beobachtungen getrennt für die einzelnen Wortlängen. Aber für einen Datensatz zu den 10- bis 15-jährigen Jugendlichen nennt er Werte für die Entwicklung der durchschnittlichen

Wortlänge. Seine Angaben beruhen auf 163 Niederschriften einer Mädchenschule in Oldesloe mit 16000 Wörtern; die festgestellten Schwankungen sind bei nur sechs Messwerten zu groß. Ergänzt man die Messungen jedoch um eines realistischen Wert für Erwachsene, lässt sich das logistische Modell

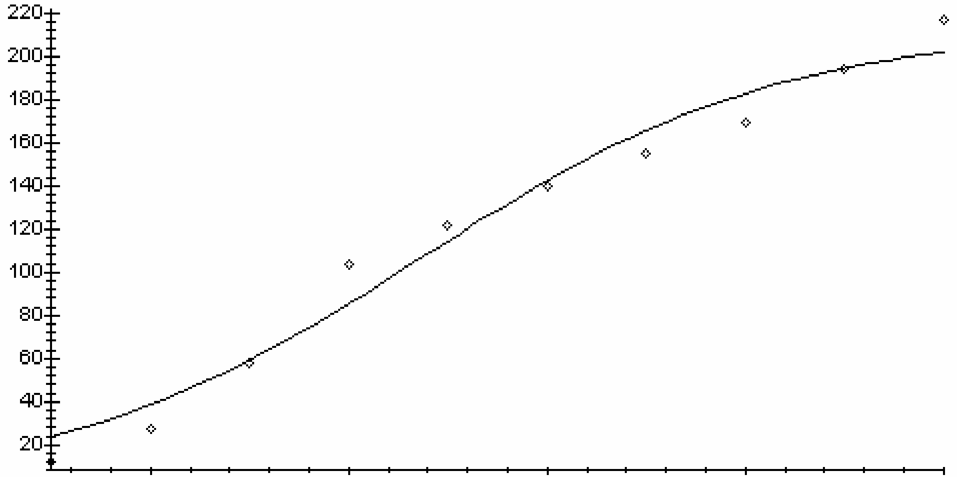
$$p_t = \frac{c}{1 + a\,e^{-bt}} \qquad (1)$$

(Altmann 1983: 61) mit sehr gutem Erfolg anpassen (Best 2006: 43).

Ein weiteres Thema ist Busemanns Untersuchung zur Entwicklung des Adjektiv-Wortschatzes, den Jugendliche benutzen, um sich selbst zu charakterisieren (Busemann 1926, 1948: 150). Diese Untersuchung des Ausbaus eines Wortschatzsegments lässt sich ebenfalls sehr gut mit dem Wachstumsgesetz (1) modellieren, wie die folgende Tabelle 1 und die Graphik dazu zeigen. Dabei sind *a*, *b* und *c* die Parameter des Modells; *D* ist der Determinationskoeffizient, der mit $D \geq 0.80$ eine gute Übereinstimmung des Modells mit den Beobachtungen bestätigt:

Tabelle 1
Zuwachs neuer, vorher nicht benutzter Adjektive zur Selbstcharakterisierung von Jugendlichen (n. Busemann 1948: 150)

t	Alter in Jahren	neue Adjektive	Adjektive kumulativ	Adjektive berechnet
1	8	12	12	23.94
2	9	15	27	38.45
3	10	31	58	59.00
4	11	46	104	85.17
5	12	18	122	114.32
6	13	18	140	142.33
7	14	15	155	165.66
8	15	14	169	183.00
9	16	25	194	194.53
10	17	23	217	201.91
$a = 13.7263$ $b = 0.5536$ $c = 212.8358$ $D = 0.97$				



Graphik: Zuwachs neuer, vorher nicht benutzter Adjektive zur Selbstcharakterisierung von Jugendlichen

Busemann hat noch eine Reihe weiterer Themen statistisch bearbeitet; so kommt vor allem zur Sprache, welche Themen die Kinder und Jugendlichen in den Niederschriften ansprechen und wie sich das mit dem Alter ändert (Busemann 1926). Auch in der *Milieukunde* findet man statistische Charakterisierungen, wobei Sprachliches aber nur am Rande auftaucht (Busemann 1927: 182). Sprachliche Daten werden dabei nicht immer so dargeboten, dass man sie für eine Modellierung der Erwerbsprozesse gut nutzen könnte. In *Krisenjahre* etwa stellt die Beobachtungen der Scupins zum Wortschatzzuwachs ihres Sohnes mit dem arithmetischen Mittel für Vierteljahreszeiträume in ganzzahligen Werten zusammen (Busemann 1953: 38); der tatsächliche Wortschatz ist daher nur näherungsweise zu bestimmen.

Busemanns Werk ist von statistischen Erhebungen zur Entwicklung von Kindern und Jugendlichen geprägt, wobei speziell sprachliche Themen außer ganz zu Anfang nicht dominieren. Charakteristisch für Busemanns spätere Einstellung sind aber resignative Bemerkungen. So wendet er sich gegen die Experimentelle Psychologie, die „nunmehr behauptet, die ganze Psychologie zu sein, und der nicht exklusiv experimentell bzw. statistisch arbeitenden den Namen der Psychologie abstreitet und das, was so ausgeschlossen wird, der Philosophie zuweist“ und fährt fort: „Eine hervorragende Sachverständige der Psychologischen Statistik hatte wohl guten Grund, in ihrem bekannten Lehrbuch zu betonen, daß die Statistik das Denken nicht überflüssig mache“ (Busemann 1967: 7).

Literatur

Altmann, Gabriel (1978). Zur Verwendung der Quotiente in der Textanalyse. *Glottometrika 1*, 91-106.

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

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

Altmann, Vivien, & Altmann, Gabriel (2005). *Erlkönig und Mathematik.* http://ubt.opus.hbz-nrw.de/volltexte/2005/325/

Best, Karl-Heinz (2006). Gesetzmäßigkeiten im Erstspracherwerb. *Glottometrics 12, 39-54.*

Busemann, Adolf (1925). *Die Sprache der Jugend als Ausdruck der Entwicklungsrhythmik. Sprachstatistische Untersuchungen.* Jena: Verlag von Gustav Fischer. Teildruck in: Helmers, Hermann (Hrsg.) (1969), *Zur Sprache des Kindes* (S. 1-59). Darmstadt: Wissenschaftliche Buchgesellschaft. (Erweiterung der Diss.)

Busemann, Adolf (1926). *Die Jugend im eigenen Urteil: eine Untersuchung zur Jugendkunde.* Langensalza: Beltz.

Busemann, Adolf (1927). *Pädagogische Milieukunde. I. Einführung in die Allgemeine Milieukunde und in die Pädagogische Milieutypologie.* Halle, Saale: Schroedel.

Busemann, Adolf (1948). *Stil und Charakter. Untersuchungen zur Psychologie der individuellen Redeform.* Meisenheim/ Glan: Westkulturverlag Anton Hain.

Busemann, Adolf (1953). *Krisenjahre im Ablauf der menschlichen Jugend.* Ratingen: Aloys Henn Verlag.

Busemann, Adolf (1967). *Weltanschauung in psychologischer Sicht. Ein Beitrag zur Lehre vom Menschen.* München/ Basel: Ernst Reinhardt Verlag.

Tuldava, Juhan (2005). Stylistics, author identification. In: Köhler, R., Altmann, G. & Piotrowski, R.G. (2005) (Hrsg.), *Quantitative Linguistik. Ein internationales Handbuch:. 368-387.* Berlin/ N.Y.: de Gruyter.

Welker, Meinrad (Bearb.) (2004). *Lexikon Greifswalder Hochschullehrer 1907-1932*. Bad Honnef: Bock. (= Buchholz, Werner (Hrsg.), *Lexikon Greifswalder Hochschullehrer*

1775-2006. Bd. 3.)

Anm.: Die Liste enthält nur die Arbeiten Busemanns, die hier zitiert wurden. Seine Bücher sind leicht zu bibliographieren und in vielen Bibliotheken vorhanden.

Über Busemann

Adolf Busemann 70 Jahre alt. *Bildung und Erziehung 10, 1957, H. 6, 370-371.*

van Dieken, Jan (1968). Professor Adolf Busemann. *Friesische Blätter, Folge 9, September 1968, 5. Jahrgang.*

Hetzer, Hildegard (1967). Zum 80. Geburtstag von Professor Dr. Adolf Busemann. Forscher und Lehrer im Dienst bedrohter und behinderter Kinder. *Lebenshilfe 6, H. 3, 113-114.*

Lexikon Greifswalder Hochschullehrer 1775-2006. Hrsg. v. Werner Buchholz. Bd. 3: *Lexikon Greifswalder Hochschullehrer 1907-1932*. Bandbearbeiter: Meinrad Welker. Bad Honnef: Bock 2004.

Für Informationen danke ich dem Archiv der Stadt Einbeck (Susanne Gerdes), dem Archiv der Universität Greifswald (Barbara Peters) und der Ostfriesischen Landschaft, Aurich (Cornelia Nath).

Karl-Heinz Best

XXXIV. Kaj Brynolf Lindgren (1822-2007)

Geb. 4.12.1922, Varkaus (Finnland). Studium der Germanistik, Nordistik und Psychologie ab 1944 in Helsinki und Zürich; Promotion 1953 in Helsinki, Habilitation 1957. Ab 1954 Lektor für Deutsch an der Wirtschaftshochschule Helsinki, ab 1962 Assoz. Prof. für Germanistik und 1964-1989 o. Prof. für Germanische Philologie am Germanistischen Institut der Universität Helsinki. (Nach: Kürschner 1994, 550.) Gest. 17.11.2007 in Helsinki.

Lindgren taucht in der Quantitativen Linguistik – soweit ich das übersehe – nur ein einziges Mal auf, in diesem Fall mit seiner Untersuchung zur e-Apokope im Deutschen, in der er am Rande auch auf die e-Epithese eingeht (Imsiepen 1983). Dies wird seiner Bedeutung nicht ganz gerecht, gehört er doch eindeutig in die Vorgeschichte des Sprachwandelgesetzes, das in der Quantitativen Linguistik seit Altmann (1983) auch unter dem Namen *Piotrowski-Gesetz* geläufig ist. Seine umfangreichen Datenerhebungen sowohl zur Apokope als auch zur Diphthtongierung in mittelhochdeutscher und anfangs der frühneuhochdeutschen Zeit gipfeln u.a. darin, dass er die Entwicklungen in Graphiken darstellt und dabei erkennt, dass sie einen prinzipiell gleichen Verlauf nehmen; diesen Verlauf stellt er dann in einer „idealisierte(n) Kurve“ (Lindgren 1961: 55) dar, die genau dem abnehmenden (Lindgren 1953: 185) oder zunehmenden Verlauf (Lindgren 1961: 56) des Piotrowski-Gesetzes für den vollständigen Sprachwandel entspricht. Er ist sich auch bewusst, dass er damit in der Linguistik auf ein Phänomen gestoßen ist, das in der Mathematik allgemein bekannt ist und dort eine Interpretation erfährt, die sich leicht auf sprachliche Entwicklungen übertragen lässt (Lindgren 1961: 57). Nachdem Lindgren so weit gekommen ist, fehlt nur noch der Versuch, solche Phänomene mathematisch zu modellieren und das dann entwickelte Modell an seinen eigenen Daten zu überprüfen. Einen Ansatz dazu, aber ohne Durchführung, findet man bei Hakkarainen (1983), der auf Lindgrens idealisierte Kurven hinweist und in (Hakkarainen 1983, 29, Fußnote 17) auf eine mathematische Herleitung des Modells unter Einbeziehung speziell sozialer Bedingungen durch Dodd (1953) verweist. Auch Hakkarainen bedient sich des

Modells mehr aus dem Wunsch heraus, seinen Vorstellungen, dass nämlich Diffusion und Sprachwandel prinzipiell gleich verlaufen, Anschaulichkeit zu verleihen; ein Test des Modells an Diffusionsdaten fehlt bei ihm jedoch ebenfalls.

Am Beispiel von Lindgrens Untersuchung zur neuhochdeutschen Diphthongierung soll gezeigt werden, dass das auch mit Erfolg geschehen kann. Dabei handelt es sich um die Ersetzung von [î] durch [ai], von [û] durch [au] und von [iu] durch [öu], Prozesse, die sich in der Zeit zwischen 1100 und 1500 abspielen. Die Daten dazu hat Lindgren durch Auszählen vieler Texte gewonnen; sie werden in einer umfangreichen Tabelle (Lindgren 1961: 15-17) in 50-Jahres-Schritten aufgeführt, getrennt nach Dialekten (Bairisch[1], Ostfränkisch, Schwäbisch, Böhmisch, Südfränkisch und Ostmitteldeutsch). Für die einzelnen 50-Jahres-Schritte werden für einen Dialekt Daten aus 1 – 6 Texten präsentiert. Speziell für das Bairische gibt Lindgren Ergebnisse aus 1 – 4 Texten an. Die Auswertung hat nun gezeigt, dass der Einfluss einzelner Texte mit vom Gesamtprozess stark abweichendem Sprachgebrauch zu sehr ins Gewicht fällt. Lindgren (1961: 24) bemerkt selbst, dass die beiden Texte der 2. Hälfte des 13. Jahrhunderts einen stark abweichenden Sprachgebrauch zeigen und vermutet daher für sie eine Herkunft von der Südseite der Alpen. Aus diesem Grund wurden alle Daten für ganze Jahrhunderte zusammengefasst. Es ergaben sich damit für das Bairische Tabellen mit Daten aus 5 Jahrhunderten, in allen anderen Fällen aus nur 3 Jahrhunderten oder noch weniger. Dies ist der Grund, weshalb hier nur bairische Daten berücksichtigt werden.

Das Modell, das hier zu prüfen ist, ist das Gesetz für den vollständigen Sprachwandel

$$p = \frac{100}{1+ae^{-bt}} \tag{1}$$

(Altmann 1983: 60). Die folgenden Tabellen enthalten Lindgrens Daten für das Bairische, auf Jahrhunderte umgerechnet, mit einer Anpassung von Modell (1). Die Anpassungen wurden mit der Software NLREG durchgeführt; die Ergebnisse zeigen, dass die Diphthongierung – wie viele andere Sprachwandel auch – gesetzmäßig verläuft. Die Graphiken zeigen, dass Lindgrens Annahme über den idealen Verlauf des Prozesses sich auch rechnerisch ergibt.

Die Ergebnisse:

Tabelle 1
Ausbreitung der Diphthongierung im Bairischen

Jh.	*t*	*[î] → [ai]*		*[û] → au]*		*[iu] → [öu]*	
		f_t	p_t	f_t	p_t	f_t	p_t
12.	1	0.00	0.84	0.15	3.48	0.07	1.45
13.	2	16.61	8.43	27.38	25.74	18.91	11.63
14.	3	45.20	50.14	75.27	76.91	48.27	54.01
15.	4	99.55	91.65	100	96.97	99.92	91.29
16.	5	100	99.17	99.88	99.68	100	98.94
		a = 1294.9015 *b* = 2.3906 *D* = 0.9819		*a* = 266.6004 *b* = 2.2630 *D* = 0.9968		*a* = 604.9947 *b* = 2.1887 *D* = 0.9805	

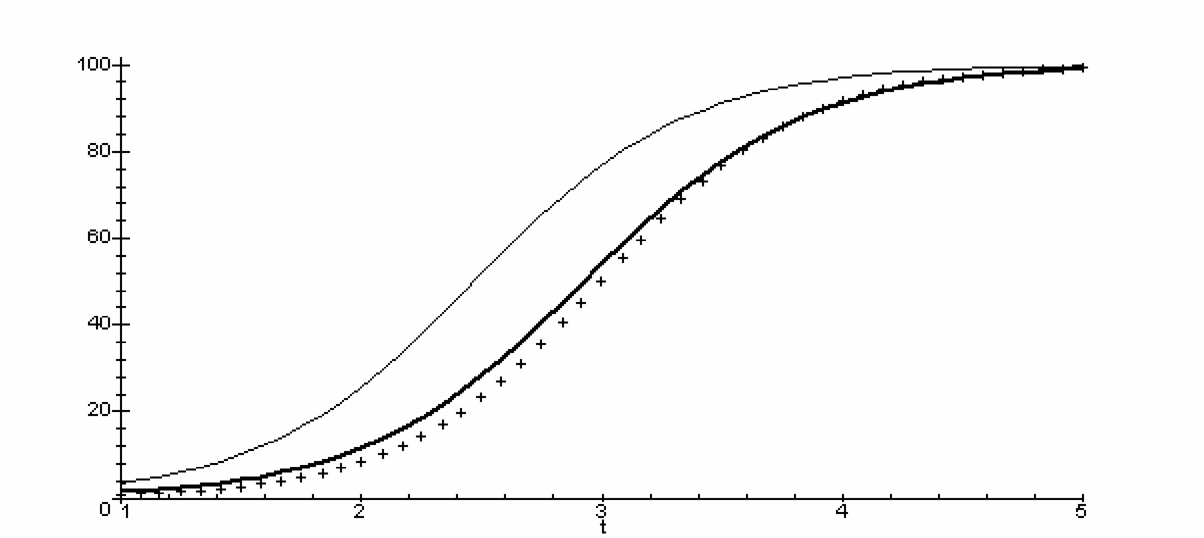
[1] Ich folge hier und auch bei den phonetischen Angaben Lindgrens Schreibweise.

Erläuterung zu den Tabellen:

f_t: beobachtete Vorkommen der betreffenden Einheit: relative Werte;
p_t: aufgrund des Modells (1) für den vollständigen Sprachwandel berechnete Vorkommen;
t: für die Berechnung festgelegter Zeitabschnitt, beginnend mit t = 1 für das 12. Jahrhundert;
a, b: Parameter;
D: Determinationskoeffizient.

Der Determinationskoeffizient soll das Testkriterium $D \geq 0.80$ erfüllen; er kann höchstens den Wert D = 1.00 erreichen und ist umso besser, je näher er an diese Grenze herankommt. Die drei in Tabelle 1 angegebenen Anpassungen des Modells erweisen sich damit als sehr gut.

Die folgende Graphik zeigt den berechneten Verlauf der drei Diphthongierungsprozesse im Vergleich zueinander:

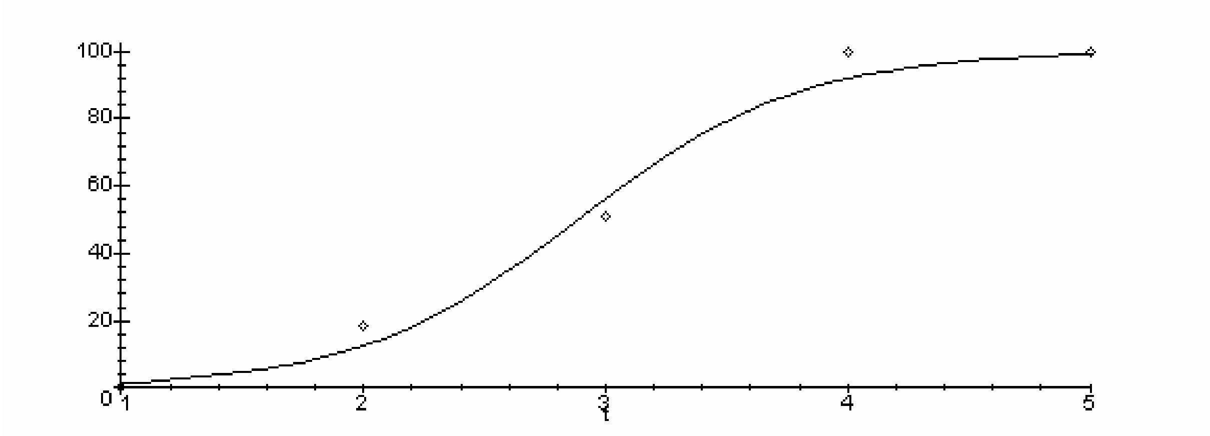


Graphik zu Tabelle 1: dünne, durchgezogene Linie: [û] → au]; starke durchgezogene Linie: [iu] → [öu]; Pluszeichenlinie: [î] → [ai]. Auf die beobachteten Werte wurde verzichtet, um die Graphik nicht zu überfrachten.

Die folgende Tabelle fasst alle drei Diphthongierungsprozesse zusammen:

Tabelle 2
Ausbreitung der Diphthongierung im Bairischen

Jh.	t	Gesamter Prozess	
		f_t	p_t
12.	1	0.04	1.60
13.	2	18.46	12.60
14.	3	51.26	56.16
15.	4	99.70	91.92
16.	5	99.98	99.02
a = 547.1003 b = 2.1841 D = 0.9854			



Graphik zu Tabelle 1: Gesamtprozess der Diphthongierung im Bairischen. Die Punkte stellen die beobachteten Werte dar.

Man kann also abschließend feststellen, dass sowohl die einzelnen Diphthongierungen als auch der Gesamtprozess sich gesetzmäßig verhalten.

Es müsste deutlich geworden sein, dass Lindgren zu den Philologen gehört, die der Sprachstatistik und der Quantitativen Linguistik dadurch einen Dienst erwiesen haben, dass sie aufwendige Datenarbeit durchführten. Er gehört auf jeden Fall zu den Vorläufern derjenigen, die das Sprachwandelgesetz herleiteten; er war sich bewusst, dass seine Forschungen im Ergebnis mit dem logistischen Modell der Mathematik übereinstimmen und brachte dies früher und deutlicher als manche andere zum Ausdruck:

„Es handelt sich um eine sog. regelmäßige Summenkurve, die in der Statistik eine grosse Rolle spielt. Sie ergibt sich prinzipiell in einem Fall folgender Art: Innerhalb einer Menge von Einzelgegenständen tritt an einem Punkt eine Änderung ein. Die Gegenstände stehen in Berührung mit ihren jeweiligen Nachbarn, so dass die an einem Einzelgegenstand vollzogene Änderung dieselbe Änderung an den benachbarten hervorruft. Diese wirken wiederum auf ihre Nachbarn ein usw., bis alle Gegenstände erfasst sind. Zuerst greift die Änderung nur langsam um sich, da sie von einem einzigen Punkt ausstrahlt, dann immer schneller, da immer mehr bewirkende Punkte vorhanden sind. Nachdem mehr als die Hälfte erfasst ist, wird die Entwicklung langsamer, weil jeweils auf einige Nachbarn schon früher von anderer Seite aus Einfluss wirkte, bis schließlich nur einige entlegene Punkte übrig bleiben, die ganz spät erfasst werden.

Wenn wir diese allgemeinen Überlegungen auf die Sprachentwicklung anwenden, kommen wir zu folgendem Bild: In einem begrenzten, einheitlichen Sprachraum tritt die Tendenz zu einer Änderung der Aussprache auf. Sie führt zunächst dazu, dass ein Wort oder eine eng zusammengehörende Wortgruppe in der neuen Weise ausgesprochen wird. Diese Wörter sind durch Analogie mit anderen verbunden, und das verursacht, dass dieselbe Änderung auch in diesen eintritt. Ausgehend von diesen breitet sich die neue Lautung weiter aus, bis schließlich alle Wörter mit den nämlichen phonetischen Bedingungen erfasst sind“ (Lindgren 1961: 57).

Die Beschreibung des Sprachwandelvorgangs findet sich in ähnlicher Weise bereits in (Lindgren 1953: 181/185), verbunden mit dem „Idealbild“ des Verlaufs. Von Hakkarainen (1983) erfährt man, dass in der Soziologie bereits vor längerer Zeit ein solches Modell mathematisch hergeleitet und überprüft wurde.

Der nächste, noch ausstehende und im Grunde abschließende Schritt, der Versuch einer mathematischen Modellierung sprachlicher Entwicklungsprozesse und einer Überprüfung des Modells unter Berücksichtigung speziell linguistischer Bedingungen, blieb Piotrovskaja &

Piotrovskij (1974) und in Weiterführung dieses Ansatzes Altmann (1983) sowie Altmann u.a. (1983) vorbehalten. Besonders Altmann (1983) mit seinen drei Modellen für unterschiedliche Sprachwandeltypen war der Auslöser für eine Vielzahl entsprechender, erfolgreicher Untersuchungen. Man darf jetzt konstatieren, dass innersprachlicher Wandel, Entlehnungen, Spracherwerb und Veränderungen im Sprachverhalten immer wieder diesen Sprachgesetzen folgen. Am Anfang dieser Entwicklung hin zum Sprachwandelgesetz stand allem Anschein nach Lindgren mit seinen Untersuchungen zum Deutschen – lange Zeit wenig bekannt für diese Pioniertat.

Literatur

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

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

Dodd, Stuart C. (1953). Testing message diffusion in controlled experiments: charting the distance and time factors in the interactance hypothesis. *American Sociological Review 18, 410-416.*

Hakkarainen, Heikki J. (1983). Sprachliche Veränderungen als Diffusion von Innovationen. *Neuphilologische Mitteilungen 84, 25-35.*

Imsiepen, Ulrike (1983). Die e-Epithese bei starken Verben im Deutschen. In: Best, K.-H., Kohlhase, J. (Hrsg.), *Exakte Sprachwandelforschung* (S. 119-141). Göttingen: edition herodot.

Lindgren, Kaj B. (1953). *Die Apokope des mhd. –e in seinen verschiedenen Funktionen.* Helsinki (= Suomalainen tiedeakatemian toimituksia/ Annales academiae scientiarum fennicae; Sarja/ Ser. B, Nide/ Tom. 78,2)

Lindgren, Kaj B. (1961). *Die Ausbreitung der nhd. Diphthongierung bis 1500.* Helsinki (= Suomalainen tiedeakatemian toimituksia/ Annales academiae scientiarum fennicae; Sarja/ Ser. B, Nide/ Tom. 123,2)

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

Zu Lindgren

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

Verzeichnis der wissenschaftlichen Schriften von K.B. Lindgren. In: *Neuphilologische Mitteilungen 84/ 1983: 1-7.*

Software

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

Für biographische Informationen danke ich Jouni Heikkinen, Germanistisches Institut der Universität Helsinki.

Karl-Heinz Best, Göttingen