Glottometrics 8
2005

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. 8 (2004), Lüdenscheid: RAM-Verlag, 2004. 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 8 (2004) **ISSN 2625-8226**

Contents

History of quantitative linguistics

Glottometrics 8, 2004, 1-11

Are the effects of vowel repetition influenced by frequencies? A corpus study on CVCVCV-structured nouns with and without vowel repetition

Katsuo Tamaoka, Hiroshima University, Japan[1]
Shogo Makioka, Osaka Women's University, Japan
Tadao Murata, Kyushu Institute of Technology, Japan

Abstract. A psychological study by Tamaoka and Murata (2001) suggested that CVCVCV-structured nonwords (e.g., /kohomo/) with the same vowel repeated showed longer naming latencies than the same-structured nonwords without vowel repetition (e.g., /kohami/). One of the possible factors for prolonging vowel repetition could be the frequency of vowel repetition in Japanese. Thus, the present study calculated token frequencies for nouns with the same vowel repeated within a CVCVCV phonological structure, based on the Japanese lexical corpus (287,792,797 words) of Amano and Kondo (2000). The results showed that vowels were repeated among Japanese nouns with a CVCVCV string more frequently than the random possibility of 4 percent. In addition, nouns with the same vowels in the first and second positions (i.e., V_1 and V_2 in the $CV_1CV_2CV_3$) showed significantly higher occurrences than the random chance of 20 percent, whereas nouns with the same vowels in the second and third positions appeared at the random level (i.e., V_2 and V_3). Since it is expected that higher frequency enhances speed and accuracy in naming, phonological structures with the same vowel repeated can be expected to be more quickly and accurately named. Conflicting results between the present corpus study and the experimental study by Tamaoka and Murata (2001) excluded the possibility of the frequency of vowel repetition affecting the speed and accuracy of phonological processing.

Keywords: vowel repetition, phonological structure, corpus study, Japanese nouns

1. Introduction

A study by Tamaoka and Murata (2001) suggested that CVCVCV-structured nonwords with the same vowel repeated like /kohomo/ showed longer naming latencies than the same-structured nonwords without vowel repetition like /kohami/. The explanation proposed for this is the 'whack-a-mole' phenomenon. The vowel in the first CV mora (C referring to 'consonant' and V 'vowel') continues to have a high activation level even when the following CV morae are activated. When the same vowel is repeated throughout the CV morae, all the CV morae will be simultaneously excited to reach the activation level. To avoid confusing the continuous order of the CV mora string, sequential morae must be inhibited so as not to be activated to the same degree as the previous CV mora. This pattern of excitation and inhibition results in the decreased speed of phonological processing for nonwords. As for naming nonwords with no repeated vow-

[1] Address correspondence to: Katsuo Tamaoka, International Student Center, Hiroshima University, 1-1, 1-Chome, Higashihiroshima, Japan 739-8524. E-mail: ktamaoka@hiroshima-u.ac.jp

els, the processing of sequential order of the CV mora string is not affected by other morae. Thus, nonwords with varying vowels are named more quickly than nonwords which repeat vowels and, concomitantly, fewer errors are observed among nonwords with non-repeated vowels.

While the 'whack-a-mole' phenomenon was a psychological explanation, some linguists provide a different explanation from a phonological perspective. The Obligatory Contour Principle (OCP) refers to a linguistic constraint on similar or same phonological features from being repeated (e.g., Fukazawa, 2000; Ito & Mester, 1986; Kubozono, 1999; Kubozono & Ota, 1998; Leben, 1973; McCarthy 1986; Yip, 1988). Kubozono and Ota (1998) suggested the possibility that vowel dissimilation in Japanese may be a result of the OCP. For example, the two Japanese morphemes /nana/ ('seven') and /ka/ ('day') combine to form the compound word /nanoka/ ('the seventh day') instead of /nanaka/, which would seem to be the likely combination. This process of vowel dissimilation occurs so as to avoid vowel repetition of /a/ in sequence within the three mora CVCVCV word structure. Thus, it would be expected that naming visually-presented Japanese words and nonwords which violate the OCP (i.e., same vowel repetition in a series of CV strings) would result in slower processing speeds and higher error rates. Yet, the linguistic explanation of OCP does not conflict with the psychological explanation of the 'whack-a-mole' phenomenon.

Despite these psychological and linguistic explanations, 'frequency of vowel repetition' in Japanese could be a possible factor for prolonging vowel repetition. Therefore, the present study calculated type and token frequencies for nouns with the same vowel repeated within a CVCVCV phonological structure, based on the Japanese lexical corpus (287,792,797 words) of Amano and Kondo (2000). Study 1 examined the same vowels in all three consecutive V positions and Study 2 in two consecutive V positions. Both Studies 1 and 2 were used to examine whether the rate of vowel repetition in CVCVCV-structured Japanese nouns appears to be significantly greater or lesser than the random chance rate of occurrence.

2. Conditions for calculating word frequency

Three conditions were established for calculating word frequency. First, the random chance rate of occurrence was established using only CVCVCV-structured words. Under this condition, the possibility of a 3-mora CVCVCV-structured word, which contains the same vowel in three consecutive positions, is calculated as 4 percent (with the 5 different Japanese vowels of /a/, /e/, /i/, /o/ and /u/ in three positions calculated as $1/5^3 \times 5 = 1/5^2$). In the same way, the possibility of these words with the same vowel in chosen two positions in their CVCVCV structure is 20 percent (with the 5 different vowels in two chosen positions calculated as $1/(5\times5) \times 5 = 1/5$, the third vowel may be arbitrary).

Secondly, only nouns were selected from the word corpus of Amano and Kondo (2000), which still served as sufficient data for the purpose of the present investigation. As Japanese verbs and adjectives have grammatical inflections, they were not included in the present corpus study. For example, the Japanese verb /ugoku/, meaning 'to move', inflects as in /ugoka(nai)/, /ugoki(masu)/, /ugoku(toki)/, /ugoke(ba)/ and /ugokoR/. It uses all the five Japanese vowels of /a/, /i/, /u/, /e/ and /o/ in its grammatical inflections.

Thirdly, both frequencies of *type* and *token* were used for the purpose of this study. In type frequency, a single word is only counted once, regardless of how many times it is repeated in the written text. On the other hand, in token frequency (i.e., accumulative word frequency) a word is counted every time it appears in the text. Since a rare word (e.g., /guNzoR/ meaning 'ultramarine') has the same type frequency of 1 as a frequently used word (e.g., /daigaku/ meaning

'university'), the present study considered token frequency as a better indicator of word frequency.

3. STUDY #1: Frequency of CVCVCV-structured nouns with the same vowels occurring in three consecutive vowel positions

Study 1 examined the occurrence of the same vowels in three consecutive V positions within CVCVCV-structured nouns. In this study, all nouns with a $CV_1CV_2CV_3$ structure had to share the same vowel in all V_1, V_2 and V_3 positions.

3.1. Lexical Corpus and Selection Procedure

As a result of the word frequency index created by Amano and Kondo (2000) from their study on accumulative word frequency (i.e., token frequency), a very large lexical corpus of 341,771 words was established from newspapers containing 287,792,797 words of accumulate frequency. All these words were taken from the *Asahi Newspaper* printed from 1985 to 1998. This is one of the largest and the most up-to-date word corpora created from calculating frequency of words in Japanese written texts. The present study utilized this corpus to investigate nouns with vowel repetition.

The programming language of MacJPerl 5.15r4J for Macintosh was used to run a calculation procedure. For Study 1, only nouns with a CVCVCV phonological structure were used. Thus, three mora nouns with VCVCV, CVVCV, CVCVV, VCVV, or VVV strings were not included. Therefore, the Japanese special long vowel /R/[1], where the same vowel appears twice without having a consonant between them, was not counted. In the same way, double vowels such as /ai/, /oi/, /ue/ were also excluded from the count. Other special sounds such as the nasal /N/ and the geminate /Q/ were also excluded as well as contrasted sounds such as /kya/, /myo/, /pyo/.

3.2. Results

Study 1 used two types of frequencies: *type* and *token* frequency. Type frequency only counts a word once and then is calculated by a simple addition of each word's frequency of '1' (ΣN_i), even though a word may appear repeatedly in a printed text. Token frequency, on the other hand, is calculated by taking the number of times each word appears in the text and adding all these frequencies together (ΣW_f). The present study took the .01 level of significance to reject the statistical null hypothesis, since the word corpus used was very large. These two frequency indexes of type and token frequency are reported separately as listed below.

3.2.1. Type Frequency

As shown in Table 1, among the five Japanese vowels, the vowel /a/ was the most frequently repeated in three consecutive positions, found in 674 nouns or 61.50 percent of the total 1,096 nouns (both general and proper nouns) with vowel repetition. The second most frequently used vowel was /o/ found in 204 different nouns or 18.61 percent of the total nouns counted.

Ranking third was the vowel /i/, repeated in 158 different nouns or 14.42 percent of the total nouns counted. The vowel /u/ came in forth, repeated in 57 nouns or 5.20 percent of all nouns counted. The least repeated vowel was /e/, repeated in only 3 different nouns or 0.27 percent of all nouns counted. Kubozono (1999) explained that the three vowels of /a/, /i/ and /u/ are most frequently found within the various languages of the world. It is therefore reasonable to expect that these vowels will be repeated more often in a single noun in Japanese than the vowel /e/. Since the vowel /o/ is ranked at the top of the 'sound hierarchy' (Murata, 1984, 1990; Tamaoka & Murata, 1999), /o/ tended to be repeated more than the vowel /e/, although both these vowels have points of articulation within the middle of the vowel space. This tendency was observed in both general and proper nouns in the same way.

Table 1
Same Vowels in Three Consecutive V Positions

	Word Frequency ($\sum N_i$)			Accumulative Word Frequency ($\sum W_f$)		
Vowels	General Nouns	Proper Nouns	Total	General Nouns	Proper Nouns	Total
/a/	197	477	674	166,823	113,351	280,174
/i/	118	40	158	36,620	1,664	38,284
/u/	25	32	57	4,710	5,364	10,074
/e/	2	1	3	136	11	147
/o/	51	153	204	183,961	13,832	197,793
Total	393	703 *	1,096 *	392,250 *	134,222 *	526,472 *
Grand Total	8,142	10,348	18,490	4,664,720	1,090,679	5,755,399
Ratio	4.83%	6.79% H	5.93% H	8.41% H	12.31% H	9.15% H

Note 1 : * $p<.01$.
Note 2 : The random possibility of 3-mora CVCVCV nouns which have the same vowels in three consecutive V positions is 4 percent.
Note 3 : The sign H refers to the frequency of nouns with vowel repetition which is significantly higher than random chance (4.00%).
Note 4 : The grand total of 18,490 refers to the total number of nouns with a CVCVCV phonological structure out of the 341,771 nouns taken from the word corpus of Amano and Kondo (2000).
Note 5 : The grand total of 5,755,399 refers to the total accumulative word frequency for the 18,490 nouns with a CVCVCV phonological structure.

The frequency of nouns with vowel repetition appearing in Japanese written texts was examined using Chebysheff's inequality theorem (see Maezono, 2002; Matsubara, Nawata & Nakai, 1994; Suzuki, 1999). The calculation of probability is provided by:

$$P(|X - m| \leq k\sigma) > 1 - 1/k^2$$

where the m is a mean of a scattered variable X and sigma (σ) is a standard deviation[2]. Using this measurement, 393 nouns (4.83%) were found to contain the same repeated vowels from among 8,142 general nouns with a CVCVCV structure, which fell within the range of the random chance of occurrence (326 times or 4.00%) at the .01 level of significance. In contrast, 703 nouns (6.79%) were found to have the same vowels repeated among 10,348 proper nouns, which was sigificantly higher than the 4 percent chance of random occurrence. A total of 1,096 nouns (5.93%) out of 18,490 repeated the same vowel in the CVCVCV phonological strings. This noun frequency was significant at the probability level of 1 percent. In short, type frequency indicated that the assimilation constraint causing vowel repetition in nouns with a CVCVCV phonological

structure which was observed in the total number of proper nouns and the total number of both general and proper nouns together, but not in the total number of general nouns alone.

3.2.2. Token frequency

As discussed in the introduction of this paper, accumulative word frequency or token frequency is considered to be more accurate in indicating occurrence of words than type frequency. Similar to type frequency, token frequency also showed a similar pattern in terms of vowels repeated in three consecutive positions. As shown in Table 1, these vowels are listed in the Japanese vowel kana order of /a/, /i/, /u/, /e/ and /o/ in all the categories of general nouns, proper nouns and the total of both.

An interesting tendency observed in token frequency is a high accumulative frequency of occurrence of Japanese nouns with vowel repetition. Although type frequency of general nouns did not show a significantly high occurrence of words with vowel repetition, token frequency in the category of general nouns was calculated as 392,250 (8.41%) out of the total of 4,664,720. According to Chebysheff's inequality theorem, this figure of token frequency was significantly higher ($p < .01$) than the random chance of occurrence of 4 percent. Therefore, these nouns with vowel repetition in three consecutive positions are often seen in written texts. For the category of proper nouns, type frequency was 134,222 (12.31%) out of 1,090,679, which was significantly higher than the random chance of occurrence ($p < .01$). The grand total of 526,472 (9.15%) out of 5,755,399 also showed significantly high occurrence of nouns with same vowel repetition ($p < .01$).

3.3. Discussion

Study 1 examined the existent to which the same vowels were repeated three times within a CVCVCV phonological structure in a corpus of Japanese nouns. Although type frequency of general nouns did not show significantly high occurrences of same vowel repetition, token frequency indicated significantly high repetition. Since type frequency only counts a word once, regardless of how often it is used, the index of token frequencies reflects actual appearance in written Japanese texts. Therefore, Study 1 concluded that vowels were repeated among Japanese general and proper nouns with a phonological CVCVCV string far more frequently (i.e., 9.15%) than the random chance level of four percent.

4. STUDY #2: Frequency of CVCVCV-structured nouns with the same vowel occurring in two consecutive vowel positions

Study 2 investigated the frequency of nouns with the same vowel occurring in two consecutive V positions within a $CV_1CV_2CV_3$ string. In this case, a vowel could be repeated in either of two ways: (1) V_1 and V_2 or (2) V_2 and V_3. The same vowels found in V_1 and V_3 were not considered to be in consecutive positions, so they were only considered for their frequency of occurrence which was used simply for comparing the other two conditions of (1) and (2). Results would then be expected to display either a lesser or a greater degree of word frequency than the random chance rate of occurrence of 20 percent among CVCVCV-structured Japanese nouns.

4.1. Lexical Corpus and Procedure

Study 2 made use of the same lexical corpus as Study 1.

4.2. Results

The number of Japanese nouns having the same vowel in two consecutive positions is shown in Table 2. Words with the same vowel in three consecutive positions discussed in Study 1 are not included in the detail counts of each vowel in Table 2. For determining the significance level of one percent, counts of three-consecutive-vowel repetitions were included. According to the totals of both type and token frequencies, the type of vowel found in two consecutive positions among $CV_1CV_2CV_3$-structured nouns was similar to the vowel found in the order of three consecutive positions, with the most frequently-repeated vowel being /a/ and the least frequently-repeated vowel being /e/.

4.2.1. Type frequency

General nouns with the same vowel in the V_1 and V_2 positions were counted 1,664 times (20.44%) out of 8,142 general nouns. Likewise, proper nouns showed a similar count of 2,138 times (20.66%) out of 10,348 proper nouns. The total of both types of nouns together indicated a percentage of 20.56 or a count of 3,802 times out of 18,490 nouns. Once frequency counts of the same vowel in three consecutive positions were included, the figures become 2,057 (25.26%) for general nouns, 2,841 (27.45%) for proper nouns and 4,898 (26.49%) for both together. As indicated by the upper arrow in Table 2, all these figures were significantly higher than the random chance rate of occurrence of 20 percent based upon the calculation from Chebysheff's inequality theorem ($p < .01$). Therefore, it is concluded that nouns with the same vowels in the first and second V positions of a CVCVCV phonological string occur more frequently than random chance.

Type frequency of general nouns which have the same vowel in the V_2 and V_3 positions was 13.68 percent or 1,114 times out of 8,142 general nouns. Including the same vowel in three consecutive positions, type frequency became 18.51 percent or 1,507 times. According to the calculation based on Chebysheff's inequality theorem, this frequency of occurrence did not significantly differ from the 20 percent random chance rate. Similarly, same vowel occurrence in V_2 and V_3 positions among proper nouns was 18.86 percent or 1.952 times (12.07 percent or 1,249 times excluding the same vowel in thee consecutive positions) out of 10,348 proper nouns, which did not significantly differ from the random chance level. The total of both general and proper nouns occurring 18.71 percent or 3,459 times (12.78 percent or 2,363 times without the same vowel in three consecutive positions) out of 18,490 nouns did not show a significantly lower or higher frequency than the random chance level. Therefore, the same vowel in the second and third positions within CVCVCV strings occurs at the random chance.

Although the V_1 and V_3 positions of a $CV_1CV_2CV_3$ string were not considered to be consecutive, type frequency was counted as a basis for comparison. General nouns with the same vowels in the first and third positions of a CVCVCV string were counted 1,740 times or 21.37 percent (1,347 times or 16.54 percent without the same vowel in the three consecutive positions) out of 8,142 general nouns. This frequency of occurrence was the same as the random chance.

However, proper nouns with such vowel repetitions appeared 2,548 times or 24.62 percent (1,845 times or 17.83 percent without the same vowel in three consecutive positions) out of 10,348 proper nouns, which was significantly higher than the random chance rate of occurrence of 20 percent ($p < .01$). The total of both general and proper nouns with this type of vowel repetition showed a frequency of 23.19 percent or 4,288 times (17.26 percent or 3,192 times without the same vowel in three consecutive positions) out of 18,490 general and proper nouns. This figure was at the random chance rate of occurrence.

Table 2
Same Vowels in Two V Positions

	Word Frequency (ΣN_i)			Accumulative Word Frequency (ΣW_f)		
Vowels	General Nouns	Proper Nouns	Total	General Nouns	Proper Nouns	Total
(1) Same vowels in V_1 and V_2						
/a/	797	1,177	1,974	689,385	144,159	833,544
/i/	188	288	476	129,141	19,437	148,578
/u/	310	319	629	169,432	56,245	225,677
/e/	32	24	56	80,639	3,638	84,277
/o/	337	330	667	268,265	29,682	297,947
Total	1,664	2,138	3,802	1,336,862	253,161	1,590,023
Ratio	20.44%	20.66%	20.56%	28.66%	23.21%	27.63%
Including V_1=V_2=V_3	2,057 *	2,841 *	4,898 *	1,729,112 *	387,383 *	2,116,495 *
Ratio	25.26% H	27.45% H	26.49% H	37.07% H	35.52% H	36.77% H
Grand Total	8,142	10,348	18,490	4,664,720	1,090,679	5,755,399
(2) Same vowels in V_2 and V_3						
/a/	368	517	885	218,740	91,836	310,576
/i/	340	320	660	136,673	14,337	151,010
/u/	179	89	268	75,007	7,238	82,245
/e/	66	17	83	55,385	707	56,092
/o/	161	306	467	131,564	13,333	144,897
Total	1,114	1,249	2,363	617,369	127,451	744,820
Ratio	13.68%	12.07%	12.78%	13.23%	11.69%	12.94%
Including V_1=V_2=V_3	1,507	1,952	3,459	1,009,619	261,673	1,271,292
Ratio	18.51%	18.86%	18.71%	21.64%	23.99%	22.09%
Grand Total	8,142	10,348	18,490	4,664,720	1,090,679	5,755,399
(3) Same vowels in V_1 and V_3						
/a/	338	643	981	194,094	96,122	290,216
/i/	562	741	1,303	266,958	68,852	335,810
/u/	193	66	259	87,718	11,936	99,654
/e/	129	22	151	22,312	1,752	24,064
/o/	125	373	498	62,130	33,314	95,444
Total	1,347	1,845	3,192	633,212	211,976	845,188
Ratio	16.54%	17.83%	17.26%	13.57%	19.44%	14.69%
Including V_1=V_2=V_3	1,740	2,548 *	4,288	1,025,462	346,198 *	1,371,660
Ratio	21.37%	24.62% H	23.19%	21.98%	31.74% H	23.83%
Grand Total	8,142	10,348	18,490	4,664,720	1,090,679	5,755,399

Note 1 : * $p < .01$.
Note 2 : The random possibility of 3-mora CVCVCV nouns which have the same vowels in the 1st and 2nd, 2nd and 3rd or 1st and 3rd V positions is 20 percent.
Note 3 : The sign H refers to frequency of words with same vowel repetition that is significantly higher than random chance while the sign L refers to frequency of words with same vowel repetition which is significantly lower than random chance (no such cases).
Note 4 : The grand total of 18,490 refers to the total number of nouns with a CVCVCV phonological structure out of the 341,771 nouns taken from the word corpus of Amano and Kondo (2000).
Note 5 : The grand total of 5,755,399 refers to the total accumulative word frequency for the 18,490 nouns with a CVCVCV phonological structure.

These results suggest that the vowels occur frequently in the first and second V positions of CVCVCV string, but occur at the random chance level in the second and third V positions. The frequency of nouns with the same vowels in the first and third V positions seems to fall between the two previous conditions. Among general nouns the frequency was equal to the random chance, among proper nouns there were significantly high occurrences, but both together the frequency returns to the random chance level.

4.2.2. Token Frequency

Token frequency of Japanese nouns with the same vowels in two consecutive positions was also examined and these figures were considered more representative of the natural occurrence of these nouns than those of type frequency.

General nouns with the same vowel in V_1 and V_2 positions of a $CV_1CV_2CV_3$ string were counted 1,729,112 times or 37.07 percent (1,336,862 times or 28.66 percent without the same vowel in three consecutive positions) out of 4,664,720 general nouns. Based upon the calculation of Chebysheff's inequality theorem, this count was significantly higher than the random chance rate of occurrence of 20 percent ($p < .01$). Proper nouns showed a relatively high frequency rate of 387,383 times or 35.52 percent (23.21 percent or 253,161 times without the same vowel in three consecutive positions) out of 1,090,679 proper nouns. This token frequency rate was also significantly higher than random chance ($p < .01$). The total of both general and proper nouns indicated a high frequency rate of 36.77 percent or 2,116,495 times (27.63 percent or 1,590,023 times without the same vowel in three consecutive positions) out of 5,755,399 general and proper nouns, which was significantly higher than random chance ($p < .01$). Overall, token frequency of the same vowels in V_1 and V_2 positions in nouns with a $CV_1CV_2CV_3$ structure indicated significantly higher occurrences than the random chance rate of 20 percent. Therefore, it is concluded that these nouns appear more frequently than just by random chance.

Token frequency of general nouns, which had the same vowel in the V_2 and V_3 positions of its $CV_1CV_2CV_3$ phonological string, was 21.64 percent or 1,009,619 times (13.23 percent or 617,369 times without the same vowel in three consecutive positions) out of 4,664,720 general nouns. This frequency of occurrence was at the random chance level of 20 percent. Similarly, the same vowel occurring in the second and third V positions among proper nouns was 23.99 percent or 261,673 times (11.69 percent or 127,451 times without the same vowel in three consecutive positions) out of 1,090,679 proper nouns, which was at the random chance level. The total frequency of both general and proper nouns was, naturally, at the random chance level, being counted 22.09 percent or 1,271,292 times (12.94 percent or 744,820 times without the same vowel in three consecutive positions) out of a total of 5,755,399 of both nouns together. Therefore, the same vowels occur at the random chance level in the second and third V positions of a CVCVCV string.

Token frequency of nouns with the same vowel in V_1 and V_3 positions of a $CV_1CV_2CV_3$ string was counted as a basis for comparison. General nouns with the same vowel in the first and third V positions were counted 1,025,462 times or 21.98 percent (633,212 times or 13.57 percent without the same vowel in three consecutive positions) out of 4,664,720 general nouns. This token frequency figure for general nouns was at the random chance level. In contrast, proper nouns with the same vowel in the first and third V positions appeared 346,198 times or 31.74 percent (211,976 times or 19.44 percent without the same vowel in three consecutive positions) out of 1,090,679 proper nouns, which was significantly higher than the 20 percent random chance

rate of occurrence. However, the total token frequency among both general and proper nouns was 1,371,660 times or 23.83 percent (845,188 times or 14.69 percent without the same vowel in three consecutive positions) out of 5,755,399 both types of nouns together. This figure was at the random chance level.

4.3. Discussion

Study 2 examined the rate of occurrence of nouns with the same vowels found in two consecutive V positions within their CVCVCV strings. As shown in Table 2, type and token frequencies showed a general trend that the same vowels were found more frequently than at the random chance level in the first and second V positions of CVCVCV-structured nouns, but basically occur at the random chance level in the second and third V positions and in the first and third V positions, excluding proper nouns of the first and third V positions.

5. General Discussion

The present study proved that the same vowels in the three consecutive V positions of CVCVCV strings of nouns are more frequently repeated than the random possibility of 4 percent in both cases of type and token frequencies. Furthermore, the same vowels occurring in two consecutive or determined V positions were also observed in type and token frequencies of CVCVCV-structured nouns. In the two consecutive or determined positions, the random chance level was 20 percent. The same vowels in the first and second V positions of CVCVCV string of nouns occur more frequently than the random chance level. On the other hand, both type and token frequency were found to be at the random chance rate of occurrence for nouns with the same vowels repeated in the second and third V positions in their CVCVCV strings. Accordingly, CVCVCV-structured nouns with the same vowels in the first and second V positions occur more frequently than random probability, whereas same vowels in the second and third V positions occur at the rate of random chance. The same vowel repetition of the first and third positions, though not consecutive but determined positions, showed mixed results; the random chance level among the general nouns, but significantly higher than the random chance level among the proper nouns. However, when both the general and proper nouns were examined together, it returned to the random chance level.

As depicted in Figure 1, an interesting trend was found when combining the results of token frequency concerning nouns with the same vowels in three consecutive and two consecutive or determined V positions (see also Tables 1 and 2). The percentage of nouns with the same vowels in three consecutive V positions overlaps with the percentage of nouns with the same vowels in two positions. In other words, the percentages of nouns with the same vowels in V_1 and V_2, V_2 and V_3, and V_1 and V_3 positions all include the same percentage of nouns with the same vowels in V_1, V_2 and V_3 positions (i.e., 9.15%). Therefore, when 9.15 percent was subtracted from all total percentages of each condition of the same vowels in two V positions, the results were 27.63 percent for V_1 and V_2 positions, 12.94 percent for V_2 and V_3 positions, and 14.69 percent for V_1 and V_3 positions.

Figure 1 uses 'H' to refer to a frequency of occurrence that is greater than random chance. Thus, according to the boxes with H, it can be seen that nouns with the same vowels occurring in V_1 and V_2 positions, as well as in V_1, V_2 and V_3 positions with a $CV_1CV_2CV_3$ structure, have a

total token frequency of 36.77 percent. This suggests that once the same vowels occur in V_1 and V_2 positions, they are likely to be repeated in the V_3 position (i.e., 9.15% for $V_1 = V_2 = V_3$).

In summary, the findings of the corpus study indicated that vowels were repeated among Japanese nouns with a CVCVCV string more frequently than the random possibility of 4 percent. In addition, nouns with the same vowels in the first and second positions (i.e., V_1 and V_2 in the $CV_1CV_2CV_3$) showed significantly higher occurrences than the random chance of 20 percent, whereas nouns with the same vowels in the second and third positions appeared at the random chance level (i.e., V_2 and V_3). Since higher frequency enhances speed and accuracy in naming, phonological structures with the same vowel repeated can be expected to be more quickly and accurately named. However, Tamaoka and Murata (2001) found in their experimental study that native Japanese speakers named CVCVCV-structured phonemic strings with same vowel repetition less quickly and less accurately than unrepeated ones. This finding conflicts with the results of the present corpus study and excludes the possibility of the frequency of vowel repetition affecting the speed and accuracy of phonological processing.

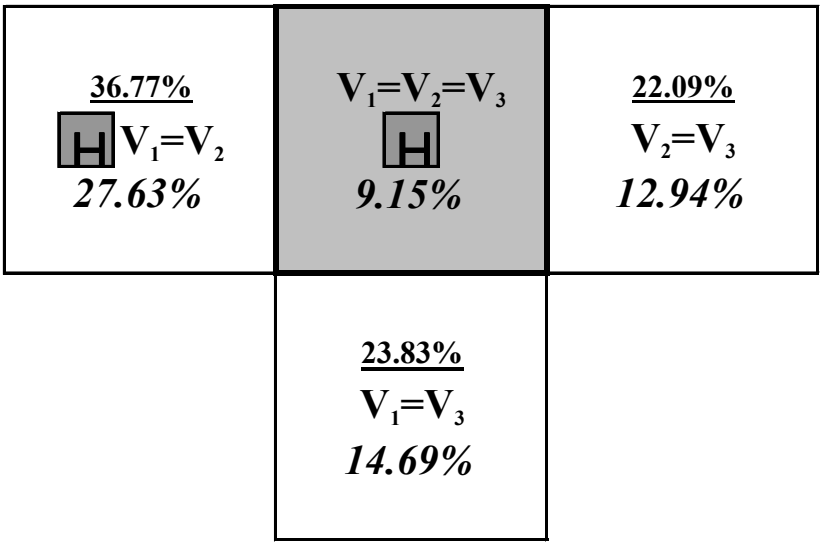


***Figure 1*.** Token frequency of Japanese nouns with vowel repetition

Note 1 : A V_1, V_2 and V_3 refers to vowels in a $CV_1CV_2CV_3$ string.
Note 2 : Percentages in Italics are the calculated frequencies of nouns out of the grand total of 5,755,399 CVCVCV-structured nouns.
Note 3 : Percentages underlined indicate token frequency of that particular box plus the token frequency of nouns with the same vowels in V_1=V_2=V_3.

Notes

[1] The pronunciation of words in this paper is transcribed using Japanese phonemic symbols which indicate three special sounds in Japanese: /N/ for nasal, /Q/ for geminate and /R/ for long vowel.

[2] The cutoff points at the 1 percent significant level for a type frequency of CVCVCV-structured general nouns is calculated as follows. The type frequency is 8,142 (see Table 1). A random probability is given by $1/5^3$ (three positions within five changeable vowels) ×5 (five different vowels) which equals 1/25. Thus, the mean m of the variable X becomes 326 ($m = 8{,}142 \times 1/25 = 325.68$). The standard deviation σ is given by m multiplied by 24/25 and squared by one half, $\sigma = (8{,}142 \times 1/25 \times 24/25)^{1/2} = 17.68$. The probability at the 1 percent level of significance was established by $1 - 1/k^2 = 0.99$. Thus, the value k was calculated as 10. Substituting all the values, the formula becomes $P(|X - 326| \leq 10 \times 17.68)$. After calculating this, the result was $P(149 \leq X \leq 503) \geq 0.99$. Since the actual type frequency is 393, this figure is within the range of 149 to 503. Consequently, the frequency 393 falls into a random chance range at the 1 percent level of significance. This calculation procedure is applied to all other frequencies in Studies 1 and 2.

References

Amano, N., & Kondo, K. (2000). *Nihongo no goi tokusei [Lexical properties of Japanese]*. Tokyo: Sanseido.

Fukazawa, H. (2000). Typology of OCP on features. *Phonological Studies 3, 121-134.*

Ito, J., & Mester, R.-A. (1986). The phonology of voicing in Japanese: Theoretical consequences for morphological accessibility. *Linguistic Inquiry 17, 49-73.*

Kubozono, H. (1999). *Nihongo no onsei [Japanese phonetics]*. Tokyo: Kenkyuusha Shuppan.

Kubozono, H., & Ota, S. (1998). *On'in koozoo to akusento [Phonological structure and accents]*. Tokyo: Kenkyuusha Shuppan.

Leben, W. (1973). *Suprasegmental Phonology*. Doctoral dissertation submitted to the Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.

Maezono, Y. (2002). *Gaisetsu kakuritsu tookei [An introduction to probability statistics]*. Tokyo: Saiensusha.

Matsubara, N., Nawata, K., & Nakai, N. (1994). *Tookeigaku nyuumon [An introduction to statistics]*. Tokyo: Tokyo University Press.

McCarthy, J. J. (1986). OCP effects: Gemination and Antigemination. *Linguistic Inquiry 17, 207-263*.

Murata, T. (1984). Jinkoo onomatope niyoru nihongo onsei haieraakii [Sound hierarchy of artificial onomatopes in Japanese]. *Linguistic Studies 85, 68-90.*

Murata, T. (1990). AB type onomatopes and reduplications in English and Japanese. In *Linguistic fiesta festschrift for professor Hisao Kakehi's sixtieth birthday: 257-272*. Tokyo: Kuroshio Shuppan.

Suzuki, G. (1999). *Joohooryoo kijun niyoru tookei kaiseki nyuumon [An introduction of information-based statistics]*. Tokyo: Kyoodansha.

Tamaoka, K., & Murata, T. (1999). Nihongo boin-no onsei haieraakii: Boin-o shotoo'on tosuru 3-paku koosei no muimi tsuzurigo-no meimei kadai kara [The hierarchical structure of Japanese vowels: A naming task investigation of 3-mora nonwords with initial vowel sounds]. *The Science of Reading 43, 79-89.*

Tamaoka, K., & Murata, T. (2001). OCP effects on Japanese phonological processing. *Phonological Studies 4, 119-126.*

Yip, M. (1988). The obligatory contour principle and phonological rules: A loss of identity. *Linguistic Inquiry 19, 65-100.*

Glottometrics 8, 2004, 12-22

Zum Gebrauch der Wortarten im Autorenstil

Viktor Levickij, Leonid Hikow[1]

Abstract. Similarities of different author's styles according to usage frequency of parts of speech were investigated. By means of quantitative analysis we established the regularities of noun over-usage (35%) and verbs over-usage (27%) in all investigated novels, which coincides with the results of similar investigations in other languages.

Keywords: author's style, quantitative analysis, combinability, frequency of usage of parts of speech, correlation

1. Ziele

Der Gebrauch der Wortarten im Text zieht immer mehr die Aufmerksamkeit der Linguisten auf sich. So erforschte B.N. Golovin (1970: 125-140) die Gebrauchshäufigkeit verschiedener Wortarten in den Werken russischer Schriftsteller als ein Merkmal des Autorenstils. Mit anderer Zielsetzung wurde die Gebrauchshäufigkeit der Wortarten von K.-H. Best (1997, 2000, 2001), R. Hammerl (1990), T.N. Jakubaitis (1981) u a. Autoren erforscht. Sie untersuchen, nach welchem statistischen Modell (z.B. nach einer Verteilung, die der normalen ähnlich ist, oder nach der geometrischen oder der negativen hypergeometrischen Verteilung) sich die Gebrauchshäufigkeiten der Wortarten in den Texten verschiedener Funktional- und Autorenstile verteilen (vgl. auch Becker 1988/1995; Judt 1955; Lindell, Piirainen 1980; Schweers, Zhu 1991).

J.A.Tuldava (1987: 124) zeigte, dass die Verteilung der Wortarten im Wörterbuch vom Umfang des Wörterbuches abhängt, „wobei der Anteil der Substantive mit Vergrößerung des Wörterbuchumfangs steigt".

Beim Gebrauch der Wortarten in den Texten eines Genres wird eine stark ausgeprägte Abhängigkeit bemerkbar. So existiert z.B. eine negative Abhängigkeit zwischen dem Gebrauch der Substantive und der Pronomen, weil die Pronomen (Personalpronomen) gerade dazu dienen, die Substantive zu vertreten (siehe Tuldava 1987:125).

A. Ziegler, K.-H. Best und G. Altmann (2001) untersuchten den zeitlichen Verlauf der Wortartenpositionierung im Text, G. Wimmer und G. Altmann (2001) entwickelten Tests für den intertextuellen Vergleich der Wortarten.

Die Aufgabe unserer Forschung besteht darin, die Verteilung einiger Wortarten im Autorenstil zu untersuchen. Das Hauptziel des Experiments besteht nicht darin, aufzuklären, ob die Gebrauchshäufigkeit der Wortarten über stildifferenzierende Eigenschaften verfügt oder nicht (solche Eigenschaften wurden in den Forschungen anderer Autoren entdeckt und werden kaum bezweifelt), sondern darin, eines der möglichen statistischen Aufdeckungsverfahren der oben genannten stilbildenden und stildifferenzierenden Wortarteneigenschaften auszuarbeiten.

---

[1] Address correspondence to: Viktor Levickij, Radiščeva Str. 6/5, UA-58000 Černivci. E-mail: leghinj@ukr.net, Leonid Hikow, Komarov Str. 31ab/51, UA-58013 Černivci. E-mail: leghinj@ukr.net

2. Materialien

Als Einheiten für die Analyse wurden 5 Wortarten – Substantiv, Adjektiv, Verb, Adverb und Personalpronomen – genommen. Die Auswahl der 5 Wortarten wurde nach folgenden Gesichtspunkten getroffen. In die Liste der analysierenden Einheiten wurden vor allem Hauptwortarten aufgenommen. Das Personalpronomen wurde als zusätzliche Kategorie aufgenommen, um eine Vermutung zu bestätigen oder ihr zu widersprechen, der Vermutung nämlich, dass der Gebrauch der Substantive auf bestimmte Weise mit dem Gebrauch des Pronomens verbunden ist.

Die grammatischen Hilfswortarten (Konjunktionen, Präpositionen, Artikel) können auch über stildifferenzierende Funktionen verfügen (siehe z.B. Golovin 1970:126), aber ihr Gebrauch ist vor allem mit den syntaktischen Besonderheiten des Textes verbunden (Präsenz des Satzgefüges oder der Satzreihe, gleichartiger Satzglieder, u. dgl.). Darum begrenzten wir die Auswahl der zu analysierenden Einheiten auf die „nominalen" Wortarten (die Aufnahme der Personalpronomina in diese Auswahl ist oben erwähnt). Die analysierten Texte wurden so ausgewählt, dass die Stichprobe mit der gleichen Anzahl der Werke jedes Autors vertreten ist. Insgesamt wurden für die Analyse 6 Autorenstile (6 Autoren, je drei Werke je Autor) ausgewählt.

Die Stichprobe enthält die Werke von H. Böll, S. Lenz, G. de Bruyn, T. Mann, M. Maron und M. Walser. Außer den Werken von T. Mann wird das Schaffen der meisten Schriftsteller durch einen relativ engen chronologischen Rahmen begrenzt. Da die Gebrauchshäufigkeit solcher Kategorien wie der Wortarten ziemlich hoch ist, wurde die Auswahl jede 10. Seite des betreffenden Buches ausgewählt. Da die Wortarten in der deutschen Sprache über deutlich ausgeprägte formale Merkmale verfügen, bereitete ihre Einteilung und Einordnung in die jeweilige qualitative Kategorie keine Schwierigkeiten. Bei der Zusammenstellung der Kartothek wurden alle aus den Texten herausgeschriebenen Wörter lemmatisiert, d.h. alle Substantive, Verben usw. werden in die in den Wörterbüchern angenommene Form gebracht (Substantive im Nominativ, Verben im Infinitiv). Insgesamt wurden 69930 Wörter herausgeschrieben. Der relative Fehler der Stichprobe δ für verschiedene Wortarten beträgt: für Substantive 0.01; für Verben 0.01; für Adjektive 0.02; für Adverbien 0.02.

3. Allgemeine Charakteristik des Gebrauchs der Wortarten

Die Ergebnisse zur Gebrauchshäufigkeit der 5 Wortarten in den Werken von 6 Autoren sind in Tabelle 1 angeführt.

Tabelle 1
Die Gebrauchshäufigkeit der 5 Wortarten in den Texten von 6 Autoren

Autor	**Subst.**	**Verb**	**Adjekt.**	**Adverb**	**Pers. Pron.**	**Insgesamt**
H. Böll	3938	3376	1329	1947	1782	12372
S. Lenz	3441	3349	1107	1562	1571	11030
G. de Bruyn	3891	3102	1269	1623	1181	11066
T. Mann	4386	2837	1705	1927	1185	12040
M. Maron	3938	3041	1291	1444	1581	11295
M. Walser	4179	3358	1010	2216	1364	12127
Insgesamt	23773	19063	7711	10719	8664	69930

Anhand dieser Tabelle wird deutlich, dass in der deutschen Prosa am häufigsten die Substantive (35%) gebraucht werden, dann folgen die Verben (27%), Adverbien (15%) und Adjektive (11%). Die Gebrauchshäufigkeit des Personalpronomens beträgt 12%. Analoge Berechnungen der Gebrauchshäufigkeit verschiedener Wortarten werden auch in anderen Sprachen verwirklicht (siehe Tiščenko 1970, Kločkova 1968, Jakubaitis 1981, u.a.). Die zusammengestellten Daten über die Gebrauchshäufigkeit der Wortarten in verschiedenen Sprachen sind bei Tuldava (1987:127) angeführt. Diese Daten werden in Tabelle 2 wiedergegeben.

Tabelle 2
Die Gebrauchshäufigkeit der 4 Wortarten in sieben Sprachen

Wortarten	Sprache						
	Ukr.	Russ.	Ung.	Lit.	Eston.	Fin.	Deutsch
Substantiv	29.2%	28.7%	30.0%	28.1%	31.7%	29.7%	35.0%
Verb	19.7%	18.3%	22.4%	23.1%	22.5%	27.6%	27.0%
Adverb	6.2%	6.0%	8.0%	10.0%	15.8%	10.9%	15.0%
Adjektiv	6.8%	7.9%	10.0%	5.4%	6.0%	7.8%	11.0%

Wie in Tabelle 2 deutlich ist, korrelieren die erhaltenen Daten zum Deutschen mit denen anderer Sprachen. Das betrifft vor allem den Anteil der Substantive und der Verben (in allen Sprachen nimmt das Substantiv den ersten Platz ein und das Verb den zweiten Platz). In mehreren Sprachen (darunter auch in der deutschen) wird das Adverb häufiger als das Adjektiv gebraucht. Ausnahmen hierzu bilden das Russische und das Ukrainische, wo das Adjektiv häufiger als das Adverb gebraucht wird. Man kann also vermuten, dass der Stil der schöngeistigen Literatur in vielen Sprachen dadurch charakterisiert wird, dass die häufigsten Wortarten das Substantiv und das Verb sind. Man kann auch vermuten, dass solch ein Verhältnis einen universalen Charakter trägt (wenigstens für die Sprachen, die einen deutlichen Unterschied zwischen den Wortarten haben).

Die grafische Darstellung der Häufigkeiten, die in Tabelle 1 angeführt werden (siehe Diagramm 1) lässt vermuten, dass zwischen der Gebrauchshäufigkeit einiger Wortarten im Autorenstil Schwankungen bemerkbar werden. So übertrifft in den Werken von T. Mann der Gebrauch des Substantivs wesentlich die Gebrauchshäufigkeit der Verben, bei S. Lenz gibt es fast keinen Unterschied zwischen den entsprechenden Häufigkeiten. Wenn man die Daten der Tabelle 1 rangiert, dann entsteht Tabelle 3.

Wenn man die Daten dieser Tabelle vergleicht, bemerkt man, dass die Substantive und die Verben bei allen Autoren die ersten und die zweiten Plätze einnehmen. Danach folgen, wie die mittleren Ränge zeigen, die Adverbien (3.3), Personalpronomen (4) und Adjektive (4.7). Die Adjektive nehmen fast bei allen Schriftstellern den 5. Platz ein. Nur in den Werken von T. Mann und G. de Bruyn hat das Adjektiv den Rang 4. Man muss aber in Betracht ziehen, dass Verteilungen wie diese nicht nur durch den Schreibstil des Autors bedingt sein können, sondern auch durch ganz andere Faktoren – Häufigkeit des Gebrauchs verschiedener Wortarten im Text. Wie man aus Tabelle 2 sieht, wird die ähnliche Verteilung der Gebrauchshäufigkeit verschiedener Wortarten auch in anderen Sprachen erkennbar.

Auf welche Weise kann man die Gebrauchsbesonderheiten verschiedener Wortarten feststellen, die ausgerechnet durch den Autorenstil bedingt sind? Oder unterscheiden sich die in Tabelle 1 aufgeführten Häufigkeiten vielleicht gar nicht voneinander?

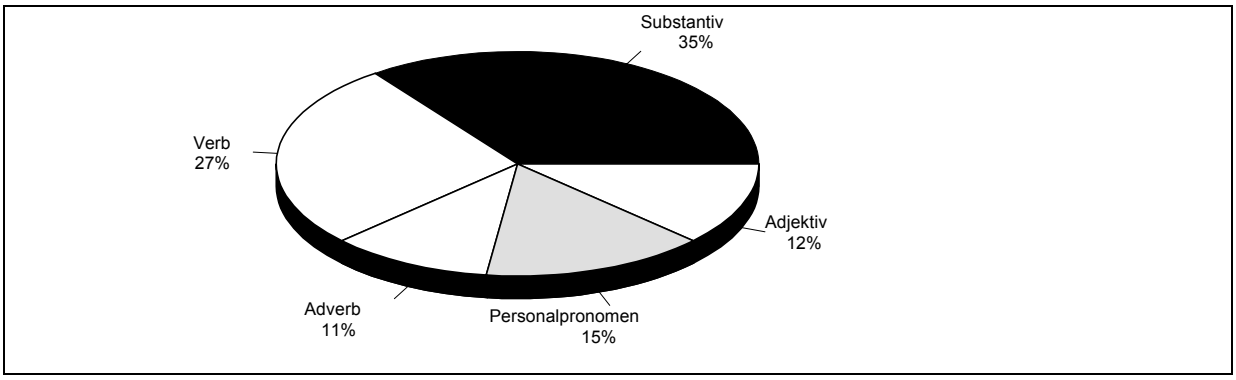


Diagramm 1. Graphische Darstellung der Wortartenhäufigkeiten

Tabelle 3
Die Ränge der Wortartenhäufigkeiten in den Texten der 5 Schriftsteller

Autor	**Subst.**	**Verb**	**Adjekt.**	**Adverb**	**Pers. Pron**
H. Böll	1	2	5	3	4
S. Lenz	1	2	5	4	3
G. de Bruyn	1	2	4	3	5
T. Mann	1	2	4	3	5
M. Maron	1	2	5	4	3
M. Walser	1	2	5	3	4
Mittelwert	1	2	4,7	3,3	4

Betrachtet man die Ränge der Tabelle 3, so zeigt ihre deutliche Übereinstimmung, dass es keine wesentlichen Unterschiede in der Rangverteilung der Gebrauchshäufigkeiten der Wortarten gibt. Das spürbarste und feinste Instrument, mit dessen Hilfe man einen Homogenitätstest für die Daten in Tabelle 1 durchführen kann, ist der Chiquadrat-Test, den man nach der Formel (1) berechnen kann.

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}, \tag{1}$$

Man muss in Betracht ziehen, dass bei sehr großen Stichprobenumfängen die Summe (1) auch sehr hoch sein wird, da das Chiquadrat mit der Stichprobengröße linear anwächst. Dieser Nachteil kann dadurch eingeschränkt werden, dass das Chiquadrat relativiert wird. In diesem Fall wird (1) mit der Zahl der Beobachtungen N und der Zahl der Freiheitsgrade (siehe Formel (2)) dividiert, um den Kontingenzkoeffizienten

$$\mathbf{K} = \sqrt{\frac{\chi^2}{\mathbf{N}\sqrt{(\mathrm{r}-1)(\mathrm{c}-1)}}} \tag{2}$$

zu bekommen (r = Zahl der Zeilen, c = Zahl der Spalten). Eine statistische Analyse erbrachte, dass sich für Tabelle 1 bei 4 × 5 = 20 Freiheitsgraden die Summe $\chi^2 \approx 707$; $\chi^2_{0.01;20} = 37.57$; $K = 0.047$ ergibt. Die Häufigkeiten in Tabelle 1 verteilen sich also nicht proportional. Der Homogenitätstest zeigt, dass das χ^2 signifikant ist. Da das χ^2 letzten Endes von den Abweichungen der empirischen Größen in jedem Tabellenfeld von den theoretisch zu erwartenden ab-

hängt, ist es erforderlich, jene Felder zu finden, in denen (a) das χ^2 statistisch signifikant ist; (b) die empirischen Häufigkeiten die theoretisch zu erwartenden übertreffen (aber nicht umgekehrt). Für die Verwirklichung des vorgesehenen Programms muss man die Vierfelder-Tabellen aufgrund der Tabelle 1 zusammenstellen. Für die Illustration vergleichen wir zwei alternative (Vierfelder)Tabellen (siehe 4 und 5).

Tabelle 4
Die Häufigkeit des Adjektivs in den Werken von H. Böll

Autor	Wortarten				Insgesamt
	Adjektiv		Andere		
H. Böll	1329	**a**	**b**	11043	12372
Andere	6382	**c**	**d**	51176	57558
Insgesamt	7711		62219		**N** 69930

Tabelle 5
Die Häufigkeit des Substantivs in den Werken von T. Mann

Autor	Wortarten				Insgesamt
	Substantiv		Andere		
T.Mann	4386	**a**	**b**	7654	12040
Andere	19387	**c**	**d**	38503	57890
Insgesamt	23773		46157		**N** 69930

Die Größe χ^2 für Tabelle 4 ist gleich 0.91; für Tabelle 5 $\chi^2 = 20.97$. Bei 1 FG im ersten Fall konstatieren wir, dass χ^2 nicht signifikant ist und keine Kontingenz zwischen den Merkmalen in Tabelle 4 anzeigt ist. Im zweiten Fall ist die Signifikanz sehr hoch, denn $\chi^2_{0,01;1} = 6.63$, und wir erhielten fast 21. Der Kontingenzkoeffizient im zweiten Fall ist gleich 0.017. Die Resultate der statistischen Analyse für alle Felder der Tabelle 1 sind in Tabelle 6 veranschaulicht.

Tabelle 6
Verbindung der Merkmale [Wortart] und [Autorenstil] (Größen χ^2 und K)

	Autor	Subst.	Verb	Adjektiv	Adverb	Pers.Pron.
1	H. Böll					$\chi^2 = 40.5$ K = 0.024
2	S. Lenz		$\chi^2 = 38.95$ K = 0.024			$\chi^2 = 30.58$ K = 0.021
3	G. de Bruyn	$\chi^2 = 4.43$ K = 0.008				
4	T. Mann	$\chi^2 = 20.97$ K = 0.017		$\chi^2 = 107.27$ K = 0.039		
5	M. Maron					$\chi^2 = 23.57$ K = 0.018
6	M. Walser				$\chi^2 = 68.62$ K = 0.031	

In dieser Tabelle sind nur die Größen χ^2 und *K* angeführt, bei denen, wie oben gesagt, die empirischen Häufigkeiten signifikant die theoretisch zu erwartenden übertreffen. In den entsprechenden Feldern der Tabelle 6 wird die statistisch signifikante Kontingenz der Merkmale [Autorenstil] und [Wortart] bemerkbar. In Tabelle 6 sieht man, dass die größte Kontingenz für die Merkmale [Mann] + [Adjektiv], [Walser] + [Adverb], [Böll] + [Personalpronomen], [Lenz] + [Verb] fixiert ist. Wie kann man nun die erhaltenen Daten interpretieren? Diese Daten zeigen, dass alle Autoren natürlich alle Wortarten gebrauchen. Im Stil mancher Schriftsteller aber bemerkt man einen Gebrauch der einen oder der anderen Wortart, der signifikant die Gebrauchshäufigkeit dieser Wortart in den Werken anderer Autoren überschreitet. So folgt aus Tabelle 3, dass das Adjektiv nach seiner Häufigkeit den letzten Platz unter anderen Wortarten einnimmt. Aber, wenn man nicht die Ränge der Wortarten vergleicht, sondern die Ränge der Schriftsteller im Rahmen einer Wortart, so erweist sich, dass T. Mann den ersten Platz beim Gebrauch der Adjektive (1705) einnimmt, H. Böll den zweiten (1329), und G. de Bruyn den dritten (1269). Die Daten der Tabelle 6 kann man nicht vom Standpunkt „Autor", sondern vom Standpunkt „Wortarten" interpretieren. In diesem Fall muss man konstatieren, dass die Substantive in den Werken von T. Mann und G. de Bruyn, das Verb nur von S. Lenz, das Adjektiv nur von T. Mann, das Adverb von M. Walser und das Personalpronomen von H. Böll, S. Lenz, und M. Maron häufiger, als zu erwarten ist, gebraucht werden.

Es ist wichtig, hier den Gebrauch des Personalpronomens und des Substantivs zusammenzustellen. In Tabelle 6 ist gut erkennbar, dass der Überschuss der Häufigkeit des Personalpronomens in jenen Zeilen der Tabelle auftritt, in denen ein Überschuss der Häufigkeit des Substantivs nicht zu sehen ist (Zeilen 1, 2, 5). Und umgekehrt, das Substantiv nimmt jene Zeilen (3,4) ein, bei denen kein Überschuss der Häufigkeiten des Pronomens zu bemerken ist. Folglich wird der Gebrauch des Personalpronomens und des Substantivs durch das Verhältnis charakterisiert, das man gewöhnlich die komplementäre Distribution nennt.

Ergebnisse wie diese legen den Schluss nahe, dass der Stil von T. Mann durch einen größeren Gebrauch von Epitheta charakterisiert ist. Die festgestellte Kontingenz der Merkmale (Verb + Lenz) zwingt die Linguisten, ihre Aufmerksamkeit auf solche Besonderheiten des Autorenstils von S. Lenz zu lenken, was bisher unbemerkt blieb.

Aber die von uns durchgeführte Analyse des Autorenstils hat mit dem Aufdecken der Kontingenzen zwischen dem Autorenstil und der Gebrauchshäufigkeit der einen oder anderen Wortart noch nicht ihr Bewenden. Außer den oben erhaltenen Daten wäre es interessant und zweckmäßig zu erfahren: (a) zwischen welchen Autoren die größte Ähnlichkeit im Gebrauch der Wortarten beobachtet wird; (b) welcher von den Autoren sich am meisten gewissen „mittleren" (theoretischen) Werten des Gebrauchs der lexikalisch-grammatischen Wortklassen nähert; (c) welcher Wortartgebrauch durch die größte Variation charakterisiert wird.

4. Messung der Unterschiede und der Ähnlichkeit zwischen den Autorenstilen

Es wäre zweckmäßiger, die Ähnlichkeit zwischen den Autorenstilen mit Hilfe des Korrelationskoeffizienten zu messen. Aber in unserem Fall erfasst die Variation nur 5 Merkmale (5 Wortarten), folglich ist die Zahl der korrelierenden Paare zu klein (FG = 5 – 2 = 3). Außerdem kann man erwarten, dass sich die Korrelationskoeffizienten nach ihrer Größe wenig voneinander unterscheiden werden (es genügt, auf den Zusammenfall der Ränge in Tabelle 3 zu schauen). Deswegen wäre es besser, für die Messung der Ähnlichkeit und der Unterschiede zwischen den Autorenstilen in unserem Fall wiederum die χ^2-Werte zu benutzen. Diese Größen steigen, je mehr sich die empirischen Häufigkeiten von den theoretischen unterscheiden. Daher ist es nicht wichtig, in welche Richtung eine solche Abweichung erfolgt – in die positi-

ve oder in die negative, weil alle Abweichungen (O - E) bei der Berechnung des χ^2 quadriert werden ($\chi^2 = \Sigma(O - E)^2/E$).

Das bedeutet, dass wir für jedes Feld der Tabelle 1 die χ^2-Werte finden und nach den Größen dieser Werte über die Ähnlichkeit oder den Unterschied der zu erforschenden Erscheinungen urteilen können. In Tabelle 7 sind die berechneten Werte des χ^2 angegeben.

Tabelle 7
χ^2-Werte bei der Häufigkeitsverteilung der 5 Wortarten

Autor	Subst.	Verb	Adjektiv	Adverb	Pers.Pron.	Gesamt
H. Böll	17.07	0	0.91	1.35	40.50	59.83
S. Lenz	25.41	38.95	9.81	9.80	30.58	114.55
G. de Bruyn	4.43	2.42	1.95	3.16	26.34	38.30
T. Mann	20.97	60.36	107.27	3.60	63.06	255.26
M. Maron	2.51	0.47	1.66	47.68	23.57	75.89
M. Walser	0.77	0.82	80.07	68.62	12.76	163.04
Insgesamt	71.16	103.02	201.67	134.21	196.81	706.87

Anhand dieser Tabelle wird deutlich, dass die summierten Abweichungen verschiedener „Stile“ und verschiedener morphologischer Einheiten nicht gleich sind. Am wenigsten weichen die Gebrauchshäufigkeiten verschiedener Wortarten von den theoretisch erwarteten Häufigkeiten im Stil von G. de Bruyn ($\chi^2 = 38.3$) ab und am meisten im Stil von T. Mann (255). Bedeutende Unterschiede werden auch für M. Walser (163) fixiert. Genauso können wir die Gebrauchshäufigkeit der Wortarten bewerten. Die niedrigste Abweichung von der theoretischen „Norm“ ist beim Substantiv (71) und beim Verb (103) zu beobachten. Im höchsten Grad weichen die Adjektive und Pronomen von der „Norm“ ab. Zur Charakteristik dieser Eigenschaften der Wortarten ist es zweckmäßig, den Begriff „Variation“ und „Stabilität“ des Gebrauchs des Textelements einzuführen. Die Indexe der Variation und der Stabilität können, wie oben gezeigt, empirisch festgestellt werden. In unserem Fall benutzten wir als entsprechende Kennziffer die Werte des χ^2. Man kann allerdings auch andere Kriterien verwenden. Die von uns erhaltenen Daten zeugen davon, dass das Substantiv und das Verb die größte Stabilität besitzen und die Adjektive und Pronomen die größte Variation zeigen.

Um die Autorenstile miteinander zu vergleichen (und nicht mit einem gewissen theoretischen Wert), muss man die empirischen Verteilungen der 5 Wortarten in den Stilen der zu erforschenden Autoren paarweise vergleichen, wie in der Tabelle 8 gezeigt,.

Tabelle 8
Gebrauchshäufigkeiten der Wortarten in den Werken von H. Böll und S. Lenz

Autor	Subst.	Verb	Adjektiv	Adverb	Pers. Pron.	Gesamt
H. Böll	3938	3376	1329	1947	1782	12372
S. Lenz	3441	3349	1107	1562	1571	11030
Gesamt	7379	6725	2436	3509	3353	23402

Der χ^2-Wert für diese Tabelle gleicht 32.5; FG = (5 - 1)(2 - 1) = 4, K = 0.03. Analog wurden die K-Werte für alle anderen Fälle erhalten. Die Ergebnisse der Analysen sind in Tabelle 9 angegeben.

Tabelle 9
Die Ähnlichkeitsstufe der Stile der 6 Autoren (Werte des Koeffizienten *K*)

Autor	H. Böll	S. Lenz	G.de Bruyn	T.Mann	M.Maron	M.Walser
H. Böll	—	0.026	0.044	0.069	0.034	0.05
S. Lenz		—	0.049	0.085	0.038	0.058
G. de Bruyn			—	0.044	0.039	0.048
T. Mann				—	0.064	0.073
M. Maron					—	0.066
M. Walser						—

Man muss unterstreichen, dass, je mehr Unterschiede zwischen zwei Autoren bemerkt werden, desto größer wird der χ^2-Wert und, folglich auch der *K*-Wert. Also der größte *K*-Wert entspricht der höchsten Stufe der Unterschiede, und der kleinste *K*-Wert der höchsten Stufe der Ähnlichkeit. Dabei kann die Ähnlichkeit dadurch bedingt sein, dass beide Autoren den Gebrauch gewisser Wortart bevorzugen, oder dadurch, dass beide Autoren seltener als erwartet die eine oder andere Wortart gebrauchen.

Folglich bezeugen, wie man in Tabelle 9 sieht, die Stile von S. Lenz und T. Mann (*K* = 0.085), H. Böll und T. Mann (*K* = 0.069), T. Mann und M. Walser (*K* = 0.073), S. Lenz und M. Walser (*K* = 0.058), H. Böll und M. Walser (*K* = 0.05) die geringste Ähnlichkeit. Die größte Ähnlichkeit wird zwischen H. Böll und S. Lenz (*K* = 0.026) beobachtet. Hätten wir beschlossen, dass es zweckmäßig ist, die Ähnlichkeit nur nach den „positiven" Abweichungen zu beurteilen, so müsste man den Koeffizienten *Φ* benutzen (dieser Koeffizient berücksichtigt positive und negative Verbindungen) und einzeln die Werte positiver und negativer Koeffizienten nach der Formel (3) zusammenstellen.

$$\Phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}, \tag{3}$$

Man könnte auch die Formel (4) benutzen.

$$u = \frac{n_{ij} - E_{ij}}{\sqrt{\frac{n_i \cdot n_j \cdot (N - n_i) \cdot (N - n_j)}{N^2 \cdot (N-1)}}}, \tag{4}$$

wobei $n_{i.}$ = die Randsummen auf der rechten Seite der Tabelle
n_j = die Randsummen unterhalb der Tabelle
N = Summe aller Häufigkeiten in der Tabelle.

Wenn man die Formel (4) benutzt, so ist für das Feld [Böll + Substantiv] in Tabelle 8 gleich $u = \frac{+37}{452} = +0{,}082$, was nicht signifikant ist (P ≈ 0.06). Das entspricht der Wahrscheinlichkeit für Chiquadrat = 0.74 (nicht signifikant). Für das Feld [H. Böll + Verb] in Tabelle 8 ergibt sich der Wert *u* nach der Formel (4) als:

$$u = \frac{3376 - 3555}{\sqrt{\frac{12372 \cdot 6725 \cdot (23402 - 12372) \cdot (23402 - 6725)}{23402^2 \cdot (23402 - 1)}}} = \frac{-179}{34.52} \approx -5.18.$$

Dieser Zahl entspricht eine sehr große Wahrscheinlichkeit (P = 0.9999). Analog hat der Wert χ^2 = 19.19 bei FG = 1 dieselbe Wahrscheinlichkeit. Aber t zeigt uns, ob die Verbindung positiv oder negativ ist. Den analogen Wert kann man auch nach der Formel (3) finden. In diesem Fall gleicht Φ für die Merkmale [Böll] + [Verb] = -0.028, und für [Lenz] + [Verb] ≈ + 0.028.

Mit Hilfe des Chiquadrat-Tests und des Kontingenzkoeffizienten kann man also die Stufe der Ähnlichkeit der Autorenstile (siehe Tabelle 6) und die Stufe der Variation der einen oder anderen Wortart (siehe Tabelle 7) bestimmen.

Für die Antwort auf die Frage, wessen Autorenstil sich in einem größeren Grad einer gewissen Norm nähert (dem invarianten Stil), muss man die empirischen Verteilungen der Häufigkeiten im Stil jedes Autors mit den summarischen Verteilungen der Häufigkeiten vergleichen (siehe Tabelle 10).

Tabelle 10
Die Gebrauchshäufigkeit der Wortarten im Stil von H Böll und im invarianten Stil

Autor	Subst.	Verb	Adjektiv	Adverb	Pers. Pron.	Gesamt
H. Böll	3938	3376	1329	1947	1782	12372
Invar.	23773	19063	7711	10719	8664	69930
Gesamt	27711	22439	9040	12666	10446	82302

Der Test ergibt χ^2 = 50.15 mit FG = 4. Analog werden die χ^2-Werte für alle anderen Autoren berechnet. In Tabelle 11 werden die entsprechenden *K*-Werte angegeben.

Tabelle 11
Abweichung des Individual-Autoren Stils von der bedingten Norm

Autor	H. Böll	S. Lenz	G. de Bruyn	T. Mann	M. Maron	M. Walser
Insgesamt	0.017	0.025	0.014	0.036	0.02	0.029

Man muss noch einmal unterstreichen, dass, je größer der *K*-Wert ist, desto größer ist die Stufe der Abweichung des Autorenstils von „der Norm“. Aus Tabelle 11 sieht man, dass sich der Stil von M. Maron (0.02) am meisten der Norm nähert, und der Stil von T. Mann (*K* = 0.036) am meisten von der Norm abweicht.

Bis jetzt betrachteten wir die Gebrauchshäufigkeit der einen oder anderen Wortart von der Position ihrer „Stabilität“ oder „Variation“. Aber die erhaltenen Daten können auch unter dem Gesichtspunkt der Stilbildung benutzt und interpretiert werden. In diesem Fall ist es zweckmäßig, die Begriffe *stildifferenzierendes Potenzial des morphologischen Textelements* und *Stufe der Selektivität des Autorenstils* einzuführen.

Tatsächlich ist es so, dass, je mehr die Gebrauchshäufigkeit der einen oder anderen Einheit (Wortart) von dem theoretischen Wert abweicht, desto mehr ist diese Wortart an der Charakteristik (an der Differenzierung) des Autorenstils beteiligt. Das bedeutet, dass sich ihr stildifferenzierendes Potenzial abhängig von der Variation der einen oder anderen Wortart vergrößern oder verkleinern wird. Offensichtlich ist das Potenzial des Substantivs und des Verbs

als zwei Grundwortarten im Vergleich nicht groß. Über ein viel größeres stilistisches Potential verfügen Adjektive und Adverbien.

Analog kann der Begriff *Selektivität des Stils* interpretiert werden. In den Werken einiger Autoren werden keine bedeutenden Abweichungen empirischer Häufigkeiten von den theoretischen bemerkbar. Der Autor verhält sich zu der einen oder anderen Wortart nicht selektiv. Er gebraucht die eine oder andere Wortart so, dass sie einer gewissen abstrakten Norm nah ist. Der Grad der Selektivität des Stils eines solchen Schriftstellers ist nicht groß. Ein anderer Autor bevorzugt umgekehrt den Gebrauch irgendeiner Wortart, so dass ihre Gebrauchshäufigkeit wesentlich von einer gewissen Norm abweicht. In diesem Fall kann man sagen, dass der Autorenstil durch eine hohe Stufe der Selektivität charakterisiert wird.

Die Stufe der Selektivität und das stildiffenzierende Potenzial können mit Hilfe einer quantitativen Kenngröße ausgedrückt werden, d.h. sie können gemessen werden. In dem von uns durchgeführten Experiment wurde eines der möglichen Verfahren der Messung, sowohl des stildifferenzierenden Potenzials als auch der Stufe der Selektivität des Autorenstils, veranschaulicht.

Genauso können auch die lexikalisch-semantischen Besonderheiten des Autorenstils erforscht werden. Dazu muss man im voraus lexikalisch-semantische Subklassen der Substantive, Verben, Adjektive und Adverbien bilden, um dann die Gebrauchshäufigkeiten jeder dieser Subklassen in den Werken verschiedener Autoren erforschen zu können.

Schlussfolgerungen

Die vorgeschlagene Methodik lässt Assoziationen zwischen den Merkmalen [Wortart] und [Autorenstil] auffinden und feststellen:

(a) welcher Wortart durch die größte Gebrauchsvariation charakterisiert wird;

(b) welcher Autors sich einem „mittleren“ Stil (einer bedingten Norm) nähert.

Es wurde festgestellt, dass das Adjektiv, das Pronomen und das Adverb durch die höchste Variation charakterisiert werden, die höchste Annäherung zur bedingten Norm wurde in den Texten von M. Maron entdeckt und im höchsten Grad weicht der Stil von T. Mann von der Norm ab.

Literaturverzeichnis

Becker, H. (1988/1995). *Die Wirtschaft in der deutschsprachigen Presse*. Frankfurt: Lang 1995.

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

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

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

Golovin B.N. (1970). *Jazyk i statistika*. Moskva: Prosveščenije.

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

Jakubaitis T.O. (1981). *Časti reči i tipy tekstov*. Riga: Zinatne.

Judt, B. (1995). *Wortartenhäufigkeiten im Deutschen und Französischen*. Göttingen: Staatsexamensarbeit.

Kločkova E. A. (1968). O raspredelenii klassov slov v nekotorych funkcionalnych stiljach russkogo jazyka. *Voprosy jazykoznanija: Saratov SGU, 109-118.*

Lindell, A., Piirainen, I.T. (1980). *Untersuchungen zur Sprache des Wirtschaftsmagazins „CAPITAL“.* Vaasa: Vaasan Kauppakorkeakoulun Julkaisuja, Tutkimuksia No 67, Philologie 5.

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

Tiščenko W. (1970). Častota častyn movy v riznych funkcionalnych styljach sučasnoji ukrajinskoji movy. – In: Perebyjnis V.S., Muravycka M. P. (Hrsg.), *Pytannja strukturnoji leksykolohiji: 215-224*. Kyjiv: Naukova dumka.

Tuldava J. (1987). *Problemy i metody kvantitativno-sistemnogo issledovanija leksiki*. Tallin: Valgus.

Wimmer, G., Altmann, G. (2001). Some statistical investigations concerning word classes. *Glottometrics 1, 109-123.*

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

Glottometrics 8, 2004, 23-41

Häufigkeiten von Satzlängen: Zum Faktor der Intervallgröße als Einflussvariable

(am Beispiel slowenischer Texte)

Emmerich Kelih, Peter Grzybek
Universität Graz

Abstract: The present study is a contribution to the study of sentence length. Specifically, the study focuses on the question of factors influencing the theoretical modeling of frequency distributions of sentence lengths. Slovenian texts are analyzed on three analytical levels: individual texts, complex texts, and a text corpus. On the basis of this material, the impact of a broadly accepted smoothing procedure (smoothing by forming specific intervals) on the adequacy of theoretical models is controlled.

Keywords: Sentence length, Slovenian

1. Theoretische Modellierung der Satzlängenverteilung:

Die theoretische Modellierung der Satzlänge – d.h. die Frage, ob die Satzlängenverteilung durch ein entsprechendes theoretisches Verteilungsmodell beschrieben werden kann – hat eine mehr als 50-jährige Geschichte. Fast ein halbes Jahrhundert nach den ersten Grundsatzüberlegungen von Sherman (1888) warf Yule (1939) in seiner Untersuchung zur Satzlänge in englischen Texten erstmals diese Frage auf; dabei nannte er zwar explizit kein theoretisches Verteilungsmodell, wies aber auf folgende Beobachtung hin: „They are not of the Poisson type but of the type in which the square of the standard deviation largely exceeds the mean" (Yule 1939: 371). In Anlehnung an diese Untersuchung schlug wenig später dann Williams (1940) vor, dass man bei der Verteilung der Satzlänge nicht die absolute Anzahl der Wörter pro Satz als Variable bestimmen solle, sondern die jeweilige logarithmisch transformierte Wortanzahl pro Satz. Die von Williams (1940) durchgeführte Re-Analyse der Daten von Yule beschränkte sich auf eine graphische Darstellung; diese zeigte seiner Ansicht nach, dass die Häufigkeitsverteilung der *x*-silbigen Wörter einer Normalverteilung folgt. Aufgrund der empirischen Untersuchung von drei Texten postulierte er sodann die genannte Lognormal-Verteilung als allgemein gültiges theoretisches Modell der Satzlängenverteilung (vgl. Williams 1940: 360).

Eine Kritik erfuhr der Ansatz von Williams erst Jahrzehnte später durch Sichel (1974), der auf einige Unzulänglichkeiten des Vorgangsweise hinweist. So kritisierte Sichel zu Recht, dass die Lognormal-Verteilung nur aufgrund einer graphischen Darstellung der Verteilung der Satzlängen postuliert wurde, ohne dass die erwähnte Verteilung einem statistischen Prüfverfahren unterzogen worden wäre, welches den Grad der Adäquatheit zwischen theoretischem Modell und empirischer Beobachtung hätte erbringen können. Sichel (1974) selbst schlug seinerseits eine zusammengesetzte Poisson-Verteilung als allgemeines Modell der Satzlängenverteilung vor, bei welcher ein Parameter wiederum als Zufallsvariable einer weiteren Verteilung folgt (vgl. Sichel 1974: 26f.). Diese Distribution wurde von ihm an acht lateini-

schen, griechischen und englischen Texten inklusive entsprechender statistischer Tests überprüft. Diese ersten Überlegungen zu einer theoretischen Modellierung wurden dann in weiterer Folge in einem allgemeinen Ansatz von Altmann aufgegriffen, der damit – wie darzulegen sein wird – eine neue Perspektive in dieser Diskussion aufzeigte.

1.1. Neuansatz in der theoretischen Modellierung der Satzlängenverteilung

Die oben einleitende dargestellten Arbeiten zur theoretischen Modellierung von Satzlängen wurden von Altmann (1988a) einer allgemeinen Kritik unterzogen: dabei wurde seinerseits darauf verwiesen, dass das Auffinden einer Verteilung (hier bezogen auf die zusammengesetzte Poisson-Verteilung von Sichel) in der Regel inhaltlich (d.h. hier: linguistisch) schwer zu begründen bzw. zu interpretieren ist und somit für die Analyse von sprachlichen Phänomenen keine nennenswerten Erkenntnisse beisteuert. Dem stellte Altmann (1988b) einen grundlegenden Neuansatz gegenüber, der darin besteht, die Distribution von sprachlichen Einheiten in einen synergetisch-linguistischen Kontext (vgl. Köhler 1986) zu stellen. Bezogen auf die Frage der Verteilung von Satzlängen werden a priori folgende systeminterne und systemexterne Faktoren als Einflussfaktoren in Betracht gezogen (Altmann (1988a: 152):

a – Wirkung des Produzenten (Stil u.a)
b – Wirkung des Rezipienten (der Sprachgemeinschaft, Rücksicht auf den Hörer)
c – Faktoren des Textes
d – Faktoren der Ebene

Ausgehend von zwei Annahmen, nämlich
1. dass jegliche Verteilung der Längen in einem Text (unter anderem, aber nicht ausschließlich also auch der Satzlängen) gesetzmäßig organisiert ist, und
2. dass es ausreichend ist, Annahmen über die Differenz zweier jeweils benachbarter Wahrscheinlichkeiten zu machen,

stellte Altmann (1988b) den folgenden Ansatz auf: Sei die Differenz benachbarter Klassen

(1) $$P_x - P_{x-1} = \Delta P_{x-1}.$$

und diese Differenz sei nicht konstant, sondern hänge vom jeweiligen P_{x-1} ab, so dass sich der Quotient

$$D = \frac{P_x - P_{x-1}}{P_{x-1}} = \frac{\Delta P_{x-1}}{P_{x-1}}$$

ergibt. Im Hinblick auf die Berechnung der Satzlänge in Anzahl der Teilsätze kommt der Faktor *d* folglich nicht zum Tragen; somit ergibt sich nach Altmann (1988b) die folgende Gleichung:

(2) $$D = \frac{b - ax}{cx}$$

Hierbei wirken die der Zipf'schen Kräfte von Unifikation bzw. Diversifikation *a* und *b* „gestaltend", während *c* „bremsend" wirkt (*a* hat ein negatives Vorzeichen, insofern der

Produzent versucht, seinen eigenen Stil in die Bindung der Textsorte einzubringen); weiterhin wirkt *b* „global“, während *a* und *c* „lokal“ wirken. Falls allerdings die Satzlänge in Wortanzahl gemessen wird, kommt der intervenierende Faktor der Sprachebene (*d*) hinzu, was zu der Quotienten (3) führt:

$$(3) \qquad D = \frac{b - ax}{cx + d}.$$

Dies führt zu den beiden Ansätzen (4) und (5)

$$(4) \qquad P_x = \frac{b - ax}{cx} P_{x-1}$$

$$(5) \qquad P_x = \frac{b - ax}{cx + d} P_{x-1}$$

Wir können uns hier die weiteren detaillierten Ableitungen ersparen und wollen uns statt dessen auf die aus ihnen (nach Reparametrisierung) resultierenden Verteilungsmodelle beschränken. So wird für die Satzlänge, gemessen in der Anzahl der Teilsätze (clauses) pro Satz, die *negative Binomialverteilung* (6) als adäquates Verteilungsmodell hergeleitet. Falls die Satzlänge in Anzahl der Worte, nicht der Anzahl der clauses pro Satz gemessen wird, kommt die Wirkung der intervenierenden Ebene des Teilsatzes (*d*) hinzu; unter dieser Bedingung wird die *Hyperpascal-Verteilung* (7) als adäquates Modell der Satzlängenverteilung postuliert.

$$(6) \qquad P_x = \binom{r + x - 1}{x} p^r q^x \qquad x = 0, 1, 2, \ldots$$

$$(7) \qquad P_x = \frac{\binom{k + x - 1}{x}}{\binom{m + x - 1}{x}} q^x P_0 \qquad x = 0, 1, 2, \ldots$$

Die Gesetzeshypothese zur Satz/Wort-Variante wurde von Altmann (1988a) an 245 Texten des Altgriechischen, Englischen und Slowakischen überprüft, wobei nur in neun Fällen keine signifikante Übereinstimmung mit dem Modell erzielt wurde. Bei den wenigen Texten, die keine Übereinstimmung mit der Hyperpascal-Verteilung zeigten, handelte es sich um solche Texte mit ungeklärter Autorschaft beziehungsweise Texte, an denen seitens der Editoren Modifikationen durchgeführt worden waren. In den anderen Fällen handelte es sich um zusammengesetzte Stichproben, die „nur unter sehr günstigen Umständen dem Gesetz folgen” (Altmann 1988a: 159). Durch die Überprüfung der clause-Variante in zehn Texten unterschiedlicher Sprachen konnte die Hypothese, dass die Satzlängenhäufigkeitsverteilung durch die negative Binomialverteilung modelliert werden kann, bestätigt werden.

Während sowohl die einleitende dargestellten Untersuchungen als auch die zuletzt referierten Überlegungen von Altmann zur theoretischen Modellierung der Häufigkeit von

Satzlängen durch theoretische Wahrscheinlichkeitsverteilungen von der sprachübergreifenden Relevanz der jeweiligen Konzepte ausgingen, haben jüngere Untersuchungen in diesem Bereich (Niehaus 2001, Best 2001 Grzybek 1999 u.a.) es als wahrscheinlicher erscheinen lassen, dass eher von verschiedenen Modellen auszugehen ist, die sich zwar vermutlich auf einen allgemeinen Ansatz wie den von Wimmer/Altmann (1996) zurückführen lassen, von denen wir aber bislang nicht wissen, unter welchen Randbedingungen sie jeweils von Relevanz sind bzw. welche Faktoren auf die Güte der Modellierung Einfluss haben.

Im vorliegenden Beitrag soll es darum gehen, auf einen solchen Einflussfaktor aufmerksam zu machen, der im Grunde genommen nicht in der spezifischen Beschaffenheit des untersuchten Sprachmaterials begründet ist, sondern sozusagen bei dessen sekundärer Bearbeitung, nämlich der Aufbereitung für die theoretische Modellierung, ins Spiel kommt. Hierbei handelt es sich um eine in den entsprechenden Untersuchungen bislang in ihrer Auswirkung nicht hinreichend systematisch reflektierte Problematik, nämlich die Zusammenfassung von Satzlängenklassen zu bestimmten Intervallen.

1.2. Der Faktor der Intervallgröße als Einflussvariable der theoretischen Modellierung

Die Zusammenfassung von Satzlängen zu Intervallen ist ein übliches Verfahren, da die Messung der Satzlänge in der Anzahl der Worte pro Satz eine erhebliche Spannweite aufweist und die Satzlänge prinzipiell keiner quantitativen Beschränkung nach oben hin unterliegt. In der Satzlängenforschung hat sich daher das Verfahren eingebürgert, die Satzlängen in 5er-Intervalle von (*1-5, 6-10, ...* Wörter pro Satz) zusammenzufassen. So fasst bereits Yule (1939) in seiner Untersuchung zu Fragen des Stils und des Nachweises der Autorschaft die Satzlängen in 5er-Intervalle Wörtern zusammen (vgl. Yule 1939: 367). Auch Altmann (1988b) greift in seiner empirischen Untersuchung der von ihm postulierten Hyperpascal-Verteilung an 236 Texten die Satzlängen in 5er-Intervallen zusammen.

Vor dem Hintergrund dieses ehemals als selbstverständlich notwendig und unproblematisch angesehenen Vorgehens erweisen sich in einer Reihe jüngerer Untersuchungen zur (in der Anzahl der Worte pro Satz gemessenen) Satzlänge diese Zusammenfassungen (Intervallbildungen) als überaus problematisch:

1. Aus der Untersuchungen von Niehaus (2001: 210) zur Verteilung der Satzlänge an 20 deutschen literarischen Texten geht hervor, dass die von Altmann (1988a) theoretisch postulierte Hyperpascal-Verteilung nur dann passt, wenn man die Satzlängenklassen zu 5er-Intervallen zusammenfasst; ohne Zusammenfassung zu Intervallen erweist sich hingegen ein anderes Modell als adäquat, nämlich die negative Binomialverteilung
2. Einen ähnlichen Befund erhält Best (2001) als Ergebnis seiner Analyse der Satzlängenverteilung von 25 Texten der deutschen Gegenwartssprache. Seiner Interpretation nach stellen sich die Ergebnisse seiner Untersuchung folgendermaßen dar: „Für Satzlängen, gemessen nach der Zahl ihrer indirekten Konstituenten (Wörter), scheint die negative Binomialverteilung das beste Modell zu sein“. In dieser Studie von Best erwies sich die von Altmann (1988a) postulierte Hyperpascal-Verteilung allerdings als gänzlich ungeeignet, insofern der von Best beobachtete Befund unabhängig davon gilt, „ob man sie zu Fünfergruppen zusammenfasst oder nicht“ (Best 2001: 198).
3. In seiner Untersuchung der Satzlänge von Sprichwörtern in einem slowenischen Sprichwortkorpus testete Grzybek (1999) die Auswirkung unterschiedlicher Intervallbildungen (*1-2, 1-3, 1-4, ...*). Im Ergebnis zeigte sich, dass die Art der Zusammenfassung (d.h. die Größe der gebildeten Intervalle) nicht unbedingt einen Einfluss auf die theoretische Modellierung haben muss: Die Satzlängenverteilung folgte nämlich in allen Arten der Zusammenfassung der Hyperpoisson-Verteilung; nur für vollkommen

ungruppierte Satzlängen konnten keine guten Ergebnisse erzielt werden (vgl. Grzybek 1999: 104). Es muss jedoch im Hinblick auf diese Untersuchung angemerkt werden, dass es sich bei einer Sprichwortsammlung um spezifisches Datenmaterial handelt, welches aufgrund seiner einem Satz-Lexikon ähnlichen Struktur vermutlich anderen Gesetzmäßigkeiten der Satzlängenverteilungen unterliegt als etwa ein üblicher Fließtext.

4. In einer Folgestudie zur Verteilung der Satzlänge in einem Korpus deutscher Sprichwörter von Grzybek/Schlatte (2002) zeigte sich hingegen, dass die konkrete Art der Zusammenfassung einen eindeutigen Einflussfaktor hinsichtlich der theoretischen Modellierung darstellt – es wirkt sich nämlich „der konkrete Umfang der zusammengefassten Klassengrößen vehement auf das anzupassende Verteilungsmodell aus" (Grzybek/Schlatte 2002: 301). Jedoch sind interessanterweise weder die schon zuvor diskutierte Hyperpascal-Verteilung noch die von Grzybek (1999) für slowenisches Sprichwortmaterial herangezogene Hyperpoisson-Verteilung geeignet, die Satzlängen im Sprichwortkorpus ohne glättende Zusammenfassung zu modellieren. Erst bei einer Zusammenfassung zu 2er-Intervallen sind die Anpassungsergebnisse für beide Modelle akzeptabel; bei einer noch weiteren Zusammenfassung der Satzlängen zu 3er-, 4er-, und 5er-Klassen ist gar nur mehr die Hyperpoisson-Verteilung in Betracht zu ziehen.

Aus den angeführten Untersuchungen zur Satzlängenverteilung wird somit deutlich, dass die Zusammenfassung von Satzlängen in Intervalle offensichtlich als ein nicht zu vernachlässigender Einflussfaktor auf die theoretische Modellierung zu berücksichtigen ist: Einerseits zeigt es sich wiederholt, dass die ursprünglich von Altmann (1988a) postulierte Hyperpascal-Verteilung nur bei einer Zusammenfassung der Satzlängen in 5er-Intervalle als adäquates Modell der Satzlängenverteilungen in Betracht gezogen werden kann.[1] Aus den genannten empirischen Untersuchungen ergibt sich auch, dass selbst bei einer Zusammenfassung von (in der Anzahl der Worte pro Satz gemessenen) Satzlängen solche Modelle in Betracht zu ziehen sind, die für die Modellierung der Teilsatz/Satz-Ebene postuliert wurden (negative Binomial-Verteilung). Darüber hinaus wird die Hyperpoisson-Verteilung diskutiert, die sich in einschlägigen empirischen Untersuchungen wiederholt als ein adäquates Modell erwiesen hat.

Es leitet sich daher die Notwendigkeit ab, die Art und Weise der Zusammenfassung zu Satzlängenintervallen als einen eigenen Einflussfaktor einer systematischen Untersuchung zu unterziehen. Um diesen Zusammenhang zwischen der Zusammenfassung von Satzlängen zu Intervallen und der Adäquatheit theoretischer Modelle näher beleuchten zu können, wird im folgenden der von Grzybek (1999) vorgeschlagene empirische Ansatz zur Lösung dieses Problems aufgegriffen: Dieser besteht darin, dass zunächst an die beobachtete Verteilung der Satzlängen ohne jegliche Zusammenfassung zu Intervallen die in Frage kommenden Modelle angepasst werden; daran anschließend werden die Satzlängen aber auch in unterschiedlichen (1er-, 2er-, 3er-, 4er-, 5er-Intervallen) Intervallen zusammengefasst, bevor auch an diese geglätteten Daten die Modelle angepasst werden.

---

[1] In einer weiteren jüngeren Arbeit zur (in der Anzahl der Worte pro Satz gemessenen) Satzlänge (vgl. Kaßel/Livesey 2001) in englischen Texten werden die Satzlängenverteilungen ohne und mit Zusammenfassungen in 5er-Intervallen durchgeführt. Interessanterweise eignet sich in dieser Arbeit insbesondere die (1-verschobene) negative Binomialverteilung und die gemischte Poissonverteilung zur Modellierung der Satzlängenverteilungen; an gegebener Stelle wird jedoch die Hyperpascal-Verteilung – aus welchen Gründen auch immer – nicht diskutiert.

2. Pilotstudie zum Einflussfaktor der Intervallbildung am Beispiel slowenischer Texte

2.1. Modellierung der Satzlängen im Slowenischen

Die Verteilung der Satzlängen und die Frage der Beschreibung von Satzlängenverteilung durch theoretische Modelle ist im Slowenischen bislang wenig erforscht. In der Untersuchung von Jakopin (2002) gibt es einige wenige Hinweise auf die prozentuale Häufigkeitsverteilung von Satzlängen in slowenischen Texten. Beispielsweise wird gezeigt, dass die am häufigsten auftretende Satzlänge (prozentualer Anteil > 7%) in einem Korpus slowenischer literarischer Texte im Bereich von 4,5 – 6 Wörtern pro Satz liegt (vgl. Jakopin 2002). Des weiteren werden auch Angaben zu den längsten im Korpus auftretenden Sätzen gemacht. Die Ergebnisse sind jedoch für die vorliegende Untersuchung nicht von unmittelbarer Bedeutung, beziehen sich doch die Angaben der Satzlängen auf das jeweilige Gesamtkorpus und aufgrund der fehlenden Rohdaten ist auch keine Modellierung der Verteilung der Satzlängen möglich.

Die offensichtlich bislang einzige Untersuchung zur Verteilung der Satzlängen in slowenischen Texten stammt von Grzybek (1999). In dieser Studien wurde die Hypothese einer gesetzmäßigen Verteilung der (in der Anzahl der Worten pro Satz gemessenen) Satzlängen auf der Basis einer slowenischen Sprichwortsammlung überprüft. An diesem Material erwies sich die *Hyperpoisson-Verteilung* als besonders geeignet, um die Satzlängen in adäquater Weise zu modellieren (vgl. Grzybek 1999: 100). Zwar gilt zu beachten, dass eine Sammlung von Sprichwörtern vermutlich anderen Gesetzmäßigkeiten der Satzlängenverteilung folgt, als die hier vorliegenden literarischen Prosatexte, dennoch aber sollte auch diese Verteilung nicht außer acht bleiben. Dass sie in unmittelbarer Beziehung zu den oben bereits dargestellten negativen Binomialverteilung und zur Hyperpascal-Verteilung steht, lässt sich recht leicht zeigen: In Analogie zu den Ansätzen (4) und (5) stellt sich die Rekursionsformel der Hyperpoisson-Verteilung wie in (8) dar:

$$P_x = \frac{b}{x+d} P_{x-1}, \tag{8}$$

was in Analogie zu den Rekursionsformeln (6) und (7) in der folgenden Verteilung resultiert:

$$P_x = \frac{b^x}{d^{(x)}\,{}_1F_1(1;d;b)}, \quad x = 0,1,2,... \tag{9}$$

Alle drei Verteilungsmodelle gehören insofern zu ein und derselben Familie, die man auch als erweiterte Katz-Familie bezeichnet: Es zeigt sich nämlich (vgl. Wimmer/Altmann 2000: 279ff., 449ff.), dass die negative Binomialverteilung ein Spezialfall der Hyperpascal-Verteilung ist (für $m = 1$), und dass die Hyperpascal-Verteilung gegen die Hyperpoisson-Verteilung konvergiert ($k \to \infty$, $q \to 0$, $kq \to b$).

Die empirische Überprüfung der Satzlängenverteilung mit Hilfe des Altmann-Fitters (2000) zeigte sich in der Tat, dass die folgenden theoretischen Modelle in Erwägung zu ziehen sind:[2]

[2] In einer einleitenden Orientierungsphase wurde mit dem Altmann-Fitter (2000) zunächst exploratorisch (d.h. ohne theoretisch geleitete Vorannahmen) untersucht, welche Verteilungen überhaupt als geeignete Modelle in Frage kommen. Nach der Reduktion auf die drei oben genannten Verteilungen wurden die Detailanalysen auf diese beschränkt.

(a) die von Altmann (1988b) für die Satz/Wort-Variante postulierte *Hyperpascal-Verteilung*;
(b) die von Grzybek (1999) diskutierte *Hyperpoisson-Verteilung*;
(c) die von Best (2001) und Niehaus (2001) nachgewiesene *negative Binomial-verteilung*.[3]

Im folgenden wird das Hauptaugenmerk auf den potentiellen Einflussfaktor der Zusammenfassung von Satzlängen zu Intervallen gerichtet; ungeachtet dieser Fokussierung lässt sich diese Frage nicht unabhängig von einer differenzierten Betrachtung der Analyseebenen (Korpus, komplexe Texte, Einzeltexte) verfolgen.

2.2. Empirische Untersuchung des Faktors Intervallgröße in slowenischen Texten

Als Basis für die empirische Untersuchung zum Faktor der Intervallgröße bei der Modellierung der Häufigkeitsverteilung von Satzlängen dient ein Korpus slowenischer Prosatexte. Diese setzte sich aus sechs Kurzromanen bzw. Kurzgeschichten (slowenisch: povest, im folgenden als Kurzerzählungen bezeichnet) von vier verschiedenen slowenischen Autoren aus dem 19. Jhd. zusammen (vgl. Tab. 1).

Tabelle 1
Slowenisches Textkorpus

Text	Autor	Titel
1	J. Kersnik	Mačkova očeta
2	J. Kersnik	Ponkrčev oča
3	I. Cankar	Hlapec Jernej in njegova pravica
4	J. Jurčic	Nemški Valpet
5	F. Levstik	Pokljuk
6	F. Levstik	Martin Krpan

Diese Texte werden im folgenden mit den entsprechenden Nummern (Text #1 bis Text #6) bezeichnet und als solche analysiert. Zum Zwecke der Qualitätskontrolle der Datenbasis – (zur Frage der Homogenität in quantitativen Untersuchungen vgl. Altmann 1992 bzw. Orlov 1982) – werden diese Texte jedoch nicht nur jeweils einzeln als komplexe Gesamttexte analysiert, sondern auf zwei weitere Arten und Weisen: Zum einen werden die genannten sechs Texte zu einem Gesamtkorpus zusammengefügt, so dass sich eine umfangreichere Textmischung ergibt; dieses Gesamtkorpus soll bedingt als „Text #7“ bezeichnet werden (vgl. Tab. 2). Zum anderen ergibt sich aufgrund der Tatsache, dass die Texte #2 und #3 jeweils aus mehreren Kapiteln bestehen, die Option, diese Kapitel als Einzeltexte zu verstehen und getrennt zu analysieren; in diesem Fall haben wir es mit den Texten #8 bis #28 (vgl. Tab. 3) zu tun.

Um die der im vorliegenden Text im Vordergrund stehende Problematik anschaulich zu illustrieren, finden sich im folgenden auf der Basis des gesamten Textkorpus die graphische

[3] Interessanterweise hatte bereits Fucks (1970) in seinen Untersuchungen zur Satzlänge im Deutschen darauf hingewiesen, dass die negative Binomialverteilung sich gut eigne, die Verteilung von Satzlängen zu erfassen, wobei er allerdings die Satzlänge einer Reihe von deutschen Texten in der Anzahl der Silben pro Satz berechnet hatte. Dabei war Fucks allerdings von ganz anderen Voraussetzungen ausgegangen und hatte die Verteilung auf ganz andere Art und Weise abgeleitet: Ausgegangen war er nämlich von der Poisson-Verteilung, die dann die negative Binomialverteilung ergibt, wenn der Parameter der Poisson-Verteilung (der Mittelwert) einer Gammaverteilung folgt.

Darstellungen 1a-1e: Während Abb. 1a die Daten ohne Zusammenfassung zu Intervallen veranschaulicht, stellen die Abb. 1b-1d das Ergebnis der unterschiedlichen Zusammenfassungen in den jeweiligen Intervallen dar, d.h. in Form einer schrittweisen Glättung.

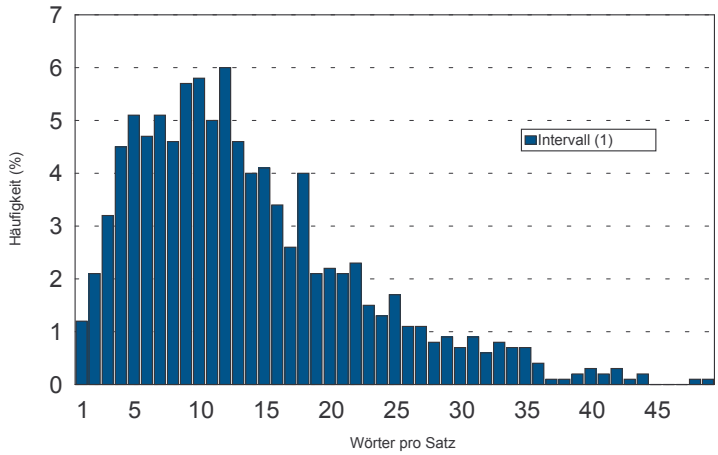


Abb. 1a: Ohne Zusammenfassung

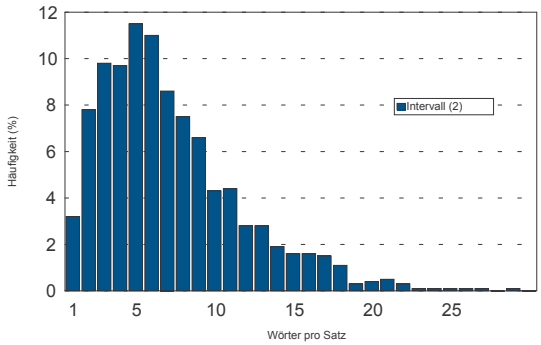


Abb. 1b: Zusammenfassung zu 2er Intervallen

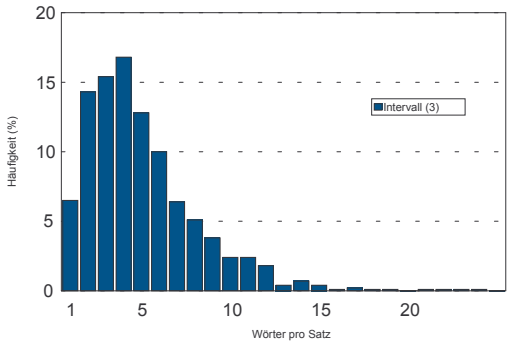


Abb. 1c: Zusammenfassung zu 3er Intervallen

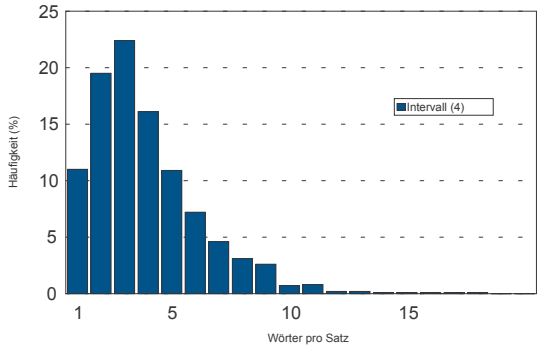


Abb. 1d: Zusammenfassung zu 4er Intervallen

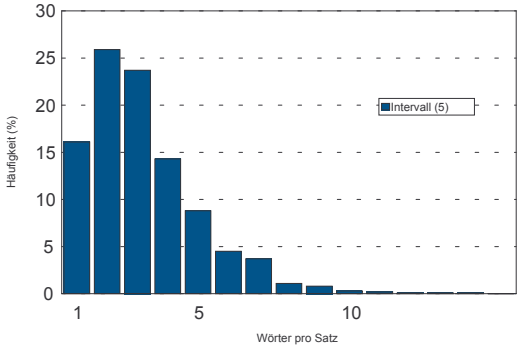


Abb. 1e: Zusammenfassung zu 5er Intervallen

Eine Zusammenfassung von Satzlängen zu Intervallen hat, wie auf den ersten Blick aus den graphischen Darstellungen ersichtlich wird, ganz offenbar folgende Auswirkungen:

a.) Einzelne Satzlängen, die sich durch eine besonders hohe Frequenz auszeichnen, machen sich mit zunehmender Intervallgröße nicht mehr so stark bemerkbar, wodurch insgesamt eine Nivellierung der Klassen bewirkt wird;

b.) Ebenso enthält mit zunehmender Intervallgröße die Häufigkeitsverteilungen immer weniger leere Satzlängenklassen (0 Worte pro Satz/Satzlängenintervall).

Dieser Befund ist als Ausgangspunkt für die folgende empirische Untersuchung des Faktors der Zusammenfassung von Satzlängen bei der theoretischen Modellierung zu verstehen.

2.3. Modellierung der Satzlängenhäufigkeit

Wie oben dargestellt, räumt die Spezifik der Textbasis die Möglichkeit ein, die Frage der Modellierung von Satzlängenhäufigkeiten unter Berücksichtigung des Einflussfaktors der Zusammenfassung zu Intervallen auf drei unterschiedlichen Analyseebenen zu untersuchen:

(1) Auf der ersten Ebene werden die genannten Kurzerzählungen zu einem Korpus zusammengefasst. Das Korpus, welches in Hinsicht auf die involvierten Textsorten als homogen zu bezeichnen ist, gibt die Möglichkeit zu prüfen, inwiefern eine Korpusanalyse gegebenenfalls anderen Gesetzmäßigkeiten folgt als die Analyse der einzelnen Texte. Insgesamt besteht das Korpus aus 2758 Sätzen mit 38966 Wörtern; die einzelnen Werte sind in der Tab. 2 zusammengefasst.[4]

Tabelle 2
Quantitative Angaben zum Textkorpus

Textnr.	**Sätze**	**Wörter**	**min.**	**max.**	$\overline{x}$
# 7	2758	39016	2	106	14.15

(2) Auf der zweiten Ebene werden die sechs Kurzerzählungen als komplexe Texte aufgefasst; anzumerken ist, dass dabei eine textinterne, von den Autoren selbst vorgenommene Kapitelgliederung nicht beachtet wird. Insgesamt handelt es sich also um sechs unterschiedliche Kurzerzählungen von vier verschiedenen Autoren, wobei Text #3 mit insgesamt 1383 Sätzen den größten Umfang aufweist. Unter dieser Voraussetzung zeichnen sich die Texte durch die in Tab. 3 zusammengefassten Charakteristika aus.

Tabelle 3
Quantitative Angaben zu den komplexen Texten

Text	**Autor**	**Titel**	**Sätze**	**Wörter**	**min.**	**max.**	$\overline{x}$
# 1	J. Kersnik	Mačkova očeta	109	1597	1	61	14.65
# 2	J. Kersnik	Ponkrčev oča	165	2178	1	42	13.20
# 3	I. Cankar	Hlapec Jernej in njegova pravica	1383	18407	1	97	13.31
# 4	J. Jurčic	Nemški Valpet	561	7921	1	63	14.21
# 5	F. Levstik	Pokljuk	169	3181	2	70	18.82
# 6	F. Levstik	Martin Krpan	371	5682	2	106	15.32

(3) Auf der dritten Ebene werden die Texte unter Berücksichtigung der von den Autoren selbst vorgenommen Kapitelgliederung jeweils individuell untersucht. Im Detail weist der

[4] Neben der jeweiligen Textnummer sowie dem Autor und Titel des Werks finden sich die Anzahl der Sätze, Minimum und Maximum der Wortanzahl pro Satz sowie die durchschnittliche Satzlänge.

Text von Janko Kersnik „Ponkrčev oča“ drei Kapitel auf; Ivan Cankars „Hlapec Jernej ...“ besteht insgesamt aus 18 Einzelkapiteln (insgesamt also 21 Einzelkapitel). Tabelle 4 resümiert die wesentlichen Charakteristika der Einzeltexte.

Tabelle 4
Quantitative Angaben zu den Einzeltexten

Text	Autor		Titel	Sätze	Wörter	min.	max.	$\overline{x}$
# 8	J. Kersnik	Ponkrčev oča	Kapitel 1	68	895	1	40	13.16
# 9			Kapitel 2	38	523	4	35	13.76
# 10			Kapitel 3	59	760	1	42	12.88
# 11	I. Cankar	Hlapec Jernej	Kapitel 1	43	602	2	41	14.00
# 12			Kapitel 2	82	977	1	36	11.91
# 13			Kapitel 3	85	1038	1	50	12.21
# 14			Kapitel 4	57	796	3	97	13.96
# 15			Kapitel 5	80	809	1	37	10.11
# 16			Kapitel 6	75	890	2	36	11.87
# 17			Kapitel 7	71	973	3	41	13.70
# 18			Kapitel 8	107	1473	1	37	13.77
# 19			Kapitel 9	60	939	5	35	15.65
# 20			Kapitel 10	113	1134	1	39	10.04
# 21			Kapitel 11	75	937	1	33	12.49
# 22			Kapitel 12	80	1203	1	78	15.04
# 23			Kapitel 13	119	1583	1	51	13.30
# 24			Kapitel 14	53	956	2	72	18.04
# 25			Kapitel 15	98	1388	1	36	14.16
# 26			Kapitel 16	77	1203	1	66	15.62
# 27			Kapitel 17	87	1203	1	33	13.83
# 28			Kapitel 18	21	303	1	38	14.43

Durch diese Gliederung auf drei unterschiedliche Analyseebenen ergeben sich insgesamt 28 Datensätze, in denen die Satzlänge bestimmt werden kann. Auf Basis dieser Texte lässt sich nunmehr – neben der Frage des Einflussfaktors Intervallgröße – auch die Frage der Datenhomogenität untersuchen.

Exkurs:
Zur Definition des Satzes und der automatischen Berechnung der Satzlänge

Die Satzlänge in den genannten Texten wurde automatisiert analysiert und berechnet (zur genaueren Bestimmung siehe Kelih/Grzybek 2004). Ausgehend von der Möglichkeit der formalen Bestimmung der Satzgrenzen aufgrund von Interpunktionszeichen wird dem Punkt, dem Ausrufe- sowie dem Fragezeichen eine satzabgrenzende Funktion zugesprochen. In Anbetracht der Tatsache, dass der Punkt im Slowenischen jedoch in einigen Positionen keine satzabgrenzende Funktion hat (z.B. als Abkürzung wie in „*c.kr*.“ oder als Auslassungszeichen wie in “*To je Sitarjevo… tam!*”) und unter Berücksichtigung des Umstandes, dass Frage- und Ausrufezeichen auch zur Kennzeichnung von Interjektionen dienen können (vgl. Pravopis 1990: 38ff.), sind einige Modifizierungen notwendig.

In Anlehnung an die grundsätzlichen Überlegungen von Grinbaum (1996) zur Automatisierung von Satzlängenuntersuchungen bietet es sich daher an, den Großbuchstaben als weiteres satzabgrenzendes Zeichen in eine formal bestimmbare Satzdefinition einzubauen. Wenn auch der Großbuchstabe im Slowenischen zur Kennzeichnung von Eigennamen, geographischen Bezeichnungen und ähnlichem dient, liegt eine weitere zentrale Funktion des Großbuchstabens darin, den Anfang von Texten, Absätzen und einzelnen Sätzen zu markieren. In dieser Funktion als Gliederungsmerkmal von Texten können Großbuchstaben bei der Bestimmung von Satzgrenzen herangezogen werden (vgl. Grinbaum 1996: 454). Somit sind die Interpunktionszeichen (.), (…), (?) und (!) in Kombination mit einem Großbuchstaben am Anfang des nächstfolgenden Satzes eindeutig als satzabschließend zu identifizieren. Da jedoch den erwähnten Interpunktionszeichen nicht in allen Fällen ein Buchstabe folgen muss (Textende, Absatzende) gelangt man zu der folgenden Satzdefinition, die auch der vorliegenden Untersuchung zur Anwendung kommt:

> **Als Satzendezeichen gelten [.], […], [!] und [?], es sein denn, ein Kleinbuchstabe ist das erste Graphem des nächsten Wortes.**[5]

Aufgrund der verwendeten Satzdefinition ist es möglich, den Satz (und damit dann auch die Satzlänge) automatisiert zu bestimmen. Das vorliegende Textkorpus und die vorgestellte Satzdefinition sind nunmehr als Ausgangspunkt für die empirische Untersuchung der Satzlängenverteilung in slowenischen Texten zu sehen.

2.4. Einflussfaktor Satzlängenintervalle bei der theoretischen Modellierung

Unter den dargestellten Voraussetzungen wird im nächsten Schritt somit zu prüfen sein, ob für die erwähnten theoretischen einzelnen Verteilungsmodelle die vorgenommene Zusammenfassung von Satzlängen in Intervalle einen Einflussfaktor darstellt. Entsprechend der obigen Ausführungen sollen nunmehr die Ergebnisse der theoretischen Modellierung – getrennt für die unterschiedenen Analyseebenen – präsentiert werden. Als Güte der Anpassung wird dabei die Überschreitungswahrscheinlichkeit P des χ^2-*Tests* angeführt: Dabei wird ein Wert von $P \geq 0.01$ als Kennwert einer akzeptablen Übereinstimmung von empirischen Daten und theoretischem Modell angesehen; für längere Texte wird der Kontingenzkoeffizient $C = \chi^2 / N$ (vgl. Grotjahn/Altmann 1993) in Betracht gezogen, wobei $C \leq 0.02$ als gute, $C \leq 0.01$ als sehr gute Anpassung des jeweiligen theoretischen Modells betrachtet wird (die jeweiligen P- und C-Werte vgl. Anhang 1-2)[6].

---

[5] Natürlich kommen bei derartigen Definition sprachspezifische bzw. kulturspezifisch-typographische Konventionen ins Spiel. So wird in der Arbeit auf russische Texte bezogenen Arbeit von Kelih/Grzybek (2004) der Satz alternativ definiert als „eine durch Punkt, Frage- und Ausrufezeichen abgegrenzte Einheit des Textes“, wobei diese einfachere Definition und die hier angeführten einem statistischen Vergleich unterzogen werden. In der genannten Arbeit kann gezeigt werden, dass die Anwendung von unterschiedlichen Satzdefinitionen zu keinen statistisch signifikanten Unterschieden führt. Auch auf der Ebene der theoretischen Modellierung von Satzlängenverteilungen spielen die beiden unterschiedlichen Satzdefinitionen keine signifikante Rolle.

[6] Die gesamten Rohdaten zur vorliegenden Untersuchung finden sich in Kelih (2002).

2.4.1. Theoretische Modellierung auf Korpusebene

Auf der Ebene des Gesamtkorpus (vgl. Tab. 2) ergibt sich erstes Ergebnis, dass die Hyperpoisson-Verteilung auf der Ebene des Gesamtkorpuses ein gänzlich unpassendes Modell darstellt. Im Gegensatz dazu weist die negative Binomialverteilung – außer bei jeglicher Nicht-Zusammenfassung von Satzlängen zu Intervallen – eine gute Übereinstimmung mit dem Modell auf (vgl. Tabelle 5 mit den *C*-Werten). Das insgesamt überzeugendste Ergebnis liefert bei Nicht-Zusammenfassung ebenso wie bei allen Arten der Intervallbildung die Hyperpascal-Verteilung, die ja von Altmann (1988a) für die (in der Anzahl der Worten pro Satz berechneten) Satzlängenverteilung postuliert worden war.

Tabelle 5
C-Werte für das Gesamtkorpus

Verteilung	**Intervall**				
	1	2	3	4	5
Negative Binomial	0.0982	0.0079	0.0102	0.0083	0.0048
Hyper-Poisson	0.0802	0.0621	0.051	0.0375	0.0262
Hyper-Pascal	0.0161	0.0130	0.0164	0.0151	0.0133

2.4.2. Theoretische Modellierung auf Ebene der komplexen Texte

In einem zweiten Schritt folgt die Analyse der sechs komplexen Texte. Auf dieser Ebene kann bereits die Frage gestellt werden, ob die Satzlängenverteilungen bzw. deren zugrunde liegenden Modelle denen des Korpus entsprechen, oder ob es hier zu Unterschieden in der Eignung der theoretischen Modelle kommt. Aufgrund der umfangreichen Daten, werden die jeweiligen *P*-Werte bzw. (für Text # 3) der *C*-Werte im Anhang (Tabelle 1.1 und 1.2) zusammengefasst.

An dieser Stelle soll lediglich eine tabellarische Übersicht des absoluten und prozentualen Anteils der Texte, die eine zufriedenstellende Übereinstimmung mit dem Modell zeigen, geboten werden.

Tabelle 6
Anteil von Texten mit $P \geq 0.01$ bzw. $C \leq 0.02$ (abs. und prozentual)

Verteilung	Intervalle									
	1		2		3		4		5	
	abs.	%	abs.	%	abs	%	abs.	%	abs	%
Negativ-Binomial	6	100	6	100	5	83	5	83	4	67
Hyper-Poisson	5	83	5	83	4	67	4	67	4	67
Hyper-Pascal	1	17	2	33	4	67	6	100	5	83

Die Analyse der theoretischen Modelle für die Satzlängen der sechs abgeschlossenen slowenischen Kurzgeschichten zeigt folgende Ergebnisse: Die Häufigkeit der Satzlängen kann in dieser Textauswahl nicht nur durch ein einziges Modell beschrieben werden, da die Ergebnisse stark in Abhängigkeit von den vorgenommenen Zusammenfassungen variieren. So ist die *negative Binomialverteilung* für diese Texte nur ohne Zusammenfassungen bzw. bei einer Zusammenfassung zu 2er-Intervallen geeignet; eine weitere Zusammenfassung bewirkt jedoch eine Verschlechterung der Ergebnisse. Die *Hyperpoisson-Verteilung* liefert auch auf dieser Ebene keine überzeugenden Ergebnisse. Für die *Hyperpascal-Verteilung* schließlich ergibt sich der interessante Befund, dass mit zunehmender Zusammenfassung der Anteil von Texten, die dieser Verteilung folgen, steigt (außer bei Zusammenfassungen zu 5er-Intervallen). Damit zeichnet sich insgesamt an dieser Stelle der Trend ab, dass sowohl die negative Binomialverteilung als auch die Hyperpascal-Verteilung als geeignete Modelle anzusehen sind. Zu beachten ist jedoch, dass die Zusammenfassung von Satzlängen zu Intervallen auf dieser Analyseebene einen Einfluss auf die Adäquatheit bestimmter theoretischer Modelle hat.

2.4.3. Theoretische Modellierung auf Ebene von Einzeltexten

Im dritten und letzten Schritt sollen nunmehr die jeweiligen im obigen Sinne definierten Einzeltexte (vgl. Tab. 4) analysiert werden. Es handelt sich hierbei um insgesamt 21 einzelne Texte, die sich aus den einzelnen Kurzgeschichten ergeben. Wie in Tab. 7 dargestellt, zeigt sich auf dieser Analyseebene ein relativ klares Bild. So gilt es als erstes allgemein festzustellen, dass die Anpassungsergebnisse für die einzelnen Texte insgesamt besser sind als für die komplexen Texte – dies ist ein starkes Argument für die Homogenität der Daten auf dieser Analyseebene, und ein Grund für die Annahme, dass wir es nicht nur auf der Ebene des Gesamtkorpus, sondern auch schon auf der Ebene der komplexen Texte mit in unterschiedlichen Regimes resultierenden Text-Mischungen zu tun haben könnten. Im Detail kann dann weiterhin festgehalten werden, dass sowohl die *negative Binomialverteilung* als auch die *Hyperpoisson-Verteilung* bei allen Arten der Zusammenfassung passende Modelle für die Häufigkeitsverteilung der Satzlängen sind.

Tabelle 7
Anteil von Kapiteln mit $P \geq 0.01$ (abs. und prozentual)

	Intervall									
	1		2		3		4		5	
Verteilung	abs.	%	abs.	%	abs	%	abs.	%	abs	%
Negativ-Binomial	21	100	21	100	21	100	21	100	21	100
Hyper-Poisson	21	100	21	100	21	100	21	100	21	100
Hyper-Pascal	2	10	11	52	14	67	14	67	16	76

Für die *Hyperpascal-Verteilung* ist festzustellen, dass dieser ohne Zusammenfassung nicht mehr als 10% der Texte folgen; erst eine Zusammenfassung zu Intervallen ergibt eine Verbesserung des Ergebnisses, wobei allerdings auch im 5er-Intervall nicht mehr als 76% der Texte dieser Verteilung folgen. Insgesamt zeigt es sich also, dass auf dieser Ebene die Hyper-

pascal-Verteilung eine erhebliche „Sensibilität“ gegenüber der Art der Zusammenfassung aufweist. Zusammenfassend ist also davon auszugehen, dass die Art der Zusammenfassung als ein nicht unerheblicher Einflussfaktor der Satzlängenmodellierung in Betracht zu ziehen ist; es zeigt sich, dass insbesondere die Hyperpascal-Verteilung davon betroffen ist.

2.4.4. Gesamtauswertung der Ergebnisse

Im Anschluss an die Durchführung der Untersuchungen auf den drei unterschiedlichen Ebenen, bei der sich bereits deutliche Tendenzen abgezeichnet haben, können nunmehr die Ergebnisse im Hinblick auf das Gesamtkorpus, die sechs Kurzgeschichten und die 21 Einzelkapitel auch zusammenfassend dargestellt werden – dies freilich nur im Sinne einer allgemeinen Orientierung, da auch weiterhin davon auszugehen ist, dass Korpusanalysen und Einzeltextanalysen nicht zu ein und denselben Ergebnissen führen (müssen). Wie der Tab. 8 zu entnehmen ist, führt die Anpassung der *negativen Binomialverteilung* insgesamt zu den besten Ergebnissen. Dies gilt grosso modo auch für die *Hyperpoisson-Verteilung*, die sich auf der Ebene des Gesamtkorpus als gänzlich unpassend erwies, aber auf der Ebene der abgeschlossenen Kurzgeschichten und der Einzelkapitel als Modell der Satzlängenverteilung als Modell in Frage kommt. Für die *Hyperpascal-Verteilung* bestätigt sich insgesamt der bereits beobachtete Trend, dass die Zusammenfassung von (in der Anzahl der Worte pro Satz gemessenen) Satzlängen zu Intervallen einen entscheidenden Einflussfaktor darstellt: Erst durch die Zusammenfassung von Satzlängen erweist sich die Hyperpascal-Verteilung als ein passendes Modell, wobei relativierend anzumerken ist, dass die Anpassungsgüte der Hyperpascal-Verteilung in keinem Fall die guten Ergebnisse der beiden anderen Modelle übertreffen kann.

Tabelle 8
Zusammenfassung der Ergebnisse
($P \geq 0.01$ bzw. $C \leq 0.02$; absolute und prozentualer Anteil)

	Intervall									
	1		2		3		4		5	
Verteilung	abs.	%	abs.	%	abs	%	abs.	%	abs	%
Negative Binomial	27	96	28	100	27	96	27	96	26	93
Hyper-Poisson	26	93	26	93	25	89	25	89	25	89
Hyper-Pascal	4	14	14	50	19	68	21	75	22	79

Das Gesamtergebnis lässt sich wie in Abb. 1 anschaulich darstellen.

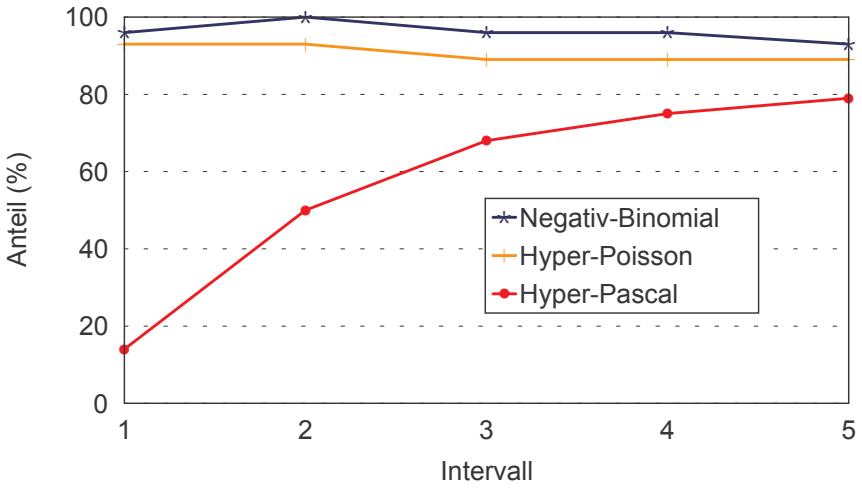


Abb. 1. Graphische Darstellung der Ergebnisse auf dem Signifikanzniveau $\alpha = 0.01$ ($P > 0.01$ bzw. $C \leq 0.02$; relativer Anteil)

3. Resümee

Die vorliegenden Untersuchungen zur Satzlänge bestätigen ein weiteres mal die hypothetisch formulierte Gesetzmäßigkeit der Häufigkeitsverteilung von Satzlängen (vgl. Altmann 1988, 1988b), gemessen in der Anzahl der Worte pro Satz. Drei in der Vergangenheit postulierte und in empirischen Arbeiten wiederholt als relevant nachgewiesene Modelle (Hyperpascal-Verteilung, Hyperpoisson-Verteilung und negative Binomialverteilung) wurden an insgesamt 28 slowenischen Datensätzen überprüft. Durch die systematische Überprüfung der Satzlängenverteilung auf drei unterschiedlichen Analyseebenen (Korpus, komplexe Texte in Form von Kurzerzählungen, sowie Einzeltexte auf der Ebene einzelner Kapitel) konnte gezeigt werden, dass von der Gattung her relativ homogene Texte unterschiedlicher Autoren im Hinblick auf die Häufigkeitsverteilung der Satzlängen durch ein einheitliches Modell beschrieben werden können.

1. Im Gesamtergebnis zeigt die *negative Binomialverteilung* insgesamt die „besten" Ergebnisse, insofern sie sich auf allen drei Analyseebenen bestens zur theoretischen Modellierung eignet und dies unabhängig von der Frage, ob und wie die Daten zu Intervallen zusammengefasst werden. Interessanterweise handelt es sich hierbei um genau jenes Modell, das Altmann (1988b) für die Verteilung der Satzlängen postuliert hat, wenn diese in der Anzahl der Teilsätze (und nicht, wie hier, in der Anzahl der Worte) pro Satz gemessen werden. Es ist daher anzunehmen, dass die Wortebene nicht prinzipiell als zusätzlicher Störfaktor bei der Modellierung der Satzlängen zu betrachten ist. Insofern bestätigt sich die auch an deutschem Sprachmaterial vorgenommene Interpretation „dass die intervenierende Ebene bei der Satz-Wort-Variante im Deutschen offenbar keine Störungen hervorruft, wie dies von Altmann noch generell vermutet wurde" (Niehaus 2001: 211).

2. Als in Frage kommendes Modell unbedingt in Betracht zu ziehen ist – zumindest (!) für slowenische Texte – auch die *Hyperpoisson-Verteilung*, die sich im Grunde genommen bei allen komplexen und individuellen Texten als bestens geeignet erwiesen hat und als Modell lediglich bei der Analyse des Gesamtkorpus nicht in Frage kam. Hier stellt sich heraus, dass ganz offenbar die Analyseebene selbst – zumindest auf der

Basis des hier untersuchten Datenmaterials – einen entscheidenden Einfluss auf die theoretische Modellierung von Satzlängenverteilungen haben kann.

3. Für die *Hyperpascal-Verteilung* schließlich konnte auf empirischen Wege gezeigt werden, dass die Zusammenfassung zu Satzlängenintervallen als ein wesentlicher Einflussfaktor bei der theoretischen Modellierung von Satzlängenverteilungen anzusehen ist. Hier gilt, dass sie unabhängig von der Analyseebene dann (und nur dann) als geeignetes Modell ins Spiel kommt, wenn die Daten zum Zwecke der Glättung zu größeren (4er- oder 5er-) Intervallen zusammengefasst werden.

Diese Schlussfolgerungen gelten zunächst einmal nur für das von uns analysierte slowenische Material.[7] Die Tragweite dieser Schlussfolgerung wird in Zukunft nicht nur an umfangreicherem Material unter Berücksichtigung auch anderer Textsorten zu überprüfen sein, sondern auch an Material aus anderen Sprachen. Dennoch sollte mit den Ergebnissen der vorliegenden Studie nachhaltig nicht nur auf die Problematik der Analyseebene, sondern auch auf die Auswirkung datenglättender Zusammenfassungen in Form von Intervallbildungen aufmerksam gemacht worden sein.

Literatur

Altmann, G. (1988a). *Wiederholungen in Texten (= Quantitative Linguistics, Vol. 36).* Bochum: Brockmeyer.

Altmann, G. (1988b). Verteilungen der Satzlängen. In: K.P. Schulz (Hrsg.), *Glottometrika 9, 147-169.*

Altmann, G. (1992). Das Problem der Datenhomogenität. In: B. Rieger (Hrsg.), *Glottometrika 13, 287-298.*

Best, K.H. (2001). Wie viele Wörter enthalten Sätze im Deutschen? Ein Beitrag zu den Sherman-Altmann-Gesetzen. In: K.H. Best (Hg.), *Häufigkeitsverteilungen in Texten: 167-201.* Göttingen: Peust & Gutschmidt.

Fucks, W. (1970). Analyse formaler Eigenschaften von Texten mit mathematischen Hilfsmitteln. In: Borck, Karl Heinz; Henss, Rudolf (Hg.), *Der Berliner Germanistentag 1968. Vorträge und Berichte: 42-52*. Heidelberg: Winter.

Grinbaum, O.N. (1996). Komp'juternye aspekty stilemetrii. In: A.S. Gerd (Hrsg.), *Prikladnoe jazykoznanie: 451-463.* Sankt Peterburg: Izdatel'stvo S.-Peterburgskogo universiteta.

Grotjahn, R., Altmann, G. (1993). Modelling the Distribution of Word length: Some Methodological Problems. In: Köhler, R. and Rieger, B.B. (eds.), *Contributions to Quantitative Linguistics: 141-153.* Dordrecht: Kluwer.

Grzybek, P. (1999). Wie lang sind slowenische Sprichwörter? Zur Häufigkeitsverteilung von (in Worten berechneten) Satzlängen slowenischer Sprichwörter. *Anzeiger für Slavische Philologie XXVII, 87-108.*

Grzybek, P., Schlatte, R. (2002). Zur Satzlänge deutscher Sprichwörter. Ein Neuansatz. In: Piirainen, E., Piirainen, I.T. (Hrsg.), *Phraseologieforschung in Raum und Zeit: 273-284.* Baltmannsweiler: Schneider.

Jakopin, F. (2002). *Entropija v slovenskih leposlovnih besedilih*. Ljubljana: ZRC SAZU.

Kaßel, A., Livesey, E. (2001). Untersuchungen zur Satzlängenhäufigkeit im Englischen: Am Beispiel von Texten aus Presse und Literatur (Belletristik). *Glottometrics 1, 27-50.*

[7] Dass die Tendenzen auch bei russischen Texten ganz ähnlich sind wie hier in Bezug auf slowenische Texte dargestellt, wurde bei Kelih (2002) nachgewiesen.

Kelih, E., Grzybek, P. (2004). Satzlänge: Definition, Häufigkeiten, theoretische Modellierung. In: A. Mehler (Ed.), *Quantitative Methoden in Computerlinguistik und Sprachtechnologie.* [= Special Issue of: LDV-Forum. Zeitschrift für Computerlinguistik und Sprachtechnologie // Journal for Computational Linguistics and Language Technology]

Kelih, E. (2002). *Untersuchungen zur Satzlänge in russischen und slowenischen Prosatexten.* Diplomarbeit in 2 Bänden. Graz.
[vgl. http://www-gewi.uni-graz.at/quanta/research/quanta_sentence.htm]

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

Niehaus, B. (2001). Die Satzlängenverteilung in literarischen Prosatexten der Gegenwart. In: L. Uhlířová, G. Wimmer, G. Altmann & R. Köhler (Eds.), *Text as a Linguistic Paradigm: Levels, Constituents, Constructs. Festschrift in honour of Luděk Hřebíček: 196-214.* Trier: WVT.

Orlov, Ju. K. (1982). Linguostatistik: Aufstellung von Sprachnormen oder Analyse des Redeprozesses? (Die Antinomie "Sprache – Rede" in der statistischen Linguistik). In: Orlov, Ju.K., Boroda, M.G., Nadarševili, I.Š. (1982), *Sprache, Text, Kunst. Quantitative Analysen: 1-55.* Bochum: Brockmeyer.

Pravopis (1990). *Slovenski pravopis. Pravila*. Ljubljana: ZRC SAZU.

Sherman, L.A. (1888). Some observations upon the sentence-length in English prose. *The University of Nebraska Studies 1, 119-130.*

Sichel, H.S. (1974). On a distribution representing sentence length in written prose. *Journal of the Royal Statistical Society A 137, 25-34.*

Williams, C.B. (1940). A note on the statistical analysis of sentence-length as a criterion of literary style. *Biometrika 31, 356-361.*

Wimmer, G., Altmann, G. (1996). The Theory of Word Length: Some Results and Generalizations. In: P. Schmidt, (ed.), *Issues in General Linguistic Theory and The Theory of Word Length: 112-133* [= Glottometrika 15.]. Trier: WVT.

Wimmer, G., Altmann, G. (2000). *Thesaurus of univariate discrete probability distributions.* Essen: Stamm.

Yule, G.U. (1939). On sentence-length as a statistical characteristic of style in prose: with application to two cases of disputed authorship. *Biometrika 30, 363-390.*

Anhang

1. Anpassungsergebnisse an die sechs komplexen Texte

1.1. *C*-Werte für Text # 3

Verteilung	**Intervalle**				
	1	2	3	4	5
Negativ Binomial	0.0086	0.0146	0.0184	0.0165	0.0121
Hyper-Poisson	0.0034	0.009	0.0199	0.0114	0.0040
Hyper-Pascal	0.0698	0.0692	0.0488	0.0038	0.0197

1.2. *P*-Werte für die Texte #1, #2, #4, #5, #6

	Negativ Binomial				
	Intervalle				
Textnr.	**1**	**2**	**3**	**4**	**5**
# 1	0.3767	0.7246	0.6146	0.8146	0.8274
# 2	0.4680	0.0610	0.0569	0.1324	0.2942
# 4	0.1845	0.2230	0.0001	0.0000	0.0020
# 5	0.4646	0.7573	0.6611	0.5292	0.2364
# 6	0.0555	0.0543	0.0532	0.0576	0.0005

	Hyperpoisson				
	Intervalle				
Textnr.	**1**	**2**	**3**	**4**	**5**
# 1	0.3041	0.5704	0.3633	0.8273	0.6818
# 2	0.0570	0.0250	0.4072	0.0681	0.0597
# 4	0.0784	0.5728	0.0000	0.0000	0.0000
# 5	0.0737	0.1756	0.1394	0.1025	0.0783
# 6	0.0000	0.0000	0.0000	0.0001	0.0000

	Hyperpascal				
	Intervalle				
Textnr.	**1**	**2**	**3**	**4**	**5**
# 1	0.000	0.000	0.485	0.504	0.0000
# 2	0.000	0.000	0.089	0.324	0.4159
# 4	0.0117	0.000	0.126	0.038	0.0167
# 5	0.0000	0.66	0.456	0.537	0.1002
# 6	0.0000	0.055	0.0000	0.061	0.1088

Anpassungsergebnisse an die 21 Einzeltexte (*P*-Werte)

Text	Negativ Binomial					Hyperpoisson				
	Intervalle					Intervalle				
	1	2	3	4	5	1	2	3	4	5
8	0.7505	0.193	0.5232	0.384	0.38	0.8114	0.181	0.5278	0.427	0.418
9	0.2347	0.444	0.2512	0.152	0.088	0.1278	0.113	0.1373	0.076	0.078
10	0.9002	0.34	0.0647	0.407	0.729	0.1555	0.102	0.3331	0.173	0.433
11	0.1834	0.201	0.3359	0.524	0.061	0.1342	0.056	0.2656	0.103	0.087
12	0.9644	0.811	0.9951	0.643	0.971	0.899	0.607	0.9595	0.696	0.92
13	0.5113	0.19	0.0634	0.558	0.072	0.2876	0.101	0.0135	0.453	0.055
14	0.1085	0.117	0.2019	0.147	0.123	0.0397	0.055	0.1271	0.241	0.041
15	0.1846	0.96	0.4997	0.852	0.426	0.1943	0.909	0.4469	0.912	0.585
16	0.6772	0.445	0.3704	0.293	0.07	0.1421	0.109	0.0776	0.089	0.427
17	0.7047	0.957	0.816	0.831	0.96	0.4694	0.708	0.5163	0.705	0.85
18	0.1201	0.425	0.0942	0.557	0.432	0.3688	0.754	0.2705	0.551	0.383
19	0.6664	0.991	0.278	0.085	0.493	0.8542	0.156	0.2861	0.166	0.371
20	0.0518	0.144	0.1543	0.06	0.045	0.1546	0.152	0.3713	0.113	0.089
21	0.1321	0.2	0.5894	0.653	0.135	0.2337	0.633	0.7811	0.815	0.251
22	0.8937	0.968	0.8915	0.688	0.885	0.7866	0.894	0.8523	0.608	0.884
23	0.8972	0.985	0.979	0.991	0.958	0.7177	0.937	0.8683	0.975	0.977
24	0.8782	0.644	0.1896	0.197	0.208	0.7422	0.385	0.1095	0.079	0.196
25	0.6377	0.663	0.8175	0.84	0.837	0.8135	0.819	0.8733	0.836	0.923
26	0.7735	0.332	0.1256	0.094	0.066	0.2599	0.078	0.453	0.453	0.799
27	0.5982	0.216	0.4004	0.271	0.79	0.8269	0.359	0.5092	0.277	0.816
28	0.6333	0.457	0.1633	0.444	0.156	0.6314	0.669	0.2288	0.471	0.181

Text	Hyperpascal					Text	Hyperpascal				
	Intervalle						Intervalle				
	1	2	3	4	5		1	2	3	4	5
8	0.0000	0.0513	0.3923	0.2690	0.0000	19	0.0000	0.0000	0.0587	0.0673	0.0688
9	0.0000	0.0117	0.4397	0.2097	0.3628	20	0.1598	0.0122	0.0477	0.0056	0.0126
10	0.0000	0.0815	0.1128	0.6353	0.8105	21	0.0000	0.0000	0.0091	0.0000	0.0002
11	0.1128	0.0000	0.0000	0.0471	0.1996	22	0.0000	0.7619	0.0413	0.1298	0.9668
12	0.0000	0.7570	0.9834	0.0000	0.9366	23	0.0000	0.4627	0.8041	0.9696	0.0000
13	0.0000	0.0531	0.0211	0.2871	0.0025	24	0.0000	0.0000	0.0000	0.0000	0.0000
14	0.0000	0.0000	0.0000	0.2351	0.5980	25	0.0000	0.0000	0.2062	0.7074	0.6542
15	0.0000	0.9266	0.2197	0.5254	0.4243	26	0.0000	0.0000	0.0000	0.1198	0.1392
16	0.0000	0.7455	0.4821	0.3198	0.3881	27	0.0000	0.0010	0.0000	0.0000	0.5115
17	0.0000	0.9890	0.8100	0.6245	0.8248	28	0.0000	0.0000	0.0830	0.0000	0.0712
18	0.0000	0.0000	0.0000	0.0000	0.2518						

Glottometrics 8, 2004, 42-67

On the acoustic elements of a poem and the formal procedures of their segmentation

A. Gumenjuk, A. Kostyshin, K. Borisov, O. Salnikova[1]

Abstract. The present paper tests the hypothesis of the element acoustic basis of a poem. Various structures of acoustic elements, selected in the composition of a poem by the pasting together some of its adjacent phonemes into specific units, which we call consonances, are presented. New modifications of one algorithm of a poetic text segmentation (Gumenjuk, Kostyshin 1999) have been described. These modifications make it possible to obtain a number of versions of the consonance vocabularies of a single poem automatically. The method for its formal comparison has been suggested. The presented variants and other possible variants of consonance vocabularies can be used by the specialists in the field of phonetic analysis as the source material for the deeper informal study of the element acoustic basis of poems and natural languages.

Keywords: phoneme, pseudo-word, acoustic element, consonance, text segmentation, segmentation algorithm, interval frequency.

On the problem of selection of elementary units in a sequence of characters

The analysis of the composition of a work of literature or a musical text on the basis of a random-probability model of a sequence of characters presented by the ideal rank-frequency distribution of Zipf-Mandelbrot implies that a researcher has at his/her disposal a chain of symbols, the elements of which are primary elements from the ontological perspective and are evident from the gnoseological perspective (i.e. are easily separated from each other) for a literate reader. In long-size literature texts these chains are words, in music texts the chains consist of elementary motives. Orlov's discovery of the exact correspondence of the Zipf-Mandelbrot's distribution to the actual rank distribution only for a complete integral text and the interval approach, which we have developed for the analysis of the chain order of a separate poem, are also realized when there are primary elementary acoustic units of a text. In general any analysis on the basis of traditional scientific approach (including the non-formal expert approach) is impossible unless an object with the complex structure is shown in the form of certain primary elements. The lack of a person's literacy and the lack of experience make it relatively difficult and even impossible to read texts due to the lack of evidence of their base of elements among other reasons of difficulties. It is known, for example, that various highly professional musical analysts independently from each other are not able to divide one and the same musical text into elementary motives, as a text of musical notes does not have intervals between 'musical words'. According to Yu. Shreider, an outstanding Russian specialist in the systems analysis, there is contradiction between the classical (element-quantity) and system methodological cognitive approaches as the details, which comprise the complex object under study, are non-obvious (Shreider, Sharov 1982). In this

[1] Address correspondence to: Alexander Gumenjuk or Alexander Kostyshin or Konstantin Borisov or Olga Salnikova, pr. Mira 11, OmGTU, Omsk, Russia, 644050. E-mail: sha@omgtu.ru, inter@omgtu.ru

case for the selection of elements it is often required to use the special procedure. The examples of such objects are literature and musical texts, for which there are no formal procedures for their segmentation into words at present. It makes it impossible for a computer to automatically read any 'unknown' text, which doesn't have spaces between words. As the graphic elements are not obvious, it is much more difficult even for professionals to "read" paintings and photographs. The computerized analysis of images is complicated for the same reason. Biochemists face the similar problem when they are able to code (write down in the form of a chain of characters) the sequence of complex organic molecules. Thus, for example, protein is a chain formed by the 'alphabet' of twenty amino acids; DNA is a long sequence, the sections of which are formed only by four different nucleotides. Recently there was information about the almost complete decoding of a human genome. In fact, there was a description of a long molecular chain in the form of characters sequence, which can be called a genetic pseudo-text, as the 'words' in this chain are yet to be defined, and the text to be written down. Applying these ideas to literature texts and musical compositions means that their primary form, written by letters, phonemes or notes is not a text. This is just a chain of characters until the words have been defined.

On the selection of consonances in poems

The direct reason for the selection of the elements-acoustic basis of a poem was an attempt to apply our interval approach to the quantitative analysis of the order of the element sequence in a poem using the same method, which used the "geometrical structure" of a simple music composition (Gumenjuk, Kostyshin, Simonova 2002). This homophone or single voice music composition in contrast to polyphonic composition is presented by a chain of separate notes, almost all the physical parameters of which are known. These are the duration and the pitch of a sound (the frequency of oscillations). Probably, the presence of these quantitative metro-rhythmic characteristics and the high musical and professional experience have enabled M. Boroda to develop the formal procedure of a single segmentation of a musical text into acoustic elements. These are the short chains of musical notes (musical words), which are called elementary motives (Boroda 1973). Many well-rhymed poems have obvious acoustic harmony. That makes it possible to suppose that the construction of such poems is similar to the one of music compositions. However, letters constituting the character sequence of a poem do not quite represent the corresponding sounds, that is why the initial acoustic basis of a text is usually represented by phonemes, which like note signs represent only the mechanical sequence of simple sounds. Since every musical texts can be presented by elementary motifs (in accordance with the procedure proposed by M. Boroda), we set up the hypothesis that a poem can be represented in the form of specific short chains of phonemes. The acoustic elements corresponding to these short chains are called consonances. So consonances can form the element acoustic basis of a poem. Words in a poem are the carriers of the meaning, but not of its acoustic basis. Separate words and word-combinations are probably constructed from consonances. Besides, words cannot be acoustic elements of a poem, as they usually appear in such short texts once, which does not correspond to the Orlov's frequency-rank criterion for the integral complete text. Also the fixed set of phonemes of a language does not meet this criterion, in accordance to which the capacity of the integral complete text's own dictionary is determined by its length (Orlov 1980).

To verify the hypothesis on the element acoustic basis of a separate poem and to select consonants in the poem there were two more modifications of the text segmentation algorithm developed, which was presented in the previous articles (Gumenjuk, Kostyshin, Simonova, 2002). One modification implies the existence of a great number of versions, each forming for

a separate poem its own vocabulary of consonances, which more or less can differ from the vocabularies of other versions of this modification of the text fragmentation algorithm.

Let us give definitions to the two critical notions used further. ***Pseudo-word*** is a random short chain of symbols (phonemes). ***An acoustic element of a poem***, ***or consonance,*** is a specific phonetic pseudo-word, which is part of the acoustic elements vocabulary of the specific poem. The capacity of the consonances vocabulary and the number of occurrences of the consonances in this particular poem meet the Orlov's criterion.

Before the computer-based experiment we had believed that the rigid limitations implied in the principle of the fragmentation algorithms would make it practically impossible to realize the correct segmentation of the phonetic representation of the poem into consonances. However, the computer experiments have shown that this fragmentation was possible to be accomplished for all the modifications and for multiple versions of the algorithm. Only in some cases the actual size of the text's own vocabulary of consonances differed by more than 1 from the size calculated using the Orlov's criterion (see Tables 3 and 8). In our opinion this fact proves the hypothesis that the element acoustic basis of a poem is in the form of consonances.

Later we will focus attention on the composition and the size of different versions of poems, whose vocabularies to some extent coincide or differ during fragmentation into consonances of one and the same poem (see Tables 1, 2, 4, 5, 6, 7, 9, 10). The wide-spread multiple-theoretic approach defines a vocabulary as an alphabet, that is a set of elements of a finite capacity with fixed composition. However, the vocabulary of any natural language is not fixed. Besides, in this research notion of a separate text's own consonances vocabulary is used and the vocabularies are obviously different for different texts. The fact is that these vocabularies also contain some general body of commonly used consonances. Is the existence of several own vocabularies acceptable in the practice of reciting of one and the same poem or do we have to admit that there exists only one "the best" vocabulary of consonances? To answer these questions it is necessary to consider the following factors: Firstly, we can assume that the "model" phonation is present by the author of the poem. However, different variants of phonation of one and the same poem are possible in practice when the poem is recited by different elocutionists. Besides, it is known that the phonation of a poem during its memorizing changes from the 'worst' emotionless variant to the 'best' expressive variant. In our opinion it shows the existence of different vocabularies belonging to one and the same poem as its rhythm is being acquired and depending on the acoustic interpretation of a poem by different elocutionists. The same phenomenon of "an authorship" is related to multivariate performance of one and the same musical composition. These facts of a human's contact with acoustic matter allow to assume that there exist various segmentations and vocabularies of consonances obtained also by the automated formal procedure of segmentation of a single poem. It is necessary to point out that the algorithm variants, which provide very similar vocabularies of consonances, probably contain the "the best" vocabulary. Besides, even the very strict Orlov's criterion (on the validity of Zipf-Mandelbrot's distribution only for an integral complete text) admits the existence of various own vocabularies for a separate text, the size of these vocabularies is determined by the text length, calculated using the same elementary units (in our case these are consonances). So there is a problem of the only possible segmentation of an obviously completed text into elementary information units.

On the basis of Zipf-Mandelbrot's distribution Ju. Orlov discovered the existence of the element basis for the complex integral chain object. Our computerized experiments show the additional corroboration of the hypothesis on the existence of element acoustic basis of a poem in the form of consonances and on Orlov's discovery of the criterion of the completeness of a text (Gumenjuk, Kostyshin, Simonova 2002).

The task setting for the segmentation of a poetic composition

Poems of Russian and English poets were the object of research. Orlov's criterion determines the integral-complete composition when there are obvious "constructive elements". We have set up the reverse task: we have tried to find the single-value segmentation of an obviously completed text into its acoustic elements called consonances.

The number of characteristics, which we call the Orlov's criterion, on the basis of which the formal identification of the integrity and the completeness of a text has been done includes the following factors in terms of priority:

1. The accuracy of the overlap of the design theoretical size and the actual size of the elements alphabet (in our case it is the vocabulary of consonances);
2. The degree of the overlap of the actual rank distribution with the one according to the Zipf-Mandelbrot's law, which can be estimated using two indexes:
    a) the maximum relative deviation of the element frequency from it's theoretical probability;
    b) the mean relative deviation of the actual frequencies of the elements occurrences from the theoretical ones.

At present the theoretical rank distribution – the Zipf-Mandelbrot's Law – is presented by the following formulas (Orlov 1980):

(1) $$p_{rT} = \frac{D}{(B+r)^{\gamma}}; \quad D, B, \gamma - const; (\gamma \approx 1);$$

(2) $$D = \frac{1}{\ln F_1};$$

(3) $$B = \frac{D}{p_1} - 1;$$

(4) $$v_T = DZ - B,$$

where p_{rT} – the design theoretical frequency of the rth word according to the word rank (in our research: consonance rank);

F_1 – the actual number of occurrences of the most frequent word; $\gamma \approx 1$;

p_1 – the relative frequency of the most frequent word;

v_T – the design theoretical size of the text's vocabulary, when the text length is *Z*.

The text length of the poem will be measured by the number of consonances used, which is similar to the fact that the length of a large-size text is determined by the number of words being used.

The methodology of conversion of graphemes into phonemes has got a new impulse in the period of the development of computer-based systems of information processing. From the beginning of the 1980s it was the focus of research by many linguists. At present there are several ways of converting written texts into phonetic texts. The selection of the optimal method of conversion of graphemes into phonemes has been a disputable issue among linguists in the world. Especially this concerns such languages as Chinese and Japanese. Among European languages the most difficult for conversion of graphemes into phonemes are English and French. For the purposes of the present research we have used the method suggested by Blatter (1980). This method uses the International Phonetic Alphabet (IPA). The advantage of using the computer-based method of segmentation of an English text into phonetic units is that it meets international standards. When the words are transcribed using

common English dictionaries, there are some differences between different variants of English language. For example, the transcription of one and the same word may differ in Oxford English Dictionary and Webster's Dictionary depending whether it is British or American English. The poetic text contained some historical words, which were not transcribed using the computer-based method, so in these cases we used Oxford English Dictionary.

The reasons why we have chosen the IPA standard instead of the standard set of phonetic symbols were the following:

1) This method allows to take into consideration the differences in the pronunciation of sounds at word boundaries, i.e. the neibouring words are taken into consideration.
2) No extra software for recording and processing phonetic texts is required.
3) Using IPA, the automated conversion of written texts into phonetic texts is possible.
4) When written texts are converted into phonetic symbols a number of subjective decisions connected with the differences between variants of English language has to be made, for example, the differences between British, American and Australian English.

The segmentation method suggested by the authors can also depend on other methods of converting written texts into phonetic texts. Linguists can be interested in the predominance of certain sounds and consonances in different poems of different authors. For example, the dominance of front and aft vowels, of voiced and unvoiced consonants in different lines of a poem, which, in combination with the rhyme and rhythm creates a certain emotional state, which has been implied by Blatter (1980).

The obtained automated phonetic compilation of poems presented in the form of the basic continuous chain was used for segmentation. For the selection of consonances in this chain the formal procedure similar to the one described by Borodovskij, Pevzner (1990) was used; short sequences of symbols with extremely high or low frequency of occurrences in the text are defined as "words" in this algorithm. In the present research a consonance is defined as an acoustic element of a poem, constructed in the form of a very short pseudo-word, which consists of phonemes.

The segmentation of a phonetic text into consonances

The procedure of splitting a phonetic representation of a text into consonances implies that the source text is a sequence of elements, which are differentiated here only as symbols having no semantic component (Gumenjuk, Kostyshin, Simonova 2002). Let B_1 be a phoneme symbol being part of a pseudo-word, which is constituted by a random chain of phonemes. The original numerical sequence of phonemes is scanned and a pseudo-word of the length k is segmented to test the role of consonance. The first pseudo-word in the text is taken $W = <B_1,...,B_n>$, and its possible occurrences in the text are determined.

The values of deviation $std(W_j)$ of actual and design characteristics of the tested pseudo-word occurrences in the text are presented in the following way:

$$(5) \qquad std(W_j) = \frac{\left| p(W_j) - p_c(W_j) \right|}{\left(p_c(W_j)\right)^{1/2}},$$

where $W_j - j^{\text{th}}$ order of tests of pseudo-word
(some short chain of symbols (phonemes) $<B_1, B_2, ...B_k>$),

$p(W_j)$ – the actual characteristics of occurrences of a pseudo-word W_j in the text,
$p_c(W_j)$ – the design characteristics of occurrences of a pseudo-word W_j.

The design characteristics of occurrences of a pseudo-word treated as a random chain of symbols-phonemes are calculated via actual characteristics of its three fragments (sub-words) in the following form:

$$p_c(B_1,...,B_k) = \frac{p(B_1,...,B_{k-1}) \times p(B_2,...,B_k)}{p(B_2,...,B_{k-1})}; \tag{6}$$

for a pseudo-word consisting of two symbols the design characteristics is determined in the following form:

$$p_c(B_1B_2) = p(B_1) \times p(B_2). \tag{7}$$

The calculation of the actual and design characteristics of pseudo-word occurrences in the text for different modifications of the segmentation algorithm is given below.

The considerable deviation of the characteristics of pseudo-word occurrences $std(W_j)$ shows the determined origin of the selected pseudo-word in contrast to the symbols randomly pasted into pseudo-words. In this case the short sequence of phonemes or a pseudo-word is considered to be an acoustic element of a text – a consonance. So, if $std(W) \geq P$, where P is the value of some threshold ($0 < P < 1$), then a pseudo-word is determined as a consonance and is put into the vocabulary of consonances. The next pseudo-word $<B_{k+1},...,B_{2k}>$ is tested if it is a consonance. If the deviation of the first pseudo-word to be tested is $std(W) < P$, than the next is the pseudo-word $<B_2,...,B_{k+1}>$, and $std(W)$ is again calculated for this chain of phonemes-symbols. These actions and tests are done for all pseudo-words of the same original sequence of phonemes.

These actions and tests are done for all pseudo-words of the original sequence of phonemes. A newly segmented consonance or its fragments cannot be part of the other consonances. The consonance entered into the vocabulary is marked through the text as an original acoustic element (Fig. 1. Marked in black). By reaching the end of the text the size of the selected pseudo-word k is decreased by 1 and there is the return to text beginning. This procedure is repeated until the complete segmentation of the symbol-phoneme sequence into consonances is comprised in the text. The remaining parts of the source sequence, which are not comprised into any of the consonances, are determined as single-symbol consonances (separate phonemes). The value of the decision point P on the forming (creating) of consonances varies after the next complete segmentation of a text until such vocabulary of consonances is formed whose volume meets Orlov's criterion. So segmentation of a phonetic representation of a text is done in steps by selecting shorter and shorter consonances.

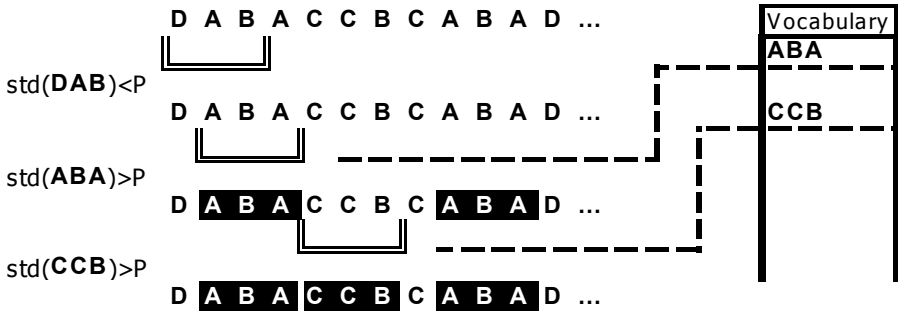


Fig. 1. The procedure of the pseudo-words selection

Figure 1 shows part of the procedure of the text segmentation as the result of which two three-symbol pseudo-words, "ABA", "CCB", are put into the vocabulary of consonances. When the further segmentation is performed, they are not changed, being shown by the black color on the Figure. The symbols comprising the selected consonances cannot be included into other pseudo-words.

Versions and modifications of the segmentation algorithm

The deviation between the design and the actual characteristic of the occurrences of a pseudo-word in a text, *std*(*W*), can be determined on the basis of the frequency characteristics of the text, as well as with the help of the here suggested interval and combined characteristics. Let us describe the methods of calculation of characteristic *std*(*W*), which corresponds to the algorithm described above and its modification.

The segmentation algorithm of the phonetic representation of a text by selecting consonances in it will be later referred to as algorithm.

1. Formula of frequency characteristics of pseudo-word occurrences, which is used in the known algorithm

In the given segmentation algorithm the characteristic of the pseudo-word is its frequency of occurrences in the text. This frequency is determined by the ratio of the number of occurrences of a pseudo-word in the text n_j to the total number of elements n, contained in this text at the moment when the next consonance is selected (10). The value n is determined by the sum of the number of consonances formed by that moment and consisting of separate phonemes (Gumenjuk, Kostyshin, Simonova 2002).

2. Determining an interval characteristic of pseudo-word occurrences used in the first modification of an algorithm

This characteristic takes into account not only the number of occurrences of a pseudo-word but also their mutual positions in the text; in some cases it is similar to the frequency characteristic but depends on the “geometrical” characteristic of a text; i.e. on the sequence of pseudo-words.

For this modification of the algorithm the numerical characteristic of the homogeneous sequence of pseudo-words W_j is determined by the formula:

(8) $$p_g(W_j) = \frac{1}{\Delta_{gj}},$$

where Δ_{gj} is the geometrical interval between the two closest occurrences of the selected W_j, which is determined by the following formula

(9) $$\Delta_{gj} = \sqrt[n_j]{\prod_{i=1}^{n_j} \Delta_{ij}},$$

where Δ_{ij} is the interval between the two closest occurrences of the selected pseudo-word Wj, determined by the number of other intervening pseudo-words plus one.

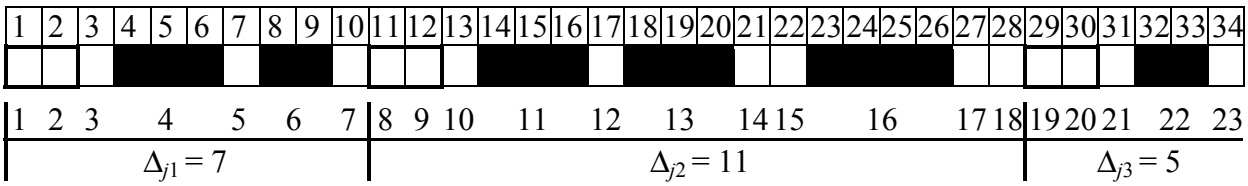
Thus $p_g(W_j)$ is the value inverse to the mean geometrical interval between the closest occurrences of the pseudo-word W_j. This takes into account the geometrical parameters of the segmented text.

Let us call this characteristic the **interval frequency** by the analogy with the ordinary frequency, which is presented by the value inverse to the simple mean interval for the selected pseudo-word in the following form

(10) $$p(W_j) = \frac{n_j}{n} = \frac{1}{n/n_j} = \frac{1}{\Delta_{aj}},$$

where n_j is the number of occurrences of a pseudo-word W_j;
n is the length of the text – the number of words used.

Below there is an example of determining the design and interval characteristics of a pseudo-word W_j occurrences in the next possible structure of a text being segmented at the stage of the next consonance.



The original length of phonetic representation here is 34 phonemes.

The black boxes are the previously segmented consonances consisting of two, three and four phonemes. There are 6 consonances. Let elements present separate phonemes.

The length of a text at the stage of selection of the next consonance is $n = 6 + 17 = 23$ elements. The number of elements of a text in the process of such segmentation is changed (is reduced).

The frame shows the next segmented pseudo-word, consisting of two phonemes. The number of its entries n_j is 3; the frequency of the pseudo-word occurrences is $p(W_j) = 3 / 23 = 0.130$; Δ_{j1}, Δ_{j2}, Δ_{j3} – intervals between the adjacent identical pseudo-words W_j; mean

geometrical interval $\Delta_{gj} = \sqrt[3]{7 \cdot 11 \cdot 5} = 7.275$; interval frequency of a pseudo-word $p_g(W_j)$ = 1 / 7.275 = 0.137.

Numerical characteristics of fragments (sub-words) of a pseudo-word, presented in formulae (6) and (7) are determined in the same way.

3. Determining of combined characteristics of pseudo-word occurrences used in the second modification of the algorithm and the number of its versions

This value is the sum of various balanced characteristics of a pseudo-word W_j, presented by

(11) $$\bar{p}(W_j) = K \cdot p_g(W_j) + (1 - K)\, p(W_j),$$

where $\bar{p}(W_j)$ – combined (frequency) characteristics;
$p_g(W_j)$ – interval frequency of a pseudo-word W_j;
$p(W_j)$ – the frequency of occurrences of a pseudo-word W_j;
K – balance factor $0 \le K \le 1$ (which sets the specific version of the present algorithm modification).

As the simple mean is not less than geometrical mean $\Delta_{aj} \ge \Delta_{gj}$, the frequency of a pseudo-word does not exceed the interval frequency of a pseudo-word:

(12) $$p(W_j) = \frac{1}{\Delta_{aj}} \le \frac{1}{\Delta_{gj}} = p_g(W_j).$$

So the combined frequency $\bar{f}(W_j)$ is larger than the frequency of the pseudo-word occurrences but less than interval frequency. The factor K determines the value of interval information in the combined characteristics and sets the corresponding version of the algorithm modification. The extreme values of the factor K give both frequency characteristics (K = 0) and interval characteristics (K = 1) .

The presented actual characteristics of the occurrences of a pseudo-word and its fragments (sub-words), $p(W_j)$, $p_g(W_j)$, $\bar{p}(W_j)$, are used in the formulae (6), (7) and (5) depending on the modification and the version of the algorithm of the text segmentation for determining the corresponding design characteristic $p_c(W_j)$ and the value of the standard deviation of characteristic, $std(W_j)$.

Determining the accuracy of segmentation

After the next complete segmentation of the text into consonances the algorithm of segmentation determines the correctness of segmentation (Gumenjuk, Kostyshin, Simonova 2002). For the evaluation of the correctness of segmentation Orlov's partial criterion is used. It requires the congruence of the design and the actual sizes of the consonances vocabularies (Orlov 1980). If it is not possible to provide the equality of vocabulary sizes, then that variant of segmentation is chosen, which corresponds to the minimal difference of these sizes.

Then with each iteration of several text segmentations the value of the threshold *P* varies until the segmentation gets the closest to the 'correct' one. To boost the algorithm processing it was considered that when the threshold value is increased, the difference of design and the actual vocabulary sizes was increased on average. That is why to find the threshold *P* the dichotomic procedure was used and then the value varied within minor limits. Similar organization of the selection of *P* was much faster (by several degrees) than the consecutive search. Thus the present algorithm is the optimal search procedure for the search of the correct segmentation of a poem using Orlov's criterion. A search simple dichotomic procedure (without final variation) cannot be applied as the dependence of the difference of the sizes of theoretical and actual vocabularies is still not monotonous.

The experiments of poem segmentations showed that some of the texts have peculiarities of the structure, such as sudden change of the size of the actual vocabulary near the sought point. For some of the texts this anomaly made it impossible to provide the 'correct' segmentation.

The study of the possibilities of the algorithm and its modifications

For the successful application of the segmentation algorithm described above, most critical is the size and the completeness of poems. The texts studied were written using the phonemes of Russian and English languages. The ends of lines were marked in this process. The segmentation was done in a way that the ends of the lines were not included in the resulting consonances. So the pause at the end of the poem's line was considered, which increases the soundness of segmentation.

Computer experiments showed that for the absolute majority of the processed texts (about 40 versions of three Russian and five English poems) the program has performed the correct segmentation based on the Orlov's criterion, i.e. the absolute value of the difference between the sizes of theoretical and actual vocabularies was equal 0 or 1 (see Tables 3 and 8). These results also prove the practical applicability of the dichotomic search procedure of the threshold value *P*.

Segmentation results, numerical characteristics and the comparative analysis of the consonances vocabularies obtained upon the different segmentations of poem texts

Each poem has been segmented using all modifications of the algorithm as well as on the basis of a number of versions of the combined characteristics. Further there is comparative analysis of the obtained consonance dictionaries presented.

To compare the possibilities of the algorithm modifications which use only frequency or interval frequency there was the study of the obtained vocabularies of consonances. The overlapped and different parts for different consonance vocabularies were obtained (see Tables 1 and 6). The degree of the overlap of each version of the actual and the design dictionaries, which prove the correctness of the text segmentation using the Orlov's criterion is presented in the left part of Tables 3 and 8. Each version of segmentation is characterized by: actual and design sizes of vocabularies (v and v_T), relative difference of vocabulary sizes δ_V, relative difference of the actual and the design distribution determined by the formula :

$$\delta_{me} = \frac{1}{v}\sum_{r=1}^{v}\frac{|p_r - p_{rT}|}{p_{rT}}, \tag{13}$$

where p_r – actual frequency of consonance occurrences of the r-th rank in the obtained vocabulary;
p_{rT} – design frequency of the consonance of r-th rank according to the Mandelbrot distribution;
v – total number of consonances in the vocabulary.

Besides, to demonstrate the truth of segmentation of poems Figures 2 and 3 show the plots of actual and theoretical distributions of consonances of the two segmentations obtained on the basis of frequency characteristics.

Frequency-rank vocabularies of consonances stating the number of occurrences of each consonance in the corresponding poem's segmentations are also plotted (Tables 2 and 7).

Segmentations of one and the same text done using different modifications and versions of the algorithm were slightly different from each other.

The contents of vocabularies obtained using frequency and interval modification of the algorithm are rather different. The number of identical consonances in such dictionaries is on the average 50–60%. At the same time the most frequent consonances for different versions of segmentation of one and the same text are slightly different (see Tables 1 and 6).

William Blake
THE GARDEN OF LOVE

Original text :	Phonetic presentation:
I went to the Garden of Love,	ie w e n t t ue th u g aar d i n u v l u v
And saw what I never had seen:	a n d s au w u t ie n e v er h a d s ee n
A Chapel was built in the midst,	u ch a p oo l w u z b i l t i n th u m i d s t
Where I used to play on the green.	w air ie y ue z d t ue p l ae aa n th u g r ee n
And the gates of this Chapel were shut	a n d th u g ae t s u v th i s ch a p oo l w er sh u t
And "Thou shalt not" writ over the door;	a n d th ou sh a l t n aa t r i t oe v er th u d or
So I turn'd to the Garden of Love	s oe ie t er n t t ue th u g aar d i n u v l u v
That so many sweet flowers bore;	th a t s oe m e n ee s w ee t f l aaw er z b or
And I saw it was filled with graves,	a n d ie s au i t w u z f i l d w i thh g r ei v z
And tomb-stones where flowers should be;	a n d t ue m s t oe n z w air f l aaw er z sh oo d b ee
And Priests in black gowns were walking their rounds,	a n d p r ee s t s i n b l a k g ou n z w er w au k ee ng th air r ou n d z
And binding with briars my joys and desires.	a n d b ie n d ee ng w i thh b r ie y er z m ie j oi z a n d d i zz ier z

The results of the segmentation of phonetic representation of a poem:

Segmentation based on frequency characteristics:

ie w e n **t** tue thu gaar d **i n uv l uv**
an d **sau** wu t ie n ev erh a d s ee n
uch ap ool wu z b i **l t i** n thu m **i** d s **t**
wair iey uez d tue p **l aeaa** n thu gr ee n
an d thu gae **t s uv th i sch ap ool** wer shu **t**
an d thou sha **l** tnaa **t r i t** oev er thu d **or**
soe ie t er n t tue thu gaar d **i n uv l uv**
tha **t soe me n ee s w ee t** fl aawer **z b or**
an d **ie sau i t wu** zf **i l d w** ithh gr eiv **z**
an d tue m **s t oe n z** wair fl aawer zsh oo **d b ee**
an d pr **ee s t s i n b l** ak gou **n z** wer **w auk eeng** thair rou n d **z**
an d **b ie** n d **eeng w** ithh **b r** iey **er z m** iej oiz an d **d** izz ier **z**

Segmentation based on the mean geometrical intervals:

ie went **t** ueth ug aard **i n uv l uv**
a nd **sau** wut iene ver ha ds een
uch ap ool wu z bi **l t i** nth um **i** ds **t**
w airie yue zd t uep **l aeaa** nth ug r een
a nd th ug ae **t s uv th i sch ap ool** w ersh u **t**
a nd th oush a **l** t n aa **t r i t** oe ver th ud **or**
soe ie t er n t t ueth ug aard **i n uv l uv**
th a **t soe me n ee s w ee t** f laaw er **z b or**
a nd **ie sau i t wu** z f **i l d w** i thhg rei v **z**
a nd t uem **s t oe n z** w air f laaw er z shoo **d b ee**
a nd p r **ee s t s i n b l** a kg ou **n z** w er **w auk eeng** th air r ou nd **z**
a nd **b ie** nd **eeng w** i thh **b r** ie y **er z m** ie joi z a nd **d** i zzier **z**

Here and in the further segmentations of poems the overlap of different modifications of vocabularies is presented.

Table 1
The compared vocabularies of consonances

Vocabulary of segmentation obtained on the basis of frequency characteristics		
aawer ak an e eiv erh ev fl gaar gae gou gr iej ier iey ithh izz oev oiz oo pr rou sha shu tha thair thou thu tnaa tue uez wair wer zf zsh	a aeaa ap auk b d ee eeng er i ie l m me n oe ool or p r s sau sch soe t th uch uv w wu z	aa aard ae air airie bi ds een ersh f ha iene joi kg laaw nd nth ou oush rei shoo thh thhg u ud uem uep ueth ug um v ver went wut y yue zd zzier
	Vocabulary of segmentation obtained on the basis of interval characteristics	

Table 2
Frequency-rank distribution of consonance vocabularies of the segmentation of the poem *The Garden of Love*

r	*K* = 0	*K* = 0.6	*K* = 0.7	*K* = 0.8	*K* = 1
1	D.....19	T.....19	T.....19	T.....18	T.....18
2	N.....15	I.....12	I.....12	I.....12	I.....12
3	T.....15	Z.....12	A.....11	A.....11	A.....11
4	I.....10	A.....11	Z.....11	Z.....11	Z.....11
5	AN.....8	N.....10	ND.....10	ND.....10	ND.....10
6	Z.....8	ND.....10	N.....8	W.....10	N.....8
7	L.....7	ER.....7	S.....8	N.....8	W.....8
8	S.....7	L.....7	IE.....7	L.....7	L.....7
9	B.....6	S.....7	L.....7	IE.....6	IE.....6
10	EE.....6	D.....6	ER.....6	TH.....6	TH.....6
11	THU.....6	W.....6	W.....6	ER.....5	B.....5
12	IE.....5	B.....5	TH.....5	S.....5	ER.....5
13	UV.....5	EE.....5	UV.....5	UV.....5	R.....5
14	W.....5	IE.....5	B.....4	B.....4	S.....5
15	TUE.....4	TH.....5	EE.....4	EE.....4	UV.....5
16	ER.....3	UV.....5	OE.....4	R.....4	EE.....4
17	M.....3	UG.....4	UG.....4	U.....4	UG.....4
18	WU.....3	F.....3	D.....3	UG.....4	D.....3
19	AAWER.....2	R.....3	F.....3	D.....3	F.....3
20	AP.....2	WU.....3	R.....3	F.....3	AARD.....2
21	EENG.....2	AARD.....2	WU.....3	AARD.....2	AIR.....2
22	FL.....2	AP.....2	AARD.....2	AIR.....2	AP.....2
23	GAAR.....2	EENG.....2	AP.....2	AP.....2	DS.....2
24	GR.....2	IEY.....2	DS.....2	DS.....2	EEN.....2
25	IEY.....2	LAAW.....2	EEN.....2	EEN.....2	EENG.....2
26	ITHH.....2	NTH.....2	EENG.....2	EENG.....2	LAAW.....2
27	OOL.....2	OOL.....2	LAAW.....2	LAAW.....2	NTH.....2
28	OR.....2	OR.....2	NTH.....2	NTH.....2	OE.....2
29	R.....2	OU.....2	OOL.....2	OE.....2	OOL.....2
30	SAU.....2	SAU.....2	OU.....2	OOL.....2	OR.....2
31	SOE.....2	SOE.....2	SAU.....2	OU.....2	OU.....2
32	WAIR.....2	UETH.....2	UETH.....2	SAU.....2	SAU.....2
33	WER.....2	WAIR.....2	VER.....2	SOE.....2	SOE.....2
34	A.....1	AA.....1	WAIR.....2	UETH.....2	UETH.....2
35	AEAA.....1	AD.....1	AA.....1	VER.....2	VER.....2
36	AK.....1	AE.....1	AE.....1	AA.....1	WU.....2
37	AUK.....1	AEAA.....1	AEAA.....1	AE.....1	AA.....1
38	E.....1	AIRR.....1	AIRR.....1	AEAA.....1	AE.....1
39	EIV.....1	AUK.....1	AUK.....1	AIRIE.....1	AEAA.....1
40	ERH.....1	BI.....1	BI.....1	AUK.....1	AIRIE.....1
41	EV.....1	CH.....1	BOR.....1	BI.....1	AUK.....1
42	GAE.....1	ERH.....1	CH.....1	BOR.....1	BI.....1
43	GOU.....1	EV.....1	DOR.....1	DOR.....1	ERSH.....1
44	IEJ.....1	IEN.....1	HA.....1	ERSH.....1	HA.....1
45	IER.....1	JOI.....1	IENE.....1	HA.....1	IENE.....1
46	IZZ.....1	KG.....1	JOI.....1	IENE.....1	JOI.....1

47	ME 1	M 1	KG 1	JOI 1	KG 1
48	OE 1	ME 1	M 1	KG 1	M 1
49	OEV 1	OE 1	ME 1	M 1	ME 1
50	OIZ 1	OEV 1	OUSH 1	ME 1	OUSH 1
51	OO 1	OUSH 1	PR 1	OUSH 1	P 1
52	P 1	PR 1	REI 1	PR 1	REI 1
53	PR 1	REI 1	SHOO 1	REI 1	SCH 1
54	ROU 1	SEE 1	SHU 1	SCH 1	SHOO 1
55	SCH 1	SHOO 1	THH 1	SHOO 1	THH 1
56	SHA 1	SHU 1	THHG 1	THH 1	THHG 1
57	SHU 1	THH 1	THU 1	THHG 1	U 1
58	TH 1	THHG 1	UCH 1	UCH 1	UCH 1
59	THA 1	THU 1	UEM 1	UEM 1	UD 1
60	THAIR 1	UCH 1	UEP 1	UEP 1	UEM 1
61	THOU 1	UE 1	UM 1	UM 1	UEP 1
62	TNAA 1	UEM 1	V 1	V 1	UM 1
63	UCH 1	UEP 1	WENT 1	WENT 1	V 1
64	UEZ 1	UM 1	Y 1	WUT 1	WENT 1
65	ZF 1	V 1	YUE 1	Y 1	WUT 1
66	ZSH 1	WENT 1	ZD 1	YUE 1	Y 1
67		ZZIER 1	ZZIER 1	ZD 1	YUE 1
68				ZZIER 1	ZD 1
69					ZZIER 1

Table 3
The characteristics of the text, obtained via segmentations using various versions of the segmentation algorithm for the poem *The Garden of Love*

№	K	Z	F_1	p_1	v	v_T	δ_v,%	δ_{me}%	Hs	G	H	g	g-H	2^{g-H}
1	0.0	198	19	0.096	66	65	1.54	17.35	1058.6	901.1	5.3	4.6	0.80	0.58
2	0.6	205	19	0.093	67	67	0.00	18.63	1106.8	914.3	5.4	4.5	-0.94	0.52
3	0.7	202	19	0.094	67	66	1.52	15.98	1095.3	892.9	5.4	4.4	-1.00	0.50
4	0.8	203	18	0.089	68	68	0.00	16.35	1106.4	899.7	5.5	4.4	-1.02	0.49
5	1.0	203	18	0.089	69	68	1.47	16.25	1112.4	905.1	5.5	4.5	-1.02	0.49

K – balance factor which determines the version of the text segmentation algorithm;
Z – length of a poem determined by the number of consonances;
F_1 – number of occurences of the most frequent consonance;
p_1 – relative frequency of the most frequent consonance;
v – number of various consonances in the present segmentation of a poem (the size of the actual vocabulary);
v_T – the size of the estimated vocabulary calculated using formula (4) of Zipf-Mandelbrot's law;
δ_v – deviation of the size of the actual vocabulary of consonances from the design one;
δ_{me} – mean relative deviation of the actual rank distribution of consonances in text from Zipf-Mandelbrot's distribution;
Hs – amount of information in a text (in bits);
H – (unconditional) entropy or the amount of information in one consonance; it is calculated in the supposed static independence of consonances by dividing Hs by Z;
G – the depth of arrangement of all consonances of a text is measured in the same way as entropy, in bits (Gumenjuk, Kostyshin, Simonova 2002);
g – mean remoteness of separate consonances in a text; it is calculated by dividing G by Z (Gumenjuk, Kostyshin, Simonova 2002);
2^{g-H} – regularity of a text; this numerical characteristic is presented for the evaluation of the regularity of all the selected elements (here consonances) in the text and is determined by the formula:

(14) $$2^{g\text{-}H} = \frac{\Delta_g}{\Delta_{gm}}$$

where Δ_g, Δ_{gm} are the mean geometrical and the maximal mean geometrical intervals of all elements of a text.

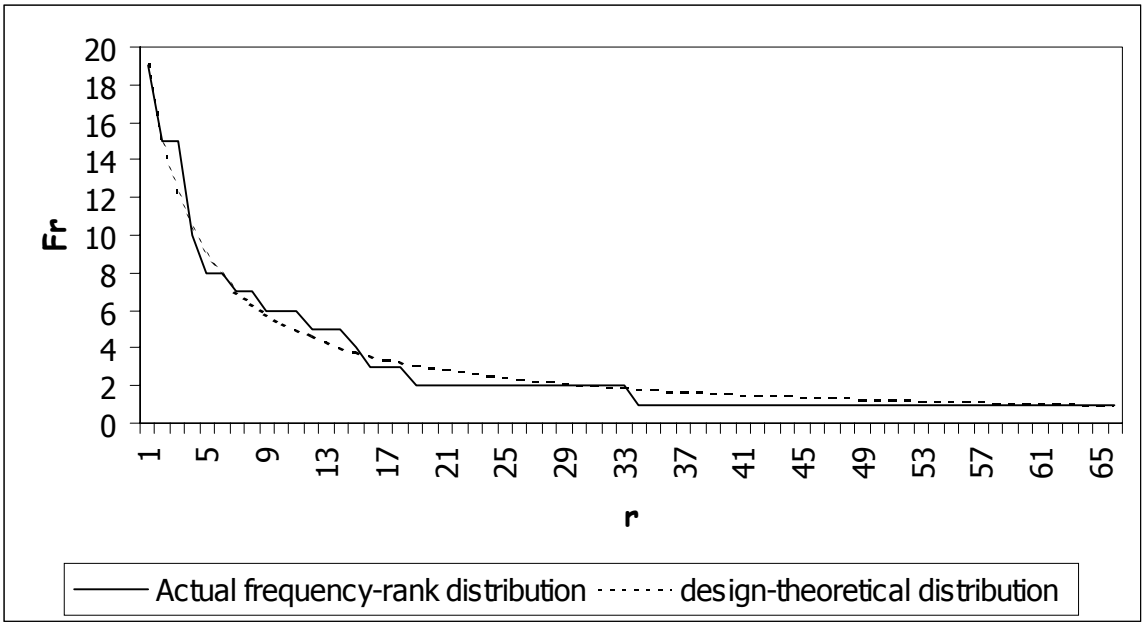


Figure 2. Actual and design theoretical (Zipf-Mandelbrot's) distributions for the frequency version of the segmentation of the poem *The Garden of Love.*
F_r – the number of occurrences of a consonance of rank r in a text;
r – the consonance rank.

Segmentation results for $K = 0.6$ *and* $K = 1$

Segmentation for $K = 0.6$

ie went t ueth ug aard i n uv l uv
a nd sau wu t ien ev erh ad see n
uch ap ool wu z bi l t i nth um i d s **t**
wair iey ue z d **t uep l aeaa nth ug r** ee n
a nd th ug ae t s uv th i s ch ap **ool w** er shu **t**
a nd th oush a l t n aa t r i t oev er thu d **or**
soe ie t er n t t ueth ug aard i n uv l uv
th a t soe me n ee s w ee t f laaw er z b or
a nd ie sau i t wu z f i l d w i thhg rei v z
a nd t uem s t oe n z wair **f laaw er z shoo d b ee**
a nd pr **ee s t s i n b l a kg ou n z w er w auk eeng th** airr **ou nd z**
a nd b ie nd eeng w i thh b r iey **er z m ie joi z a nd d i zzier z**

Segmentation for $K = 1$

ie went t ueth ug aard i n uv l uv
a nd sau wut iene ver ha ds een
uch ap ool wu z bi l t i nth um i ds **t**
w airie yue zd **t uep l aeaa nth ug r** een
a nd th ug ae t s uv th i sch ap **ool w** ersh u **t**
a nd th oush a l t n aa t r i t oe ver th ud **or**
soe ie t er n t t ueth ug aard i n uv l uv
th a t soe me n ee s w ee t f laaw er z b or
a nd ie sau i t wu z f i l d w i thhg rei v z
a nd t uem s t oe n z w air **f laaw er z shoo d b ee**
a nd p r **ee s t s i n b l a kg ou n z w er w auk eeng th** air r **ou nd z**
a nd b ie nd eeng w i thh b r ie y **er z m ie joi z a nd d i zzier z**

Table 4
The compared vocabularies of consonances

Segmentation vocabulary obtained for $K = 0.6$		
ad airr ch erh ev ien iey oev pr see shu thu ue wair	a aa aard ae aeaa ap auk b bi d ee eeng er f i ie joi kg l laaw m me n nd nth oe ool or ou oush r rei s sau shoo soe t th thh thhg uch uem uep ueth ug um uv v w went wu z zzier	air airie ds een ersh ha iene p sch u ud ver wut y yue zd
	Segmentation vocabulary obtained on the basis of interval characteristic.	

Segmentation results for K = 0.7 *and K* = 0.8

Segmentation for *K* = 0.7

ie went t ueth ug aard i n uv l uv
a nd sau wu t **iene ver ha ds een**
uch ap ool wu **z bi l t i nth um i ds t**
wair ie **yue zd t uep l aeaa nth ug r een**
a nd th ug ae t s uv th i s ch **ap ool** w er shu **t**
a nd th oush a l t n aa t r i t oe ver thu **dor**
s oe **ie t er n t t ueth ug aard i n uv l uv**
th a t s oe **me n ee s w ee t f laaw er z bor**
a nd ie sau i t wu **z f i l d w i thhg rei v z**
a nd t uem s t oe n z wair **f laaw er z shoo d b ee**
a nd pr ee s t s i n b l a kg ou n z w er w auk eeng th airr **ou nd z**
a nd b ie nd eeng w i thh b r ie y er z m ie joi z a nd d i zzier z

Segmentation for *K* = 0.8

ie went t ueth ug aard i n uv l uv
a nd sau wut **iene ver ha ds een**
uch ap ool w u **z bi l t i nth um i ds t**
w airie **yue zd t uep l aeaa nth ug r een**
a nd th ug ae t s uv th i sch **ap ool** w ersh u **t**
a nd th oush a l t n aa t r i t oe ver th u **dor**
soe **ie t er n t t ueth ug aard i n uv l uv**
th a t soe **me n ee s w ee t f laaw er z bor**
a nd ie sau i t w u **z f i l d w i thhg rei v z**
a nd t uem s t oe n z w air **f laaw er z shoo d b ee**
a nd pr ee s t s i n b l a kg ou n z w er w auk eeng th air r **ou nd z**
a nd b ie nd eeng w i thh b r ie y er z m ie joi z a nd d i zzier z

Table 5
The compared vocabularies of consonances

Segmentation vocabulary obtained for *K* = 0.7		
airr ch shu thu wair wu	a aa aard ae aeaa ap auk b bi bor d dor ds ee een eeng er f ha i ie iene joi kg l laaw m me n nd nth oe ool ou oush pr r rei s sau shoo t th thh thhg uch uem uep ueth ug um uv v ver w went y yue z zd zzier	air airie ersh sch soe u wut
	Segmentation vocabulary obtained for *K* = 0.8	

George Gordon Byron
ON THIS DAY I COMPLETE MY THIRTY-SIXTH YEAR

Original text:	Phonetic presentation:
This time this heart should be unmoved,	th i s t ie m th i s h aar t sh oo d b ee a n m u v d
Since others it hath ceased to move;	s i n t s u th er z i t h a thh s ee s d t ue m ue v
Yet, though I cannot be beloved.	y e t th oe ie k a n aa t b ee b ee l u v i d
Still let me love!	s t i l l e t m ee l u v
My days are in the yellow leaf;	m ie d ae z aar i n th u y e l oe l ee f
The flowers and fruits of love are gone:	th u f l aaw er z a n d f r ue t s u v l u v aar g au n
The worm, the canker, and the grief	th u w er m th u k ae n k er a n d th u g r ee f
Are mine alone!	aar m ie n u l l oe n
The fire that on my bosom prey	th u f ier th a t aan m ie b oo z i m p r ae
Is lone as some volcanic isle;	i z l oe n a z s u m v aa l k k a n i k ie l
No torch is kindled at its blaze -	n oe t or ch i z k i n d oo l d a t i t s b l ae z
A funeral pile	u f y ue n e r oo l p ie l
The hope, the fear, the jealous care,	th u h oe p th u f eer th u j e l i s k air
The exalted portion of the pain	th u i g zz au l t i d p or sh i n u v th u p ae n
And power of love, I cannot share,	a nd p aa w er u v l u v ie k a n aa t sh a i r
But wear the chain.	b u t w air th u ch ae n
But 'tis not thus - and 'tis not here -	b u t t i s n aa t th u s a n d t i s n aa t h eer
Such thoughts should shake my soul, nor now,	s u ch thh au t s sh oo d sh ae k m ie s oe l n or n ou
Where glory decks the hero's bier,	w air g l oo r i d e k s th u h er ou z b eer
Or binds his brow.	or b ie n d z h i z b r ou
The sword, the banner, and the field,	th u s or d th u b a n er a n d th u f ee l d
Glory and Greece, around me see!	g l o r i a n d g r ii s ur r ou n d m ee s ee
The Spartan, borne upon his shield,	th u s p aar t i n b oor n u pp aa n h i z sh ee l d
Was not more free.	w u z n aa t m or f r ee
Awake! (not Greece - she is awake!)	u ww ae k n aa t g r ii s sh ee i z u ww ae k
Awake, my spirit! Think through whom	u ww ae k m ie s p ee r i t thh ee n k thh r ue h ue m
Thy life-blood tracks its parent lake,	th ie l ie f b l u d t r a k s i t s p ai r i n t l ae k
And then strike home.	a n d th e n s t r i e k h oe m
Tread those reviving passions down,	t r e d th oe z r i vv ie v ee ng p a sh i n z d ou n
Unworthy manhood! - unto thee	u n ww er th ee m a n h oo d u n t ue th ee
Indifferent should the smile or frown	i n d d i f r i n t sh oo d th u s m ie l or f r ou n

Of beauty be.
If thou regrett'st thy youth, why live?
The land of honourable death
Is here: - up to the field, and give
Away thy breath!

u v b y ue t ee b ee
i f th ou r i g r e t s t th ie y ue thh w ie l i v
th u l a n d u v o n a r a b l d e thh
i z h eer u p t ue th u f ee l d a n d g i v
u ww ae th ie b r e thh

Segmentation results

Segmentation on the basis of frequency characteristic :

th is t ie m th is h aart shoo d b ee an m uv d
s in ts uther z i t hathh s ee s d t uem ue **v**
ye t thoe ie **k an** aat b eeb eel uv i d
s t i l l e t m eel **uv**
m iedae z aar **in** thu ye loe l eef
thu flaaw er **z and** fr ue ts uv l **uv** aar gau n
thu wer m thu k aen k **er and** thu gr eef
aar mie nu l loe n
thu f iertha taan **mie** b oo zim prae
iz loe na **z** s u m v aalk **k** a nik iel
n oetor ch **iz k in d ool d** a t **i** ts bl ae **z**
u fy uene r ool p iel
thu **hoe p** thu f eerthuj e l **is k air**
th uig zz ault **i d** p orsh **in uv** thu **p aen**
and p aa wer uv l uv ie k an aat shai **r**
but **wair** thu ch **aen**
but **t is n aat** thu **s and t is n** aat heer
s uch thhau ts shoo d shae k **mie s oe l n ornou**
wair gl **oo ri d** ek **s** thu her **ou** z beer
or b ie n d z h iz br **ou**
thu **s or d** thu **b an er and** thu **f eel d**
gl **o ri and** gr iis **urr** oun **d m ee s ee**
thu s p aart **in boor** nu pp aanh iz sh eel **d**
wuz n aat m or fr ee
uww aek n aat gr iis **shee iz uww aek**
uww aek mie s peer i **t** thheen k thhr **ue h** uem
th iel **ie f** bl **u d** tr **a k s i** ts pai r **in t l aek**
and th ens t ri ek **hoe m**
tr e **d** thoe **z ri vvie** veeng pash **in** zdou **n**
unww er thee **m an** hoo dun tue thee
in ddi **fr in t** shoo **d** thu **s mie l or fr** oun
uv by ue t eeb **ee**
i f thou **ri** gr e ts **t** thie yue thhw iel **iv**
thu **l and uv o** na ra bl **d** ethh
iz h eerup tue thu **f eel d and g iv**
uww ae thie br e **thh**

Segmentation on the basis of mean geometrical intervals:

thist iem thish aartsh oo dbeea nmu vd
sin tsu th erzi th a thhsee sdt uemue **v**
y etth oeie **k an** aa tbee beel u vid
sti lle tmee l **uv**
mie dae zaar **in** th uy el oelee f
th u fl aawer **z and** frue tsu vl **uv** aarg aun
th uw erm th u k aenk **er and** th u gree f
aar mie n ull oen
th u fier th at aan **mie** boo z im p r ae
iz l oen a **z** sum vaa l k **k** an i k ie l
n oe t orch **iz k in d ool d** at **i** t s blae **z**
u fy uene r ool p ie l
th u **hoe p** th u feer th uj el **is k air**
th uig zzau l t **i d** por sh **in uv** th u **p aen**
and p aa wer uv l uv ie k an aat sh a i **r**
bu t **wair** th uch **aen**
bu t **t is n aat** th u **s and t is n** aa th eer
s uch thhau t s sh oo d sh aek **mie s oe l n ornou**
wair g l **oo ri d** e k **s** th u h er **ou** zb eer
or b ie n d z hi zb r **ou**
th u **s or d** th u **b an er and** th u **f ee l d**
g l **o ri and** g rii s **urr** ou n **d m ee s ee**
th u sp aar t **in boor** n u ppaa n hi z shee l **d**
wuz n aat m or fr ee
uww aek n aat g rii s **shee iz uww aek**
uww aek mie sp ee ri **t** thh eenk thh r **ue h** ue m
th ie l **ie f** b l **u d** t r **a k s i** t sp air **in t l aek**
and th ens t ri e k **hoe m**
t re **d** th oe **z ri vvie** v ee ngp a sh **in** z d ou **n**
u n wwer th ee **m an** h oo d u n t ue th ee
in d d i **fr in t** sh oo **d** th u **s mie l or fr** ou n
uv b y uetee b **ee**
i f th ou **ri** g re t s **t** th ie y ue thh w ie l **iv**
th u **l and uv o** n a r a b l **d** e thh
iz h eer u p t ue th u **f ee l d and g iv**
uww ae th ie b re **thh**

Table 6
The compared vocabularies of consonances

Segmentation vocabulary obtained on the basis of frequency characteristics		
aalk aanh aart ault beer bl br but by ch ddi dun eeb eef eel eerthuj eerup ek ethh flaaw gau gl gr	a aa aar aat ae aek aen air an and b boor d e ee ens er f fr fy g h hoe i	aan aarg aartsh aawer aenk at aun beel blae boo bu dae dbeea eenk eer el erm erzi etth feer fier fl frue gree

hathh heer her hoo iedae iel iertha iis loe na nik nu oetor orsh oun pai pash peer pp prae ra shae shai shoo taan thee thheen thhr thhw thie thoe thou thu tr ts tue uem unww uther veeng ye yue zdou zim zz	ie in is iv iz k l m mie n o oe oo ool or ornou ou p r ri s sh shee t th thh thhau u uch ue uene uig urr uv uww v vvie wair wer wuz z	hi iem im lle ngp nmu oeie oelee oen orch por ppaa re rii sdt sin sp sti sum tbee thhsee thish thist tmee tsu uemue uetee uj ull uw uy vaa vd vid vl w wwer y zaar zb zzau
	Segmentation vocabulary obtained on the basis of interval characteristics	

Table 7
Frequency-rank distribution of consonance vocabularies of the segmentation of the poem
On This Day I Complete My Thirty-Sixth Year

r	*K* = 0	*K* = 0.6	*K* = 0.7	*K* = 0.8	*K* = 1
1	D ... 19	TH ... 33	TH ... 33	TH ... 33	TH ... 33
2	THU ... 19	N ... 24	N ... 24	N ... 24	U ... 23
3	S ... 15	L ... 21	L ... 21	D ... 20	L ... 20
4	T ... 15	D ... 20	D ... 20	I ... 20	T ... 20
5	L ... 11	S ... 17	T ... 19	T ... 19	D ... 19
6	K ... 10	U ... 16	U ... 17	L ... 18	N ... 16
7	M ... 10	T ... 15	I ... 16	U ... 17	S ... 13
8	N ... 10	EE ... 11	EE ... 11	EE ... 11	K ... 11
9	UV ... 10	K ... 11	K ... 11	K ... 11	EE ... 10
10	AND ... 9	I ... 10	S ... 11	S ... 11	AND ... 9
11	IN ... 9	R ... 10	Z ... 11	Z ... 11	IE ... 9
12	I ... 8	A ... 9	AND ... 9	AND ... 9	IN ... 8
13	Z ... 8	AND ... 9	IE ... 9	IE ... 9	Z ... 8
14	AAT ... 6	B ... 9	R ... 9	R ... 9	A ... 7
15	EE ... 6	IE ... 9	B ... 8	RI ... 8	B ... 7
16	F ... 6	P ... 8	RI ... 8	UV ... 8	I ... 7
17	IZ ... 6	RI ... 8	UV ... 8	A ... 7	R ... 7
18	P ... 6	UV ... 8	A ... 7	B ... 7	UV ... 7
19	TS ... 6	Z ... 7	MIE ... 6	F ... 7	F ... 6
20	AN ... 5	MIE ... 6	P ... 6	MIE ... 6	G ... 6
21	B ... 5	UF ... 6	UF ... 6	P ... 6	MIE ... 6
22	E ... 5	AEK ... 5	AEK ... 5	UF ... 6	OU ... 6
23	EEL ... 5	AN ... 5	AN ... 5	AEK ... 5	P ... 6
24	IE ... 5	E ... 5	E ... 5	AN ... 5	RI ... 6
25	IS ... 5	M ... 5	M ... 5	E ... 5	SH ... 6
26	MIE ... 5	OO ... 5	OO ... 5	M ... 5	AEK ... 5
27	RI ... 5	SH ... 5	SH ... 5	OO ... 5	AN ... 5
28	TH ... 5	TI ... 5	AAT ... 4	SH ... 5	M ... 5
29	AEK ... 4	AAT ... 4	EER ... 4	AAT ... 4	OO ... 5
30	ER ... 4	EER ... 4	UE ... 4	EER ... 4	THH ... 5
31	FR ... 4	IZ ... 4	UWW ... 4	OR ... 4	UE ... 5
32	GR ... 4	UE ... 4	AA ... 3	UE ... 4	AAT ... 4
33	H ... 4	UWW ... 4	AE ... 3	UWW ... 4	H ... 4
34	IEL ... 4	AA ... 3	F ... 3	V ... 4	IZ ... 4
35	OR ... 4	AE ... 3	G ... 3	AA ... 3	OR ... 4
36	UE ... 4	G ... 3	GR ... 3	GR ... 3	UWW ... 4
37	UWW ... 4	GR ... 3	H ... 3	H ... 3	AA ... 3
38	A ... 3	H ... 3	IS ... 3	IS ... 3	E ... 3
39	AAR ... 3	OE ... 3	OE ... 3	OE ... 3	EER ... 3
40	AEN ... 3	OU ... 3	OU ... 3	OU ... 3	ER ... 3
41	BL ... 3	ROU ... 3	ROU ... 3	ROU ... 3	FR ... 3
42	LOE ... 3	THH ... 3	SP ... 3	SP ... 3	IS ... 3
43	R ... 3	AAR ... 2	THH ... 3	THH ... 3	OE ... 3
44	SHOO ... 3	AEN ... 2	AAR ... 2	AAR ... 2	RE ... 3
45	U ... 3	BUT ... 2	AEN ... 2	AE ... 2	SP ... 3
46	AART ... 2	EL ... 2	AT ... 2	AEN ... 2	Y ... 3
47	AE ... 2	ER ... 2	BUT ... 2	AT ... 2	AAR ... 2
48	BR ... 2	F ... 2	EL ... 2	BUT ... 2	AE ... 2
49	BUT ... 2	HI ... 2	ER ... 2	EL ... 2	AEN ... 2
50	CH ... 2	HOE ... 2	HI ... 2	ER ... 2	AIR ... 2
51	EEB ... 2	IF ... 2	HOE ... 2	GL ... 2	AT ... 2
52	EEF ... 2	IIS ... 2	IF ... 2	HI ... 2	BU ... 2
53	EK ... 2	IV ... 2	IIS ... 2	HOE ... 2	EL ... 2

54	GL 2	O 2	IV 2	IIS..... 2	HI 2
55	HOE 2	OEN 2	O 2	O 2	HOE..... 2
56	IIS..... 2	OOL 2	OEN 2	OEN 2	IV 2
57	IV 2	OR 2	OOL 2	OOL 2	O 2
58	NA 2	ORF..... 2	OR 2	SHEE..... 2	OEN..... 2
59	NU 2	SHEE 2	ORF 2	TSU..... 2	OOL..... 2
60	O 2	TSU..... 2	SHEE..... 2	UCH..... 2	RII 2
61	OO..... 2	UCH 2	TSU..... 2	WAIR 2	SHEE 2
62	OOL 2	V 2	UCH 2	Y 2	TSU 2
63	OU 2	WAIR 2	V 2	YUE 2	UCH..... 2
64	OUN..... 2	Y 2	WAIR 2	ZB 2	V 2
65	THEE 2	YUE 2	Y 2	AAN..... 1	WAIR 2
66	THIE 2	ZB..... 2	YUE 2	AARG..... 1	ZB 2
67	THOE 2	AARG..... 1	ZB..... 2	AARTSH..... 1	AAN..... 1
68	TR 2	AARTSH 1	AAN..... 1	AENK..... 1	AARG 1
69	TUE 2	AENK 1	AARG..... 1	AI 1	AARTSH..... 1
70	UEM 2	AIR 1	AARTSH 1	AIR 1	AAWER..... 1
71	V 2	AUN..... 1	AENK..... 1	AUN..... 1	AENK..... 1
72	WAIR 2	BEEL..... 1	AI 1	BEEL..... 1	AUN..... 1
73	WER 2	BOO 1	AIR 1	BLAE 1	BEEL 1
74	YE 2	DAE 1	AUN..... 1	BOO..... 1	BLAE 1
75	AA..... 1	DBEEA 1	BEEL..... 1	BOOR..... 1	BOO..... 1
76	AALK..... 1	EEF 1	BOO 1	DAE 1	BOOR..... 1
77	AANH 1	EENK 1	BOOR..... 1	DBEEA 1	DAE..... 1
78	AIR 1	ENS..... 1	DAE 1	EENK 1	DBEEA..... 1
79	AULT 1	ERM..... 1	DBEEA 1	ENS..... 1	EENK 1
80	BEER 1	ERZ..... 1	EENK 1	ERM..... 1	ENS 1
81	BOOR..... 1	ERZI..... 1	ENS..... 1	ERZ..... 1	ERM..... 1
82	BY 1	ETHH 1	ERM..... 1	ERZI 1	ERZI 1
83	DDI..... 1	ETTH..... 1	ERZ..... 1	ETHH 1	ETTH 1
84	DUN..... 1	FRUE..... 1	ERZI..... 1	ETTH 1	FEER..... 1
85	EERTHUJ..... 1	HER 1	ETHH 1	FRUE 1	FIER..... 1
86	EERUP 1	IEM 1	ETTH..... 1	G 1	FL..... 1
87	ENS 1	IER..... 1	FRUE..... 1	GREE 1	FRUE 1
88	ETHH 1	LAAW 1	GREE 1	HER 1	FY 1
89	FLAAW 1	LLE 1	HER 1	IEM..... 1	GREE 1
90	FY 1	MP 1	IEM..... 1	IER 1	IEM..... 1
91	G 1	NGP 1	IER..... 1	LAAW 1	IM..... 1
92	GAU..... 1	NM..... 1	LAAW 1	LLE..... 1	LLE..... 1
93	HATHH 1	OEIE..... 1	LLE..... 1	MP..... 1	NGP 1
94	HEER 1	OELEE..... 1	MP 1	NGP 1	NMU 1
95	HER 1	OOR 1	NGP 1	NM 1	OEIE 1
96	HOO..... 1	ORCH 1	NM..... 1	OEIE 1	OELEE 1
97	IEDAE 1	ORNOU 1	OEIE 1	OELEE 1	ORCH..... 1
98	IERTHA..... 1	ORSH 1	OELEE..... 1	ORCH..... 1	ORNOU 1
99	NIK 1	PAI..... 1	ORCH..... 1	ORNOU 1	POR 1
100	OE 1	PPAA..... 1	ORNOU 1	ORSH 1	PPAA 1
101	OETOR 1	SDT..... 1	ORSH 1	PPAA 1	SDT 1
102	ORNOU 1	SIN..... 1	PPAA..... 1	SDT..... 1	SIN 1
103	ORSH 1	STI 1	SDT..... 1	SIN 1	STI..... 1
104	PAI 1	SUM..... 1	SIN..... 1	STI..... 1	SUM..... 1
105	PASH 1	TAAN 1	STI..... 1	SUM..... 1	TBEE 1
106	PEER..... 1	TBEE..... 1	SUM..... 1	TBEE..... 1	THHAU 1
107	PP 1	THHAU..... 1	TBEE..... 1	THHAU 1	THHSEE 1
108	PRAE 1	THHSEE..... 1	THHAU 1	THHSEE 1	THISH 1
109	RA 1	THHW 1	THHSEE 1	THHW 1	THIST 1
110	SH 1	THISH 1	THHW 1	THISH 1	TMEE..... 1
111	SHAE 1	THIST..... 1	THISH 1	THIST..... 1	UEMUE 1
112	SHAI 1	TMEE 1	THIST..... 1	TMEE 1	UENE 1
113	SHEE..... 1	UEMUE 1	TMEE 1	UEMUE 1	UETEE 1
114	TAAN..... 1	UENE 1	UEMUE 1	UENE 1	UIG..... 1
115	THH 1	UG 1	UENE 1	UIG 1	UJ 1
116	THHAU 1	UIG 1	UIG 1	UJ 1	ULL 1
117	THHEEN..... 1	UJ 1	UJ 1	ULL 1	URR..... 1
118	THHR..... 1	ULL..... 1	ULL 1	UR 1	UW 1
119	THHW 1	UR 1	UR 1	UW 1	UY 1
120	THOU..... 1	UW 1	UW 1	UY 1	VAA..... 1
121	UCH..... 1	UY 1	UY 1	VAA..... 1	VD..... 1
122	UENE 1	VAA 1	VAA 1	VID 1	VID..... 1
123	UIG 1	VID 1	VID 1	VL 1	VL 1
124	UNWW 1	VL..... 1	VL 1	VVIE 1	VVIE 1
125	URR 1	VVIE 1	VVIE 1	WER 1	W 1
126	UTHER 1	WER..... 1	WER..... 1	WUZ 1	WER 1
127	VEENG..... 1	WUZ..... 1	WUZ 1	WWER 1	WUZ 1
128	VVIE 1	WWER..... 1	WWER..... 1	ZAAR 1	WWER 1

129 130 131 132 133	WUZ ..... 1 YUE ..... 1 ZDOU..... 1 ZIM..... 1 ZZ..... 1	ZAAR ..... 1 ZZAU 1	ZAAR ..... 1 ZZAU 1	ZZAU 1	ZAAR ..... 1 ZZAU 1

Table 8
The characteristics of the text, obtained via segmentations using various versions of the segmentation algorithm for the poem *On This Day I Complete My Thirty-Sixth Year*

№	K	Z	F_1	p_1	v	v_T	δ_v,%	δ_{me}%	Hs	G	H	g	g-H	2^{g-H}
1	0.0	411	19	0.046	133	134	0.75	14.56	2648.5	2243.5	6.4	5.5	-0.99	0.51
2	0.6	461	33	0.072	130	129	0.78	16.55	2840.5	2287.8	6.2	5.0	-1.20	0.44
3	0.7	463	33	0.071	130	130	0.00	16.28	2847.7	2290.9	6.2	4.9	-1.20	0.43
4	0.8	465	33	0.071	129	130	0.77	17.65	2852.0	2295.8	6.1	4.9	-1.20	0.44
5	1.0	460	33	0.072	130	129	0.78	16.50	2852.4	2292.0	6.2	5.0	-1.22	0.43

See the explanation of all the variables below Table 2.

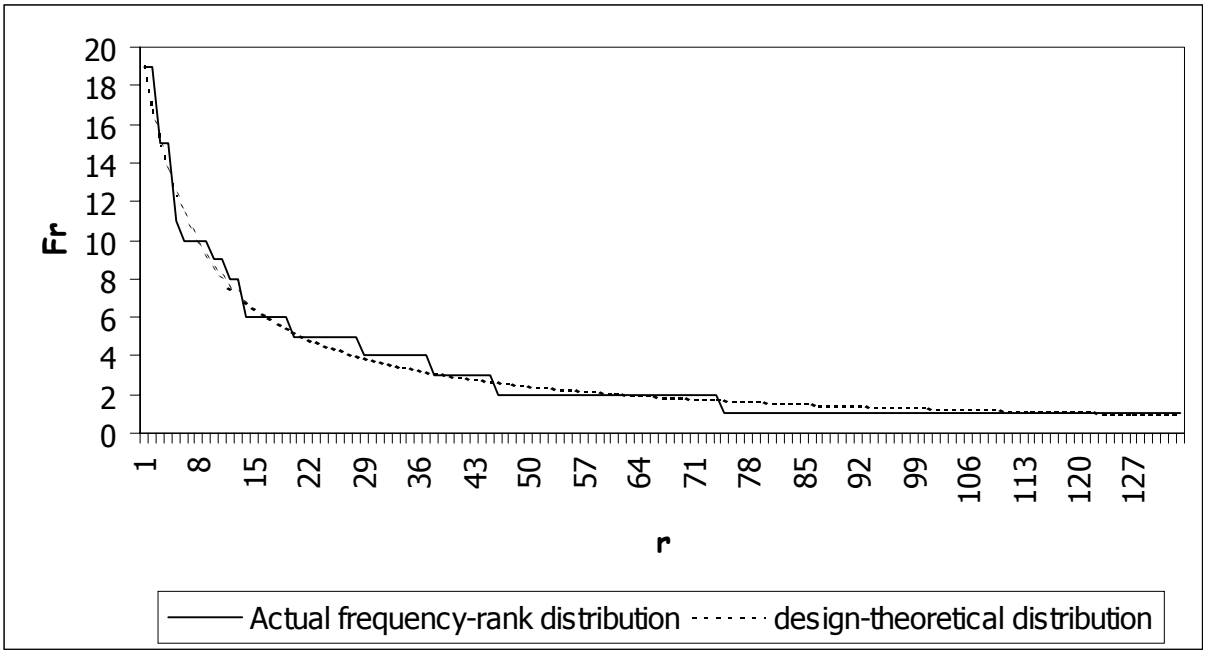


Figure 3. Actual and design theoretical (Zipf-Mandelbrot's) distributions for the frequency version of the segmentation of the poem *On This Day I Complete My Thirty-Sixth Year*
F_r , r – see comments to Figure 2.

Segmentation results for K = 0.6 and K = 1

Segmentation for K = 0.6:

thist iem thish aartsh oo dbeea nm uv d
sin tsu th erzi th a thhsee sdt uemue v
y etth oeie k an aa tbee beel u vid
sti lle tmee l uv
mie dae zaar i n **th uy el oelee f**
th uf laaw erz **and frue tsu vl uv aarg aun**

Segmentation for K = 1:

thist iem thish aartsh oo dbeea nmu vd
sin tsu th erzi th a thhsee sdt uemue v
y etth oeie k an aa tbee beel u vid
sti lle tmee l uv
mie dae zaar in **th uy el oelee f**
th u fl aawer z **and frue tsu vl uv aarg aun**

th uw erm th u k aenk er and th ug r eef
aar mie n ull oen
th uf ier **th** a taan **mie boo z** i mp **r ae**
iz l oen a z sum vaa l k k an i k ie l
n oe t orch iz k i n **d ool d** a ti t s b l ae **z**
uf y **uene r ool p ie l**
th u hoe p th uf eer **th uj el** i s **k air**
th uig zzau l ti **d** p orsh i n **uv th u p aen**
and p aa wer uv l uv ie k an aat sh a i r
but **wair th uch aen**
but ti s **n aat th u s and** ti s **n aa th eer**
s uch thhau t s sh oo d sh aek mie s oe l n ornou
wair g l oo ri d e k s th u her **ou zb eer**
or b ie n d z hi zb rou
th u s or d th u b an er and th uf **ee l d**
g l o ri and gr iis ur rou **n d m ee s ee**
th u s p **aar** ti n b oor **n u ppaa n hi z shee l d**
wuz n aat m orf r **ee**
uww aek n aat gr iis **shee iz uww aek**
uww aek mie s p **ee ri t thh eenk thh r ue h ue m**
th ie l ie f b l u d t r a k s i t s pai ri n **t l aek**
and th ens t ri e k hoe m
t r e **d th oe z ri vvie v ee ngp a sh** i n **z d ou n**
u n wwer th ee m an h oo d u n t ue th ee
i n **d d** if ri n **t sh oo d th u s mie l** orf rou **n**
uv b yue t ee **b ee**
if **th ou ri** gr e **t s t th ie** yue thhw **ie l iv**
th u l and uv o n a r a b l d ethh
iz h eer u p t ue th uf **ee l d and g iv**
uww ae th ie b r e **thh**

th uw erm th u k aenk er and th u gree f
aar mie n ull oen
th u fier **th** at aan **mie boo z** im p **r ae**
iz l oen a z sum vaa l k k an i k ie l
n oe t orch iz k in **d ool d** at i t s blae **z**
u fy **uene r ool p ie l**
th u hoe p th u feer **th uj el** is **k air**
th uig zzau l t i **d** por sh in **uv th u p aen**
and p aa wer uv l uv ie k an aat sh a i r
bu t **wair th uch aen**
bu t t is **n aat th u s and** t is **n aa th eer**
s uch thhau t s sh oo d sh aek mie s oe l n ornou
wair g l oo ri d e k s th u h er **ou zb eer**
or b ie n d z hi zb r ou
th u s or d th u b an er and th u f **ee l d**
g l o ri and g rii s urr ou **n d m ee s ee**
th u sp **aar** t in boor **n u ppaa n hi z shee l d**
wuz n aat m or fr **ee**
uww aek n aat g rii s **shee iz uww aek**
uww aek mie sp **ee ri t thh eenk thh r ue h ue m**
th ie l ie f b l u d t r a k s i t sp air in **t l aek**
and th ens t ri e k hoe m
t re **d th oe z ri vvie v ee ngp a sh** in **z d ou n**
u n wwer th ee m an h oo d u n t ue th ee
in **d d** i fr in **t sh oo d th u s mie l** or fr ou **n**
uv b y uetee **b ee**
i f **th ou ri** g re **t s t th ie** y ue thh w **ie l iv**
th u l and uv o n a r a b l d e thh
iz h eer u p t ue th u f **ee l d and g iv**
uww ae th ie b re **thh**

Table 9
The compared vocabularies of consonances

Segmentation vocabulary for $K = 0.6$		
but eef erz ethh gr her ier if iis laaw mp nm oor orf orsh pai rou taan thhw ti uf ug ur yue	a aa aar aarg aartsh aat ae aek aen aenk air an and aun b beel boo d dae dbeea e ee eenk eer el ens er erm erzi etth f frue g h hi hoe i ie iem iv iz k l lle m mie n ngp o oe oeie oelee oen oo ool or orch ornou ou p ppaa r ri s sdt sh shee sin sti sum t tbee th thh thhau thhsee thish thist tmee tsu u uch ue uemue uene uig uj ull uv uw uww uy v vaa vid vl vvie wair wer wuz wwer y z zaar zb zzau	aan aawer at blae boor bu feer fier fl fr fy gree im in is nmu por re rii sp uetee urr vd w
	Segmentation vocabulary for $K = 1$	

Segmentation results for $K = 0.7$ and $K = 0.8$

Segmentation for $K = 0.7$:

thist iem thish aartsh oo dbeea nm uv d
sin tsu th erzi th a thhsee sdt uemue v
y etth oeie k an aa tbee beel u vid
sti lle tmee l uv
mie dae zaar i n th uy el oelee f
th uf laaw erz and frue tsu vl uv aarg aun
th uw erm th u k aenk er and th u gree f
aar mie n ull oen

Segmentation for $K = 0.8$:

thist iem thish aartsh oo dbeea nm uv d
sin tsu th erzi th a thhsee sdt uemue v
y etth oeie k an aa tbee beel u vid
sti lle tmee l uv
mie dae zaar i n th uy el oelee f
th uf laaw erz and frue tsu vl uv aarg aun
th uw erm th u k aenk er and th u gree f
aar mie n ull oen

th uf ier th at aan mie boo z i mp r ae
i z l oen a z sum vaa l k k an i k ie l
n oe t orch i z k i n d ool d at i t s b l ae **z**
uf y uene r ool p ie l
th u hoe p th uf eer th uj el is k air
th uig zzau l t i d p orsh i n uv th u p aen
and p aa wer uv l uv ie k an aat sh a i r
but wair th uch aen
but t is n aat th u s and t is n aa th eer
s uch thhau t s sh oo d sh aek mie s oe l n ornou
wair g l **oo ri d e k s th u her ou zb eer**
or b ie n d z hi zb rou
th u s or d th u b an er and th uf ee l d
g l **o ri and gr iis ur rou n d m ee s ee**
th u sp aar t i n boor n u ppaa n hi z shee l d
wuz n aat m orf **r ee**
uww aek n aat gr iis shee i z uww aek
uww aek mie sp ee ri t thh eenk thh r ue h ue m
th ie l ie f b l u d t r a k s i t sp ai ri n t l aek
and th ens t ri e k hoe m
t r e d th oe z ri vvie v ee ngp a sh i n z d ou n
u n wwer th ee m an h oo d u n t ue th ee
i n d d if **ri n t sh oo d th u s mie l** orf **rou n**
uv b yue t ee b ee
if **th ou ri gr e t s t th ie yue thhw ie l** iv
th u l and uv o n a r a b l d ethh
i z h eer u p t ue th uf ee l d and g iv
uww ae th ie b r e thh

th uf ier th at aan mie boo z i mp r ae
i z l oen a z sum vaa l k k an i k ie l
n oe t orch i z k i n d ool d at i t s blae **z**
uf y uene r ool p ie l
th u hoe p th uf eer th uj el is k air
th uig zzau l t i d p orsh i n uv th u p aen
and p aa wer uv l uv ie k an aat sh a i r
but wair th uch aen
but t is n aat th u s and t is n aa th eer
s uch thhau t s sh oo d sh aek mie s oe l n ornou
wair gl **oo ri d e k s th u her ou zb eer**
or b ie n d z hi zb rou
th u s or d th u b an er and th uf ee l d
gl **o ri and gr iis ur rou n d m ee s ee**
th u sp aar t i n boor n u ppaa n hi z shee l d
wuz n aat m or f **r ee**
uww aek n aat gr iis shee i z uww aek
uww aek mie sp ee ri t thh eenk thh r ue h ue m
th ie l ie f b l u d t r a k s i t sp ai ri n t l aek
and th ens t ri e k hoe m
t r e d th oe z ri vvie v ee ngp a sh i n z d ou n
u n wwer th ee m an h oo d u n t ue th ee
i n d d i f **ri n t sh oo d th u s mie l** or f **rou n**
uv b yue t ee b ee
i f **th ou ri gr e t s t th ie yue thhw ie l** i v
th u l and uv o n a r a b l d ethh
i z h eer u p t ue th uf ee l d and g i v
uww ae th ie b r e thh

Table 10
The compared vocabularies of consonances

Segmentation vocabulary for $K = 0.7$		
if iv orf	a aa aan aar aarg aartsh aat ae aek aen aenk ai air an and at aun b beel boo boor but d dae dbeea e ee eenk eer el ens er erm erz erzi ethh etth f frue g gr gree h her hi hoe i ie iem ier iis is k l laaw lle m mie mp n ngp nm o oe oeie oelee oen oo ool or orch ornou orsh ou p ppaa r ri rou s sdt sh shee sin sp sti sum t tbee th thh thhau thhsee thhw thish thist tmee tsu u uch ue uemue uene uf uig uj ull ur uv uw uww uy v vaa vid vl vvie wair wer wuz wwer y yue z zaar zb zzau	blae gl
	Segmentation vocabulary for $K = 0.8$	

The combined evaluation of the difference between theoretical and combined characteristics of the chain is calculated using the formulae (5 and 11) and depends on the balance factor K, $0 \le K \le 1$. This factor determines the influence of the interval frequency and the value $(1 - K)$ determines the influence of the ordinary frequency of occurrences on the resulting evaluation. The extreme values of the factor provide the other two modifications of the segmentation algorithm: when using the combined evaluation (Formula 11), $K = 1$ is determined only by the mean geometrical interval (8 and 9), and when $K = 0$, this evaluation is determined only by the frequency.

The variants of the factor K in the formula (11) provide different evaluations of the difference between theoretical and combined characteristics of the chain, which results in different segmentation versions of one and the same text. One and the same sequence of phonemes in the poem was segmented using different values of K factor to study and to compare the opportunities of the various versions of the segmentation algorithm based on the

combined evaluation. The balance factor varied from 0 to 1 with the fixed step equal 0.1; so 11 values of the balance factor were used

(15) $K_i = 0.1i$, where $i = 0,1,2,\ldots,10$.

So there were obtained 11 segmentations of the source sequence. The obtained segmentations differed from each other both by the sizes of the obtained vocabularies and by its content (see Tables 2 and 7, 4 and 5, 9 and 10). For quantitative comparison of the obtained vocabularies by the degree of their overlap or their similarity a measure suggested by A.N. Florensov was used (Florensov 2000).

(16) $$r(V_1,V_2) = \frac{\left| V_1 \cap V_2 \right|}{\left| V_1 \cup V_2 \right|},$$

where V_1, V_2 – compared vocabularies viewed as a number of consonances, i.e. the value $r(V_1, V_2)$ shows the part of common consonances among the total number of consonances of the two vocabularies.

For each pair of vocabularies there was a comparison done, the results of which are shown in the square table of comparative characteristics of the consonances vocabularies (Tables 11 and 13), with the size of 11x11 (upon the number of versions). The table is symmetrical in relation to the diagonal. At the intersection of line i and column j there are four values presented in the following form:

$$\begin{matrix} a & b \\ c & d \end{matrix},$$

where $a = \left| V_i \cap V_j \right|$, $b = \left| V_i \cup V_j \right|$, $c = \left| V_i \right|$, $d = \left| V_j \right|$, where V_i, V_j – i-th and j-th vocabulary of consonances. It is obvious that $b = c + d - a$. The most interesting is the value of the common part of the compared vocabularies a, for determining their "stability" in relation to each other.

The tables of normalized comparative characteristics (Tables 12 and 14) have the same format, but instead of the absolute values the parts of the combined vocabularies are shown in the table:

$$\begin{matrix} a/b & 1 \\ c/b & d/b \end{matrix}$$

The cell element a/b contains the value of the measure $r(V_i, V_j)$, representing the overlap of vocabularies. This characteristic shows that when most of the combined characteristic is the interval information, then the degree of the overlap of vocabularies is more stable to changes of the balance factor. The area of such stability is put in the black frame in Tables 11–14. Elements of cells located along the diagonal are presented by 1 as in this case the compared dictionaries are identical.

Table 11

The comparative characteristics of vocabularies obtained by different versions of the segmentation algorithm for the poem *The Garden of Love*

K	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
1.0	31 104 69 66	36 97 69 64	38 97 69 66	39 96 69 66	38 97 69 66	56 79 69 66	53 83 69 67	59 77 69 67	65 72 69 68	56 83 69 70	69 69 69 69
0.9	33 103 70 66	31 103 70 64	32 104 70 66	33 103 70 66	31 105 70 66	46 90 70 66	55 82 70 67	51 86 70 67	55 83 70 68	70 70 70 70	56 83 70 69
0.8	29 105 68 66	34 98 68 64	37 97 68 66	38 96 68 66	37 97 68 66	58 76 68 66	52 83 68 67	61 74 68 67	68 68 68 68	55 83 68 70	65 72 68 69
0.7	31 102 67 66	36 95 67 64	38 95 67 66	39 94 67 66	37 96 67 66	61 72 67 66	57 77 67 67	67 67 67 67	61 74 67 68	51 86 67 70	59 77 67 69
0.6	37 96 67 66	37 94 67 64	37 96 67 66	38 95 67 66	36 97 67 66	57 76 67 66	67 67 67 67	57 77 67 67	52 83 67 68	55 82 67 70	53 83 67 69
0.5	34 98 66 66	38 92 66 64	40 92 66 66	41 91 66 66	39 93 66 66	66 66 66 66	57 76 66 67	61 72 66 67	58 76 66 68	46 90 66 70	56 79 66 69
0.4	46 86 66 66	57 73 66 64	63 69 66 66	62 70 66 66	66 66 66 66	39 93 66 66	36 97 66 67	37 96 66 67	37 97 66 68	31 105 66 70	38 97 66 69
0.3	49 83 66 66	58 72 66 64	65 67 66 66	66 66 66 66	62 70 66 66	41 91 66 66	38 95 66 67	39 94 66 67	38 96 66 68	33 103 66 70	39 96 66 69
0.2	48 84 66 66	59 71 66 64	66 66 66 66	65 67 66 66	63 69 66 66	40 92 66 66	37 96 66 67	38 95 66 67	37 97 66 68	32 104 66 70	38 97 66 69
0.1	50 80 64 66	64 64 64 64	59 71 64 66	58 72 64 66	57 73 64 66	38 92 64 66	37 94 64 67	36 95 64 67	34 98 64 68	31 103 64 70	36 97 64 69
0.0	66 66 66 66	50 80 66 64	48 84 66 66	49 83 66 66	46 86 66 66	34 98 66 66	37 96 66 67	31 102 66 67	29 105 66 68	33 103 66 70	31 104 66 69

Table 12

The normalized comparative characteristics of vocabularies obtained by different versions of the segmentation algorithm for the poem *The Garden of Love*.

K	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
1.0	0.30 1.00 0.66 0.63	0.37 1.00 0.71 0.66	0.39 1.00 0.71 0.68	0.41 1.00 0.72 0.69	0.39 1.00 0.71 0.68	0.71 1.00 0.87 0.84	0.64 1.00 0.83 0.81	0.77 1.00 0.90 0.87	0.90 1.00 0.96 0.94	0.67 1.00 0.83 0.84	1.00 1.00 1.00 1.00
0.9	0.32 1.00 0.68 0.64	0.30 1.00 0.68 0.62	0.31 1.00 0.67 0.63	0.32 1.00 0.68 0.64	0.30 1.00 0.67 0.63	0.51 1.00 0.78 0.73	0.67 1.00 0.85 0.82	0.59 1.00 0.81 0.78	0.66 1.00 0.84 0.82	1.00 1.00 1.00 1.00	0.67 1.00 0.84 0.83
0.8	0.28 1.00 0.65 0.63	0.35 1.00 0.69 0.65	0.38 1.00 0.70 0.68	0.40 1.00 0.71 0.69	0.38 1.00 0.70 0.68	0.76 1.00 0.89 0.87	0.63 1.00 0.82 0.81	0.82 1.00 0.92 0.91	1.00 1.00 1.00 1.00	0.66 1.00 0.82 0.84	0.90 1.00 0.94 0.96
0.7	0.30 1.00 0.66 0.65	0.38 1.00 0.71 0.67	0.40 1.00 0.71 0.69	0.41 1.00 0.71 0.70	0.39 1.00 0.70 0.69	0.85 1.00 0.93 0.92	0.74 1.00 0.87 0.87	1.00 1.00 1.00 1.00	0.82 1.00 0.91 0.92	0.59 1.00 0.78 0.81	0.77 1.00 0.87 0.90
0.6	0.39 1.00 0.70 0.69	0.39 1.00 0.71 0.68	0.39 1.00 0.70 0.69	0.40 1.00 0.71 0.69	0.37 1.00 0.69 0.68	0.75 1.00 0.88 0.87	1.00 1.00 1.00 1.00	0.74 1.00 0.87 0.87	0.63 1.00 0.81 0.82	0.67 1.00 0.82 0.85	0.64 1.00 0.81 0.83
0.5	0.35 1.00 0.67 0.67	0.41 1.00 0.72 0.70	0.43 1.00 0.72 0.72	0.45 1.00 0.73 0.73	0.42 1.00 0.71 0.71	1.00 1.00 1.00 1.00	0.75 1.00 0.87 0.88	0.85 1.00 0.92 0.93	0.76 1.00 0.87 0.89	0.51 1.00 0.73 0.78	0.71 1.00 0.84 0.87
0.4	0.53 1.00 0.77 0.77	0.78 1.00 0.90 0.88	0.91 1.00 0.96 0.96	0.89 1.00 0.94 0.94	1.00 1.00 1.00 1.00	0.42 1.00 0.71 0.71	0.37 1.00 0.68 0.69	0.39 1.00 0.69 0.70	0.38 1.00 0.68 0.70	0.30 1.00 0.63 0.67	0.39 1.00 0.68 0.71
0.3	0.59 1.00 0.80 0.80	0.81 1.00 0.92 0.89	0.97 1.00 0.99 0.99	1.00 1.00 1.00 1.00	0.89 1.00 0.94 0.94	0.45 1.00 0.73 0.73	0.40 1.00 0.69 0.71	0.41 1.00 0.70 0.71	0.40 1.00 0.69 0.71	0.32 1.00 0.64 0.68	0.41 1.00 0.69 0.72
0.2	0.57 1.00 0.79 0.79	0.83 1.00 0.93 0.90	1.00 1.00 1.00 1.00	0.97 1.00 0.99 0.99	0.91 1.00 0.96 0.96	0.43 1.00 0.72 0.72	0.39 1.00 0.69 0.70	0.40 1.00 0.69 0.71	0.38 1.00 0.68 0.70	0.31 1.00 0.63 0.67	0.39 1.00 0.68 0.71
0.1	0.63 1.00 0.80 0.82	1.00 1.00 1.00 1.00	0.83 1.00 0.90 0.93	0.81 1.00 0.89 0.92	0.78 1.00 0.88 0.90	0.41 1.00 0.70 0.72	0.39 1.00 0.68 0.71	0.38 1.00 0.67 0.71	0.35 1.00 0.65 0.69	0.30 1.00 0.62 0.68	0.37 1.00 0.66 0.71
0.0	1.00 1.00 1.00 1.00	0.63 1.00 0.82 0.80	0.57 1.00 0.79 0.79	0.59 1.00 0.80 0.80	0.53 1.00 0.77 0.77	0.35 1.00 0.67 0.67	0.39 1.00 0.69 0.70	0.30 1.00 0.65 0.66	0.28 1.00 0.63 0.65	0.32 1.00 0.64 0.68	0.30 1.00 0.63 0.66

Table 13

The comparative characteristics of vocabularies obtained by different versions of the segmentation algorithm for the poem *On This Day I Complete My Thirty-Sixth Year*

K	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
1.0	65 198 130 133	74 190 130 134	77 184 130 131	87 178 130 135	103 152 130 125	104 154 130 128	106 154 130 130	111 149 130 130	111 148 130 129	114 145 130 129	130 130 130 130
0.9	67 195 129 133	78 185 129 134	82 178 129 131	91 173 129 135	110 144 129 125	114 143 129 128	117 142 129 130	123 136 129 130	124 134 129 129	129 129 129 129	114 145 129 130
0.8	68 194 129 133	78 185 129 134	82 178 129 131	92 172 129 135	110 144 129 125	116 141 129 128	120 139 129 130	127 132 129 130	129 129 129 129	124 134 129 129	111 148 129 130
0.7	68 195 130 133	78 186 130 134	83 178 130 131	92 173 130 135	110 145 130 125	119 139 130 128	123 137 130 130	130 130 130 130	127 132 130 129	123 136 130 129	111 149 130 130
0.6	70 193 130 133	81 183 130 134	86 175 130 131	93 172 130 135	113 142 130 125	120 138 130 128	130 130 130 130	123 137 130 130	120 139 130 129	117 142 130 129	106 154 130 130
0.5	67 194 128 133	81 181 128 134	87 172 128 131	96 167 128 135	117 136 128 125	128 128 128 128	120 138 128 130	119 139 128 130	116 141 128 129	114 143 128 129	104 154 128 130
0.4	68 190 125 133	86 173 125 134	87 169 125 131	101 159 125 135	125 125 125 125	117 136 125 128	113 142 125 130	110 145 125 130	110 144 125 129	110 144 125 129	103 152 125 130
0.3	77 191 135 133	104 165 135 134	115 151 135 131	135 135 135 135	101 159 135 125	96 167 135 128	93 172 135 130	92 173 135 130	92 172 135 129	91 173 135 129	87 178 135 130
0.2	82 182 131 133	110 155 131 134	131 131 131 131	115 151 131 135	87 169 131 125	87 172 131 128	86 175 131 130	83 178 131 130	82 178 131 129	82 178 131 129	77 184 131 130
0.1	93 174 134 133	134 134 134 134	110 155 134 131	104 165 134 135	86 173 134 125	81 181 134 128	81 183 134 130	78 186 134 130	78 185 134 129	78 185 134 129	74 190 134 130
0.0	133 133 133 133	93 174 133 134	82 182 133 131	77 191 133 135	68 190 133 125	67 194 133 128	70 193 133 130	68 195 133 130	68 194 133 129	67 195 133 129	65 198 133 130

Table 14

The normalized comparative characteristics of vocabularies obtained by different versions of the segmentation algorithm for the poem *On This Day I Complete My Thirty-Sixth Year.*

K	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
1.0	0.33 1.00 0.66 0.67	0.39 1.00 0.68 0.71	0.42 1.00 0.71 0.71	0.49 1.00 0.73 0.76	0.68 1.00 0.86 0.82	0.68 1.00 0.84 0.83	0.69 1.00 0.84 0.84	0.74 1.00 0.87 0.87	0.75 1.00 0.88 0.87	0.79 1.00 0.90 0.89	1.00 1.00 1.00 1.00
0.9	0.34 1.00 0.66 0.68	0.42 1.00 0.70 0.72	0.46 1.00 0.72 0.74	0.53 1.00 0.75 0.78	0.76 1.00 0.90 0.87	0.80 1.00 0.90 0.90	0.82 1.00 0.91 0.92	0.90 1.00 0.95 0.96	0.93 1.00 0.96 0.96	1.00 1.00 1.00 1.00	0.79 1.00 0.89 0.90
0.8	0.35 1.00 0.66 0.69	0.42 1.00 0.70 0.72	0.46 1.00 0.72 0.74	0.53 1.00 0.75 0.78	0.76 1.00 0.90 0.87	0.82 1.00 0.91 0.91	0.86 1.00 0.93 0.94	0.96 1.00 0.98 0.98	1.00 1.00 1.00 1.00	0.93 1.00 0.96 0.96	0.75 1.00 0.87 0.88
0.7	0.35 1.00 0.67 0.68	0.42 1.00 0.70 0.72	0.47 1.00 0.73 0.74	0.53 1.00 0.75 0.78	0.76 1.00 0.90 0.86	0.86 1.00 0.94 0.92	0.90 1.00 0.95 0.95	1.00 1.00 1.00 1.00	0.96 1.00 0.98 0.98	0.90 1.00 0.96 0.95	0.74 1.00 0.87 0.87
0.6	0.36 1.00 0.67 0.69	0.44 1.00 0.71 0.73	0.49 1.00 0.74 0.75	0.54 1.00 0.76 0.78	0.80 1.00 0.92 0.88	0.87 1.00 0.94 0.93	1.00 1.00 1.00 1.00	0.90 1.00 0.95 0.95	0.86 1.00 0.94 0.93	0.82 1.00 0.92 0.91	0.69 1.00 0.84 0.84
0.5	0.35 1.00 0.66 0.69	0.45 1.00 0.71 0.74	0.51 1.00 0.74 0.76	0.57 1.00 0.77 0.81	0.86 1.00 0.94 0.92	1.00 1.00 1.00 1.00	0.87 1.00 0.93 0.94	0.86 1.00 0.92 0.94	0.82 1.00 0.91 0.91	0.80 1.00 0.90 0.90	0.68 1.00 0.83 0.84
0.4	0.36 1.00 0.66 0.70	0.50 1.00 0.72 0.77	0.51 1.00 0.74 0.78	0.64 1.00 0.79 0.85	1.00 1.00 1.00 1.00	0.86 1.00 0.92 0.94	0.80 1.00 0.88 0.92	0.76 1.00 0.86 0.90	0.76 1.00 0.87 0.90	0.76 1.00 0.87 0.90	0.68 1.00 0.82 0.86
0.3	0.40 1.00 0.71 0.70	0.63 1.00 0.82 0.81	0.76 1.00 0.89 0.87	1.00 1.00 1.00 1.00	0.64 1.00 0.85 0.79	0.57 1.00 0.81 0.77	0.54 1.00 0.78 0.76	0.53 1.00 0.78 0.75	0.53 1.00 0.78 0.75	0.53 1.00 0.78 0.75	0.49 1.00 0.76 0.73
0.2	0.45 1.00 0.72 0.73	0.71 1.00 0.85 0.86	1.00 1.00 1.00 1.00	0.76 1.00 0.87 0.89	0.51 1.00 0.78 0.74	0.51 1.00 0.76 0.74	0.49 1.00 0.75 0.74	0.47 1.00 0.74 0.73	0.46 1.00 0.74 0.72	0.46 1.00 0.74 0.72	0.42 1.00 0.71 0.71
0.1	0.53 1.00 0.77 0.76	1.00 1.00 1.00 1.00	0.71 1.00 0.86 0.85	0.63 1.00 0.81 0.82	0.50 1.00 0.77 0.72	0.45 1.00 0.74 0.71	0.44 1.00 0.73 0.71	0.42 1.00 0.72 0.70	0.42 1.00 0.72 0.70	0.42 1.00 0.72 0.70	0.39 1.00 0.71 0.68
0.0	1.00 1.00 1.00 1.00	0.53 1.00 0.76 0.77	0.45 1.00 0.73 0.72	0.40 1.00 0.70 0.71	0.36 1.00 0.70 0.66	0.35 1.00 0.69 0.66	0.36 1.00 0.69 0.67	0.35 1.00 0.68 0.67	0.35 1.00 0.69 0.66	0.34 1.00 0.68 0.66	0.33 1.00 0.67 0.66

Conclusions

1. In our point of view the present research corroborates the hypothesis on the existence of poems' elementary acoustic units named consonances. It is important to note that these units were revealed both for Russian and English. The obtained frequency vocabularies of consonances can be a source material for specialists in the field of linguistics for further informal study of the phonetic nature of languages.

2. Studying the versions of the algorithm using the combined evaluation it was possible to find the area of relative stability of the close in content vocabularies of one and the same poem for this algorithm modification. It is possible to suppose that "the best", i.e. the most probable vocabulary of consonances is contained within this area.

3. The measure of similarity suggested by A.N. Florensov and used in this study proves the usefulness and the effectiveness of application of this type of evaluation in this field. Besides, the new method of visual presentation of comparative characteristics of vocabulary versions for the poem under study was used.

4. The convergence of the algorithm, its modifications and versions were experimentally corroborated regardless the following strong limitations:
- considerable deviation of the frequency of a consonance from the design value;
- correspondence to Orlov's criterion in the vocabulary value;
- taking into account pauses at the end of the poem lines.

5. The research done reveals the potential practical application of the suggested algorithm, which makes the created program a new tool for the study of symbol sequences.

6. Segmentation of texts in natural languages into consonances makes it possible to study the acoustic nature of the poetic language and can be a step to the more complete understanding of the principles of human cognition and may allow to develop new algorithms of text processing.

References

Blatter, A. (1980). *Instrumentation/ Orchestration.* New York: Longman.

Boroda, M.G. (1973). On the concept of the elementary methrorythmic unit in music. In: *Bulletin of the Academy of Sciences of the Georgian SSR. Volume 71, № 3, 745-748.*

Borodovskij, M.Ju., Pevzner, P.A. (1990). Statističeskie metody analiza genetičeskich tekstov. In: *Kompjuternij analiz tekstov: 33-80.* Moskva: Nauka.

Florensov, A.N. (2000). Postroenie abstraktnogo prostranstva dlja semantičeskoj teorii informacii. In: *Doklady otdelenija Akademii nauk vysšej školy 2, 94–101.* Moskva.

Gumenjuk, A., Kostyshin, A., Simonova, S. (2002). An approach to the analysis of text structure. *Glottometrics 3, 61-89.*

Gumenjuk, A.S., Kostyshin, A.S. (1999). O kompjuternom analize tekstov i odnom formalizme segmentacii stichotvorenii russkoj literatury na sočetanija fonem. In: *Kvantitativnaja lingvistika i semantika. Vypusk 10, 3-18* . Novosibirsk: NGPU.

Orlov, Ju.K. (1980). Nevidimaja garmonija. In: Č*islo i mysl'. Vypusk 3, 70-106*. Moskva: Znanie.

Shreider, U.A., Sharov, A.A. (1982). *Sistemy i modeli*. Moskva: Radio i svjaz'.

Glottometrics 8, 2004, 68-74 *To Walter A. Koch at the occasion of his seventieth birthday*

Script complexity

Gabriel Altmann[1]

Abstract. This article describes a simple method for measuring script complexity by weighting the form of the script's symbols and their connections.

Keywords: Script, complexity

1. Any measurement of complexity is based on a set of criteria. Complexity is not an inherent property of things; things simply exist. Rather, complexity is a property of how we perceive and interpret their structure or how we construct their shape. Though there is a very extensive research in complexity in different sciences (cf. Peak, Frame 1994; Lewin 1992; Coveney, Highfield 1995; Mainzer 1997 etc., cf. also specialized journals), our unique problem here is to devise a procedure to ascribe grades of complexity to written signs. Different methods could be used, based on different aspects of the script; for instance ease of retention, ease of writing, the number of movements performed without interruption, a psychological scaling, and so on. Here, a graphical criterion is proposed.

It must be noted that this is not an exercise in *describing* or *characterizing* individual signs, whether using a special descriptive system or a unique description for computer use (cf. Watt 1975, 1980, 1981, 1988, 2002; Eden 1961, Eden, Halle 1961; Gibson et al. 1963, Gibson 1965, 1969; Koch 1971; Mounin 1970; Althaus 1973; Volockaja et al. 1964), but rather a method for capturing their complexity. It can be considered a generalization of Grzybek´s (2004) approach.

Intuitively, most people would say that Japanese signs (*kanji*) are more complex than Tibetan letters, which in turn are more complex than Latin letters. At the same time, perhaps most would agree that ´A´ is simpler than ´A´ or ´A´. To devise a procedure to help differentiate between these complexities, we require criteria that

(1) are applicable to all scripts;
(2) are simple to use;
(3) can be adapted to idiosyncrasies of individual scripts or styles.

Requirement (3) is especially important, because in writing (for instance) a Japanese sign, we can consider a horizontal stroke connected at the right side with a downward stroke as 1 line with an abrupt change of direction – as it is written in Japanese – or as 2 lines with a touching point. Again, intuitively one would say that ´≡´ is simpler than ´≠´ or ´A´, all of which have 3 strokes. Thus the system of criteria must take into account not only the number of strokes, but also their mutual relations. But *the direction of writing is irrelevant.*

The first problem is simple. There are just three kinds of elementary units of which any script is made up:

[1] Address correspondence to: G. Altmann, e-mail: 02351973070-0001@t-online.de.

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

The **points** have, of course, two dimensions; they are full circles, full squares, full triangles, and so on (only geometrical points are dimensionless). Thus, it is also possible to consider such shapes as filled figures. In that case, the contour outlines are evaluated as lines and the filling is assigned a value of 1 point.

The **straight lines** can be of any thickness (even variable thickness), any length and any direction. If the thickness is relevant, then again, the filling is assigned a value of 1 point and the contours are evaluated as lines.

The **arches** are simple, or prolonged by straight lines (*hooks*), but never closed. If the arches are thick or irregular, then the same considerations discussed above for points and straight lines apply.

Thus, ´O´ consist of two arches, ´U´ of one arch, whereas ´j´ consists of a point and an arch. How the components are evaluated may also depend on the size of the printed letters. For example Russian ж (12 point) appears to be no more than a number of straight lines; but if one magnifies it to 48 point Ж it becomes clear that the parts may be interpreted in different ways. There can be points, straight lines, arches and fillings, and there is no unique interpretation. Fortunately, this problem exists only with computer writing systems where the size can be manipulated.

To these values, those of **juncture forms** between the above elements must be added. There are as follows:

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

A **continuous contact** or **change** of direction means that the ends of arches are juxtaposed (joined with 0 angle). ´O´ consists of 2 arches joined by 2 continuous transitions. A wave consists of as many parts as there are turning points on it, plus the connections.

---

[2] Eden (1961) and Eden and Halle (1961) distinguish *hook*, *arch* and *loop*, but for the purposes of measuring complexity they need not be distinguished, as their complexity is equal.

Two straight lines joined at their ends can be considered as two straight lines (2+2) with 1 **crisp contact** (2) yielding 2+2+2 = 6. The Japanese prefer to consider ´┐´ as one stroke but when calculating complexity it should be treated as two strokes. However, even an arch and a straight line can be joined in a crisp way. The same holds for greater points.

Of course, one could modify this scale of complexity in many different ways. The system must allow anyone to evaluate the complexity of any script. It is very probable that a native writer of Burmese or Tibetan would propose another scaling system.

For the sake of illustration let us present some analyses using different scripts:

	Types	**Connections**	**Total**	**Comments**
Latin Arial				
A	2 2 2	2 2 2	12	3 straight lines; 3 crisp contacts
B	2 3 3	2 2 2 2	16	1 straight line; 2 arches; 4 crisp contacts
C	3 3	1	7	2 arches; 1 crisp contact
D	2 3	2 2	9	
O	3 3	1 1	8	
X	2 2	3	8	
Y	2 2 2	2	8	
Z	2 2 2	2 2	10	
a	3 3 3	2 2	13	
b	2 3	2 2	9	
c	3 3	1	7	
d	2 3	2 2	9	
y	2 3	2	7	
Ò	3 3 2	1 1	10	
Ô	3 3 2 2	1 1 2	14	
Õ	2 2 3 3	1 1	12	
Ö	3 3 1 1	1 1	10	
ß	3 3 3	1 1	11	3 arches; 2 continuous contacts
æ	3 3 3 3 2	1 1 2 2 2 2	24	4 arches; 1 straight line; 2 continuous contacts; 4 crisp contacts
Ķ	2 2 2 3	2 2	13	
Japanese *kanji* (20 pt)				
亜	8(2)	8(2)+4(3)	44	
圧	4(2)+3	2(2)+3	18	
悪	11(2)+3	9(2)+4(3)	55	
握	12(2)+2(3)	3(3) +8(2)	52	

扱 5(2) +4(3) 5(2)+3(3) 41

Arabic (48 pt) (considering the lines and points without contours):

ئ 2 2 3 3 3 2 2 2 19 (there are 2 straight lines)

ت 1 1 3 3 2 10

غ 1 3 3 2 9

ظ 1 1 2 3 3 2 2 2 3 19 2 points, one of them crossed by an arch

ي 1 1 2 3 3 2 2 14

ق 1 1 3 3 2 2 12

If the thick lines are treated as two-dimensional bodies, the complexity will be higher. The same holds for serifs which can be treated in three ways: (a) ignoring them because they are merely thicker continuations of strokes; (b) considering them as points (circles or triangles); (c) considering them points (the circles) and straight lines.

For Cyrilic (Times New Roman) we would obtain:

ж 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 30 The ends are points (6), 5 straight lines, 7 crisp contacts

Looking at a larger variant it is also possible to take into account all the contours because the lines are two-dimensional. Thus in Ж we have:

- for the left upper part ,four arches [4(3)], 3 continuous connections [3(1)], and two crisp ones [2(2)], yielding 19; the same for the right upper part, yielding 38;
- the left lower part contains 3 arches [3(3)] and one straight line [2], connected continuously in 1 case and crisply in 3 cases [1+3(2)], yielding 17; and the same for the right lower part, yielding 34 altogether;

- the middle part contains 8 straight lines [8(2)], joined with one another and with the other parts crisply in 10 cases [10(2)], yielding 36;
- for the filling we have 1 point,
- thus, in total, the value for this symbol is 38 + 34 + 36 + 1 = 109.

2. The next question is to calculate the complexity of a whole alphabet (or script) and compare this mechanical evaluation with the intuitive evaluations of test persons who did not previously know the script. A European confronted with Latin script would consider all letters probably equally complex even though, in computational terms, the number of commands for writing them is different. Our aim is not a vector description or characterization of signs but an objective method for evaluating a kind of complexity. This intuition-based test would consist of establishing X ranks of complexity and asking the test subjects to ascribe individual signs to these ranks.

Since modern scripts are very stable, the printed signs or letters do not depend on other properties; however, in handwritten texts or stenography, the letters can look different. The only script taking account of letter frequency is the Morse alphabet. On the other hand, the complexity of a sign can be linked with its frequency. Simpler signs can be used more frequently. In Japanese the stroke number of a kanji is correlated with its frequency and *eo ipso* with other properties of the morpheme (cf. Grzybek 2004).

All scripts developed slowly and in some cultures exist in multiple varieties, especially if they pre-date the printing press. Thus one possibility is to analyze the diversity of a writing system. Another would be to compare the development of the complexity of individual signs or of a whole alphabet. A very interesting question is whether the slow process from icon to symbol accompanies a simplification or an increase in complexity. An associated problem is the rate of change of complexity over time. Another problem is to compare different printed fonts – for example, Arial differs in complexity from Times New Roman, or from Vivaldi which is very complex – or even the same font at different sizes.

Even individuals differ in their writing, but our problem is not a graphological one. However, it is possible that a greater complexity of writing is linked with some character problems of the writer or of the style or of the kind of document, as is well known. With handwritten letters there is a further problem of connections and of the problematic existence of straight lines.

3. For the sake of comparison let us analyze the *Arial* and the *Courier New* fonts as shown in Table 1 below.

Table 1
Complexities of the Arial (12 pt) and Courier New (14 pt) fonts

Arial		Courier New	
Letter	Complexity	Letter	Complexity
A	12	A	22
B	16	B	16
C	7	C	11
D	9	D	9
E	14	E	26
F	10	F	22
G	15	G	15
H	10	H	26

I	2	I	10
J	3	J	7
K	10	K	26
L	6	L	14
M	14	M	26
N	10	N	20
O	8	O	8
P	9	P	13
Q	13	Q	21
R	14	R	22
S	15	S	23
T	6	T	18
U	3	U	11
V	6	V	14
W	14	W	22
X	7	X	23
Y	8	Y	20
Z	10	Z	18

Adding a serif (considered a straight line) makes the script much more complex. The mean complexity of Arial (12 pt) is 251/26 = 9.65 while Courier New (14 pt) yields 463/26 = 17.81, that is, it is almost twice as complex. However, in the case of certain individual letters, the value is the same for both fonts, for instance D, where the Courier New serif is simply considered as a part of the arch.

References

Althaus, H.P. (1973). Graphetik. In: Althaus, H.P., Henne, H., Wiegand, H.E. (eds.), *Lexikon der germanistischen Linguistik: 105-110*. Tübingen: Niemeyer.

Coveney, P. and Highfield, R. (1995). *Frontiers of Complexity. The search for order in a chaotic world.* New York: Ballantine.

Eden, M. (1961). On the formalization of handwriting. In: Jakobson, R. (ed.), *Structure of langugae and its mathematical aspects*. Providence: American Mathematical Society.

Eden, M. and Halle, M. (1961). The characterization of cursive handwriting. In: Cherry, C. (ed.), *Information Theory: Fourth London Symposium*. Washington, D.C.: Butterworths.

Gibson, E.J. (1965). Learning to read. *Science 148, 1066-1072.*

Gibson, E.J. (1969). *Principles of perceptual learning and development*. New York: Appleton Xentury Crofts.

Gibson, E.J., Osser, H., Schiff, W., Smith, J. (1963). An analysis of critical features of letters, tested by a confusion matrix. In: *Cooperative Research Project No. 639: A Basic Research program on Reading.* Washington: U.S. Office of Education. et al

Grzybek, P. (2004). A study of Russian graphemes (to appear in *Festschrift für T.M. Nikolaeva*).

Koch, W.A. (1971). *Taxologie des Englischen*. München: Fink.

Peak, D. and Frame, M. (1994). *Chaos under Control. The Art and Science of Complexity*. New York: Freeman.

Lewin, R. (1992). *Complexity. Life at the Edge of Chaos*. New York: Macmillan.

Mainzer, K. (1997). *Thinking in Complexity. The complex dynamics of matter, mind, and mankind.* Berlin: Springer.

Mounin, G. (1970). *Introduction à la sémiologie*. Paris : Les Editions de Minuit.

Volockaja, Z.M., Mološnaja, T.N., Nikolaeva, T.M. (1964). *Opyt opisanija russkogo jazyka v ego pis´mennoj forme.* Moskva.

Watt, W.C. (1975). What is the proper characterization of the alphabet? I. Desiderata. *Visual Language 9, (4): 293-327.*

Watt, W.C. (1980). What is the proper characterization of the alphabet? II: Composition. *Ars Semiotica 3 (1), 3-46.*

Watt, W.C. (1981). What is the proper characterization of the alphabet? III. Appearance. *Ars Semiotica 4 (3), 269-313.*

Watt, W.C. (1988). What is the proper characterization of the alphabet? Part 4: Union. *Semiotica 70 (3/4), 199-241.*

Watt, W.C. (2002). What is the proper characterization of the alphabet? Part V: Transcendence. *Semiotica 138, 131-178.*

Glottometrics 8, 2004, 75-78

Zur Ausbreitung von Wörtern arabischer Herkunft im Deutschen

Karl-Heinz Best

Abstract. This study presents a further support of the logistic law, known in linguistics as Piotrowski law, using data which can be gathered from Tazi's monograph *Arabismen im Deutschen: Lexikalische Transferenzen vom Arabischen ins Deutsche* (1998).

Keywords: Arabisms, German, Piotrowski-law

Das logistische Gesetz als Modell für Entlehnungsprozesse

Altmann (1983), Altmann u.a. (1983) sowie Best & Altmann (1986) haben die Hypothese entwickelt, dass Sprachwandel generell und damit auch Entlehnungsprozesse einem Sprachgesetz des logistischen Typs entsprechend verlaufen sollten. Hinzu kam der sog. „reversible" Sprachwandel, der auf die gleiche Weise einsetzt, dann aber irgendwann wieder zurückgenommen wird. Eine ganze Reihe von Untersuchungen haben gezeigt, dass dieses Modell in seinen unterschiedlichen Formen sich bewährt, wenn nur ausreichend Daten verfügbar sind (Best 2003; Körner 2004; und viele andere mehr). Der Verlauf der Entlehnung von Arabismen konnte noch nicht getestet werden. Bisher wurden solche Prozesse daraufhin untersucht, wie viele Wörter aus einer bestimmten Sprache oder Sprachgruppe direkt ins Deutsche gelangten. Dies betrifft nur wenige arabische Wörter, darunter Atlas (Seidenstoff), Fakir, Haschisch, Kadi, Scheich und Sultan (Tazi 1998: 311ff.). Solche unmittelbar aus dem Arabischen stammenden Wörter erreichen in etymologischen Wörterbüchern lediglich einen Anteil von 0.10% (Duden. Herkunftswörterbuch ³2001) bzw. 0.11% (Duden. Etymologie 1963) und im Fremdwörterbuch (Kirkness [Hrsg.] 1988) von 0.23% aller Übernahmen (Best 2001; Körner 2004), zu wenig, um die Gesetzmäßigkeit des Übernahmeprozesses zu testen. Die meisten arabischen Wörter erreichten das Deutsche jedoch über Vermittlersprachen, meist über Französisch und Italienisch. Berücksichtigt man auch diesen Wortschatz, dann erhält man genügend Belege, um das Piotrowski-Gesetz noch einmal zu prüfen.

Wörter arabischer Herkunft im Deutschen

Thema der folgenden Ausführungen sind also alle Wörter, die aus dem Arabischen stammen und direkt oder über Vermittlersprachen das Deutsche erreicht haben. Grundlage der Datengewinnung ist die Untersuchung von Tazi (1998); alle von ihr erfassten Wörter wurden in eine Liste eingetragen, mit Angabe des Zeitpunkts der Übernahme. Bei den Wörtern, für die bei Tazi keine Zeitangaben zu finden waren, wurden diese aus den etymologischen Wörterbüchern *Duden. Herkunftswörterbuch* (³2001), *Kluge* ([24]2002) und *Pfeifer* ([Ltg.] ²1993) ergänzt, sofern das möglich war. Alle die Wörter, die Tazi auflistet, die aber weder in *Duden.*

Fremdwörterbuch ([7]2001) noch in *Duden. Das Große Wörterbuch der deutschen Sprache in 8 Bänden* ([2]1993-95) enthalten sind, wurden als im Deutschen nicht mehr vorhanden eingestuft und aus der Liste gestrichen. Übrig blieben genau 150 datierbare Wörter arabischer Herkunft im Deutschen, von denen angenommen wird, dass sie noch zum Bestand des deutschen Wortschatzes gehören. Diese sind in Tabelle 1 als „beobachtet" aufgeführt, getrennt nach den Jahrhunderten ihrer Übernahme. Die beobachteten Werte wurden für die Untersuchung noch in kumulierte Werte überführt. Tabelle 1 beginnt mit dem 14. Jahrhundert; hier sind alle Wörter zusammengefasst, die entweder einem der Jahrhunderte bis zum 14. einschließlich zugeordnet oder allgemeiner als „mhd.", „spätmhd." eingestuft wurden. Die folgenden Zeiträume betreffen immer nur das angegebene Jahrhundert.

Als Nächstes wurde mit einem entsprechenden Programm in NLREG geprüft, ob auch der Zuwachs von Wörtern arabischer Herkunft dem logistischen Gesetz

$$(1) \quad p_t = \frac{c}{1 + ae^{-bt}}$$

entspricht. Das Ergebnis findet sich in der Tabelle 1 unter „berechnet". Es handelt sich um die Werte, die man erhält, wenn man die Formel (1) an die kumulierten Werte anpasst. Das Ergebnis ist hervorragend, wie der Testwert D = 0.996 und die folgende Graphik (Abb. 1) zeigen.

Tabelle 1
Arabismen im Deutschen

Jhd.	t	beobachtet	kumuliert	berechnet
14.	1	38	38	34.0934
15.	2	14	52	56.3274
16.	3	32	84	83.4508
17.	4	26	110	109.8020
18.	5	21	131	130.3095
19.	6	14	145	143.6839
20.	7	5	150	151.4299
a = 7.4099 b = 0.6964 c = 160 D = 0.996				

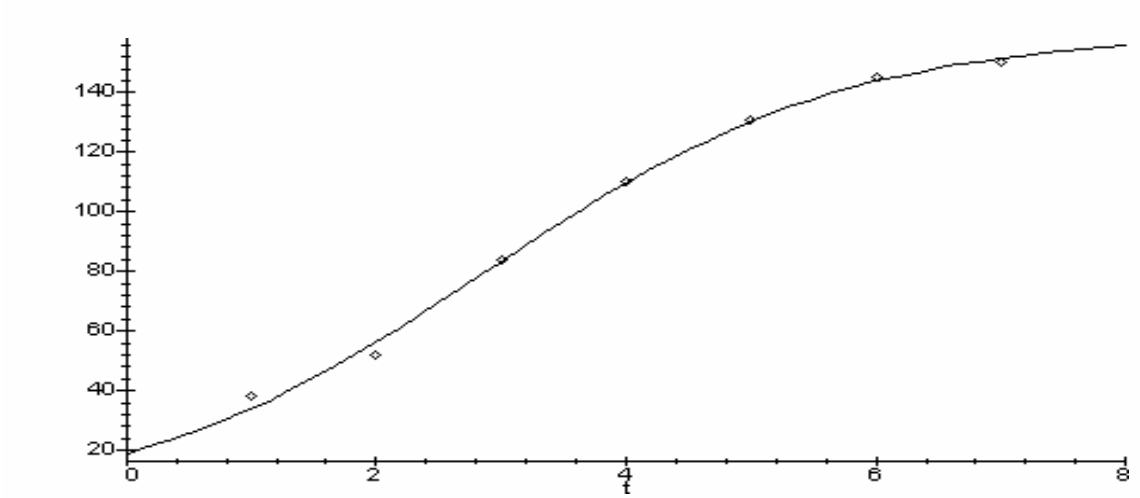


Abb.1 Die Entwicklung der Arabismen im Deutschen. (In dieser Graphik steht t = 1 für die Entlehnungen bis zum 14. Jahrhundert einschließlich, t = 2 für das 15. Jahrhundert; etc.)

a, *b* und *c* sind die Parameter des Modells; *c* gibt den Zielwert an, auf den nach der Berechnung der Prozess hinausläuft. *D* ist der Determinationskoeffizient, der höchstens den Wert 1 erreichen kann.

Auch dieser Sprachwandel verläuft also gemäß dem logistischen Gesetz, wie Tabelle und Graphik belegen; er befindet sich offenbar in seiner Endphase. Der Determinationskoeffizient $D = 0.996$ signalisiert, dass das Modell hervorragend geeignet ist, um diesen Entlehnungsprozess in seinem Verlauf zu erfassen.

Schlussbemerkung

Die Untersuchung hat ergeben, dass die Entlehnungen aus dem Arabischen ins Deutsche dem sog. Piotrowski-Gesetz in der Form des unvollständigen Sprachwandels (Formel 1) folgen. Dieses Gesetz wird damit einmal mehr als ein sehr valides Modell für jegliche Art von Sprachwandel bestätigt.

Abschließend sei angemerkt, dass diese Darstellung die Arabismen im Deutschen nicht vollständig erfasst. Es gibt in der älteren deutschen Literatur, vor allem in der Reiseliteratur, etliche weitere Wörter arabischer Herkunft, die wie „Sacker(falk)“ (eine Falkenart) in den konsultierten Wörterbüchern nicht mehr aufgeführt sind. Hinzu kommen solche, die nicht datiert werden konnten („Alkazar“: Burg, Schloss; „Quintal“: Gewichtseinheit); wieder andere Wörter wie „materaz“ (Matratze) sind von einer neueren Form (in diesem Fall: „Matratze“) verdrängt worden. Nicht zu vergesssen auch solche Ausdrücke, die wie „Dschihad“ (heiliger Krieg) und „Falafel“ (Bällchen aus Kichererbsen und Linsen) wohl erst nach Tazis Untersuchung eine größere Verbreitung im Deutschen gewonnen haben.

Nach Auswertung des Fremdwörterbuchs (Kirkness 1988) ist das Arabische unter insgesamt 35 Sprachen, aus denen das Deutsche Wörter entlehnt hat, an 8. Stelle platziert (Best 2001: 14); nach *Duden. Herkunftswörterbuch* (^{3}2001) an 16. Stelle unter 32 „Geber“-Sprachen (Körner 2004: 30). Es gehört jedoch zu den Sprachen, die hauptsächlich über Vermittlersprachen das Deutsche erreicht haben; berücksichtigt man diesen Aspekt, so wird der Einfluss des Arabischen auf den deutschen Wortschatz als wesentlich größer einzuschätzen sein.

Literatur

Altmann, Gabriel (1983). Das Piotrowski-Gesetz und seine Verallgemeinerungen. In: Best, Karl-Heinz, & Kohlhase, Jörg (Hrsg.), *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, Barron (ed.), *Historical linguistics: 104-115*. Bochum: Brockmeyer.

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

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

Best, Karl-Heinz, & Altmann, Gabriel (1986). Untersuchungen zur Gesetzmäßigkeit von Entlehnungsprozessen im Deutschen. *Folia Linguistica Historica 7, 31-41.*

Duden. Etymologie (1963). Mannheim: Bibliographisches Institut - Dudenverlag.

Duden. Fremdwörterbuch. (^{7}2001). 7., neu bearbeitete und erweiterte Auflage. Mannheim/ Leipzig/ Wien/ Zürich: Dudenverlag.

Duden. Herkunftswörterbuch (^{3}2001). 3., völlig neu bearbeitete und erweiterte Auflage. Mannheim/ Wien/ Zürich: Dudenverlag.

Duden. Das große Wörterbuch der deutschen Sprache in 8 Bänden. (²1993-95). 2., völlig neu bearbeitete und stark erweiterte Auflage. Mannheim/ Leipzig/ Wien/ Zürich: Dudenverlag.

Kirkness, Alan (Hrsg.) (1988). *Deutsches Fremdwörterbuch* (1913-1988). Begründet v. Hans Schulz, fortgeführt v. Otto Basler, weitergeführt im Institut für deutsche Sprache. Bd. 7: Quellenverzeichnis, Wortregister, Nachwort. Berlin/ New York: de Gruyter.

Kluge. Etymologisches Wörterbuch der deutschen Sprache. ([24]2002). Bearb. v. Elmar Seebold. 24., durchgesehene und erweiterte Auflage. Berlin/ New York: de Gruyter.

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

Pfeifer, Wolfgang [Ltg.] (²1993/1995). *Etymologisches Wörterbuch des Deutschen.* München: dtv.

Tazi, Raja (1998). *Arabismen im Deutschen: Lexikalische Transferenzen vom Arabischen ins Deutsche.* Berlin/ New York: de Gruyter. (= Diss. phil., Heidelberg 1994).

Verwendete Software

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

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

Glottometrics 8, 2004, 79-98

History of Quantitative Linguistics

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

V. Dmitrij Nikolaevič Kudrjavskij (1867-1920) – ein Wegbereiter quantitativer Methoden in der russischen Sprachwissenschaft

D.N. Kudrjavskij (1867-1920)

Къ статистикѣ глагольныхъ формъ въ Лаврентьевской лѣтописи.

Erste statistische Arbeit von D.N. Kudrjavskij zur Statistik von Verb-Formen in der Laurentiuschronik aus dem Jahr 1909

Im Rahmen von wissenschaftsgeschichtlich orientierten Arbeiten zur Entwicklung quantitativer Verfahren in der russischen Sprachwissenschaft wird in mehreren Fällen auf die Pionierarbeiten von Dmitrij Nikolaevič Kudrjavskij (1867-1820) verwiesen (vgl. Papp 1966, Kempgen 1995, Grzybek/Kelih 2004). Die Originalität und die inhaltliche Breite seiner Arbeiten legen es nahe, näher auf die sprachwissenschaftlichen und insbesondere statistischen Arbeiten dieses russischen Wissenschaftlers einzugehen.

Vor einer detaillierten inhaltlichen Darstellung der Arbeiten Kudrjavskijs sind vorweg einige bibliographische Eckpunkte seines wissenschaftlichen Werdeganges (vgl. Smirnov 1971) zu nennen: Nach dem Studium an der historisch-philologischen Fakultät in Sankt Peter-

burg (1885-1891) und Aufenthalten in Deutschland wird Kudrjavskij 1898 als Ordinarius an die Universität von Jur'ev (ehemals Dorpat, heute: Tartu) auf den Lehrstuhl für deutsche und vergleichende Sprachwissenschaft berufen. Diese westlichste russische Universitätsstadt stellte in diesen Jahren einen Fokus von quantitativ orientierten linguistischen Arbeiten dar, lehrten und arbeiteten doch bedeutende Linguisten wie Baudouin de Courtenay und A.S. Budilovič an dieser Universität. Für Kudrjavskij sind die Jahre seiner Professur in Jur'ev von einer hohen wissenschaftlichen Produktivität geprägt: Neben einer Reihe von Monographien wie beispielsweise *Psichologija i jazykoznanie* [Psychologie und Sprachwissenschaft] (Kudrjavskij 1904) und *Vvedenie v jazykoznanie* [Einführung in die Sprachwissenschaft] (Kudrjavskij 1912/1913)[1] verfasst Kudrjavskij drei Artikel, die explizit auf der Anwendung statistischer Methoden basieren (vgl. Kudrjavskij 1909, 1911, 1912). Abschließend zu seiner Biographie ist anzumerken, dass Kudrjavskij die Universität Jur'ev (Tartu) im Zuge der Wirren des Ersten Weltkrieges 1918 verlassen musste und mit der gesamten Belegschaft nach Voronež evakuiert wurde. Dort führte er seine Tätigkeit im Rahmen der neu geschaffenen Staatlichen Universität Voronež bis zu seinem Tod im Jahre 1920 fort.

In dem offensichtlich ersten von Kudrjavskij verfassten Artikel aus dem Jahr 1909 mit dem Titel *„K statistikě glagol'nych form v Lavrent'evskoj lětopisi" [Zur Statistik von Verbformen in der Laurentiuschronik]* zeichnet sich bereits eine bestimmte wissenschaftliche Grundlinie ab, die durch die explizite Notwendigkeit der Anwendung quantitativer Verfahren in der Sprachwissenschaft gekennzeichnet ist. Inhaltlich geht es um die Evolution der Verwendung von bestimmten Tempusformen im Altrussischen und Russischen. Wichtig erscheinen in diesem Zusammenhang vor allem die methodologischen Reflexionen zur Anwendung von statistischen Methoden in der Sprachwissenschaft:

> „[...] čto statističeskij metod daet vozmožnost' otmětit' javlenija, obyknovenno uskol'zajuščija ot vnimanija islědovatelja. Meždu těm po charakteru svoemu ėti javlenija otličajutsja universal'nost'ju, tak kak massovyja nabljudenija zachvatyvajut samuju atmosferu žizni jazyka." [… dass die statistische Methode die Möglichkeit einräumt, Phänomene zu registrieren, die gewöhnlich der Aufmerksamkeit des Forschers entgehen. Unterdessen zeichnen sich jedoch diese Phänomene durch Universalität aus, da eine Massenbeobachtung doch die ganze Atmosphäre der Sprache umfasst] (Kudrjavskij 1909: 53).

Ausgehend von dieser programmatischen Aussage zur Anwendung von Statistik, die eben darauf hinausläuft, dass mit der Hilfe von Statistik sprachliche Phänomene sichtbar gemacht werden können, sieht Kudrjavskij einen weiteren Vorteil quantitativer Methoden darin, dass die untersuchten Phänomene nicht durch subjektive Wertungen erfasst werden, sondern einzig und allein durch „leidenschaftslose Zahlen" (vgl. ebd. 1909: 54).

Inhaltlich wird in dieser Untersuchungen auf folgende Fragestellungen eingegangen: Auf der Basis eines von ihm erstellten vollständigen Verzeichnisses von Verbalformen aus der 1377 erschienenen Laurentius-Chronik (*Lavrent'evskaja letopis'*) geht er der Vorkommenshäufigkeit von Aorist-, Imperfekt- und Partizipialformen nach. Die Häufigkeit dieser Formen wird schrittweise pro 100 Zeilen der Chronik[2] angeben, wobei der prozentuale Anteil

[1] Interessant ist die Tatsache, dass dieses Standardwerk der russischen "vorrevolutionären" Sprachwissenschaft offensichtlich Stalin als inhaltliches Vorbild beim Verfassen seiner Beiträge zur Sprachwissenschaft im Jahr 1950 diente (Ende der 'Neuen Lehre' von N.Ja. Marr). Diese Hypothese, die sich indirekt auch in einer verstärkten Wiederkehr zu historisch-vergleichenden Arbeiten in der sowjetischen Linguistik der Jahre 1950-1956 nachweisen lässt, wird in Alpatov (1991: 185) unter Bezug auf V.A. Zvegincev vertreten.

[2] Die Auszählungen sind – wie sich Kudrjavskij (vgl. 1909: 49) zurecht beklagt – sehr zeitintensiv und aufwendig: dazu investierte er nach eigenen Angaben eine Stunde pro Tag nur für das Auszählen und die Erstellung seines Verb-Index der gesamten Laurentiuschronik. Die Zeitintensivität von Auszählungen und Berechnungen

im Anhang in der Form von Diagrammen dargestellt wird. Die unterschiedliche Häufigkeit von Temporalformen wird in Bezug zum jeweiligen Inhalt der Chronik gesetzt (Details dazu vgl. Kempgen 1995: 87f.). Seine umfangreichen Auszählungen bestätigen seine a priori formulierte Hypothese des Verschwindens der Aoristformen nicht; vielmehr zeigt sich, dass diese Form in der von ihm bearbeiteten Chronik mit einer konstanten Häufigkeit nachzuweisen ist. Insgesamt muss Kudrjavskij eingestehen, dass die von ihm präsentierten Resultate mehr Fragen als Antworten aufwerfen. Dennoch ist er von der Richtigkeit der von ihm angewandten Methode überzeugt, denn diese erste Untersuchung stellt dann den Ausgangspunkt für zwei weitere, ähnlich ausgerichtete Analysen dar.

In dem 1911 publizierten Artikel *„K istorii russkago prošedšago vremeni« [Zur Geschichte des russischen Präteritum]* untersucht Kudrjavskij auf der Basis von der nunmehr um weitere altrussische Schriftdenkmäler (Slovo o polku Igorevě, Russkaja pravda u.a.) erweiterten Textbasis, ob in einer chronologischen Perspektive die Vorkommenshäufigkeit bestimmter Tempusformen zu beobachten ist. Wiederum versteht er seine quantitativ orientierte Untersuchung als eine Möglichkeit, eine auf umfangreiches sprachliches Material gestützte Untersuchung durchzuführen, die Hinweise auf die Tendenz von sprachlichen Prozessen geben soll. Insbesondere geht es um die Verwendung von Verbalpartizipien, welche sprachhistorisch im Altrussischen mit bzw. ohne das Kopulativverb „byti“ für die Bildung der Vergangenheitsform herangezogen wurden. Genau um die Verwendungshäufigkeit dieser Verbalform mit oder ohne Kopulativverb geht es Kudrjavskij. Als Datenbasis für diese Rekonstruktion des Verschwindens des Kopulativverbes wurden beachtliche 33.000 unterschiedliche Tempusformen ausgezählt, die als zufriedenstellende Stichprobe für die Lösung der zugrunde gelegten Fragestellung (vgl. Kudrjavskij 1911: 121) angesehen wird. Die Interpretation der einzelnen analysierten Handschriften bringt das Ergebnis, dass für das Altrussische ein sukzessives Verschwinden des Kopularverbums bei der Bildung der Vergangenheitsform nachzuweisen ist, wobei eine Ausnahme für die 1. und 2. Person Singular zu gelten scheint (vgl. Kudrjavskij 1911: 137ff.).

Eine weitere statistische Untersuchung ist in Kudrjavksij (1912) zu finden, wo die Vorkommenshäufigkeit des Partizips Präsens Aktiv mit der Endung *–a* (bzw. *ja*, *y*) im Altrussischen (Textbasis: Laurentiuschronik) untersucht wird. Die erhaltenen statistischen Auszählungen ergeben nach Kudrjavskij (1912: 397) folgendes Ergebnis: Das Partizip Präsens Aktiv mit der Endung auf *–a* zeigt im untersuchten Textmaterial eine äußerst geringe Vorkommenshäufigkeit (ca. 21%), während die palatalisierte Endung *–ja* im Laufe der Sprachgeschichte aufgrund von phonetisch motivierten Veränderungen häufiger vertreten ist (ca. 78%).

Insgesamt sind die Arbeiten von Kudrjavskij als ein wichtiger Teilbereich der Vorgeschichte der Anwendung quantitativer Methoden in der russischen Sprachwissenschaft am Anfang des 20. Jahrhundert anzusehen. Es ist davon auszugehen, dass die Arbeiten von Kudrjavskij – im Gegensatz zu anderen Arbeiten in diesem Zeitraum – keineswegs methodologisch unbedarft sind, sondern als mustergültige sprachgeschichtliche Analysen zum altrussischen Verbsystem anzusehen sind. Folgende Merkmale sind dabei von hervorragender Bedeutung:

a.) Die Anwendung statistischer Verfahren ist auf die (einfache) Zählung von Tempusformen und auf die graphische Darstellung der Ergebnisse beschränkt. Diese

---

ist wohl in der Tat als hemmender Faktor bei der Etablierung von quantitativen Verfahren anzusehen. Zumindest hat dies seine Gültigkeit bis zur Möglichkeit einer computer-gestüzten Analyse, die erst ab den sechziger Jahren einsetzte.

Form der Anwendung von statistischen Methoden stellt für Kudrjavskij keinen Selbstzweck dar, sondern dient ihm – nach einleitender theoretischer Begründung der Notwendigkeit, statistische Verfahren in der Sprachwissenschaft anzuwenden – als Ausgangspunkt für die Überprüfung von a priori formulierten linguistische Hypothesen. Die gewählte Textbasis ist nicht nur aufgrund der systematischen Auswahl, sondern vor allem auch aufgrund des Umfangs (in einem Fall von ca. 33.000 Verbformen) als einer der ersten russischen 'Korpusuntersuchungen' innerhalb der Sprachwissenschaft zu verstehen.

b.) Insgesamt sieht Kudrjavskij seine Analysen nicht nur als eine Bestandsaufnahme sprachlicher Fakten, sondern es wird der Häufigkeit von linguistischen Formen eine erklärende Kraft zugesprochen: Diese Argumentation ist jedoch nicht in den oben erwähnten Analysen zu finden, sondern wird an anderer Stelle in die Diskussion eingebracht. Im Rahmen der Untersuchung von russischen Adverbialpartizipien wird festgestellt (vgl. Kudrjavskij 1915: 12f.), dass Adverbialpartizipien sich vor allem aus Partizipien gebildet haben, die im Altrussischen eine hohe Verwendungshäufigkeit aufweisen. In diesem Sinne kann Kudrjavskij auch der Verdienst zugeschrieben werden, explizit das Häufigkeitskriterium (vgl. Meier 1961: 55) als erklärenden Faktor für Prozesse des Sprachwandels in die Diskussion eingebracht zu haben.

Abschließend lässt sich festhalten, dass diese ersten statistisch orientierten sprachwissenschaftlichen Arbeiten auf keine breitere Rezeption gestoßen sind. Trotzdem stellen diese Arbeiten eine bemerkenswerte empirisch-quantitative Herangehensweise bei der Lösung von Fragen der Sprachevolution und des Sprachwandels des Altrussischen dar. Darüber hinaus sind seine Überlegungen in jeder Weise als Pionierarbeiten der russischen quantitativen Linguistik zu sehen.

Literatur

Alpatov, V.M. (1991). *Istorija odnogo mifa. Marr i Marrizm.* Moskva: Nauka.

Grzybek, P., Kelih, E. (2004). Zur Vorgeschichte quantitativer Ansätze in der russischen Sprach- und Literaturwissenschaft. In: *Quantitative Linguistik - Quantitative Linguistics. Ein internationales Handbuch/An International Handbook* (Herausgegeben von G. Altmann, R. Köhler, R. Piotrowski). New York: de Gruyter, 2004. [= Handbücher zur Sprach- und Kommunikationswissenschaft]

Kempgen, S. (1995): *Russische Sprachstatistik. Systematischer Überblick und Bibliographie.* München. [= Vorträge und Abhandlungen zur Slavistik, Band 26]. München.

Kudrjavskij, D.N. (1904). *Psichologija i jazykoznanie.* Sankt Peterburg.

Kudrjavskij, D.N. (1912). *Vvedenie v jazykoznanie.* Jur'ev.

Kudrjavskij, D.N. (1913). *Vvedenie v jazykoznanie. Izdanie 2. Ispravlennoe i dopolnennoe.* Jur'ev.

Kudrjavskij, D.N. (1909). K statistike glagol'nych form v Lavrent'evskoj letopisi, in: *Izvestija otdelenija russkago jazyka i slovesnosti Imperatorskoj Akademii Nauk, t. XIV, č. 2, 49-56.*

Kudrjavskij, D.N. (1911). K istorii russkago prošedšago vremeni. *Russkij filologičeskij vestnik LXV, 119-139.*

Kudrjavskij, D.N. (1912). Vvedenie v jazykoznanie. Jur'ev.

Kudrjavskij, D.N. (1912). Drevne-russkija pričastija nastojaščago vremeni dejstvitel'nago zaloga na –a. *Russkij filologičeskij vestnik LXVIII*; 119-139.

Kudrjavskij, D.N. (1916). K istorii russkich děepričastij. Vypusk 1. Děepričastija prošedšago vremeni. [= Učenye zapiski Imperatorskago Jur'evskogo Universiteta, XXIV]

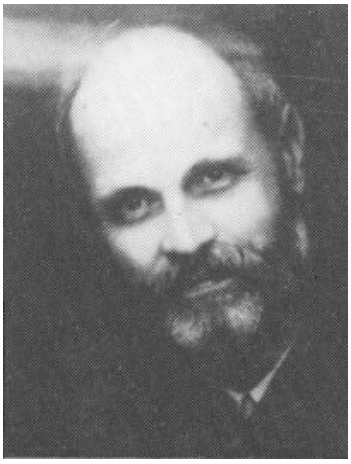
Meier, G. F. (1961). Das Zéro-Problem in der Linguistik. Kritische Untersuchung zur strukturalistischen Analyse der Relevanz sprachlicher Form. Berlin: Akademie Verlag. [= Schriften zur Phonetik, Sprachwissenschaft und Kommunikationswissenschaft, Nr. 2]

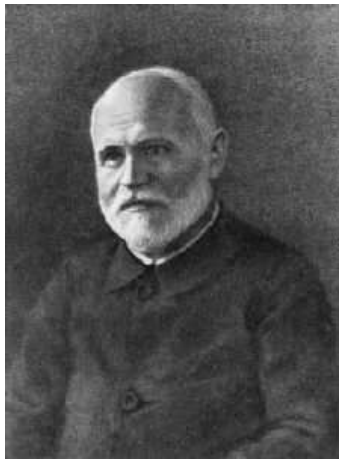
Papp, F. (1966), *Mathematical Linguistics in the Soviet Union*. [= Janua Linguarum, Series Minor, XL]. The Hague: Mouton.

Smirnov, S.V. (1971). Iz istorii jazykoznanija. Dmitrij Nikolaevič Kudrjavskij (1867-1920), *Russkaja reč' 3, 137-145*.

Emmerich Kelih, Graz

VI. Wincenty Lutosławski (1863 – 1954)
A Forgotten Father of Stylometry





Wincenty Lutosławski (1863 – 1954)

The origin and development of modern quantitative linguistics is associated with the structuralist revolution of the first decades of the 20th century. Support for this notion can be found in the words of one of the creators of structuralism, J. N. Baudouin de Courtenay (1845–1929), who in fact did not apply mathematical methods himself, but who did, while conducting field studies, realise the virtues of a quantitative description of language and foresaw the advent of rigorous investigations into the laws of language. Citing J. Rozwadowski's concept of the quantitative rules of language development (Rozwadowski 1909), he presented his view on the emerging relationships between the realm of numbers and "linguistic thought" (Baudouin de Courtenay 1927 [1990]: 549). His concept principally involves the semantic, syntactic, and morphologic representation of the number, dimensions, and intensities of attributes, and thus does not touch upon the concept of statistical linguistics operating with frequencies or other expressly numerical features of language elements. Nonetheless, this scholar perceived analogies between the physical domain, defined by precise and formalised laws, and language. He realised that the contemporary level of linguistic and mathematical knowledge was inadequate for the formulation of exact linguistic laws. "I, personally, having considered the rigour and functional dependency of the laws of the world of physics and chemistry, would hesitate to call that a 'law' which I consider merely an exceptionally skilful generalisation applied to phenomena at large" (*ibid*. 547). However, he anticipated such laws

also being formulated for linguistic relationships in future, "[...] the time for genuine laws in the psycho-social realm in general, and first and foremost in the linguistic realm, is approaching: laws which can stand proudly beside those of the exact sciences, laws expressed in formulae of the absolute dependency of one quantity on another" (*ibid*. 560).

The roots of Polish quantitative linguistics go back further, though, to the period prior to this revolution. The scholar who may be recognised as its forerunner and one of the creators of stylometry was Wincenty Lutosławski (1863–1954). A graduate of the Technical University of Riga and the University of Dorpat, he was a lecturer at the University of Kazan and professor at the universities of Vilnius and Cracow (Jadacki 1998: 54–87; Chyl 1999: 12; Lutosławski 1933 [1994]). Having been educated in a German secondary school (Mitau/ Mitawa/Jeglava in Latvia), lecturing at a Russian university, and being a classical philologist, in addition to being a Pole experiencing Poland's own peculiar form of Diaspora (Poland did not regain formal statehood until 1918), he had command of most of the European languages[3]. His main field of interest was Platonic philosophy, he was also fascinated in messianic teachings, spiritualism, and Polish national movement.

The issues which today associate the work of Lutosławski with quantitative linguistics, and precisely the methodology of stylometry, arose from his studies of Plato. One of the classical problems of Hellenism, unresolved to this day, is the periodisation of Plato's *Dialogues*. This is of vital significance for the interpretation of his legacy, as the chronological proximity (or remoteness) of the texts may suggest relationships in content (or the possible lack of such), which would consequently determine a reconstruction of the complete Platonic philosophical system (cf. Pawłowski, Pacewicz 2005).

Lutosławski decided to solve the problem of platonic chronology. Inspired by the ideas of the Scottish philosopher L. Campell (Lutosławski 1933 [1994]: 219–220), he worked out his own method based on the comparison of a great number of stylistic text characteristics. He was convinced that it would be possible to reconstruct the true order of platonic writings solely using their stylistic features: "If an exact definition be possible of the notes which distinguish Plato's style from the style of other writers, or by which a work written contemporaneously with the Laws differs from a work written at the time when Plato founded the Academy, then we may hope to ascertain the true order of Platonic dialogues according to the stylistic variations observed in them." (Lutosławski 1897a [1983]: 65–66) A concise formulation of his method is the *law of stylistic affinity* which states that: "Of two works of the same author and of the same size, that is nearer in time to a third, which shares with it the greater number of stylistic peculiarities, provided that their different importance is taken into account, and that the number of observed peculiarities is sufficient to determine the stylistic character of all the three works." (*ibid*. 152)

We shall introduce the fundamentals of Lutosławski's method below and then mention the origin of stylometry in light of his achievement. The novelty of this idea, compared with earlier work, is the attention to Lutosławski's role. Investigation indicates it was most probably he who first introduced the term "stylometry" into scientific use ("This future science of s t y l o m e t r y [emphasis mine – AP] may improve our methods beyond the limits of imagination [...]" – Lutosławski 1897a [1983]: 193, cf. also Lutosławski 1896, 1897b and 1898) and, despite being unfamiliar with modern statistical tools and research on the quantitative structure of lexicon and text, he defined the majority of its cardinal rules.

Lutosławski's method rests on a few premises, not always directly articulated, which he accepted on the basis of observation, research results, and intuition. The effect of these efforts is surprisingly good compared with the assumptions of modern stylometry, all the more so as

[3] During his university years he claimed speaking 9 languages (Lutosławski 1933 [1994]: 118–19). We might add here that his first wife was Sofía Pérez Eguía Y Casanova Lutosławska, a Spanish journalist, poet and novelist from Galicia.

the author was primarily interested in sequencing the works of Plato, while the question of their authenticity (thus authorship) was secondary (cf. Lutosławski 1933 [1994]: 225, cf. discussion in Pawłowski, Pacewicz 2005). In Lutosławski's view, the most important premises of the method of stylomertry are:

1) Reliable information about dating of some writings by the controversial author (e.g. *Laws*, considered as Plato's last text). It allows working out and verifying the hypotheses concerning the evolution of his style and the application thereof to the litigious works.
2) Existence of individual style in the texts of every author and its independence of contents: "Now the external form of a writer is his style, and it betrays him even if he for some reason may be professing thoughts very different from those which we usually associate with his name." (Lutosławski 1897a [1983]: 64)
3) Possibility of solving the question of author's arguable identity on the ground of stylistic proprieties of his texts, considered as external characteristics: „There is no exaggeration in this pretension, since questions of identification are generally settled by purely external tests." (*ibid*. 65)
4) Analogy between stylometry and graphology indicating potential effectiveness of the stylometric analysis. Lutosławski argued that if the uniqueness of handwriting is uncontroversial and officially recognised in the legal practice, a similar distinctive power should be associated with the characteristics of style: „The identity of handwriting, consisting in many minute signs difficult of definition, is held to be so far ascertainable, that on an expert's decision in such matters a man's life may sometimes depend. The limited number of marks of identity contained in a signature is sufficient to decide its authenticity for all purposes. […] If handwriting can be so exactly determined as to afford certainty as to its identity, so also with style, since style is more personal and characteristic than handwriting." (*ibid*. 65)
5) Large but limited set of relevant stylistic features (peculiarities): „It may be objected that, since science style has an almost infinite number of characteristic notes, it cannot be reduced to one external formula. The answer is, that a like infinity of characteristics exists in every object of natural science, and that science is possible only through the distinction of essential marks from those which are unessential." (*ibid*. 66). These features should appear in all the compared texts: „the number of observed peculiarities is sufficient to determine the stylistic character of all the three works." (*ibid*. 152).
6) Hierarchy of importance of the analysed stylistic features: „In order to draw our conclusions, we begin by recognising four degrees of importance, distinguishing stylistic peculiarities." (*ibid*.146).
7) Possibility of quantification and measurement of the degree of similarity of texts based on the number of shared stylistic features. Lutosławski formulated this principle as a *law of stylistic affinity*: „Of two works of the same author […] that is nearer in time to a third, which shares with it the greater number of stylistic peculiarities […]."(*ibid*. 152).
8) Superiority of techniques synthesizing complex information about text and style: "[…] we needed a greater number of facts than has been known heretofore to any single author; but we found that five hundred peculiarities, selected at random from the special investigation, were sufficient for our purpose." (*ibid*. 145). Text and style are considered here as very complex objects and this kind of synthesis cannot be obtained with a traditional methodology: "But the definition of style requires a deeper study, because style is not, like handwriting, accessible to the senses." (*ibid*.) It is worth emphasising that many

years after Lutosławski's publication certain coefficients of lexical richness and the methods of multidimensional scaling became fully operational as tools of the effectively synthesising information about text.

9) Unidirectional evolution of personal style during the whole period of authorial creativity (concerns only chronological research): „[…] that the style of some writers has changed in the course of years is a patent fact" (*ibid*. 64).

10) Necessity of comparing samples of equal length: „Of two works of the same author and of the same size […]" (*ibid*. 152).

Armed with the above premises, Lutosławski believed that by comparing those dialogues whose dates were beyond dispute with disputable texts whose similarities with the former were numerically expressed, it was possible to establish a complete chronology of Plato's works. Drawing on studies by other authors, he defined 500 characteristics of Plato's style and conducted a sequencing of the questionable dialogues on the basis of their appearance in 58'000 fragments (*ibid*. 74–139). Despite criticism, recent advances in Platonic studies (see: Brandwood 1990), and fundamental doubts as to any sort of periodisation of antique texts, Lutosławski's proposition enjoys recognition in some Hellenistic circles to this day. "Lutosławski's sequence [...] was widely accepted in the twentieth century. [...] Today, Lutosławski's canon is still functional, although it is being challenged by more recent research." (Kubikowska 1999: 6; cf. also Zaborowski 2000: 50)

From the perspective of modern quantitative linguistics, Lutosławski's technique is inadequate with respect to statistics, and the very question of periodising texts is in itself disputable. Multidimensional methods have supplanted most of the traditional solutions (cf. Wishart, Leach 1970), also connectionnist techniques employing artificial neural networks are proving to be increasingly effective (cf. Tweedie et al. 1996). Nevertheless, certain elements of his work present a permanent part of the development of not only classical philology, but quantitative linguistics as well. These involve the fact that it was Lutosławski who first introduced the term "stylometry", used until today, and defined some of its general principles (presented above). As a typical representative of the positivistic view of science, he put great hope in it for the future. "This exceptional importance of one particular case will enable us to decide questions of authenticity and chronology of literary works with the same certainty as palaeographers now know the age and authenticity of manuscripts." (Lutosławski 1897a [1983]: 193).

When analysing the actual influence of Lutosławski's work on the development of stylometry as a division of quantitative linguistics, it is worth turning our attention to the absence of his work in modern published studies. If he is cited, the philological or material aspects are emphasised, while his methodology is wholly absent (e.g. Herdan 1966: 1). For A. Kenny, of value was only Lutosławski's treatment on the state of research into the chronology of Plato's works: "The work of theses and other scholars was magisterially synthesized by the Polish scholar W. Lutosławski in his work of 1897, *The Origin and Growth of Plato's Logic*." (Kenny 1982: 3). Probably the only discussion of his method available so far appears in B. Pindlowa's study (1994: 18–20, 161), but there, too, there is no synthesis of his methodological premises and the postulates.

D. Holmes (1988), in a comprehensive and inspiring article on the history of stylometry, passes over Lutosławski's contribution in silence, observing, "Mendenhall's labours seemed to deter statisticians from following him, and there was a gap of some 30 years before G. Udny Yule In England and the American linguist George Zipf worked on alternative

features of style."[4] Despite all the reservations concerning Lutosławski, the expression "gap of some 30 years" stands in sharp contrast to that portion of his work devoted to the stylometric method. It's hard to explain the causes of this state of affairs. As a comment one could just invoke the words of Lutosławski himself, referring to the platonic studies at the end of the XIXth century: "As a Pole, the author may possibly be more impartial than the representatives of other nations more active in Platonic research. The works of British scholars are little known in Germany, and, on the other hand, many special German investigations are overlooked in France and Great Britain." (Lutosławski 1897a [1983]: vii)".

A provisional definition of the beginnings of stylometry would necessitate accepting a polygenetic concept consisting of a gradual crystallisation of the premises of this discipline based on the work of various researchers. Although it is difficult in a short article to discuss in detail all the studies which had a part in the process, we can mention the pioneers and most important creators of stylometry. In our opinion, this group would include, in chronological order, A. De Morgan[5], W. Lutosławski, G.U. Yule (1938, 1944), J.K. Zipf (1935, 1949), F. Mosteller, and D. Wallace (1978, 1984), as well as the authors of the first studies applying multidimensional analysis (cf. Holmes, Forsyth 1995). None of these can be assigned absolute priority in calling stylometry into being. Each though had a, lesser or greater part in this process.

SHORT CURRICULUM VITAE OF WINCENTY LUTOSŁAWSKI[6]

1863	Wincenty Lutosławski is born in Warsaw
1877–1881	secondary school (gymnasium) in Mitau/Mitawa (Jelgava), Latvia
1881–1883	studies at the Riga Polytechnic (Wilhelm Ostwald's class)
1883–1884	travels in Europe (Switzerland, France, Italy, Austria)
1884–1885	studies of chemistry at the Dorpat (Tartu) University, candidate's degree in chemistry
1884–1886	studies of philosophy at the Dorpat (Tartu) University under Gustav Teichmüller's supervision, candidate's degree in philosophy
1885–1886	studies of French philology at the École de Hautes Études in Paris, travel to Portugal, Spain and Morocco
1887–1888	studies of Plato under G. Teichmüler's supervision in Dorpat (Tartu)
1887	master's degree in philosophy at the Dorpat (Tartu) University
1888–1889	stay in Moscow, discovery of two unknown manuscripts of Giordano Bruno
1889–1890	stay in London

[4] This is with reference to a 1887 article by T. C. Mendenhall devoted to the empirical frequency distributions of words. The term or the notion of stylometry do not appear in this article (Mendenhall 1887).

[5] In a letter from 1851 he mentioned the relationship between the authorship of a text and mean word length. He also presented initial calculations of mean word length in the letters of St. Paul. He finally arrived at the conclusion that style allows distinguishing the author of a text even when the texts vary in topic. "I would have Greek, Latin or English tried and I should expect to find that one man writing on two different subjects agrees more closely with himself than two different men writing on the same subject." (De Morgan, 1882, quotation from a study by Williams (1970: 5)).

[6] This selective CV is based on an extensive biography presented by J.J. Jadacki (1998: 54–57).

1890–1893	"dozent" at the Kazan University, lectures in logic, psychology and history of philosophy
1893–1894	stay in Spain, USA and England
1894–1895	stay in Drozdowo near Łomża (Poland)
1895–1898	stay in Spain and in England
1898–1899	stay in Finland, Sweden, Denmark and Germany
1898	doctor's degree in philosophy at the Helsinki University
1889–1901	stay in Cracow
1889–1900	"Privatdozent" at the Jagiellonian University in Cracow
1901–1902	stay in Switzerland, lectures in Lausanne and Geneva
1904–1906	lectures at the University College in London
1907–1908	stay in the USA, numerous lectures
1908–1910	stay in Warsaw
1912–1916	lectures at the Geneva University
1916–1919	lectures at the University of Paris
1920–1933	professorship at the University of Vilna
1921	lectures on philosophy in Poznań
1923	lectures on philosophy in Warsaw and Lvov
1929	retirement at the University of Vilna
1931–1932	stay in France
1946–1948	lectures at the Jagiellonian University in Cracow
1954	Wincenty Lutosławski dies in Cracow

References

Baudouin de Courtenay, Jan (1927). *Ilościowość w myśleniu językowym* [Quantity as a dimension of thought about language]. In: *Symbolae Grammaticae in honorem Ioannis (Jan) Rozwadowski* v.1. (Festschrift) Cracoviae: Gebethner & Wolff, 3–18. Reprint: Baudouin de Courtenay, J. (1990), *Dzieła wybrane* t.IV [Selected Writings, v.4]. Warszawa: PWN, *546–563.*

Brandwood L. (1990). *The Chronology of Plato's Dialogues*. Cambridge: Cambridge University Press.

Chyl S. (1999). *Lutosławscy* [The Lutosławski family]. Drozdowo-Zambrów: PWSM Zambrów.

De Morgan S.E. (1882), *Memoir of Augustus de Morgan, by his wife Sophia Elisabeth de Morgan, with Selection of his Letters*. London: Longmans, Green, and co.

Herdan G. (1966). *The Advanced Theory of Language as Choice and Chance*. Berlin, Heidelberg, New York: Springer Verlag.

Holmes D. (1998). The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing* 13/3,111–117.

Holmes D., Forsyth R.S. (1995). The Federalist Revisited: New Directions in Authorship Attribution. In: *Literary and Linguistic Computing 10/2, 111–127.*

Jadacki J.J. (1998). *Wincenty Lutosławski, rozdział z dziejów myśli polskiej* [Wincenty Lutosławski, a chapter from the history of Polish science]. In: Klukowski B., *Lutosławscy w kulturze polskiej: 54–87.* Drozdowo: Towarzystwo Przyjaciół Muzeum Przyrody,

Kenny A. (1982). *The Computation of Style*. London: Pergamon Press.

Kubikowska E. (1999). *Od redakcji* [From the editors]. In: Platon, *Dialogi*, v.1 (przekład W. Witwicki). Kęty: Wydawnictwo Antyk, 3–8.

Lutosławski W. (1896). *Sur une nouvelle méthode pour déterminer la chronologie des dialogues de Platon* (mémoire lu le 16 mai 1896 à l'Institut de France). Paris: H. Welter.

Lutosławski W. (1897a). *The origin and growth of Plato's logic*. London, New York, Bombay: Longmans, Green and Co. Reprint: *The origin and growth of Plato's logic*. Hildersheim: Georg Olms Verlag, 1983.

Lutosławski W. (1897b). On stylometry. Abstract of a paper read at the Oxford Philological Society on May 21st by Dr. W. Lutosławski, of Drozdowo, near Lomza, Poland. *Classical Review 11, 284–286.*

Lutosławski W. (1898). Principes de stylométrie. *Revue des études grécques 41, 61–81.*

Lutosławski W. (1933). *Jeden łatwy żywot* [One easy existence]. Warszawa: Hoesick. Reprint: *Jeden łatwy żywot* [One easy existence]. Kraków: Fundacja im. Wincentego Lutosławskiego 1994.

Mendenhall T.C. (1887). The characteristic curves of composition. *Science 11, 237–249.*

Mosteller F., Wallace D. (1978). *Deciding authorship*. In: Tanur J.M., Lehmann E.L. et al (eds.) (1978), *Statistics: a guide to the unknown* (2nd ed.): 207-219. San Francisco: Holden-Day.

Mosteller F., Wallace D.L. (1984). *Applied Bayesian and Classical Inference*. New York, Berlin etc.: Springer Verlag.

Pawłowski A., Pacewicz A. (2005). Wincenty Lutosławski – philosophe, helléniste ou fondateur sous-estimé de la stylométrie? *Historigraphia Linguistica* [to be published].

Pindlowa W. (1994), *Infometria w nauce o informacji* ['Infometry' in the science of information]. Kraków: Universitas.

Rozwadowski J. (1909). Ein quantitatives Gesetz der Sprachentwickelung. *Indogermanische Forschungen XXV, 38–50.*

Tweedie F.J., Singh S., Holmes D.I. (1996). Neural Network Applications in Stylometry: The Federalist Papers. *Computers and the Humanities 30, 1–10.*

Williams C.B. (1970). *Style and vocabulary: numerical studies*. London: Griffin.

Wishart D., Leach S. (1970). A multivariate analysis of Platonic prose rhythm. *Computer Studies in the Humanities and Verbal Behaviour 3, 90–99.*

Yule G.U. (1938). On ssentence-length as a statistical characteristic of style in prose: with application to two cases of disputed authorship. *Biometrika 30, 363–390.*

Yule G.U. (1944). *The Statistical Study of literary Vocabulary*. Cambridge: Cambridge University Press.

Zaborowski R. (2000). Platon w ujęciu Wincentego Lutosławskiego (1863–1954) i Adama Krokiewicza (1890–1977) [Wincenty Lutosławski's (1863–1954) and Adam Krokiewicz's (1890–1977) views on Plato]. In: Zaborowski R. (Ed.), *Filozofia i mistyka Wincentego Lutosławskiego: 47–84.* [The philosophy and mysticism of Wincenty Lutosławski]. Warszawa: Stakroos.

Zipf G.K. (1935). *The psycho-biology of language; an introduction to dynamic philology*. Boston: Houghton Mifflin Company.

Zipf G.K. (1949). *Human behavior and the principle of least effort; an introduction to human ecology*. Cambridge, Mass.: Addison-Wesley Press.

Adam Pawłowski, Wroclaw

VII. Jerzy Woronczak (1923–2003)

The Founder of Polish Quantitative Linguistics*

Jerzy Woronczak (1923–2003)

Jerzy Woronczak (1923-2003) is one of the most important Polish scholars in the field of quantitative linguistics, and, in fact, one of the founders of this discipline in Poland. There are three areas in which one must examine Jerzy Woronczak's work in the field of quantitative linguistics (cf. Pawłowski, Sambor 2004, Kamińska-Szmaj 2004): The first pertains to the scientific merits of his own studies, the second to the knowledge he imparted to his students (master's and doctor's degree candidates), and the third to the value of the popularisation of his achievements against the background of the rather traditionalistic currents predominant in Polish linguistics.

The subject of this article will be a general description of the first area. The Professor's knowledge contained in the works of his students, conveyed in the form of lectures, consultations, and private conversations, often several hours long, is rather a topic for a separate treatise. Here it should be sufficient to state that during his active scientific career, J. Woronczak (as a participant, patron, reviewer, or adviser) took part in practically all the scientific initiatives in Poland connected with mathematical applications in linguistic research. With regard to the third area (popularisation), we must admit that the studies on the application of mathematics to problems of history and literary theory, which he undertook in the 1960's,

* This is a deeply modified version of the Festschrift paper Pawłowski, Sambor (2004).

testified not only to his vast knowledge, which defied precise labelling, but also to his civil courage. One should not forget that for the traditional representatives of the humanities of the time, combining the poetics and aesthetics of literature with mathematics was a sign of unaccountable intellectual bravado, with a dash of humbug. Such studies were treated as peculiarly eccentric, based on the appropriation of a methodology foreign to the discipline, leading in the best case to reductionism, and therefore a simplification of complex linguistic material. Referring to the issue of numerical methods in prosody and versification (one of Woronczak's favourite topics), M.R. Mayenowa remarked that "Statistical methods of versification analysis often arouse hostility even where there are no such basic reservations as to the principle; what is more, they arouse objections also there where, in their simplest form, they are always applied. [...] It seems that the basis of the protest is sheer psychological, and one should not wonder. The professional who has devoted many years to mastering the traditional language of his discipline can only with difficulty and great humility accept a situation in which someone discusses his discipline in another tongue." (Mayenowa 1965: 170) One of the reasons for the success of statistical linguistics in Poland was the fact that Woronczak was able to speak "in another tongue" about the traditional issues of philology and linguistics and set an example for the younger generation of scholars.

Below we shall discuss the most important quantitative works of Jerzy Woronczak according to thematic groups. It is worth adding that in the 1970's, he gradually turned to his original interests, namely early mediaeval history, the antiquity, biblical studies and, most of all, Hebrew studies and the history of Polish Jews. From this time on, quantitative themes appeared mostly in the subjects of the theses and dissertations of his students, predominantly involving, by the way, the Bible and/or ancient texts.

WORONCZAK'S STUDIES ON STYLOMETRY

It is worth reminding that the tradition of stylometric research in Poland goes back to the end of the XIXth century. One of the fathers of stylometry was Polish Hellenist W. Lutosławski, who coined the term "stylometry" and defined its general rules (Lutosławski 1897 [1983]). In the 1950's, W. Kuraszkiewicz, an expert in Slavonic studies, suggested using numerical measures of lexical richness (Kuraszkiewicz, Łukaszewicz 1951). His coefficient, like the one of Guiraud to which it is similar, has no practical significance today, but it played an important role in promoting mathematical methods among Polish linguists.

Woronczak turned to the problems of stylometry at the beginning of the 1960's. In contrast to his predecessors, though, he applied significantly more refined and effective mathematical tools. It should be emphasized that his work had both a theoretical (the analytical derivation of estimators) and a practical (applications in the solving of real problems in linguistics) aspect. It was his goal to discover unbiased estimators of indices of lexical richness which were sensitive to lexical variety, but independent of the length of the text fragment under investigation (Woronczak 1965b). Starting from the so-called G o o d ' s m e a s u r e s (Good 1953), which express the probability of randomly selecting *m* elements belonging to one and the same class in *m* independent samplings from a general population,

(1) $$c_m = \sum_i p_i^m$$

Woronczak derived equations for the estimators c_m for $m = 2$ and $m = 3$:

(2) $$\bar{c}_2 = \frac{\sum f_i^2 - N}{N^2 - N}$$

(3) $$\bar{c}_3 = \frac{\sum f_i^3 - 3 \cdot \sum f_i^2 + 2N}{N(N-1)(N-2)}$$

where f_i is the frequency of the *i*-th word-form, and *N* the length of the sample[7].

Equation (4) is a generalisation of equations (2) and (3). Its author, though, did not recommend calculating its value for $m > 3$ (Woronczak 1976).

(4) $$\bar{c}_m = \sum_i \frac{f_i(f_i - 1)...(f_i - m + 1)}{N(N-1)...(N - m + 1)}$$

Applying the parameters *B* and ρ from Mandelbrot's equation, Woronczak then derived equations for the expected length of a given text's vocabulary and the expected size of the class of words of a given frequency (*ibid.*).

The estimators (2) and (3) were initially verified by their author (Woronczak 1965b), but he admitted in another article that the set-up of the test was not entirely satisfactory (Woronczak 1976: 167). That is also why they were submitted to further verification on an extensive corpus (Pawłowski 1994). The dynamics of change in the values of several indices of lexical richness were compared in a corpus of French literary texts (prose by Romain Gary), the length of which was gradually increased from 20'000 to 600'000 words. The measure for evaluating an index was the dispersion[8] of its value with increasing length of text. One already observes a significant improvement in index stability of log TTR, though the Dugast and Yule indices, as well as those of Woronczak, proved to be the most stable (Tab.1).

Table 1
Indices of lexical richness and their dispersions

Index	Coefficient of dispersion	Index	Coefficient of dispersion
TTR	0.610	c_3 (Good)	0.022
Kuraszkiewicz	0.202	c_2 (Good)	0.006
Guiraud	0.202	*K* (Yule)	0.006
V_1/V_2 (Guiraud)	0.179	*UBER* (Dugast)	0.001
log TTR (Herdan)	0.038		

Woronczak (1976) also showed that there is a connection between the values of the estimators c_2 and c_3 and the lexical c o h e s i o n of a text. Analyzing the dynamics of the mean variations of these estimators with ever-increasing sample length of a continuous text (for $N = 2, 4, 8...$), he noticed that the estimators first increased in value with increasing *N*, but then stabilized, despite the geometric progression of *N*. The value of *N* at which the relative stabilization of the indices c_2 and c_3 takes place (or their maximum values), marks precisely the limit of the lexical cohesion of the text, indicating at the same time the average length of the fragments, which are closed to a certain degree with respect to vocabulary and theme.

[7] Equation (2) was also derived by G. Herdan. Both scholars demonstrated the similarity of c_2 and Yule's *K*-characteristic.
[8] Standard deviation divided by the mean.

The test which Woronczak conducted on texts by St. Fulgentius and St. Augustine confirmed this hypothesis. The Augustinian text, which was addressed to an uneducated social class and therefore written in a simple manner, produced an *N* limit of *ca*. 45 words, while that for the more difficult and literary Fulgentius text was *ca*. 128 words.

CORPUS RESEARCH

The beginnings of research using corpora in Poland must be associated with the preparation of the Frequency Dictionary of Modern Polish (*Słownik Frekwencyjny Polszczyzny Współczesnej*, hereinafter SFPW) in the 1960's and 70's, modelled on the Juilland dictionaries (Kurcz et al. 1990). Woronczak was, next to J. Sambor, one of the chief initiators and authors of this undertaking of several years' duration (Lewicki, Sambor 1969). The SFPW was compiled on the basis of a sampling of 500'000 words encompassing five functional styles (genres): scientific texts, small press items, commentary on current affairs, literary prose, and drama. The fundamental indicators describing the frequency distribution of a lexeme in the stylistic categories were the Juilland measures *F*, *D*, and *U*. The empirical data contained in the SFPW became the basis for several analyses of Polish (see: Kamińska-Szmaj 1988, 1989, 1990; Sambor 1971; Hammerl 1989; Pawłowski 1999a, 1999b). It is worth mentioning that the current SFPW corpus has been converted to digital form and is available on the Internet (http://www.mimuw.edu.pl/polszczyzna/, cf. Ogrodniczuk 2003).

MULTIDIMENSIONAL ANALYSIS

Although Woronczak did not extensively apply this type of methodology, he was fully aware of the possibilities which multidimensional analysis had to offer in the taxonomy of textual objects. He knew the works of J. Czekanowski[9], whom he met in Wrocław on several occasions during seminars on applied mathematics organized by H. Steinhaus. In 1962 he published a study, where multidimensional scaling in a rudimentary form was applied to establish the origin and filiations of *Bogurodzica* (the oldest Polish literary text). Using 56 text features, he classified all the *Bogurodzica*'s remaining versions (coming from the period of XV–XVII century). This helped him conclusively settle the perennial dispute over the originality and chronology of *Bogurodzica*'s stanzas (Woronczak 1962 [1993]). Woronczak also mentioned his discussions with A. Kolmogorow[10] on the topic of spatial representations of "linguistic objects" and encouraged the author of these lines to conduct a taxonomy of Polish poetic texts.

THE STATISTICAL LAWS OF LANGUAGE IN WORONCZAK'S RESEARCH

The American linguist J. K. Zipf is recognized as the initiator of studies on the statistical laws of language. Other scholars, such as J.N. Baudouin de Courtenay (Baudouin de Courtenay 1927 [1990]: 549), also anticipated their appearance. The dependencies Zipf discovered between the frequencies of expressions, their lengths, number of meanings, and rank are generally known as "Zipf's laws". They stimulated the search for other linguistic laws within the

[9] In the 40s Jan Czekanowski introduced multivariate methods in anthropology and linguistics (for further information see: Adam Pawłowski, *Jan Czekanowski (1882–1965) – a pioneer of multidimensional taxonomy*. To be published in one of the forthcoming issues of *Glottometrics*).

[10] Most likely during a conference on the versification of Slavic languages organised in Warsaw by the Institute of Literary Research of the Polish Academy of Sciences in August of 1964 (see Mayenowa 1965).

framework of a broad paradigm of systems theory or cognitive science (Hammerl, Sambor 1993).

Woronczak studied Zipf's fundamental law, which describes the relationship between the rank of a word in a list and its frequency (Woronczak 1967). Starting with the equations of Estoup, Joos, and Mandelbrot, he developed an analytical description of the quantitative structure of the vocabulary of a complete text, treating it as a sampling from the general population of the language, and derived equations for the expected size of the vocabulary of a text with a length of N word-forms and for the expected number of words with an assigned frequency (*ibid*. 2259). He also considered generalising the equations he obtained for an infinite text of length $N \to \infty$ and rank $r \to \infty$. It must be added that Woronczak's above-mentioned generalisations had never been the subject of empirical verification and were of deductive-theoretical nature.

STUDIES ON VERSIFICATION AND POETICS

As an expert on the literature, versification, and musical notation of the Middle Ages, Woronczak devoted many of his studies to research into texts in Old Polish (1958 [1993], 1960, 1965a, 1993) and in Old Czech (1963 [1993]). He approached this topic in his typical manner, i.e. both from a philological and a quantitative perspective. The statistical models he elaborated and the tests he employed were never goals in themselves, nor, consequently, were the linguistic materials he used merely a pretext for the abstract solutions often encountered in formalistic approaches. It is certainly this balance between philological-linguistic content and mathematical formalism which resulted in this aspect of Woronczak's work becoming an especially valuable element of his scientific legacy. We will discuss here just some of his most representative works devoted to versification.

In 1960 his analysis of the distributions of the verse lengths of asyllabic Slavonic poetry of the $15^{th} - 16^{th}$ centuries appeared (Woronczak 1960). For the sake of comparison he described the numerical distribution of the length of sentences in Polish prose; this proved to be a gamma distribution with a large right-sided asymmetry. He then found that the variance in length of asyllabic verses was less than that of sentences in prose and decreased with time, which was an indication of the gradual formation of the Polish syllabic system. This gradual formation process of Polish syllabic verse was the leitmotif of Woronczak's studies of the Biernat from Lublin's writings (1958 [1993]).

While in controversy with the theses of Czech mediaevalists over the structure of the Old Czech versification in the *Dalimil Chronicles* (org. *Staročeská Kronika tak Řečeneho Dalimila*), Woronczak submitted the hypothesis that if one proceeded from the opening chapters of the chronicles towards its end, one would be able to observe the process of its development into prose, in that the structures of its versification and rhythm would gradually become less rigid (Woronczak 1963 [1993]). He maintained that the beginning fragments of the chronicles, which speak of the pre-Christian era, temporally remote and unknown to the annalist, would be versified in a more orderly manner. He explained this phenomenon through two causes. First, the beginning chapters may have contained quotations from a surviving oral literary tradition introduced into the text. One must remember that the majority of medieval texts were originally transmitted orally, these being easy to remember by their regular, formulaic structure, which served not only an aesthetic, but also, and perhaps foremost, a mnemonic function. Secondly, one could imagine that the author, writing of events remote in time and not familiar to the contemporary audience, might, as the need arose, alter the content to fit the linguistic form rather than the form to fit the content, making it in this way more splendid. The opposite situation would prevail in the last fragments, presenting contemporary

events which have not yet been consolidated into an oral tradition and which demanded adherence to facts, the rules of correct versification being of secondary importance.

Woronczak conducted the verification of this hypothesis employing the *test of runs*, which is a technique which allows defining the degree of randomness of a numerical series. The data he used were the lengths of subsequent verses. The tests confirmed the agreement of the hypothesis with the structure of the *Dalimil Chronicles*.

CONCLUSIONS

If one were to consider the number of his publications as the only criterion in evaluating Jerzy Woronczak's achievements in the field of quantitative linguistics, the result would be modest. His determination to promote his achievements in international journals was also, by today's standards, too slight and not proportional to their scientific value. But do these strictly utilitarian measures embrace the totality of scientific output? Time has shown that the main distinguishing feature of Woronczak's work is its depth, quality and originality. For in the overwhelming number of cases, the Professor was able to find the optimal point of balance at which philological and linguistic issues do not disappear in a thicket of mathematical formalism, but preserve their cognitive value and freshness even for the demanding specialists in the given discipline. And that is perhaps the last lesson which he taught his students.

SHORT CURRICULUM VITAE OF JERZY WORONCZAK[11]

9 Nov. 1923	Jerzy Woronczak is born in Radomsko;
1936–1939	secondary school in Radomsko, studies of Hebrew with Jakub Fajner (until 1940);
1945	matura (secondary-school certificate) at the gymnasium of Radomsko, arrival to Wrocław;
1945–1947	studies at the Faculty of Human Sciences, University of Wrocław (history);
1948–1952	studies at the Faculty of Human Sciences, University of Wrocław (Polish philology);
1945–1953	employment at the Wrocław University Library;
1953–2003	employment at the Institute of Literary Research, Polish Academy of Sciences, Wrocław section (Instytut Badań Literackich Polskiej Akademii Nauk, henceforth IBL PAN);
1952	master's degree at the University of Wrocław in Polish philology;
1959	doctor's degree at the IBL PAN, dissertation "Studia nad wierszem polskiego średniowiecza" (Studies of Polish medieval verse);
1961–2003	member of the editorial committee of a dictionary of 16th century Polish;
1962	Woronczak's study is published which conclusively settles the perennial dispute over the originality and chronology

[11] Cf. also http://www.staropolska.gimnazjum.com.pl/sredniowiecze/opracowania/J_Woronczak.html

	of *Bogurodzica*'s stanzas [*Bogurodzica* is the oldest Polish literary text], as well as indicates the filiations of its numerous versions.
1965	member of the editorial board of "Biblioteka Pisarzów Polskich" (Library of Early Polish Writers);
1965–1970	beginning of works on the *Frequency dictionary of modern Polish* (J. Woronczak was leading the project);
1968	chief of the editorial board of "Biblioteka Pisarzów Polskich" (Library of Early Polish Writers);
1966	habilitation at the IBL PAN, dissertation "Rękopis nr 149 Biblioteki Kapitulnej w Gnieźnie (*Missale Plenarum* z przełomu XI I XII w.). Opracowanie filologiczne" (Manuscript no. 149 from the Capitular Library in Gniezno, *Missale Plenarum*, XI–XII c. A Philological Analysis.);
1967–1978	head of the Section of Medieval Texts at the IBL PAN;
1975–2001	lectures at the Institute of Polish Philology, University of Wrocław (European and Polish medieval literature and culture, linguistics, Judaism);
1978	beginning of Woronczak's *opus vitae* – the edition of the *Complete Works* (Dzieła Wszystkie) of Jan Kochanowski;
1984	extraordinary professor at the IBL PAN;
1991	ordinary professor at the IBL PAN;
1993–1996	founder and chief of the Research Centre for the Culture and Languages of the Polish Jews at the Institute of Polish Philology, University of Wrocław;
1993	officially retired, but scientifically active until the end of his days;
winter 2003	Woronczak's last lecture at a meeting of the Wrocław Philological Society;
6 Mar. 2003	death of Jerzy Woronczak in Wrocław.

REFERENCES

Baudouin de Courtenay J. (1927). Ilościowość w myśleniu językowym [Quantity as a dimension of thought about language]. In: *Symbolae Grammaticae in honorem Ioannis (Jan) Rozwadowski* v. 1. (Festschrift) Cracoviae: Gebethner & Wolff, 3–18. Reprint: Baudouin de Courtenay J. (1990), *Dzieła wybrane t. IV, 546–563* [Selected Writings, v. 4]. Warszawa: PWN.

Good I.J. (1953). On the Population frequencies of Species and estimation of population parameters. *Biometrika 40, 237–264.*

Hammerl R. (1989), Metoda wyodrębniania słownika minimum (na materiale słownika frekwencyjnego polszczyzny współczesnej) [A method of establishing the minimum dictionary of Polish (on the data from the frequency dictionary of contemporary Polish)]. *Poradnik Językowy 1989*, *614-628.*

Hammerl R., Sambor J. (1993). *O statystycznych prawach językowych* [On statistical laws of language]. Warszawa: Polskie Towarzystwo Semiotyczne.

Kamińska-Szmaj I. (1988). Części mowy w słowniku i tekście pięciu stylów funkcjonalnych polszczyzny pisanej (na materiale słownika frekwencyjnego) [[Parts of speech in the lexicon and text of five functional styles (genres) of written Polish (on the material of the frequency dictionary)]. *Biuletyn Polskiego Towarzystwa Językoznawczego 41, 127-136.*

Kamińska-Szmaj I. (1989). Charakterystyka statystyczno-stylistyczna części mowy [Stylo-statistical characteristics of parts of speech]. *Polonica 14, 87-120.*

Kamińska-Szmaj I. (1990). *Różnice leksykalne między stylami funkcjonalnymi polszczyzny pisanej. Analiza statystyczna na materiale słownika frekwencyjnego* [Lexical differences between the styles (genres) of written Polish. Statistical analysis based on the frequency dictionary of Polish]. Wrocław: Wydawnictwo Uniwersytetu Wrocławskiego.

Kamińska-Szmaj I. (2004) (Ed.). *Od starożytności do współczesności. Księga poświęcona pamięci profesora Jerzego Woronczaka* [From antiquity to contemporary times. Festschrift for Jerzy Woronczak]. Wrocław: Wydawnictwo Uniwersyttetu Wrocławskiego.

Kuraszkiewicz W., Łukaszewicz J. (1951). Ilość różnych wyrazów w zależności od długości tekstu [The frequency of different words as a function of text length]. *Pamiętnik Literacki 42(1), 168–182.*

Kurcz I., Lewicki A., Sambor J., Szafran K., Woronczak J. (1990). *Słownik frekwencyjny polszczyzny współczesnej*, t. 1-2 [Frequency dictionary of contemporary Polish, v. 1-2]. Kraków: PAN, Instytut Języka Polskiego.

Lewicki A., Sambor J. (1969). Projekt słownika frekwencyjnego współczesnego języka polskiego [The project of the frequency dictionary of present-day Polish]. *Sprawozdania PAN 12/4, 90-103*.

Lutosławski W. (1897). *The origin and growth of Plato's logic*. London, New York, Bombay: Longmans, Green and Co. Reprint: *The origin and growth of Plato's logic*. Hildersheim: Georg Olms Verlag, 1983.

Mayenowa M.R. (1965). Granice matematyzacji (w opisie wiersza) [The limits of mathematization (in verse description)]. *Kultura i społeczeństwo 24, 170–173*.

Ogrodniczuk M. (2003). *Nowa edycja wzbogaconego korpusu słownika frekwencyjnego* [A new enhanced edition of the frequency dictionary corpus]. In: Gajda S. (Ed.) (2003), *Językoznawstwo w Polsce. Stan i perspektywy: 181–190*. [Linguistics in Poland. Its present state and perspectives.]. Opole: PAN, Komitet Językoznawstwa.

Pawłowski A. (1994). Ein Problem der klassischen Stilforschung: Die Stabilität einiger Indikatoren des Lexikonumfangs. *Zeitschrift für Empirische Textforschung 1, 67–74.*

Pawłowski A. (1999a). *Metodologiczne podstawy wykorzystania słowników frekwencyjnych w badaniu językowego obrazu świata* [Methodological foundations for the use of frequency dictionaries in investigating the linguistic image of the world]. In: A. Pajdzińska, P. Krzyżanowski, *Przeszłość w językowym obrazie świata: 81-99* [Past in the linguistic image of the world]. Lublin: wyd. UMCS.

Pawłowski A. (1999b). The Quantitative Approach in Cultural Anthropology: Application of Linguistic Corpora in the Analysis of Basic Color Terms. *Journal of Quantitative Linguistics 6/3, 222–234.*

Pawłowski A., Sambor J. (2004). *Jerzy Woronczak – twórca polskiej lingwistyki kwantytatywnej* [Jerzy Woronczak – the founder of Polish quantitative linguistics]. In: Kamińska-Szmaj I. (Ed.), *Od starożytności do współczesności. Księga poświęcona pamięci profesora Jerzego Woronczaka* [From antiquity to contemporary times. Festschrift for Jerzy Woronczak]. Wrocław: Wydawnictwo Uniwersyttetu Wrocławskiego.

Sambor J. (1971). Z zagadnień gramatyki w słowniku frekwencyjnym współczesnego języka polskiego [On the problems of grammar in the frequency dictionary of modern Polish]. *Biuletyn Polskiego Towarzystwa Językoznawczego 29, 117-129.*

Woronczak J. (1960). *Statistische Methoden in der Verslehre*. In: *Poetics – poetyka – poetika: 607–627*. Warszawa: PWN, IBL.

Woronczak J. (1962 [1993]). *Wstęp filologiczny do Bogurodzicy* [Philologica introduction to „Bogurodzica"] In: M.R. Mayenowa (1962), *Liryka średniowiczna* t.1. BPP, Seria A, 1. Worcław etc.: Ossolineum, 7–25. [Reprint in the volume: Woronczak J. (1993), *Studia o literaturze średniowiecza i renesansu* [Papers on the literature of Middle Ages and Renaissance]. Wrocław: Wydawnictwo Uniwersytetu Wrocławskiego, 76–94].

Woronczak J. (1965a). *Rytmika akcentowa sylabowca*. [Accentual rhythm of the syllable verse]. In: M.R. Mayenowa (red.), *Poetyka i matematyka: 72-78* [Poetics and mathematics]. Warszawa: PIW.

Woronczak J. (1965b). *Metody obliczania wskaźników bogactwa słownikowego* [Methods of calculating indices of the lexical richness of texts]. In: M.R. Mayenowa (red.), *Poetyka i matematyka: 145-165* [Poetics and mathematics]. Warszawa: PIW.

Woronczak J. (1967), *On an attempt to generalize Mandelbrot's distribution*. In: *To Honor Roman Jakobson vol. II, 2254-2268*. The Hague: Mouton.

Woronczak J. (1976). *O statystycznym określeniu spójności tekstu* [On the statistical definition of test coherence]. In: M.R. Mayenowa (red.), *Semantyka tekstu i języka*: *165-173* [Semantics of text and language]. Wrocław: Ossolineum.

Woronczak J. (1993). *Studia o literaturze średniowiecza i renesansu* [Papers on the literature of Middle Ages and Renaissance]. Wrocław: Wydawnictwo Uniwersytetu Wrocławskiego.

Woronczak J. (1963 [1993]). *Zasada budowy wiersza Kroniki Dalimila* [The principle of construction of Dalimil Chronicles' verse]. *Pamiętnik Literacki 2, 1963, 469–478*. Reprint in the volume: *Studia o literaturze średniowiecza i renesansu* [Papers on the literature of Middle Ages and Renaissance]. Wrocław: Wydawnictwo Uniwersytetu Wrocławskiego, 67–75.

Woronczak J. (1958 [1993]). *Z badań nad wierszem Biernata z Lublina* [The research of the Biernat from Lublin verse]. *Pamiętnik Literacki* 3, 1958, 97–118. Reprint in the volume: *Studia o literaturze średniowiecza i renesansu: 139–156*. [Papers on the literature of Middle Ages and Renaissance]. Wrocław: Wydawnictwo Uniwersytetu Wrocławskiego.

Adam Pawłowski, Wroclaw