Glottometrics 30
2015

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. 30 (2015), Lüdenscheid: RAM-Verlag, 2015. 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 30 (2015) **ISSN 2625-8226**

Contents

Glottometrics 30, 2015, 1-18

Phonosemantic features of English and German consonants

Hanna Gnatchuk[1]

Abstract. The problem of the connection between sounds and meanings has been a point of debate among linguists throughout centuries. In this project, we are intended to confirm and establish semantic features for English and German consonants in the human mind. In order to achieve the objective, we undertake a psycholinguistic experiment. Then we treat the data with the help of quantitative methods — Osgood's semantic differential and the chi-square test. As a result, we have confirmed and established the semantic features for English and German consonants. Moreover, the outcomes of the psycholinguistic experiment have shown that the meanings of the sounds bear a close resemblance with their acoustic features: voiced and sonorant phonemes were evaluated as "kind" and "smooth" while voiceless – as "rough" and "fast" (in English and German). The practical application of the results may be of great use in creating brand names for industrial goods with a special emphasis on the semantics of the selected sounds.

Key words: phonosemantics, sound (phonetic) symbolism, quantitative methods.

1. Introduction

In order to do a systematic analysis of semantic features for both English and German consonants, it is necessary at first to have a look at two important types of classifications of sound symbolism. In particular, J.J. O'Hala, L. Hinton and D. Nichols (1994) suggested classifying phonetic symbolism into four categories (according to the direct linkage between sounds and their meanings): *corporeal* (interjection, cry), *imitative* (onomatopoeic words), *synesthetic* (separate sounds) and *conventional* (combination of sounds). I. Taylor and M. Taylor (1965) distinguished *subjective* and *objective* sound symbolism. *Subjective sound symbolism* deals with the connection of certain sounds and their semantics in the human mind (consciousness). This linkage can be revealed in an experimental way. *Objective sound symbolism* investigates the connection of certain sounds and their meanings in the words of a particular language. Such researchers as Lvova N. (2005), Uznadze (1924), Levitskij (2008), Kushneryk (2004), Sapir (1929), Newman (1933), Zhuravlov (1974) dealt with subjective phonetic symbolism. In particular, Lvova (2005) investigated semantic functions of English initial consonants. Kushneryk (2004) dealt with the meanings of sounds in Germanic and Slavic languages whereas Levitskij (2008) was engaged with the research of both objective and subjective sound symbolism in Finno-Ugric languages. The Russian researcher Zhuravlov (1974) did experimental research in order to reveal the symbolic meanings of Russian sounds according to 25 scales of Osgood's semantic differential. Moreover, he calculated the obtained meanings according to his own formula in which he paid attention to the position of stressed and unstressed sounds. The focus of our research is on the investigation of *subjective synesthetic phonetic symbolism*, namely on a systematic analysis of semantic features for both English and German consonants (which belong to the West-Germanic language group) using Osgood's Semantic Differential and the chi-square test.

[1] agnatchuk@gmail.com

It is to be remarked that any classification of this kind is merely a play with concepts created in the course of evolution in the given domain of science. Nevertheless, one must begin somewhere and test all possibilities.

2. Application of Osgood's semantic differential

The purpose of the investigation is to determine the semantic features of 24 English and 24 German consonants in the human mind with the help of the method of *semantic differential*. In order to achieve the given aim, we have conducted a psycholinguistic experiment in which 30 English (USA, Great Britain, Australia and Republic of Ireland) and 30 German (Klagenfurt, Austria) native speakers participated.

The number of the informants. The given number of informants (30) is considered to be minimal in any psycholinguistic experiment. Moreover, it is worth bearing in mind that similar experiments have been conducted with different numbers of informants – beginning from 20 ending in 300. The number of 20-50 respondents is considered to be enough for receiving objective results. A substantial increase in the number of informants, e.g. 300, did not lead to the improvement of the results of the experiment. Taking into account this fact, we have decided to choose 30 informants for our experiment. The informants were students (20—30 years old) from different faculties. In this case, we took into consideration the fact shown by Edward Sapir (1929) and Stanley Newman (1933) that age and gender as sociolinguistic factors might not affect the results of the research.

The questionnaire. All the consonants were printed in the form of phonetic transcription on the sheets of paper. In such a way, the respondent received the questionnaire in the written form with the necessary instruction.

The instruction contained the following text: "This experiment is aimed at studying semantic (meaningful) features of English (German) consonants. On this sheet of paper you will see the sounds which you should evaluate. Your task is as follows: look at the consonant, pronounce it and try to determine what this sound may mean (i.e. the consonant [b] according to the scale of potency – is it strong or weak or neutral, etc)".

The procedure. In such a way, the task of the respondents was to determine the semantic features of consonants according to six scales of Osgood's semantic differential:

the scale of activity (*slow – fast*),
the scale of potency (*weak – strong*),
the scale of roughness (*rough – smooth*),
the scale of size (*small – big*),
the scale of evaluation (*pleasant – unpleasant*),
the scale of kindness (*cruel – kind*).

The answers were represented by three variants: neutral and two contrary qualities. The consonants were given to the native speakers in the written form in so far as the graphical transcription of the sound was supposed to help them to reproduce the consonants in the human mind more accurately and with fewer faults. Then the answers were counted and treated with the help of *semantic differential*. According to Charles Osgood, "by semantic differential we mean the successive allocation of a concept to a point in the multidimensional semantic space by selection from among a set of given scaled semantic alternatives" (Osgood, 1957:26). In general, Semantic Differential belongs to a psycholinguistic method aimed at detecting symbolic meanings of sounds in phonosemantics.

3. Methods and results

In order to evaluate the answers one can take various ways. (1) One can ascribe the answers to the three individual classes separately in each of the six property dimensions mentioned above and test them for uniformity, e.g. using the chi-square test. This method only shows that there is a kind of neutrality or a tendency to associate the sound with some property. Consider for example: the associations of 30 test persons in German with the sound [b] in the "weak-strong" dimension (cf. Table 1).

Table 1
Reactions of 30 German speakers to the sound [b] in the weak-strong dimension

Category	**1. weak**	**2. neutral**	**3. strong**
No. of speakers	21	3	6

Since we have 3 categories, the expected number in each of them is 30/3 = 10. Considering 10 the expected value we obtain the chi-square as

(1) $$X^2 = \sum_{j=1}^{3} \frac{(f_j - 10)^2}{10} = \frac{1}{10}[(21-10)^2 + (3-10)^2 + (6-10)^2] = 18.6$$

The result is distributed as a chi-square with 2 degrees of freedom. It simply says that there is no equidistribution, hence one must seek the class which strongly deviates. Though in this case, an intuitive evaluation is possible, we are interested rather in the strength of the deviation. (2) To this end we consider the deviation in individual classes from the expectation and compute its probability. The expected proportion in each class is 10/30 = 0.3333, hence we compute the probability that the class acquires the given or still more extreme value, i.e. we compute the sum of binomial probabilities defined as

(2) $$P(X \geq f_x) = \sum_{j=f_x}^{n} \binom{n}{j} 0.3333^j 0.6667^{n-j}$$

where, in our case, n = 30. Since the greatest contribution to the chi-square for [b] in German is given by the "weak" category (f_x = 21), we compute

$$P(X \geq 21) = \sum_{j=21}^{30} \binom{30}{j} 0.3333^j 0.6667^{30-j} = 0.000044.$$

Since this probability is much smaller than, say 0.025, we may consider [b] as a sound associated with weakness. Computing the probability for the "neutral" class we obtain $P(X \leq 3) = 0.0033$, indicating that it deviates from neutrality: here one can say that [b] displays significant association in some direction. For the "strong" class we obtain $P(X \leq 6) = 0.08$, i.e. no tendency. Performing this test for all the consonants and all dimensions we obtain the results presented in Table 2 for German consonants and Table 3 for English consonants. Since we have to do with fixed parameters (n = 30, p = 0.3333) it can easily be shown that if the number of speakers in a category is smaller than 4 or equal to 4, the sum of probabilities (from x = 0,…,4) is 0.0122, i.e. the given class is significantly deviating. If the number of speakers is greater or equal to 16, the class is significantly preferred because the sum of probabilities from 16 to 30 is 0.0188. These results are illustrated in Tables 2 and 3, namely, the class that is significantly preferred obtains a "+", the class that is significantly avoided obtains a "–", and a class where no significant deviation can be observed remains empty.

Table 2
Significant associations of extreme classes for German consonants

	weak	neutral	strong	unpleasant	neutral	pleasant	slow	neutral	fast
[b]	+		–		–	+	+	–	–
[p]			+	–	+	–			+
[t]			+						+
[d]	+					–	+		
[k]		–	+	–	+			–	+
[g]	+		–	–	–	+	+	–	–
[m]	–	+			–	+	+	–	
[n]	+				–	+	+	–	
[ŋ]	+		–			+	+	–	
[f]		–	+		+	–			+
[v]	–		+		+	–	–		+
[s]			+	+		+		–	+
[z]		–	+	+		–		–	+
[ʃ]		–	+	–	+	+	+	–	–
[ç]	+		–	+	–		–	–	+
[x]	–		+	+			+		–
[h]					+	–	+		–
[j]	+		–						
[l]	+		–		–	+	+	–	
[ʧ]	–		+		–		+		–
[ʤ]	–		+	–	+		+		–
[pf]		–	+	+	–		+		–
[ts]	–		+	+				–	+
[r]	–		+	+	–		+	–	

Table 2 (cont.)
Significant associations of extreme classes for German consonants

	rough	neutral	smooth	cruel	neutral	kind	small	neutral	big
[b]		–	+		+	+	–	+	
[p]		+	–	–	+	+		+	–
[t]	–	+			+		–	–	+
[d]		–	+		–	+	+		+
[k]	–	+		–	+	–	–	+	–
[g]	–		+		+	–	+	+	+
[m]		–	+		–	+		+	–
[n]		+	+		+	–	–	+	
[ŋ]		–	+		+	–	–	+	

[f]	+	–	+	–	+		–	+	+
[v]		–	+	–	+			+	–
[s]		+	–	–	+	+	–		+
[z]	–	+		+	–			+	–
[ʃ]	–	+			+	–		–	+
[ç]					+	–	+	+	
[x]			+	+	–				
[h]	–		+	–	+	+		+	–
[j]		+			+	–	–	+	+
[l]	–		+		–	+		+	
[ʧ]	+	–			+		–	+	
[ʤ]		–	+		+	–		+	–
[pf]	+	–			+	–		+	
[ts]	+	–		–	+		–	+	
[r]	+	–		+	–			+	–

Table 3
Significant associations of extreme classes for English consonants

	weak	neutral	strong	unpleasant	neutral	pleasant	slow	neutral	fast
[b]	–		+		+	+	+	–	–
[d]	+		–		–	+		+	–
[f]	+		–	+		–	+	–	
[g]	–		+	–	+			–	+
[h]	+		–	+	–	+	+	–	
[j]	+		–	+	+	+	+	–	+
[k]		–	+	–	+	–		–	+
[m]		–	+	–	–	+	+		
[n]		–	+		–	+	+	–	–
[ŋ]	+		–	–		+	+	–	
[l]	–	+	–		–	+	+		
[p]	–		+		–	+	–		+
[r]	+	–	+	+	–	+	+	–	
[s]	–		+	+	–		–	+	
[z]		–	+	–	+	+	+		–
[t]		–	+	–	–		–	+	–
[ʧ]	–		+	–	–	+	+		–
[Ө]	–	–	+	+	–			+	–
[ð]	–	+	–		+	–	–	+	
[v]	+		–	–	+	+	+		–
[w]	+	+	–	–	+	–	+		–
[z]	–	–	–	–		+	+		–
[ʒ]	+		–	–	–	+	–		+
[ʤ]		–	+	–	–	–	+	–	–

Table 3 (cont.)
Significant associations of extreme classes for English consonants

	rough	neutral	smooth	cruel	neutral	kind	small	neutral	big
[b]	–	+	+		–	+		+	+
[d]		–	+		+				
[f]	–	–	–	+		–	+		–
[g]	–		+		+			+	
[h]	+		–	+		–	+		+
[j]		+			+		+		
[k]	+	–		+	+			–	+
[m]		–	+		–	+	–		+
[n]	–	–	+	–		+	–	–	+
[ŋ]		–	+		+		–	+	
[l]		–	+		–	+	–	+	+
[p]		–	+		+	+	–		+
[r]	+	–		+	–		–	+	+
[s]	–		+	+	–	–	–	–	+
[z]	+	+	–		+		–	+	+
[t]	+	–		+	+	–	+		–
[ʧ]	+	–		–	+	–		+	–
[Ɵ]		+	–		+	–	+	+	–
[ð]		+	–				–	+	
[v]	–		+	–	+		+	+	–
[w]		–	+		+	+		+	
[z]	+		–	+	+	–	+	–	+
[ʒ]		–	+	+	+	–	+	–	+
[ʤ]			+		–	+	–	+	+

In our research the semantic differential has been simplified. We used only 3 classes but, as a matter of fact, one can use any number of them. If one would use, say, 10 degrees and more informants, one would obtain curves having a special character. Theoretical insight useful for setting up linguistically or psychologically substantiated differential equations could be obtained only applying such a procedure. If one has merely 3 classes, one could use the trinomial distribution but the computation of cumulative probabilities would be very laborious.

The procedure for determining the grades of semantic differential was as follows. The grades in Appendix A, B show that 21 informants evaluated sound [b] as weak whereas 6 respondents as strong, and the rest (3 native speakers) as neutral. The semantic features are arranged under the following letters : A – weak, B – strong, C – unpleasant, D – pleasant, E – slow, F – fast, G – rough, H – smooth, I – cruel, J – kind, K – small, L – big (Appendix A, B). These results (for sound [b]) are also given in Table 1. The calculation is done in the following way:

1×21 = 21 (1 stands for weak × 21 speakers)
2×3 = 6 (2 stands for neutral × 3 speakers)
3×6 = 18 (3 stands for strong × 6 speakers)
21+6+18 = 45
45 : 30 (30 the total number of native speakers) = 1.5

In such a way, Table 4 (for German consonants) and Table 5 (for English consonants) contain the grades of Osgood's semantic differential. These grades can be explained in the following way: the grade 2 of Semantic Differential means that the consonant is devoid of any semantic feature. The grades 1.5 and lesser denote that the consonant is "small", "cruel", "weak", "unpleasant", "rough/even", "slow", whereas 2.5 and higher express such features as "big", "kind", "strong", "fast", "pleasant", "smooth/even". For example, the marked grade **1.5** for the sound [b] indicates that the given consonant is weak according to the scale of potency.

Table 4
The grades of Osgood's semantic differential for German consonants

	weak- strong	pleasant - unpleasant	slow - fast	uneven - smooth	cruel-kind	small - big
[b]	**1.5**	**2.8**	**1.4**	**2.6**	**2.5**	1.8
[p]	**2.8**	2.1	**2.5**	2.2	1.9	1.9
[t]	**2.8**	1.9	**2.8**	1.7	2.0	2.0
[d]	**1.2**	**2.8**	**1.5**	2.2	**2.9**	1.9
[k]	**2.7**	1.8	2.3	**1.5**	2.0	2.2
[g]	**1.3**	2.3	**1.1**	2.3	2.0	1.9
[m]	2	**2.8**	**1.3**	**2.6**	**2.6**	2.2
[n]	1.8	**2.6**	**1.4**	**2.6**	2.3	1.7
[ŋ]	**1.2**	**2.7**	**1.2**	**2.6**	**2.5**	1.7
[f]	2.3	1.9	2.4	1.7	1.6	1.7
[v]	2.2	2.2	2.2	2.4	1.9	2.1
[s]	**2.7**	2.1	2.3	2.1	2.2	2.3
[z]	**2.8**	**1.5**	2.3	**1.4**	1.6	2.1
[ʃ]	**2.6**	2.1	1.8	1.8	2.1	**2.6**
[ç]	1.8	1.7	2.2	**1.8**	2.2	1.6
[x]	**2.5**	1.6	1.9	**1.3**	1.8	2.0
[h]	1.8	1.8	1.7	2.4	2.3	2.0
[j]	1.7	2.1	1.8	2.0	2.2	1.7
[l]	1.6	**2.6**	**1.3**	**2.5**	**2.5**	2.0
[ʧ]	**2.7**	1.8	1.8	**1.5**	2.0	2.2
[ʤ]	2.2	1.9	1.6	1.9	2.3	2.3
[pf]	**2.7**	1.7	2.0	1.6	1.7	2.0
[ts]	**2.5**	**1.5**	**2.7**	1.7	1.6	1.7
[r]	**2.7**	1.6	1.8	**1.1**	**1.4**	**2.6**

Table 5
The grades of Osgood's semantic differential for English consonants

	weak-strong	unpleasant - pleasant	slow-fast	rough - smooth	cruel - kind	small - big
[b]	2.3	**2.5**	1.9	**2.6**	**2.5**	**2.6**
[d]	1.9	**2.9**	2.4	2.2	2.2	2
[f]	1.9	1.9	**1.5**	1.7	1.8	1.8
[g]	2.3	1.7	**2.8**	2.2	2.0	2.0
[h]	**1.3**	1.9	**1.5**	1.8	1.6	1.7
[j]	1.8	2.0	1.6	1.9	2.0	1.9

[k]	2.9	2.0	2.9	**1.3**	1.6	**2.7**
[l]	2	**2.9**	**1**	**2.9**	**2.9**	1.9
[m]	**2.6**	**2.7**	**1**	**2.9**	**2.6**	2.3
[n]	2	1.7	**1**	**2.9**	**2.9**	1.9
[ŋ]	**1.3**	**2.6**	**1.4**	**2.6**	2.3	1.6
[p]	2.2	1.9	2.4	**2.7**	2.4	2.3
[r]	2	2	**1.5**	1.8	2.2	2.3
[s]	**2.8**	1.9	**1.5**	2	1.9	2.1
[ʃ]	**2.7**	1.9	2.2	2	2.4	2.1
[t]	**2.9**	1.6	2.3	**1.5**	1.7	2
[ʧ]	**2.9**	2.4	1.7	2.1	2.0	2.4
[Ɵ]	2.4	1.9	2.0	1.8	2.1	2.0
[ð]	1.7	2	1.6	2.3	2.0	**1.3**
[w]	1.6	2.1	1.6	**2.5**	2.2	2.1
[z]	2.3	1.6	**1.2**	1.7	1.6	2
[ʒ]	**1.3**	**2.7**	2.1	**2.7**	1.9	2.4
[ʤ]	1.7	2.4	1.6	1.9	1.9	1.9
[v]	1.7	2.4	**1.4**	2.2	1.6	**1.5**

4. Discussion

Judging from the results of Semantic Differential (cf. Table 4), it is possible to observe that such German consonant sounds as [p] (**2.8**), [t] (**2.8**), [k] (**2.7**), [z] (**2.8**), [s] (**2.7**), [ʃ] (**2.6**), [x] (**2.5**), [ʧ] (**2.7**), [pf] (**2.7**), [ts] (**2.5**), [r] (**2.7**) are evaluated as *strong*; consonants [b] (**1.4**), [d] (**1.2**), [g] (**1.3**), [ŋ] (**1.2**) – *weak* (voiced sounds); consonants [b] (**2.8**), [d] (**2.8**), [m] (**2.8**), [n] (**2.6**), [ŋ] (**2.7**) – *pleasant* (voiced and sonorants); consonants [z] (**1.5**), [ts] (**1.5**) – *unpleasant*; consonants [p] (**2.5**), [t] (**2.8**), [ts] (**2.7**) – *fast* (voiceless); consonants [b] (**1.4**), [d] (**1.5**), [g] (**1.1**), [m] (**1.3**), [n] (**1.4**), [ŋ] (**1.2**), [l] (**1.3**) – *slow* (voiced and sonorants); consonants [b] (**2.6**), [m] (**2.6**), [n] (**2.6**), [ŋ] (**2.6**), [l] (**2.5**) – *smooth*; consonants [k] (**1.5**), [z] (**1.4**), [x] (**1.3**), [ʧ] (**1.5**), [r] (**1.1**) – *rough*; consonants [b] (**2.5**), [m] (**2.6**), [ŋ] (**2.5**), [l] (**2.5**) – *kind* (sonorant and voiced); consonant [r] (**1.4**) – *cruel*; consonants [ʃ], [r] – *big*. Therefore, the results of Semantic Differential have shown that the respondents evaluate German voiced and sonorant sounds as “weak”, “kind”, “smooth”, “pleasant” and “slow” whereas voiceless – as “fast”.

Having studied the results of semantic differential for English consonants (cf. Table 5), it is possible to state that English consonants [k] (**2.9**), [m] (**2.6**), **[ʃ]** (**2.7**), [t] (**2.9**), **[ʧ]** (**2.9**) are *strong*; consonants [h] (**1.3**), [ŋ] (**1.3**), [ʒ] (**1.3**) are *weak* (voiced); consonants [b] (**2.5**), [d] (**2.9**), [l] (**2.9**), [m] (**2.7**), [ŋ] (**2.6**), [ʒ] (**2.7**) are *pleasant* (voiced and sonorants); no *unpleasant* consonant has been revealed; consonants [k] **(2.9)**, [g] (**2.8**) are *fast* (voiceless); consonants [f] (**1.5**), [h] (**1.5**), [s] (**1.5**), [v] (**1.4**), [r] (**1.5**), [z] (**1.2**), [l] (**1**), [m](**1**), [n] (**1**), [ŋ] (**1.4**) are *slow*; consonants [k] (**1.3**), [t] (**1.5**) are *rough*_(voiceless); consonants [b] (**2.6**), [l] (**2.9**), [m] (**2.9**), [n] (**2.9**), [p] (**2.7**), [w] (**2.5**) are *smooth* (sonorant and voiced, except [p]); consonants [b] (**2.5**), [l] (**2.9**), [m] (**2.6**), [n] (**2.9**) are *kind*_(sonorant and voiced sounds); no *cruel* sound was detected; consonants [b] (**2.6**), [k] (**2.7**) are *big*; consonants [ð] (**1.3**), [v] (**1.5**) are *small* (voiced consonants).

Thus, the results of Semantic Differential for English consonants have shown that the respondents evaluate English voiced and sonorant sounds as “pleasant”, “weak” (only voiced), “kind”, “smooth” and “small” (only voiced) while voiceless – as “fast” and “rough”. In such a way, both English and German native speakers turned out to appreciate voiced and voiceless consonants as “weak”, “kind”, “smooth” and “pleasant” whereas voiceless as “fast”.

5. Chi-square tests and conclusions

The chi-square test is a statistical method aimed at measuring the degree of the correspondence of the actual data with theoretically expected. With the help of this method it is possible to confirm or refute the hypothesis about the connection of a sound with its meaning. The reason for the usage of the chi-squared test is that the outcomes of the previous investigation need to be more accurate and systematized.

The aim of the research is to reveal which semantic features the consonant is able to express to the full extent. In such a way, the hypothesis about the existence of symbolic meanings for English and German consonants may or may not be confirmed with the help of the chi-square test. Moreover, we are intended to reveal a) the semantic features of German and English consonants; b) the sound which has the highest and the lowest symbolic potential; c) the most active scale of semantic differential; d) the most active pole of semantic features.

The procedure of the investigation consists of arranging the data (cf. Table 2; Table 3) into the alternative tables for each consonant and for each scale. It is relevant in this case to give the example of the English consonant [b]:

Table 6
The frequency distribution of the English consonant [b]
according to the scale of potency

	Weak	Strong	Total
[b]	21 (a)	6 (b)	27
Other consonants	193 (c)	388 (d)	581
Total	214	394	608 N

After making alternative tables, the calculation of the chi-square was done according to the formula

$$(3) \qquad X^2 = \frac{(ad - bc)^2 N}{(a+b)(a+c)(b+d)(c+d)}$$

a, b, c, d – the empirical values in the alternative table
N – the total amount of observations.

a) German consonants

The results of the chi-square test for German consonants are given in Table 7. In such a way, the semantic features are arranged under the following letters: A – weak, B – strong, C – unpleasant, D – pleasant, E – slow, F – fast, G – rough, H – smooth, I – cruel, J – kind, K – small, L – big.

Table 7
The values of the chi-square test for German consonants

	A	B	C	D	E	F	G	H	I	J	K	L
[b]	**18.6**			**12.9**	**5.1**			**14.5**		**9.8**	2.2	
[p]		**9.6**	0.018			**24.1**	0.615		3.0			**4.2**
[t]		**10.7**	0.67			**21.7**	**4.2**		1.9			0.04
[d]	**34.7**			**16.8**	**5.4**			**9.1**		**18.4**		0.04
[k]		**14.1**	**9.0**			**6.8**	**4.8**		**8.6**			**16.5**
[g]	**10.9**			0.54	**11.4**			3.2		0.1	0.2	
[m]	3.0			**20.2**	**13.8**			**14.1**		**15.9**		**22.2**
[n]	**4.6**			**12.3**	**10.2**			**14.0**		**8.6**	2.8	
[ŋ]	**4.7**			**11.3**	**20.8**		**16.5**		**6.6**	**4.9**		0.09
[f]		**24.7**		3.1		**7.1**	0.31		**2.1**			0.09
[v]	0.002			1.8		1.9		**3.9**	2.9			0.00
[s]		**5.1**	0.76			**5.4**		**14.3**				**11.1**
[z]		**14.1**	**19.1**			**12.4**	**11.2**		**6.2**			0.09
[ʃ]		**6.8**		**4.7**	**8.4**					**5.2**		**6.9**
[ç]	**7.0**		**7.5**		2.1		2.8			2.4	**9.0**	
[x]	**5.1**		**8.6**		0.088		**10.6**		**4.9**		0.2	
[h]	**9.6**			0.092	**4.5**			**16.0**		2.8		**7.5**
[j]	**8.8**		6.7			0.13	**4.7**			0.7	**5.4**	
[l]	**7.1**			**29.1**	**18.8**			**4.3**		**6.8**	0.2	
[ʧ]		**4.8**		0.5	**5.1**		**14.8**		**5.3**			3.6
[ʤ]		**5.2**	3.1		**13.9**			1.9		1.4		
[pf]		**12.9**	**7.5**		**4.5**		**25.8**		**6.3**			**15.0**
[ts]		**38.7**	**16.5**			**13.6**	**7.7**		**12.6**		2.4	
[r]		**8.8**	**18.2**		0.013		**17.3**		**22.1**			**16.5**

Table 7 includes the values of the chi-square for German consonants. If the value of the chi-square for the consonant is higher than 3.84, it means that there is a significant statistical linkage between the sound and its semantic feature. Judging from Table 7, it is possible to state that each German consonant is characterized by specific semantic features. In particular, we have received the following semantic features for German consonant [b]: weak (X^2 = 22.4), pleasant (X^2 = 12.9), slow (X^2 = 5.1), smooth (X^2 = 14.3) and kind (X^2 = 9.8). In this case, it would be relevant to arrange the semantic features for this consonant in decreasing order according to the value of the chi-square: i.e. [b] – weak (X^2 = 22.4), smooth (X^2 = 14.3), pleasant(X^2 = 12.9), kind (X^2 = 9.8), slow (X^2 = 5.1). The analogical list of semantic features is made for each German consonant:

[b] – weak, smooth, pleasant, kind, slow
[p] – fast, strong, big
[t] – fast, strong, rough
[d] – weak, kind, pleasant, smooth, slow
[k] – big, strong, unpleasant, cruel, fast
[g] – slow, weak
[m] – big, pleasant, kind, smooth, slow
[n] – smooth, pleasant, slow, kind, weak
[ŋ] – slow, smooth, pleasant, kind, small, weak
[f] – strong, fast

[v] – smooth
[s] – smooth, big, fast, strong
[z] – unpleasant, strong, fast, rough, cruel
[ʃ] – slow, big, strong, kind, pleasant
[ç] – small, unpleasant, weak
[x] – rough, unpleasant, weak, cruel
[h] – smooth, weak, big, slow
[j] – weak, unpleasant, small, rough
[l] – pleasant, slow, weak, kind, smooth
[ʧ]– rough, cruel, slow, strong
[pf] – rough, big, strong, unpleasant, cruel, slow
[ʤ] – slow, strong
[ts] – strong, unpleasant, fast, cruel, rough
[r] – cruel, unpleasant, rough, big, strong

The given analysis is of great use in order to find out the strongest and the weakest German consonant, the smallest and the biggest, etc. Table 7 shows that the strongest German consonant is [ts] (38.7), the weakest – [d] (34.7), the slowest – [ŋ] (20.8), the farthest – [p] (24.1), the most unpleasant – [z] (19.1), the most pleasant – [m] (20.2), the roughest – [pf] (25.8), the smoothest – [h] (16.0), the smallest – [ç] (9.0), the biggest – [m] (22.2), the kindest – [d] (18.4), the cruelest – [r] (22.1).

The next step is to determine a) the symbolic potential of German consonants and b) the symbolic activity of scales. These notions were coined and introduced by V. Levitskij . According to V. Levitskij, symbolic potential is understood as "the ability of the sound to symbolize a certain notion", whereas symbolic activity of the scales – as "the ability of the notions or a group of notions to be symbolized by a certain sound" (Levitskij, 1998 :39).

In order to find symbolic potential of German consonants, all the values of chi-square for each consonant (cf. Table 7) are added within all scales. The results are shown in Table 8.

Table 8
The total values of X^2 (German consonants)

	The total value		The total value
[b]	66.9	[z]	63.09
[p]	41.53	[ʃ]	32.00
[t]	39.21	[ç]	30.8
[d]	84.4	[x]	29.48
[k]	59.8	[h]	40.49
[g]	25.85	[j]	41.33
[m]	89.2	[l]	66.3
[n]	52.5	[ʧ]	34.1
[ŋ]	64.8	[ʤ]	25.5
[f]	37.4	[pf]	72.0
[v]	10.50	[ts]	47.6
[w]	36.2	[r]	82.91
[s]	36.74		

As a result, we have found that the German consonant [m] (89.2) has the highest symbolic potential while the sound [v] (10.50) – the lowest. In such a way, we have arranged the consonants in the following decreasing order (starting with the sound that has the highest symbolic potential and ending in the lowest one): [m] (89.2), [d] (84.4), [r] (82.91), [pf] (72.9), [b] (66.95), [l] (66.3), [ŋ] (64.8), [z] (63.09), [k] (59.8), [n] (52.5), [ts] (47.6), [p] (51.53), [j] (41.33), [h] (40.49), [t] (39.21), [f] (37.4), [s] (36.74), [w] (36.2), [ʧ] (34.1), [ʃ] (32.00), [ç] (30.8), [g] (25.85), [ʤ] (25.5), [v] (10.50). The highest semantic potential is characterized by the German consonants [m] (**89.2**), [d] (**84.4**), [r] (**82.91**) whereas the lowest semantic potential is characteristic of [g] (**25.85**), [ʤ] (**25.5**), [v] (**10.50**).

The next task of this investigation is to reveal the most semantically active scale. In order to do the given objective, all the values of the chi-square for all consonants in Table 7 are added within one scale. Finally, we have obtained the following results for the German consonants: the scale of potency – 285.6; the scale of activity – 230.2; the scale of roughness – 223.4; the scale of evaluation – 211.1; the scale of cruelty – 158.1; the scale of size – 131.24.

Table 9
The total values of x^2 for all consonants within one scale

The scale of potency	285.6	the scale of evaluation	211.1
The scale of activity	230.2	the scale of cruelty	158.1
The scale of roughness	223.4	the scale of size	131.24

These values mean that the highest symbolic activity is characteristic of the scale of strength while the scale of size possesses the lowest activity. Then we added the values of the chi-square for both positive (strong, pleasant, fast, kind, smooth and big) and negative (weak, unpleasant, slow, cruel, rough and small) features. The results are given in Table 10.

Table 10
The total values according to positive and negative poles

Weak	130.1	**Rough**	104.8
Strong	155.5	**Smooth**	118.6
Pleasant	98.2	**Cruel**	75.9
Unpleasant	112.8	**Kind**	82.2
Slow	135.0	**Small**	27.3
Fast	95.2	**Big**	103.7

The results for German consonants are as follows: the positive “strong” quality proved to be active within the scale of strength (“strong” = 155.5); a negative “slow” quality - within the scale of speed (“slow” = 135.0); a positive “smooth” quality within the scale of roughness (“smooth” = 118.6); a positive “pleasant” quality within the scale of evaluation (“pleasant” = 112.8); a positive “big” quality within the scale of size (“big” = 103.7); a positive “kind” quality within the scale of roughness (“kind” = 82.2).

In such a way, it is possible to make the following conclusions concerning the semantic features of German consonants:

- The presence of the phonosemantic connection for German consonants is statistically confirmed;
- The semantic features for each German consonant were determined and arranged with the help of semantic differential and the chi-square test;

- The statistical analysis of the German consonants has shown a direct linkage between the acoustic features of the consonants with its semantics. In particular, voiced sounds were evaluated as “kind”, “pleasant” and “smooth” while voiceless as “fast”
- The German consonant [m] turned out to possess the highest symbolic potential whereas [v] – the lowest;

b) English consonants

Similar analyses have been made for English consonants. In particular, Table 11 includes the values of chi-square for English consonants with the following explanations: A – weak, B – strong, C – unpleasant, D – pleasant, E – slow, F – fast, G – rough, H – smooth, I – cruel, J – kind, K – small, L – big. The value of more than 3.84 shows a significant connection between a consonant and its meaning.

Table 11
The values of chi-square for English consonants

	A	B	C	D	E	F	G	H	I	J	K	L
[b]		0.4		**9.9**		**8.8**		**12.2**		**4.1**		**11.8**
[d]	**8.3**			**18.9**		1.0		**10.8**		**10.6**	3.4	
[f]	**4.7**		**5.9**		**18.0**		**5.4**		**5.7**		**22.7**	
[g]		0.6	1.2			**36.6**		**8.4**	0.3		0.09	
[h]	**7.8**		1.0		**17.3**		**6.5**		**7.7**		**7.7**	
[j]	**6.7**		1.2			1.5	0.6			**6.0**	**8.8**	
[k]		**16.8**	0.7			**42.0**	**23.1**		**17.4**			**20.0**
[m]		**5.8**		**5.3**	**18.9**			**20.9**		**12.6**		**16.5**
[n]		2.1	0.2		**5.2**			**8.7**		**6.7**		**8.5**
[ŋ]	**3.8**			**14.7**	**7.7**			**15.9**		**4.4**		3.47
[l]	**9.0**			**18.9**	**37.0**			**20.1**		**27.6**		3.0
[p]	0.2			2.2		**36.7**		**16.7**		**9.5**		**14.6**
[r]	1.3		1.0		**14.2**		2.4		**4.9**			3.0
[s]		**6.7**	2.4		**10.2**			0.2	2.3			0.3
[ʃ]		**10.9**		1.5	**4.7**		3.2			**5.1**		2.9
[t]		**16.9**	0.02		1.2		**12.9**		**7.8**			**24.1**
[ʧ]		**6.8**		0.2	**5.8**		**7.0**			0.00		0.02
[v]	**10.6**			2.1	**17.9**			0.2		**19.2**	**13.9**	
[w]	**12.2**		0.6		**13.0**			**6.8**		3.5	0.14	
[z]		**10.2**		0.00	**37.3**		**15.4**		**11.5**		**5.0**	
[ʒ]	**20.7**		**12.3**		**6.7**			**5.2**	1.3			**5.6**
[ʤ]		2.4		**6.2**	**12.2**			3.4		3.6		**4.6**
[ð]	0.05		0.02		3.2			**8.6**	0.7		0.3	
[Ɵ]		1.8	3.2			3.0		3.3		1.8	**8.5**	

The results of the research. Table 11 indicates that each English consonant possesses its specific semantic features. For example, the following semantic features are obtained for the English consonant [b]: pleasant (X^2 = 9.9), fast (X^2 = 8.8), smooth (X^2 = 12.2), kind (X^2 = 4.1) and big (X^2 = 11.8). In this case, it would be relevant to arrange the semantic features for this consonant in decreasing order according to the value of the chi-square: i.e. [b] – smooth, big, unpleasant, fast, kind. The analogical list of semantic features is made for each English consonant:

[b] – smooth, big, unpleasant, fast, kind
[d] – pleasant, smooth, kind, weak
[f] – small, slow, unpleasant, cruel, rough, weak
[g] – fast, smooth
[h] – slow, weak, cruel, small, rough
[j] – small, weak, kind
[k] – fast, rough, big, cruel, strong
[m] – smooth, slow, big, kind, strong, pleasant
[n] – smooth, big, kind, slow, strong
[ŋ] – smooth, pleasant, slow, kind
[l] – slow, kind, smooth, pleasant, weak
[p] – fast, smooth, big, kind
[r] – slow, kind
[s] – slow, strong
[ʃ] – strong, kind, slow
[t] – big, unpleasant, strong, rough, cruel
[ʧ] – rough, strong, slow
[v] – kind, slow, small, weak
[w] – slow, weak, smooth
[z] – slow, weak, smooth
[ʒ] – weak, pleasant, slow, big, smooth
[ʤ] – strong, slow
[ð] – smooth
[Ɵ] – weak

Judging from Table 11, it is possible to state that the weakest English consonant is [ʒ] (X^2 = 20.7), the strongest – [t] (X^2 = 16.9), the most unpleasant – [ʒ] (X^2 = 12.3), the most pleasant – [l] (X^2 = 18.9), [d] (X^2 = 18.9), the slowest – [z] (X^2 = 37.3), the farthest – [k] (X^2 = 42.0), the roughest – [k] (X^2 = 23.1), the smoothest – [m] (X^2 = 20.9), the cruelest – [k] (X^2 = 17.4), the kindest [l] (X^2 = 27.6), the smallest – [f] (X^2 = 22.7), the biggest – [t] (X^2 = 11.8).

The next step is to determine a) the symbolic potential of English consonants and b) the symbolic activity of scales. To find symbolic potential, all the values of chi-square for each consonant (cf. Table 11) are added within all scales. The results are given in Table 12.

Table 12
The total values of the chi-square for English consonants

	the total value		**the total value**
[b]	47.2	[p]	79.9
[d]	53.0	[r]	26.8
[f]	62.4	[ʃ]	28.3
[g]	47.19	[t]	62.6
[h]	48.02	[ʧ]	19.8
[j]	24.8	[ð]	12.8
[k]	12.0	[Ɵ]	21
[m]	8.0	[v]	45.9
[n]	31.4	[w]	36.2
[ŋ]	50	[z]	79.4

[l]	115.6	[ʒ]	51.8
[ʤ]	52.4	[s]	22.1

Table 12 shows that the English consonant [l] (115.6) has the highest symbolic potential whereas [m] (8.0) is the lowest one. In such a way, the consonants have been arranged in decreasing order beginning with the consonant that has the highest symbolic potential and ending in the lowest one: [l] (115.6), [p] (79.9), [z] (79.4), [t] (62.6), [f] (62.4), [d] (53.0), [ʤ] (52.4), [ʒ] (51.8), [ŋ] (50), [h] (48.0), [b] (47.2), [g] (47.1), [v] (45.9), [w] (36.2), [n] (31.4), [ʃ] (28.3), [r] (26.8), [j] (24.8), [s] (22.1), [Ɵ] (21), [ʧ] (19.8), [ð] (12.8), [k] (12.0), [m] (8.0). The English consonants [l] (115.6), [p] (79.9), [z] (79.4) have the highest symbolic potential whereas [ð] (12.8), [k] (12.0), [m] (8.0) the lowest.

The next task of this investigation is to reveal the semantically most active scale. In order to do the given objective, all the values of chi-square for all consonants in Table 11 are added within one scale. The results are given in Table 13,

Table 13
The total values of X^2 for all consonants within one scale

The scale of activity	360.1	**The scale of potency**	186.7
The scale of roughness	217.0	**The scale of cruelty**	179.7
The scale of size	188.9	**The scale of evaluation**	109.6

In such a way, the highest symbolic activity is characteristic of the scale of activity and the scale of evaluation turned out to have the lowest one. Then we added the values of chi-square for both positive (strong, pleasant, fast, kind, smooth and big) and negative (weak, unpleasant, slow, cruel, rough and small) properties. The results are given in Table 14.

Table 14
The total numbers according to positive and negative poles

Weak	85.35	**Rough**	76.2
Strong	101.4	**Smooth**	141.4
Unpleasant	49.28	**Cruel**	59.6
Pleasant	79.90	**Kind**	145.8
Slow	193.5	**Small**	70.5
Fast	166.6	**Big**	118.8

The outcomes for English consonants are as follows: a “strong” positive pole (101.4) proved to be active within the scale of strength; within the scale of evaluation – a positive “pleasant” pole (79.90); within the scale of speed – a negative “slow” pole (199.5); in the scale of roughness - a positive “smooth” pole (141.4); within the scale of cruelty – a positive “kind” pole (145.8); within the scale of size – a positive “big” pole (118.3).

Thus, it is possible to make the following conclusions on the basis of the given investigation:

- The presence of the phonosemantic linkage for English consonants is confirmed statistically;
- The symbolic features for each English consonant were established and arranged with the help of semantic differential and the chi-square test;

- The statistical analysis of the English consonants has shown a direct linkage between the acoustic features of the consonants with their semantics. In particular, voiced sounds were evaluated as “kind”, “pleasant” and “smooth” while voiceless – as “strong”, “fast” and “rough”
- The English consonant [k] proved to have the highest symbolic potential while [ð] – the lowest.
- The highest symbolic activity is characteristic of the scale of speed (activity), the lowest is the scale of evaluation.

Further perspectives of the research

It will be relevant in further research on this topic to investigate the semantic features of consonants in the literary texts, namely, to observe the existence or the absence of the connection between the emotional mood of the texts and the usage of consonants in English and German literal texts.

References

Kushneryk, V. (2004). *Fonosemantyzm y hermanskuh i slovjanskuh movah [Phonosemantism in Germanic and Slavic languages]* – Chernivtsi: Ruta, 370 (in Ukrainian).

Levickij, V.V. (1998). *Zvykovoj simvolizm. Osnovnije itogi: Monografija.* [Sound symbolism. Basic results] Chernovtsy: Ruta (in Russian).

Levickij, V. V. (2008). *Zvukovoj simvolizm: Mify i real’nost* [Sound symbolism: myths and reality], Chernivtsi National University (in Russian).

Lvova, N. (2005). Semantic functions of English initial clusters. *Glottometrics* 9, 21-27.

Newman, S. (1933). Further experiments in phonetic symbolism. *American Journal of Psychology 45, 53 – 75.*

Ohala, John J., Hinton, L., Nichols J. (1994). *Sound Symbolism*. Cambridge: Cambridge University Press.

Osgood, Ch., Suci, G.J., Tannenbaum P. H. (1957). *The measurement of meaning.* Urbana: University of Illinois Press

Sapir, E. (1929). A study in phonetic symbolism. *Journal of Experimental Psychology 12(3), 225 – 239.*

Taylor, I.K., Taylor, M.M. (1965). Another look at phonetic symbolism. *Psychological Bulletin 64(6), 117-124.*

Uznadze, D. (1924). Ein experimenteller Beitrag zum Problem der psychologischen Grundlagen der Namengebung. *Psychologische Forschung 5, 25-43.*

Zhuravlov, A.P. (1974). *Foneticheskoje znachenije*. [Phonetic meaning]. Leningrad University Press (in Russian).

Appendix A

The frequencies of semantic features for German consonants (according to the psycholinguistic experiment)

	POTENCY		EVALUAT		SPEED		ROUGH		CRUELTY		SIZE	
	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**
/b/	21	6	0	23	21	4	4	21	0	15	8	4
/p/	2	25	6	8	2	25	4	8	11	8	6	10
/t/	2	27	10	10	2	23	10	4	4	2	10	11
/d/	25	4	0	27	19	6	2	17	0	27	10	11
/k/	0	25	10	2	4	17	11	4	8	8	4	10
/g/	19	11	8	13	23	4	10	15	6	11	10	10
/m/	8	6	0	27	21	2	2	23	0	23	4	11
/n/	11	8	0	17	17	2	0	17	0	13	10	5
/ŋ/	13	10	3	25	23	0	0	20	0	10	8	2
/f/	4	17	4	8	6	17	11	11	8	6	8	11
/v/	10	18	3	10	10	15	3	15	8	5	8	10
/s/	4	23	11	11	6	15	4	10	8	11	8	15
/z/	0	25	17	2	4	19	13	2	15	3	8	11
/ ʃ /	2	20	6	12	14	8	10	8	0	8	4	18
/ç/	15	10	13	5	8	13	10	10	8	10	13	3
/x/	8	18	18	18	13	10	3	23	13	8	10	10
/h/	10	10	6	10	15	10	8	17	3	13	6	11
/j/	15	8	10	10	13	13	6	5	3	8	13	5
/l/	15	10	0	23	21	0	6	19	2	17	8	8
/tʃ/	5	25	8	8	18	12	18	3	5	5	10	13
/ʤ/	10	15	10	6	18	10	8	12	4	12	8	10
/pf/	0	23	13	5	15	10	15	3	4	10	8	8
/ts/	5	23	20	5	3	23	15	5	13	3	10	3
/r/	3	27	18	3	12	9	18	2	19	3	4	10

Appendix B

The frequencies of semantic features for English consonants (according to the psycholinguistic experiment)

	POTENCY		**EVALUAT.**		**SPEED**		**ROUGH.**		**CRUELTY**		**SIZE**	
/b/	9	18	0	15	15	10	0	16	4	16	0	15
/d/	14	11	0	26	0	6	6	13	5	12	10	10
/f/	15	10	16	9	15	5	14	14	16	9	17	13
/g/	8	17	8	0	0	21	9	17	4	4	8	9
/h/	17	9	13	13	16	5	17	8	18	9	13	4

/j/	16	9	11	10	11	11	8	7	4	9	15	5
/k/	0	25	9	9	0	26	17	0	14	1	0	25
/m/	4	21	5	21	30	0	0	26	0	17	8	18
/n/	5	16	8	14	21	4	4	22	8	16	4	22
/ŋ/	14	10	2	27	20	2	0	20	2	12	10	2
/l/	9	9	0	26	30	0	0	25	0	26	4	13
/p/	13	17	4	13	9	21	0	21	0	13	8	17
/r/	13	13	13	13	16	5	12	8	13	7	4	13
/s/	5	25	12	9	9	5	11	17	12	9	8	13
/ ʃ /	3	21	5	13	13	9	12	7	3	7	5	14
/t/	0	25	13	0	8	8	17	4	12	4	18	7
/tʃ/	5	25	8	14	16	11	15	6	4	6	7	10
/Ɵ/	6	17	11	7	7	9	5	9	4	12	12	11
/ð/	3	4	6	8	9	6	4	9	10	10	9	5
/v/	17	7	4	13	17	4	9	14	13	0	16	3
/w/	13	3	8	8	14	4	3	17	0	5	4	4
/z/	5	13	9	14	26	4	21	5	13	4	12	12
/ʒ/	17	2	6	18	10	12	5	19	11	9	12	10
/ʤ/	4	14	2	15	17	7	8	14	7	13	3	14

Glottometrics 30, 2015, 19-44

A simplified lambda indicator in text analysis

Ioan-Iovitz Popescu
Gabriel Altmann

Abstract. The aim of the article is to show an alternative, easier computation of the Lambda indicator which displays the frequency structuring of the text. Here not all frequencies need to be taken into account; it is sufficient to consider the first and the last values and the *h*-point. The article brings a survey of many texts in several languages.

Keywords. Lambda indicator, rank-frequency, text similarity

In some previous publications (see esp. Popescu, Čech, Altmann 2011) the lambda indicator has been defined as a normalized arc length of the rank-frequency distribution of words or other entities in a text. Since arc length increases with text size, it was proposed to normalize it as

$$\Lambda = \frac{L(\log_{10} N)}{N} \qquad (1)$$

where N is the text size (given in the number of words or other respective entities), and L is the arc length defined as

$$L = \sum_{x=1}^{V-1}\left[(f_x - f_{x+1})^2 + 1]^{1/2}\right] \qquad (2)$$

where V is the highest rank (vocabulary) and f_x are the frequencies at ranks x.

Unfortunately, the variance of L given in this form is quite complex (cf. Popescu, Mačutek, Altmann 2010) and every comparison of texts, setting up classes, confidence intervals, etc. is associated with extensive computations.

In Popescu, Mačutek, Altmann (2009: 68) an indicator has been defined which took into account both L_{max} and the h-point in the form

$$p = \frac{L_{max} - L}{h-1} \qquad (3)$$

where $L_{max} = (V - 1) + f_1 - f(V)$. Now since $f(V)$ is usually 1, one can define

$$L_{max} = V - 1 + f_1 - 1 .$$

On the other hand, from (3) we have

$$L = L_{max} - p(h - 1).$$

or, since *p* converges to 1, we can finally get an approximate arc length as

$$L^* = V + f_1 - (h+1) \tag{4}$$

Hence we obtain for (1) an approximate lambda in the form

$$\Lambda^* = \frac{L^*(\log N)}{N} = \frac{(V + f_1 - h - 1)(\log N)}{N} \tag{5}$$

Considering *V* a constant and *h* a fixed point, we obtain the variance of the above indicator as simple as

$$Var(\Lambda^*) = \frac{Var(f_1)(\log N)^2}{N^2} = \frac{f_1(N - f_1)(\log N)^2}{N^3} \tag{6}$$

In order to exemplify the formulas we consider twenty short Slovak texts by S. Svoráková concerning art criticism and obtain the results presented in Table 1.

The significance of the difference between the approximate lambdas of two texts (lower case 1 and 2 in formula (7)) can be computed by means of the usual asymptotic normal test in form

$$u = \frac{|\Lambda_1^* - \Lambda_2^*|}{\sqrt{Var(\Lambda_1^*) + Va(\Lambda_2^*)}} \tag{7}$$

If we perform this operation for each pair of texts, we obtain the results presented in Table 2. Here u <= −1,96 and u => 1.96 are significant. Hence texts whose similarity expressed by *u* varies in (−1.96, 1.96) have some common frequency background. Needless to say, the frequencies can be directly compared using e.g. a chi-square test; but with short texts one meets problems because of many small frequencies (hapax legomena). If we link the texts having non-significant difference – as shown in Table 2 – we obtain a matrix in which the crosses represent the similarity. The matrix is presented in Table 3.

Now every matrix of this kind can be presented in form of a graph which displays the similarities visually. It can be seen in Figure 1.

The centrality of individual texts can be given simply as the number of all other texts having similar lambda, i.e. for which *u* ε (−1.96, 1.96). We obtain the sequence

T16	T15	T17	T4	T5	T6	T1	T7	T9	T13	T12	T14	T20	T2	T3	T11	T19	T18	T8	T10
12	11	10	9	9	9	8	8	8	8	7	7	7	6	5	5	4	3	2	2

and the graph of similarities visualized in Figure 1.

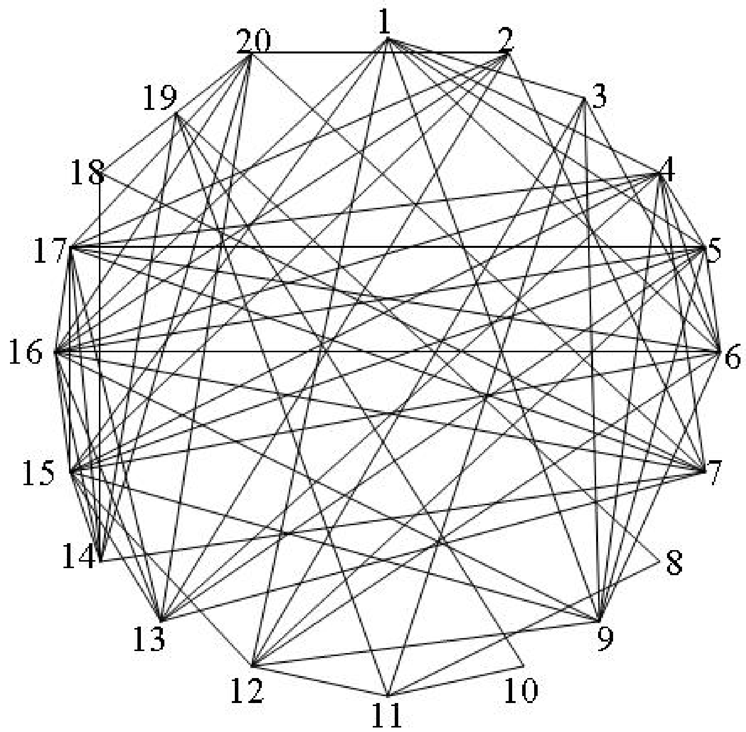

Figure 1. Graph of text similarities (Svoráková)

However, the centrality may also be computed as the sum of the absolute values of the criterion *u* in Table 2. We obtain the ordering as follows:

T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
46.95	64.06	63.23	45.43	47.01	44.66	64.99	93.88	56.19	107.5

T11	T12	T13	T14	T15	T16	T17	T18	T19	T20
77.58	48.83	57.79	55.88	38.51	34.17	62.12	102.14	121.41	101.6

Hence the texts according to *decreasing weighted centrality* are

16, 15, 6, 4, 1, 5, 12, 14, 9, 13, 17, 3, 2, 7, 11, 8, 20, 18, 10, 19.

The left side of the sequence shows the texts having more features characteristic of the style of Svoraková than those on the right hand side. Further research could help us to go a step deeper.

Table 1
The lambda indicator and its approximations in texts by S. Svoráková
Notice the close coincidence of Λ and Λ* (up to a few per-mille)

Text	**N**	**V**	**f_1**	**h**	**L**	**L***	**Λ**	**Λ***	**Var(Λ*)**
T1	750	501	38	7.0000	530.6654	531.0000	2.0343	2.0355	0.000530
T2	1084	672	39	11.0000	698.6078	699.0000	1.9560	1.9571	0.000295
T3	971	653	32	9.0000	673.9880	675.0000	2.0735	2.0766	0.000293
T4	783	486	61	8.0000	536.4371	538.0000	1.9825	1.9883	0.000768
T5	618	429	24	7.0000	443.6461	445.0000	2.0036	2.0097	0.000470
T6	765	501	44	6.5000	535.4395	537.5000	2.0183	2.0261	0.000589
T7	594	401	22	7.0000	414.3057	415.0000	1.9347	1.9379	0.000462
T8	1094	743	37	7.0000	769.6943	772.0000	2.1381	2.1445	0.000276
T9	807	555	24	7.6667	568.9179	570.3333	2.0493	2.0544	0.000302
T10	701	522	22	7.0000	534.7316	536.0000	2.1707	2.1759	0.000351
T11	448	353	11	4.8000	358.0160	358.2000	2.1188	2.1198	0.000376
T12	382	297	17	6.2000	307.6729	306.8000	2.0797	2.0738	0.000742
T13	748	496	25	8.0000	510.7673	512.0000	1.9624	1.9672	0.000357
t14	249	189	13	5.0000	195.4062	196.0000	1.8805	1.8862	0.001141
T15	402	299	18	4.6667	310.1175	311.3333	2.0090	2.0169	0.000722
T16	228	184	13	4.3333	190.8138	191.6667	1.9734	1.9822	0.001311
T17	397	289	18	5.5000	299.5405	300.5000	1.9608	1.9671	0.000736
T18	461	311	20	6.5000	321.9613	323.5000	1.8603	1.8692	0.000639
T19	2075	1285	82	11.0000	1350.8956	1355.0000	2.1595	2.1661	0.000201
T20	1218	730	41	10.5000	757.8380	759.5000	1.9199	1.9241	0.000254

Table 2
Differences between texts (Svoráková)

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	T16	T17	T18	T19
T2	2.81	0.00																	
T3	-1.39	-4.98	0.00																
T4	1.36	2.73	2.73	0.00															
T5	0.79	-2.02	2.35	-0.69	0.00														
T6	0.35	-2.33	1.75	-1.01	-0.40	0.00													
T7	3.30	0.84	5.24	1.56	2.59	2.85	0.00												
T8	-3.66	-7.72	-2.69	-4.73	-4.71	-3.95	-7.65	0.00											
T9	-0.49	-3.88	1.06	-1.93	-1.40	-0.87	-4.27	3.74	0.00										
T10	-4.59	-8.52	-3.80	-5.53	-5.62	-4.84	-8.42	-4.78	-4.78	0.00									
T11	-2.89	-6.48	-1.82	-4.02	-3.85	-3.19	-6.60	0.67	-2.80	1.84	0.00								
T12	-0.99	-3.59	0.15	-2.17	-1.72	-1.29	-4.00	2.15	-0.65	3.05	1.55	0.00							
T13	2.53	-0.20	4.53	0.79	1.76	2.08	-0.98	7.12	3.49	7.94	6.00	3.33	0.00						
T14	3.99	2.18	5.35	2.61	3.44	3.64	1.49	7.07	4.66	7.73	6.41	4.56	2.27	0.00					
T15	0.73	-1.73	2.07	-0.61	0.02	0.39	-2.28	4.12	1.25	4.96	3.42	1.59	-1.52	-0.10	0.00				
T16	1.42	-0.48	2.52	0.27	0.86	1.13	-1.00	4.13	1.87	4.82	3.60	2.12	-0.35	-0.07	0.79	0.00			
T17	2.14	-0.13	3.62	0.71	1.48	1.78	-0.78	5.65	2.80	6.42	4.89	2.90	0.03	-0.06	1.33	0.33	0.00		
T18	4.59	2.48	6.46	2.86	3.96	4.15	1.59	8.64	5.58	9.33	7.68	5.16	2.56	-0.01	3.56	2.14	2.15	0.00	
T19	-4.78	-9.43	-4.03	-5.72	-5.95	-5.04	-9.07	-1.15	-5.14	0.31	-1.76	-3.06	-8.66	-0.21	-5.12	-4.88	-6.69	-9.91	0.00
T20	4.15	1.53	6.69	3.40	3.40	3.62	0.48	9.55	5.53	10.25	8.11	4.80	1.65	-0.03	2.92	1.39	1.26	-1.33	11.51

Table 3
Lambda-similarities between texts (Svoráková)

T	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
2																			
3	X																		
4	X																		
5	X			X															
6	X		X	X	X														
7		X		X															
8																			
9	X		X	X	X	X													
10																			
11			X					X		X									
12	X		X		X	X			X		X								
13		X		X	X		X												
14							X												
15	X	X		X	X	X			X			X	X	X					
16	X	X		X	X	X	X		X				X	X	X				
17		X		X	X	X	X						X	X	X	X			
18							X							X					
19								X		X	X			X					
20		X					X						X	X		X	X	X	

The ordering of texts according to years does not bring any regularity as can be seen in Figure 2.

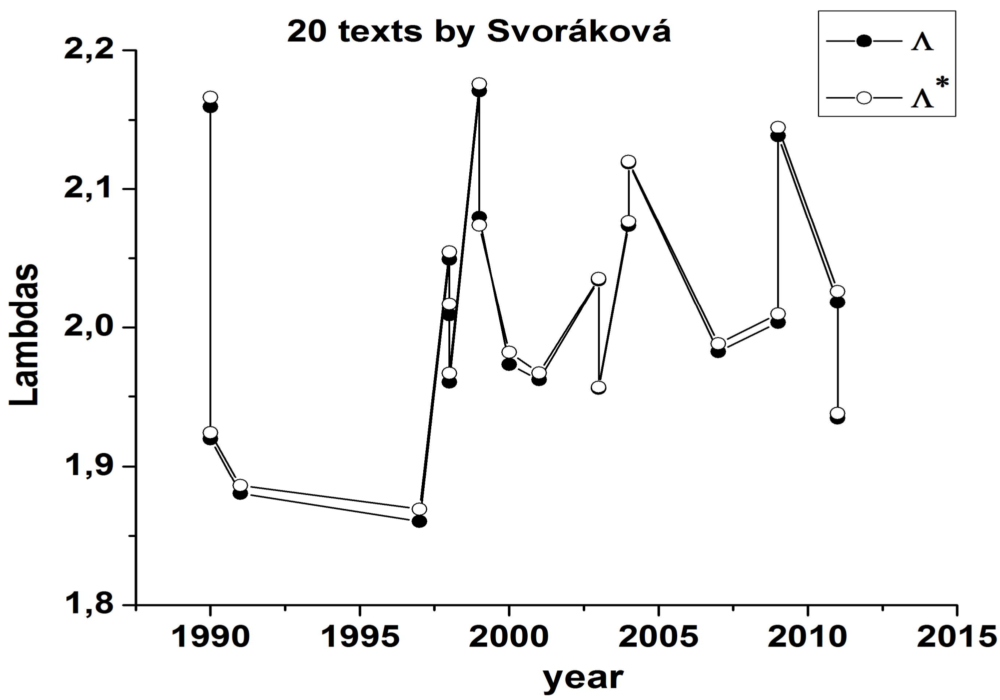

Figure 2. Lambdas in terms of years (Svoráková)
Notice the close coincidence of Λ and Λ* (up to a few per-mille)

No trend can be observed. The mean lambda converges against 2.00. Notice the clos3e coincidence of Λ and Λ* (up to a few pro-mille).

The writer can be characterized by his lambda in form of an indicator. Though writers cannot consciously control the frequency distribution in their texts, they have, perhaps, an intuitive image of it depending on the given language, on the "prescriptions" for a good style, on the aim, etc. Many texts do not have significantly different lambdas as can be seen in Table 1. An indicator of the unity of the material style can be proposed in form of a ratio between the number of similarities *S* (= non-significant lambda differences) and the number of all text pairs, $n(n - 1)/2$, that is as

$$SI = \frac{2S}{n(n-1)} \tag{8}$$

where *S* is the number of similarities and *n* is the number of texts. The number of non-significantly differing pairs (*S* = similar pairs) is given in Table 3. The number of crosses in the lower triangle of the matrix is *S* = 70. The number of all possible pairs in the lower triangle of the matrix is $n(n - 1)/2 = 20(19)/2 = 190$ hence for the 20 texts by Svoráková we obtain *SI*(Svoraková) = 70/190 = 0.3684. The result is a simple proportion which can, again, be compared with texts of other writers using either the binomial or the asymptotic normal test.

The greater is *SI*, the more a writer uses an unconscious background model of frequencies. That is, if all lambdas had the same (non-significantly different) value, the graph would be complete. If *SI* is smaller than 0.5, then there are groups of texts having a similar frequency structure.

As can be seen, there is no unique structuring with Svoraková. If one expects the similarity of two texts with $p = 0.5$ used as the parameter of the binomial distribution, then the probability that up to 70 pairs out of 190 have a similar lambda structure ($X <= 70$) is 0.0002, i.e. a quite variegated rank-frequency structure. The most prominent structure is the one represented by text T16 having similarities with 12 other texts. Thus some texts follow the same background tendency which must still be deciphered.

Needless to say, this is only one aspect of style considering the rank-frequencies of words. But *Λ** can be computed for any property, hence the search for the property which is either constant with the writer or converges with years towards a specific value opens a new domain of research.

In the sequel we present tabular results displaying the modified lambda for various data, perform the tests for similarity and present the resulting similarities

For Latin we considered some works by Horace and Vergil as presented in Table 4

Table 4
Modified lambda for some Latin texts

	Text	**Λ***	**Var(Λ*)**
1	Horatius, Carmen Saeculare	2.3191	0.000917
2	Horatius, Ars Poetica	2.4640	0.000180
3	Horatius, Epodes	2.4322	0.000101
4	Horatius, Carmina Liber I	2.5252	0.000096
5	Horatius, Carmina Liber II	2.6476	0.000163

6	Horatius, Carmina Liber III	2.6579	0.000119
7	Horatius, Carmina Liber IV	2.5738	0.000139
8	Vergilius, Georgicon Liber I	2.4774	0.000145
9	Vergilius, Georgicon Liber II	2.4149	0.000142
10	Vergilius, Georgicon Liber III	2.3954	0.000132
11	Vergilius, Georgicon Liber IV	2.4255	0.000127
12	Vergilius, Aeneid I	2.2660	0.000092
13	Vergilius, Aeneid II	2.2162	0.000099
14	Vergilius, Aeneid III	2.3468	0.000118
15	Vergilius, Aeneid IV	2.2805	0.000088
16	Vergilius, Aeneid V	2.2074	0.000076

After having tested the similarities of individual authors we obtained

SI(Horatius) = 2(10)/[7(6)] = 0.4762
SI(Vergilius) = 2(5)/[9(8)] = 0.1389

Hence Horatius is more concentrated than Vergilius.

The lambdas of the End-of-Year speeches of Czech presidents are presented in Table 5.

Table 5
End-of-Year speeches of Czech presidents

President	year	Λ*	Var(Λ*)		President	year	Λ*	Var(Λ*)
Gottwald	1949	1.9663	0.000308		Husák	1981	1.9102	0.000307
Gottwald	1953	1.9090	0.000293		Husák	1978	1.7998	0.000331
Gottwald	1952	1.8949	0.000298		Klaus	2007	2.0198	0.000405
Gottwald	1950	1.8074	0.000205		Klaus	2006	1.9276	0.000468
Gottwald	1951	1.7963	0.000179		Klaus	2009	1.9462	0.000383
Havel	1997	1.9502	0.000634		Klaus	2011	1.9964	0.000458
Havel	1998	1.7966	0.000266		Klaus	2010	1.9066	0.000395
Havel	2001	1.9109	0.000265		Klaus	2004	1.8191	0.000450
Havel	1999	2.0116	0.000240		Klaus	2008	1.8992	0.000373
Havel	2002	1.9254	0.000223		Klaus	2005	1.9475	0.000392
Havel	2003	1.9532	0.000215		Novotný	1961	1.8645	0.000232
Havel	2000	1.9235	0.000203		Novotný	1958	1.8655	0.000293
Havel	1990	1.8540	0.000163		Novotný	1963	1.7448	0.000210
Havel	1991	1.8530	0.000187		Novotný	1959	1.8016	0.000228
Havel	1994	1.7920	0.000144		Novotný	1965	1.7071	0.000180
Havel	1996	1.8460	0.000151		Novotný	1968	1.7989	0.000178
Havel	1995	1.8734	0.000170		Novotný	1967	1.7051	0.000156
Havel	1992	1.9080	0.000172		Novotný	1962	1.7462	0.000161
Husák	1988	2.0501	0.000526		Novotný	1960	1.7648	0.000192
Husák	1989	1.9723	0.000548		Novotný	1964	1.6652	0.000111
Husák	1984	2.0923	0.000519		Novotný	1966	1.7003	0.000116
Husák	1983	1.9131	0.000396		Svoboda	1974	2.0004	0.000786

Husák	1982	1.9038	0.000381		Svoboda	1972	1.8201	0.000748
Husák	1977	1.7757	0.000342		Svoboda	1973	1.9175	0.000674
Husák	1986	1.9703	0.000391		Svoboda	1971	1.9120	0.000243
Husák	1979	1.9793	0.000375		Svoboda	1969	1.8497	0.000206
Husák	1980	1.9912	0.000396		Svoboda	1970	1.7943	0.000208
Husák	1985	1.9410	0.000340		Zápotocký	1955	1.9033	0.000384
Husák	1976	1.8898	0.000353		Zápotocký	1957	1.9368	0.000181
Husák	1987	1.9106	0.000306		Zápotocký	1954	1.8318	0.000219
Husák	1975	1.7783	0.000301		Zápotocký	1956	1.8740	0.000191

For the Czech presidents we obtain 537 similarities between 62 texts.

SI(Czech presidents) = 2(537)/[62(61)] = 0.2840.

For the individual presidents we obtain

SI(Klaus) = 2(13)/[8(7)] = 0.4483
SI(Zápotocký) = 2(2)/[4(3)] = 0.3333
SI(Havel) = 2(31)/[15(14)] = 0.2952
SI(Gottwald) = 2(2)/[5(4)] = 0.2000
SI(Novotný) = 2(11)/[11(10)] = 0.2000
SI(Svoboda) = 2(3)/[6(5)] = 0.2000
SI(Husák) = 2(15)/[31(30)] = 0.0323

For the episodes in *Finnegans Wake* by James Joyce (in his special English) we obtain the results in Table 6, and *SI* = 2(24)/[17(16)] = 0.1765

Table 6
Finnegans Wake by J. Joyce

Text	**Λ***	**Var(Λ*)**
FW Episode 01	1.9120	0.00009865
FW Episode 02	1.9750	0.00013841
FW Episode 03	1.9940	0.00009003
FW Episode 04	1.9602	0.00009225
FW Episode 05	1.8622	0.00010627
FW Episode 06	1.8508	0.00005766
FW Episode 07	1.9456	0.00008813
FW Episode 08	1.8772	0.00009362
FW Episode 09	1.9751	0.00005528
FW Episode 10	2.0103	0.00005512
FW Episode 11	1.9741	0.00004526
FW Episode 12	1.7342	0.00015782
FW Episode 13	1.8336	0.00007823
FW Episode 14	1.7052	0.00004930
FW Episode 15	1.8422	0.00003257

FW Episode 16	1.8659	0.00005619
FW Episode 17	1.8805	0.00006718

The Latin data concerning the Metamorphoses by Apuleius are presented in Table 7.

Table 7
Modified lambdas for Apuleius' prose

Title	**Λ***	**Var(Λ*)**
Metamorphoses, Liber I	2.24972	0.000147
Metamorphoses, Liber II	2.32547	0.000146
Metamorphoses, Liber III	2.26398	0.000108
Metamorphoses, Liber IV	2.39426	0.000091
Metamorphoses, Liber V	2.28823	0.000121
Metamorphoses, Liber VI	2.39123	0.000122
Metamorphoses, Liber VII	2.41834	0.000104
Metamorphoses, Liber VIII	2.39675	0.000086
Metamorphoses, Liber IX	2.31362	0.000071
Metamorphoses, Liber X	2.38133	0.000077
Metamorphoses, Liber XI	2.34833	0.000082

The similarity is SI(Apuleius) = 2(14)/[11(10)] = 0.2545.

For the poems by H. Heine the results are presented in Table 8.

Table 8
Modified lambdas for Heine's poems

ID	**Poem title**	**Λ***	**Var(Λ*)**
1	An eine Saengerin	1.7169	0.001188
2	Belsazar	1.6562	0.001346
3	Das Lied von den Dukaten	1.2116	0.002158
4	Das Liedchen von der Reue	1.6614	0.001221
5	Der arme Peter	1.6010	0.001607
6	Der Traurige	1.6738	0.002754
7	Der wunde Ritter	1.4844	0.003084
8	Die Bergstimme	1.2804	0.002475
9	Die Botschaft	1.4989	0.003531
10	Die Fensterschau	1.3058	0.002676
11	Die Grenadiere	1.5253	0.000927
12	Die Heimfuehrung	1.6814	0.001405
13	Die Minnesaenger	1.5953	0.002086
14	Don Ramiro	1.6392	0.000534
15	Gespraech auf der Paderborner Heide	1.4734	0.000906

16	Lebensgruss	1.5306	0.003096
17	Lied des Gefangenen	1.4338	0.001989
18	Wahrhaftig	1.4111	0.004331
19	Wasserfahrt	1.4744	0.003127
20	Zwei Brueder	1.7203	0.001108

The resulting *SI* is *SI*(Heine) = 2(78)/[20(19)] = 0.4105.

The results for 7 poems by Goethe are presented in Table 9

Table 9
Modified lambda for some poems by Goethe

Text	**Λ***	**Var(Λ*)**
Der Gott und die Bajadere	1.7349	0.000686
Elegie 19	1.7200	0.000532
Elegie 13	1.7372	0.000541
Elegie 15	1.7516	0.000564
Elegie 2	1.6826	0.001208
Elegie 5	1.6371	0.001433
Der Erlkönig	1.3381	0.001143

The resulting *SI* is *SI*(Goethe) = 2(12)/[7(6)] = 0.5714.

The data for the poems by Schiller given alphabetically are presented in Table 10

Table 10
Modified lambda for poems by Schiller

Poem title	**Λ***	**Var(Λ*)**
Abschied vom Leser	1.63709	0.0016654
Amalia	1.56339	0.0022265
An den Fruehling	1.18404	0.0025974
An die Astronomen	1.38338	0.0035301
An einen Moralisten	1.77892	0.0012644
Bittschrift	1.72785	0.0012213
Das Geheimnis	1.78819	0.0011794
Das Glueck der Weisheit	1.61459	0.0017723
Das Lied von der Glocke	1.78486	0.0002398
Das Maedchen aus der Fremde	1.46877	0.0016274
Das Maedchen von Orleans	1.59199	0.0019622
Das Spiel des Lebens	1.68363	0.0023222
Das verschleierte Bild zu Sais	1.60798	0.0004396
Der Abend	1.67933	0.0035509
Die Antiken zu Paris	1.50499	0.0042235
Die schoenste Erscheinung	0.90268	0.0084043
Die Weltweisen	1.77038	0.0011538
Epigramme Friedrich Schiller	1.60172	0.0015133

Forum des Weibes	1.12886	0.0053097
Odysseus	1.37725	0.0039142
Sehnsucht	1.69123	0.0015812
Spinoza	1.33445	0.0039782
Thekla	1.65335	0.0009662
Triumph der Liebe	0.97592	0.010194
Weibliches Urteil	1.21145	0.0042913
Winternacht	1.83867	0.000615
Zum Geburtstag der Frau Griesbach	1.67568	0.0015782

The resulting similarity is *SI*(Schiller) = 2(115)/[27(26)] = 0.3276.

The results for the poetry by Droste are given in Table 11.

Table 11
Modified lambda for the poems by Droste-Hülshoff

Poem title (alphabetically)	**Λ***	**Var(Λ*)**		**Poem title (alphabetically)**	**Λ***	**Var(Λ*)**
Ungastlich oder nicht?	1.802953	0.000567		Stammbuchblätter	1.690145	0.000572
Die Stadt und der Dom	1.766735	0.000495		Nachruf an Henriette von Hohenhausen	1.726039	0.001053
Die Verbannten	1.675264	0.000424		Vanitas Vanitatum!	1.677657	0.001001
Der Prediger	1.812078	0.000581		Instinkt	1.759228	0.000613
An die Schriftstellerinnen in Deutschland und Frankreich	1.783095	0.000732		Die rechte Stunde	1.590940	0.001864
Die Gaben	1.847654	0.000766		Der zu früh geborene Dichter	1.777213	0.001043
Vor vierzig Jahren	1.753502	0.001199		Not	1.579473	0.003173
An die Weltverbesserer	1.656243	0.000736		Die Bank	1.745004	0.000895
Alte und neue Kinderzucht	1.796161	0.0004		Clemens von Droste	1.696942	0.000694
Die Schulen	1.652365	0.001607		Guten Willens Ungeschick	1.722175	0.001175
Die Lerche	1.829529	0.00065		Der Traum	1.722504	0.000738
Die Jagd	1.69345	0.000505		Locke und Lied	1.679074	0.001694
Die Vogelhütte	1.696461	0.000347		An Levin Schücking	1.738357	0.001265
Der Weiher	1.521921	0.003202		An denselben	1.669921	0.000493
Das Schilf	1.538325	0.002021		Poesie	1.765049	0.001168
Die Linde	1.656367	0.001546		An Levin Schücking	1.617261	0.001563
Die Wasserfäden	1.672923	0.002083		An Elise	1.731476	0.000803
Kinder am Ufer	1.627271	0.002223		Ein Sommertagstraum	1.693294	0.00085
Der Hünenstein	1.757728	0.000513		Das Autograph	1.772874	0.000694
Die Steppe	1.801789	0.002037		Der Denar	1.749692	0.00084
Die Mergelgrube	1.670071	0.000351		Die Erzstufe	1.744440	0.000838
Die Krähen	1.667236	0.000302		Die Muschel	1.744578	0.000767
Das Hirtenfeuer	1.686183	0.000922		Die junge Mutter	1.763445	0.000687
Der Heidemann	1.714451	0.000903		Meine Sträuße	1.803367	0.000882

Das Haus in der Heide	1.681602	0.001578		Das Liebhabertheater	1.679659	0.000881
Der Knabe im Moor	1.604106	0.001091		Die Taxuswand	1.635019	0.001147
Die Elemente	1.819348	0.000552		Nach fünfzehn Jahren	1.668216	0.000823
Die Schenke am See	1.782087	0.000492		Der kranke Aar	1.406127	0.001962
Am Turme	1.674223	0.001756		Sit illi terra levis!	1.676783	0.000466
Das öde Haus	1.740705	0.000734		Die Unbesungenen	1.531045	0.002555
Im Moose	1.651482	0.000761		Das Spiegelbild	1.626478	0.001121
Am Bodensee	1.782606	0.000657		Neujahrsnacht	1.830559	0.000536
Das alte Schloß	1.642105	0.001093		Der Todesengel	1.661097	0.001415
Der Säntis	1.813262	0.000618		Abschied von der Jugend	1.560498	0.001223
Am Weiher	1.771844	0.000635		Was bleibt	1.799409	0.001348
Mein Beruf	1.710763	0.000622		Dichters Naturgefühl	1.777573	0.000531
Meine Toten	1.7064	0.000726		Der Teetisch	1.851103	0.000626
Katharine Schücking	1.622416	0.000713		Die Nadel im Baume	1.669545	0.000814
Nach dem Angelus Silesius	1.498151	0.00056		Die beschränkte Frau	1.670499	0.00056
Gruß an Wilhelm Junkmann	1.686478	0.000723		Die Stubenburschen	1.674037	0.00089
Junge Liebe	1.687609	0.001387		Die Schmiede	1.684100	0.001259
Das vierzehnjährige Herz	1.587469	0.001173		Des alten Pfarrers Woche	1.636108	0.000196
Blumentod	1.593652	0.00186		Der Strandwächter am deutschen Meere und sein Neffe vom Lande	1.742256	0.000555
Brennende Liebe	1.58624	0.001046		Das Eselein	1.813313	0.00079
Der Brief aus der Heimat	1.743498	0.001266		Die beste Politik	1.715257	0.00109
Ein braver Mann	1.776374	0.000468				

The resulting similarity is *SI*(Droste-Hülshoff) = 2(2164)/[91(90)] = 0.5284.
For the Slovak poetic texts by E. Bachletová one obtains the results presented in Table 12.

Table 12
Modified lambda for Slovak poetry by E. Bachletová

Text	**Λ***	**Var(Λ*)**		**Text**	**Λ***	**Var(Λ*)**
Aby spriesvitnela	1.513731	0.003056		Neopusť ma...	1.232073	0.009344
Bez rozlúčky	1.367603	0.00367		Nepoznateľné	1.545153	0.003276
Čakáme šťastie...	1.488599	0.00345		Podobnosť bytia	1.770526	0.004146
Čakanie na Boží jas	1.567992	0.004771		Pravidlá odpúšťania	1.352899	0.005338
Čas pre nádych vône	1.731773	0.001604		Precitnutie	1.540161	0.003165
Dielo Stvoriteľa	1.835471	0.001853		Prvotný sen	1.778897	0.003084
Dnešný luxus	1.232073	0.006645		Rozdelená bytosť	1.681442	0.001665
Do večnosti beží čas	1.339271	0.004132		Rozťatá prítomnosť	1.455457	0.002233
Dovoľ mi slúžiť	1.463914	0.003819		Som iná	1.408714	0.00569
Ešte raz	1.27541	0.005406		Spájania	1.488599	0.00345
Hľadanie odpovedí	1.580781	0.002129		Stály smútok pre šesť písmen	1.482433	0.00242
Iba neha	1.546876	0.002606		Tá Láska	1.279371	0.00367

Iba život	1.593652	0.0039		Tak málo úsmevu	1.685097	0.005697
Idem za Tebou	1.676759	0.001913		Ťažko pokoriteľní	1.280172	0.006546
Ihly na nebi	1.251173	0.003812		Tiché verše	1.347036	0.00433
Istota	1.317762	0.005586		To všetko je dar	1.145216	0.004224
Keď dohorí deň	1.468503	0.004921		Ulomené zo slov	1.202711	0.006271
Kým ich máme	1.494048	0.005073		Vďaka Pane!	1.44105	0.003399
Len áno	1.171131	0.003819		Vďaka za deň	1.427878	0.003158
Malé modlitby	1.406234	0.003165		Večerná ruža	1.602512	0.003664
Malý ošiaľ	1.293536	0.00456		Večerné ticho	1.482176	0.003973
Miesto pre Nádej	1.386757	0.004735		Vo večnosti slobodná	1.705637	0.001467
Moje určenie	1.808569	0.002235		Vrátili sa	1.540161	0.003165
Nado mnou Ty sám...	1.455064	0.006186		Vyznania	1.550505	0.00284
Náš chrám	1.743298	0.004077		Z neba do neba	1.553526	0.004058
Naše mamy	1.545273	0.00362		Zasľúbenie jasu	1.386003	0.004021
Naše svetlo	1.280978	0.003437		Zbytočné srdce	1.296919	0.009344

The similarity for Bachletová is *SI*(Bachletová) = 2(701)/[54(53)] = 0.4899.

The results the Hungarian poems written by E. Ady are presented in Table 13.

Table 13
Modified lambda for Hungarian poetry by E. Ady

Text	**Λ***	**Var(Λ*)**
A Rákóczi vén harangja	1.7586	0.001821
Dal a rózsáról	1.7085	0.003408
Divina Comoedia	1.6318	0.004258
E eéhány dalban...	1.4060	0.002176
Egy csókodért	1.5616	0.003914
Egy szép leányhoz	1.7303	0.001213
Eltagadom	1.3970	0.003351
Én szép világom...	1.3834	0.003530
Epilógok	1.7260	0.003278
Érted	1.6316	0.002920
Karácsony	1.5386	0.001960
Láttalak...	1.0714	0.003165
Milyen az ösz?...	1.4758	0.001896
Mutamur	1.7921	0.001167
Nem élek én tovább...	1.5040	0.001156
Ösz felé	1.1561	0.002976
Sirasson meg	1.5850	0.000697
Sorsunk	1.7141	0.001556
Színházban	1.2884	0.003978
Temetetlenül	1.8161	0.002009
Válasz	1.7202	0.001762
Válaszúton	1.6190	0.001799
Van olyan perc...	1.5998	0.003518

The similarity with E. Ady is *SI*(Ady) = 2(98)/[23(22)] = 0.3874.

The values for the poems by the Romanian writer M. Eminescu are presented in Table 14.

Table 14
Modified lambda for the Romanian poems by M. Eminescu

ID	Poem title (alphabetically)	Λ*	Var(Λ*)		Id #	Poem title (alphabetically)	Λ*	Var(Λ*)
1	Adânca mare…	1.5751	0.002917		74	La moartea lui Heliade	1.7637	0.000773
2	Adio	1.5784	0.001628		75	La moartea lui Neamțu	1.7065	0.000736
3	Ah, mierea buzei tale	1.5306	0.001023		76	La moartea principelui Ştirbey	1.5904	0.001478
4	Amicului F.I.	1.8192	0.000431		77	La mormântul lui Aron Pumnul	1.7264	0.001594
5	Amorul unei marmure	1.7047	0.000723		78	La o artistă (Ca a nopții poezie)	1.6142	0.001529
6	Andrei Mureşanu	1.7440	0.000170		79	La o artistă (Credeam ieri)	1.6992	0.001295
7	Atât de frageda…	1.7798	0.001679		80	La Quadrat	1.5125	0.002257
8	Aveam o muză	1.8077	0.000634		81	La steaua	1.5905	0.001953
9	Basmul ce i l-aş spune ei	1.7833	0.000733		82	Lacul	1.5525	0.00264
10	Când	1.7086	0.001837		83	Lasă-ți lumea…	1.7929	0.001044
11	Când amintirile…	1.6591	0.001989		84	Lebăda	1.4425	0.004302
12	Când crivățul cu iarna…	1.7712	0.000480		85	Lida	1.5990	0.002856
13	Când marea…	1.4795	0.002139		86	Locul aripelor	1.6213	0.000673
14	Când priveşti oglinda mărei	1.6670	0.002223		87	Luceafărul	1.6581	0.000278
15	Care-i amorul meu în astă lume	1.7326	0.000809		88	Mai am un singur dor	1.7279	0.001090
16	Călin (file de poveste)	1.7600	0.000201		89	Melancolie	1.8016	0.001262
17	Ce e amorul?	1.6545	0.002133		90	Memento mori	1.6212	6.75E-05
18	Ce te legeni…	1.5655	0.002859		91	Miradoniz	1.7984	0.000728
19	Ce-ți doresc eu ție, dulce Românie	1.5989	0.001169		92	Misterele nopții	1.5827	0.001335
20	Cine-i?	1.5653	0.002009		93	Mitologicale	1.9428	0.000559
21	Copii eram noi amândoi	1.8258	0.001058		94	Mortua est!	1.6908	0.000631
22	Crăiasa din poveşti	1.6588	0.001930		95	Mureşanu	1.6649	0.000220
23	Criticilor mei	1.4800	0.001025		96	Murmură glasul mării	1.7790	0.001733
24	Cu mâne zilele-ți adaogi…	1.6233	0.001335		97	Napoleon	1.7653	0.001297
25	Cugetările sarmanului Dionis	1.9649	0.000600		98	Noaptea,,,	1.6574	0.001232
26	Cum negustorii din Constantinopol	1.6868	0.001872		99	Nu e steluță	1.2512	0.002916
27	Cum oceanu-ntărâtat…	1.6537	0.002276		100	Nu mă-nțelegi	1.7565	0.000527
28	Dacă treci râul Selenei	1.7631	0.001015		101	Nu voi mormânt bogat	1.8169	0.001578
29	De câte ori, iubito…	1.7034	0.002190		102	Numai poetul	1.3835	0.003450
30	De ce nu-mi vii	1.4612	0.002408		103	O arfă pe-un mormânt	1.6924	0.001485
31	De ce să mori tu?	1.6227	0.001028		104	O călărire în zori	1.8016	0.000967

32	De-aş avea	1.3123	0.002515		105	O stea prin ceruri	1.5525	0.001697
33	De-aş muri ori de-ai muri	1.6015	0.000840		106	O, adevăr sublime…	1.7908	0.000972
34	Demonism	1.7533	0.000385		107	O, mamă…	1.5329	0.001563
35	De-oi adormi (variantă)	1.7956	0.001131		108	Odă în metru antic	1.6220	0.001468
36	De-or trece anii…	1.4491	0.003199		109	Odin şi poetul	1.6912	0.000262
37	Departe sunt de tine	1.6885	0.001653		110	Ondina (Fantazie)	1.8877	0.000383
38	Despărțire	1.7070	0.000891		111	Oricâte stele…	1.6570	0.001491
39	Din Berlin la Potsdam	1.6682	0.001793		112	Pajul Cupidon…	1.7450	0.001818
40	Din lyra spartă…	1.4732	0.003165		113	Pe aceeaşi ulicioară…	1.6282	0.001598
41	Din noaptea	1.5091	0.002083		114	Pe lângă plopii fără soț	1.6288	0.001147
42	Din străinătate	1.7123	0.001178		115	Peste vârfuri	1.4409	0.005655
43	Din valurile vremii…	1.5072	0.001376		116	Povestea codrului	1.8367	0.000979
44	Dintre sute de catarge	1.4044	0.004249		117	Povestea teiului	1.8071	0.000758
45	Doi aştri	1.5420	0.003048		118	Prin nopți tăcute	1.3835	0.003450
46	Dorința	1.7394	0.002528		119	Privesc oraşul furnicar	1.8241	0.001577
47	Dumnezeu şi om	1.9564	0.000517		120	Pustnicul	1.8636	0.000536
48	Ecò	1.8823	0.000477		121	Replici	1.1879	0.002928
49	Egipetul	1.9282	0.000394		122	Revedere	1.5624	0.001335
50	Epigonii	1.9021	0.000368		123	Rugăciunea unui dac	1.8591	0.000688
51	Făt-Frumos din tei	1.8326	0.000649		124	S-a dus amorul	1.6672	0.001090
52	Feciorul de impărat fără de stea	1.5380	7.86E-05		125	Sara pe deal	1.8182	0.001140
53	Floare-albastră	1.8611	0.001071		126	Scrisoarea I	1.8097	0.000282
54	Foaia veştedă (dupa Lenau)	1.7919	0.001536		127	Scrisoarea II	1.8011	0.000479
55	Freamăt de codru	1.8249	0.001065		128	Scrisoarea III	1.8271	0.000227
56	Frumoasă şi jună	1.5444	0.002454		129	Scrisoarea IV	1.8522	0.000375
57	Ghazel	1.7966	0.00067		130	Scrisoarea V	1.7140	0.000378
58	Glossă	1.3605	0.000832		131	Se bate miezul nopții…	1.4695	0.003779
59	Horia	1.8237	0.001306		132	Singurătate	1.7546	0.000978
60	Iar când voi fi pământ (variantă)	1.7455	0.001496		133	Somnoroase păsărele…	1.4714	0.003714
61	Iubind în taină…	1.7054	0.001440		134	Sonete	1.7923	0.000649
62	Iubită dulce, o, mă lasă	1.6201	0.000599		135	Speranța	1.5213	0.001667
63	Iubitei	1.5708	0.000646		136	Steaua vieții	1.4497	0.002620
64	Împărat şi proletar	1.8895	0.000235		137	Stelele-n cer	1.6577	0.002190
65	În căutarea Şeherezadei	2.0002	0.000343		138	Sus în curtea cea domnească	1.7286	0.001302
66	Înger de pază	1.5608	0.002190		139	Şi dacă…	1.2037	0.003914
67	Înger şi demon	1.8072	0.000327		140	Te duci…	1.7955	0.001358
68	Îngere palid…	1.4995	0.002331		141	Trecut-au anii	1.6462	0.001864
69	Întunericul şi poetul	1.7418	0.000974		142	Unda spumă	1.3807	0.002565
70	Junii corupți	1.8795	0.000737		143	Venere şi Madona	1.6944	0.000709
71	Kamadeva	1.6611	0.002111		144	Veneția (de Gaetano Cerri)	1.6935	0.001665
72	La Bucovina	1.7478	0.001020		145	Viața mea fu ziuă	1.6747	0.001764
73	La mijloc de codru…	1.3344	0.008811		146	Vis	1.7844	0.001084

The similarity value for M. Eminescu is *SI*(Eminescu) = 2(3734)/[146(145)] = 0.3528.

The modified lambdas for the Russian poetry by Pushkin is presented in Table 15.

Table 15
Modified lambda for the Russian poetry by A. Pushkin

ID	Poem title	Λ*	Var(Λ*)		ID	Poem title	Λ*	Var(Λ*)
1	Анчар	1.8822	0.002532		19	Няне	1.5510	0.003079
2	Арион	1.5552	0.002620		20	О, дева-роза, я в оковах...	1.4411	0.004952
3	Бесы	1.6468	0.000821		21	Поэт	1.5938	0.002150
4	Брожу ли я вдоль улиц шумных...	1.8313	0.001512		22	Признание	1.6724	0.001420
5	Вакхическая песня	1.6353	0.003437		23	Пробуждение	1.4746	0.002630
6	Во глубине сибирских руд...	1.6256	0.003294		24	Пророк	1.8314	0.004380
7	Десятая заповедь	1.5186	0.002197		25	Птичка	1.3963	0.003819
8	Если жизнь тебя обманет…	1.2027	0.004330		26	Свободы сеятель пустынный...	1.5609	0.001879
9	Зимнее утро	1.9020	0.001690		27	Старик	1.4300	0.004161
10	Зимний вечер	1.3422	0.001882		28	Стихи, сочиненные ночью во время бессонницы...	1.5457	0.003359
11	Зимняя дорога	1.7632	0.000862		29	Талисман	1.4762	0.001860
12	К ***	1.5243	0.003448		30	Туча	1.4707	0.004123
13	К морю	1.8802	0.000882		31	Узник	1.5190	0.002676
14	К Чаадаеву	1.7775	0.001665		32	Утопленник	1.9034	0.000825
15	Когда в объятия мои...	1.6894	0.003199		33	Что в имени тебе моем?	1.6545	0.004553
16	Красавица	1.5263	0.002794		34	Я вас любил: любовь еще, быть может...	1.2572	0.005196
17	На холмах Грузии лежит ночная мгла...	1.3633	0.005073		35	Я пережил свои желанья...	1.5411	0.002351
18	Ночь	1.4565	0.004132					

The similarities in A. Pushkin are expressed by *SI*(Pushkin) = 2(251)/[35(34)] = 0.4128.

The modified lambdas for the Russian poetry by Lermontov is presented in Table 16.

Table 16
Modified lambda for the Russian poetry by M. Lermontov

ID	Poem title	Λ*	Var(Λ*)		ID	Poem title	Λ*	Var(Λ*)
1	Баллада	1.8143	0.001697		16	Незабудка	1.9568	0.00058
2	Бородино	1.8613	0.000528		17	Одиночество	1.5190	0.002276
3	Валерик	2.1011	0.000336		18	Предсказание	1.7526	0.002859
4	Видение	1.9831	0.000907		19	Пророк	1.6847	0.001793

5	Воля	1.6569	0.002009		20	Разлука	1.7395	0.001731
6	Гроза	1.7639	0.001999		21	Раскаянье	1.7377	0.001958
7	Гусар	1.7274	0.001896		22	Ребенку	1.7429	0.000948
8	Дары Терека	1.9287	0.000834		23	Русалка	1.6546	0.002446
9	Два Великана	1.6376	0.002367		24	Св. Елена	1.7621	0.002073
10	Договор	1.6285	0.00264		25	Сентября 28	1.6637	0.001084
11	Дума	1.9408	0.001605		26	Смерть Поэта	2.0450	0.001145
12	Желание	1.6096	0.001874		27	Совет	1.7119	0.002947
13	Листок	1.7708	0.00207		28	Сон	1.6331	0.002555
14	Мой Демон	1.5783	0.004756		29	Соседка	1.8040	0.000835
15	Наполеон	1.8935	0.001217		30	Счастливый Миг	1.8345	0.001107

For Lermonotov we obtain SI(Lermontov) = 2(194)/[30(29)] = 0.4460.
For the Hawaiian texts we obtained the results presented in Table 17.

Table 17
Modified lambda for the texts of Hawaiian Romance of Laieikawai by Anonymous

ID	Title	Λ*	Var(Λ*)
1	I. The birth of the Princess	0.6510	0.000304
2	II. The flight to Paliuli	0.5722	0.000241
3	III. Kauakahialii meets the Princess	0.6459	0.000267
4	IV. Aiwohikupua goes to woo the Princess	0.5681	0.000180
5	V. The boxing match with Cold-nose	0.6839	0.000289
6	VI. The house thatched with bird feathers	0.7258	0.000351
7	VII. The Woman of the Mountain	0.6533	0.000387
8	VIII. The refusal of the Princess	0.6787	0.000371
9	IX. Aiwohikupua deserts his sisters	0.5574	0.000207
10	XI. Abandoned in the forest	0.6134	0.000334
11	XII. Adoption by the Princess	0.5926	0.000295
12	XIII. Hauailiki goes surf riding	0.6816	0.000352
13	XIV. The stubbornness of Laieikawai	0.6183	0.000267
14	XV. Aiwohikupua meets the guardians of Paliuli	0.6906	0.000381
15	XVI. The Great Lizard of Paliuli	0.7121	0.000423
16	XVII. The battle between the Dog and the Lizard	0.7035	0.000434
17	XVIII. Aiwohikupua's marriage …	0.6438	0.000294
18	XIX. The rivalry of Hina and Poliahu	0.6281	0.000339
19	XX. A suitor is found for the Princess	0.6274	0.000308
20	XXI. The Rascal of Puna wins the Princess	0.6359	0.000372
21	XXII. Waka's revenge	0.6389	0.000304
22	XXIII. The Puna Rascal deserts the Princess	0.6341	0.000311
23	XXIV. The marriage of the chiefs	0.6192	0.000351
24	XXV. The Seer finds the Princess	0.6738	0.000347
25	XXVI. The Prophet of God	0.7094	0.000376
26	XXVII. A journey to the Heavens	0.6786	0.000373
27	XXVIII. The Eyeball-of-the-Sun	0.6982	0.000289
28	XXIX. The warning of vengeance	0.7409	0.000500

29	XXX. The coming of the Beloved	0.8553	0.000507
30	XXXI. The Beloved falls into sin	0.6530	0.000300
31	XXXII. The Twin Sister	0.6490	0.000360
32	XXXIII. The Woman of Hana	0.6489	0.000279
33	XXXIV. The Woman of the Twilight	0.6293	0.000315

The similarity for Hawaiian texts is *SI*(Hawaiian) = 2(268)/[33(32)] = 0.5076.

The results for the poems by Byron are presented in Table 18.

Table 18

Modified lambda in the poems by Byron

ID	Text	Λ*	Var(Λ*)		ID	Text	Λ*	Var(Λ*)
1	And Wilt Thou Weep When I Am Low?	1.3709	0.001349		21	Stanzas to the Po	1.3621	0.000599
2	Farewell to the Muse	1.5584	0.001105		22	There Was A Time, I Need Not Name	1.5230	0.000928
3	Love's Last Adieu	1.6047	0.000963		23	To Caroline	1.5531	0.001222
4	On a Distant View of Harrow	1.7022	0.001515		24	To Mary, On Receiving Her Picture	1.7034	0.00115
5	Remind Me Not, Remind Me Not	1.5484	0.00106		25	To Romance	1.6728	0.000541
6	Sonnet --- to Genevra	1.6952	0.001309		26	When We Two Parted	1.4981	0.00148
7	Stanzas for Music	1.5634	0.003437		27	So, We'll Go no More a Roving	1.2931	0.005762
8	Stanzas to Jessy	1.5572	0.000913		28	from Childe Harold's Pilgrimage	1.4641	0.001562
9	The Tear	1.5797	0.00094		29	And Thou art Dead, as Young and Fair	1.5227	0.000637
10	To A Lady	1.5636	0.000757		30	The Destruction of Sennacherib	1.5511	0.003116
11	To M.S.G.	1.5268	0.001021		31	The Eve of Waterloo	1.7687	0.001133
12	To M.	1.7255	0.000805		32	On this Day I Complete my Thirty-Sixth Year	1.7870	0.002043
13	To Time	1.6201	0.000992		33	Prometheus	1.5073	0.00116
14	Darkness	1.6466	0.000966		34	There be none of beauty's daughters	1.5715	0.003382
15	I Saw Thee Weep	1.5743	0.002855		35	Many Are Poets Who Have Never Penn'd	1.6530	0.001321
16	Ode To Napoleon Buonaparte	1.6837	0.000476		36	Kirke White	1.7916	0.002499
17	Remember Him, Whom Passion's Power	1.6495	0.000748		37	Crabbe	1.5925	0.002806
18	She Walks In Beauty	1.5940	0.001711		38	England! with all thy faults I love thee still	1.4268	0.001778

19	Stanzas Composed During a Thunderstorm	1.7842	0.000743		39	Adieu, adieu! my native shore	1.5019	0.000577
20	Stanzas To A Lady, On Leaving England	1.4938	0.00096		40	America	1.6733	0.001165

The similarity for Byron is *SI*(Byron) = 2(362)/[40(39)] = 0.4641.

For the End-of-Year speeches of Italian Presidents we obtain the results presented in Table 19.

Table 19
Modifed lambda in the End-of-Year speeches of Italian Presidents

ID	**Text**	**Λ***	**Var(Λ*)**		**ID**	**Text**	**Λ***	**Var(Λ*)**
1	1949 Einaudi	1.6982	0.001319		34	1982 Pertini	1.2620	0.000162
2	1950 Einaudi	1.5813	0.001781		35	1983 Pertini	1.1848	0.000104
3	1951 Einaudi	1.7662	0.000912		36	1984 Pertini	1.2625	0.000222
4	1952 Einaudi	1.8501	0.001065		37	1985 Cossiga	1.3711	0.000229
5	1953 Einaudi	1.7510	0.001102		38	1986 Cossiga	1.4186	0.000333
6	1954 Einaudi	1.7369	0.000988		39	1987 Cossiga	1.5825	0.000260
7	1955 Gronchi	1.7081	0.000683		40	1988 Cossiga	1.3855	0.000234
8	1956 Gronchi	1.6725	0.000500		41	1989 Cossiga	1.4502	0.000239
9	1957 Gronchi	1.6238	0.000447		42	1990 Cossiga	1.4287	0.000164
10	1958 Gronchi	1.6267	0.000433		43	1991 Cossiga	1.5990	0.000820
11	1959 Gronchi	1.6766	0.000523		44	1992 Scalfaro	1.3368	0.000174
12	1960 Gronchi	1.6767	0.000508		45	1993 Scalfaro	1.3948	0.000171
13	1961 Gronchi	1.6700	0.000388		46	1994 Scalfaro	1.3208	0.000158
14	1962 Segni	1.5739	0.000504		47	1995 Scalfaro	1.2826	0.000127
15	1963 Segni	1.6022	0.000360		48	1996 Scalfaro	1.4916	0.000214
16	1964 Saragat	1.6636	0.000660		49	1997 Scalfaro	1.1391	8.8E-05
17	1965 Saragat	1.5771	0.000408		50	1998 Scalfaro	1.1602	0.000108
18	1966 Saragat	1.6089	0.000279		51	1999 Ciampi	1.4890	0.000183
19	1967 Saragat	1.6178	0.000398		52	2000 Ciampi	1.5496	0.000211
20	1968 Saragat	1.5804	0.000365		53	2001 Ciampi	1.5327	0.000214
21	1969 Saragat	1.5401	0.000332		54	2002 Ciampi	1.5429	0.000224
22	1970 Saragat	1.4970	0.000236		55	2003 Ciampi	1.5636	0.000252
23	1971 Leone	1.6060	0.000976		56	2004 Ciampi	1.5717	0.000237
24	1972 Leone	1.5609	0.000434		57	2005 Ciampi	1.4906	0.000343
25	1973 Leone	1.6599	0.000389		58	2006Napolitano	1.5715	0.000271
26	1974 Leone	1.6240	0.000404		59	2007Napolitano	1.5922	0.000314
27	1975 Leone	1.6016	0.000332		60	2008Napolitano	1.5744	0.000256
28	1976 Leone	1.5769	0.000264		61	2009 Napolitano	1.5612	0.000238
29	1977 Leone	1.5626	0.000304		62	2010 Napolitano	1.5864	0.000226
30	1978 Pertini	1.3635	0.000231		63	2011 Napolitano	1.6364	0.000242
31	1979 Pertini	1.2387	0.000144		64	2012 Napolitano	1.6539	0.000231
32	1980 Pertini	1.3133	0.000256		65	2013 Napolitano	1.6074	0.000209
33	1981 Pertini	1.2067	0.000139					

For the comparison of all texts of Italian Presidents with all we would obtain *SI*(Italian Presidents) = (2(502)/[65(64)] = 0.2413. However, we need the values for “internal” similarities of individual presidents which can be obtained as

SI(Einaudi) = 2(7)/[6(5)] = 0.4667
SI(Gronchi) = 2(19)/[7(6)] = 0.9048
SI(Segni) = 2(1)/[2(1)] = 1.0000
SI(Saragat) = 2(11)/[7(6)] = 0.5238
SI(Leone) = 2(15)/[7(6)] = 0.7143
SI(Pertini) = 2(5)/[7(6)] = 0.2381
SI(Cossiga) = 2(6)/[7(6)] = 0.2857
SI(Scalfaro) = 2(2)/[7(6)] = 0.0952
SI(Ciampi) = 2(12)/[7(6)] = 0.5714
SI(Napolitano) = 2(17)/[8(7)] = 0.6071

The Slavic data concern the translations of the novel *Kak zakaljalas` stal`* by Ostrovskij from Russian to 11 Slavic languages. The data are given in Table 20, then individual within-language comparisons yield the final *SI*-s.

Table 20
Modified lambdas for the translations of the Russian novel *Kak zakaljajas` stal`* by Ostrovskij

Chapter	**Λ***	**Var(Λ*)**		**Chapter**	**Λ***	**Var(Λ*)**		**Chapter**	**Λ***	**Var(Λ*)**
Bel_01	1.8075	0.000128		Mac_01	1.3813	0.000109		Slk_01	1.7475	0.000128
Bel_02	1.9192	0.000111		Mac_02	1.5043	0.000101		Slk_02	1.8747	0.000124
Bel_03	1.8264	0.000075		Mac_03	1.4038	0.000073		Slk_03	1.8338	0.000095
Bel_04	2.1020	0.000111		Mac_04	1.7071	0.000103		Slk_04	2.0617	0.000133
Bel_05	1.8498	0.000108		Mac_05	1.4809	0.000102		Slk_05	1.8610	0.000135
Bel_06	1.8020	0.000048		Mac_06	1.3850	0.000060		Slk_06	1.7484	0.000075
Bel_07	1.9253	0.000060		Mac_07	1.5589	0.000087		Slk_07	1.8849	0.000080
Bel_08	2.0210	0.000069		Mac_08	1.6264	0.000095		Slk_08	1.9314	0.000090
Bel_09	1.9670	0.000103		Mac_09	1.6125	0.000122		Slk_09	1.9482	0.000148
Bel_10	2.0585	0.000070		Mac_10	1.6721	0.000107		Slk_10	1.9777	0.000086
Bul_01	1.4811	0.000116		Pol_01	1.7658	0.000108		Sln_01	1.6689	0.000192
Bul_02	1.5996	0.000099		Pol_02	1.8985	0.000100		Sln_02	1.7490	0.000176
Bul_03	1.5090	0.000075		Pol_03	1.8270	0.000072		Sln_03	1.7224	0.000134
Bul_04	1.8073	0.000106		Pol_04	2.0824	0.000103		Sln_04	1.9088	0.000169
Bul_05	1.5554	0.000104		Pol_05	1.8749	0.000108		Sln_05	1.7825	0.000215
Bul_06	1.4653	0.000056		Pol_06	1.8326	0.000064		Sln_06	1.6889	0.000117
Bul_07	1.6452	0.000084		Pol_07	1.7773	0.000038		Sln_07	1.8087	0.000124
Bul_08	1.7043	0.000084		Pol_08	2.0129	0.000071		Sln_08	1.8860	0.000127
Bul_09	1.6732	0.000111		Pol_09	1.9736	0.000112		Sln_09	1.8079	0.000089
Bul_10	1.7354	0.000091		Pol_10	2.0249	0.000068		Sln_10	1.8888	0.000136
Cro_01	1.6539	0.000117		Rus_01	1.8081	0.000125		Sor_01	1.6640	0.000130
Cro_02	1.7608	0.000103		Rus_02	1.9416	0.000112		Sor_02	1.7899	0.000117

Cro_03	1.6911	0.000078	Rus_03	1.8612	0.000074	Sor_03	1.6883	0.000082
Cro_04	1.9184	0.000102	Rus_04	2.1683	0.000112	Sor_04	2.0070	0.000141
Cro_05	1.7428	0.000127	Rus_05	1.8854	0.000109	Sor_05	1.7870	0.000121
Cro_06	1.6360	0.000074	Rus_06	1.8138	0.000047	Sor_06	1.6864	0.000067
Cro_07	1.7827	0.000075	Rus_07	1.9535	0.000063	Sor_07	1.8172	0.000081
Cro_08	1.9002	0.000085	Rus_08	2.0428	0.000066	Sor_08	1.8807	0.000081
Cro_09	1.8614	0.000157	Rus_09	1.9673	0.000110	Sor_09	1.8798	0.000114
Cro_10	1.8719	0.000092	Rus_10	2.0722	0.000078	Sor_10	1.8579	0.000083
Cze_01	1.7687	0.000144	Ser_01	1.6535	0.000117	Ukr_01	1.7508	0.000090
Cze_02	1.8902	0.000121	Ser_02	1.7600	0.000103	Ukr_02	1.8774	0.000073
Cze_03	1.8272	0.000097	Ser_03	1.6814	0.000077	Ukr_03	1.8098	0.000050
Cze_04	2.0872	0.000147	Ser_04	1.9051	0.000098	Ukr_04	2.0797	0.000071
Cze_05	1.8589	0.000135	Ser_05	1.7457	0.000129	Ukr_05	1.8417	0.000079
Cze_06	1.7669	0.000073	Ser_06	1.6441	0.000075	Ukr_06	1.7701	0.000041
Cze_07	1.8902	0.000087	Ser_07	1.7778	0.000075	Ukr_07	1.9410	0.000061
Cze_08	1.9643	0.000089	Ser_08	1.8964	0.000085	Ukr_08	2.0181	0.000054
Cze_09	1.9930	0.000143	Ser_09	1.8604	0.000159	Ukr_09	1.9955	0.000092
Cze_10	2.0241	0.000090	Ser_10	1.8668	0.000094	Ukr_10	2.0471	0.000066

The inner-language similarities in decreasing order are as follows

SI(Slovenian) = 2(8)/[10(9)] = 0.1778
SI(Slovak) = 2(7)/[10(9)] = 0.1556
SI(Sorbian) = 2(6)/[10(9)] = 0.1333
SI(Croatian) = 2(5)/[10(9)] = 0.1111
SI(Russian) = 2(4)/[10(9)] = 0.0889
SI(Serbian) = 2(4)/[10(9)] = 0.0889
SI(Belorussian) = 2(4)/[10(9)] = 0.0889
SI(Czech) = 2(3)/[10(9)] = 0.0667
SI(Macedonian) = 2(3)/[10(9)] = 0.0667
SI(Polish) = 2(3)/[10(9)] = 0.0667
SI(Ukrainian) = 2(2)/[10(9)] = 0.0444
SI(Bulgarian) = 2(1)/[10(9)] = 0.0222

Evidently, the geographic distance does not play any role here. The result depends both on the evolution of language and on the style of translators.

Comparing the results in evaluated texts we obtain the *SI* indicator as presented in Table 21. As can be seen, the indicator says something about the person and style, but not about language or text sorts. Of course, many individual investigations are necessary in order to set up hypotheses containing the forces, boundary conditions and links to other properties. Here only the first approximation is presented. Most concentrated is the German poetry and Slovak texts. Latin shows the smallest similarities. But, perhaps, a text sort like "presidential speeches" is quite heterogeneous to yield reliable results. But at least the first step has been done.

Table 21
Summary of similarities in texts

Individual texts	**n**	**S**	**SI descending**
German, poetry, Goethe	7	12	0.5714
German, poetry, Droste-Hülshoff	91	2164	0.5284
Hawaiian, Romance of Laieikawai , Anonymous	33	268	0.5076
Slovak, poetry, Bachletová	54	701	0.4899
English, poetry, Byron	40	362	0.4641
Russian, poetry, Lermontov	30	194	0,4460
Russian, poetry, Pushkin	35	251	0.4218
German, poetry, Heine	20	78	0.4105
Hungarian, poetry, Ady Endre	23	98	0.3874
Slovak, prose, Svoráková	20	70	0.3684
Romanian, poetry, Eminescu	146	3734	0.3528
German, poetry, Schiller	27	115	0.3276
Latin, prose, Apuleius	11	14	0.2545
English, prose, Joyce, *Finnegans Wake*	17	24	0.1765
Latin, poetry, Horatius	7	10	0.4762
Latin, poetry, Vergilius	9	5	0.1389
Czech, Presidential speeches			
Klaus	8	13	0.4483
Zápotocký	4	2	0.3333
Havel	15	31	0.2952
Gottwald	5	2	0.2000
Novotný	11	11	0.2000
Svoboda	6	3	0.2000
Husák	31	15	0.0324
Italian, Presidential speeches			
Einaudi	6	7	0.4667
Gronchi	7	19	0.9048
Segni	2	1	0.9048
Saragat	7	11	0.5238
Leone	7	15	0.7143
Pertini	7	5	0.2381
Cossiga	7	6	0.2857
Scalfaro	7	2	0.0952
Ciampi	7	12	0.5714
Napolitano	8	17	0.6071

Tramslations into Slavic languages *Kak zakaljalas` stal`* by Ostrovskij			
Belorussian	10	4	0.0889
Bulgarian	10	1	0.0222
Croatian	10	5	0.1111
Czech	10	3	0.0667
Macedonian	10	3	0.0667
Polish	10	3	0.0667
Russian	10	4	0.0889
Serbian	10	4	0.0889
Slovak	10	7	0.1556
Slovenian	10	8	0.1778
Sorbian	10	6	0.1333
Ukrainian	10	2	0.0444

Conclusions

Here we merely displayed computed data in order to show the first image of the situation. It must be emphasized that everything that has been stated for the indicator lambda holds also for the modified lambda, both individually (individual texts) and as a whole, i.e. for the *SI*-values.

There is a number of problems that could/should be scrutinized in the future. Here we list only some of them:

(1) Does modified lambda or *SI* develop with time? A writer cannot create a new structure each time he writes, hence the hypothesis may be conjectured: the similarity of works increases with time. The testing should be performed on very productive writers. Unfortunately, our data seldom corroborate this hypothesis.

(2) Can the divergence of languages be studied using modified lambda or *SI*? The null hypothesis is: There is no change of modified lambda or *SI* with increasing geographic distance.

(3) Does areal distance influence the style of the authors? The respective hypothesis cannot easily be tested because only (at least) bilingual writers can be taken into account but it is not easy to obtain relevant texts.

(4) Is there a relationship between modified lambda and other properties of texts? This is rather a long-lasting problem. It can be managed only stepwise, restricted to one language and to one other property. The main aim is to set up a control cycle analogous to that by R. Köhler (2005) in which modified lambda is a property among many others.

(5) Simple examples of (4) are the relations of modified lambda to vocabulary richness, to entropy, to Gini's coefficient, to text sort, to writer's personality, to the morphological complexity of the given language.

(6) Is there a clear trend of evolution? Comparing Latin texts with texts in Roman languages or Old Church Slavic with modern Slavic languages could, perhaps, serve to setting up a substantiated hypothesis.

(7) Is modified lambda relevant for language typology? Unfortunately, the number of texts processed for this purpose would be very large. But using corpuses would

allow us to touch this problem, too. Of course, typology is possible only if other indicators already exist, hence this problem is a continuation of problem (4) above.

Solving any of these problems would create further hypotheses or questions.

Acknowledgements

We are very obliged to Emmerich Kelih who computed the rank-frequencies of word forms in all Slavic texts of the Russian novel *Kak zakaljajas` stal`* by N. Ostrovsky. Further, to Reinhard Köhler who computed the rank-frequencies of word forms in all Slovak texts, and to Radek Čech who computed the rank-frequencies of word forms in E. Ady's Hungarian poems and in Pushkin's and Lermontov's Russian poems.

References

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

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

Popescu, I.-I., Mačutek, J., Altmann, G. (2010). Word forms, style and typology. *Glottotheory 3(1), 89-96.*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Glottometrics 30, 2014, 45-72

Quantifying Joyce's *Finnegans Wake*

C. George Sandulescu, Monaco
Lidia Vianu, Bucharest
Ioan-Iovitz Popescu, Bucharest
Andrew Wilson, Lancaster
Róisín Knight, Lancaster
Gabriel Altmann, Lüdenscheid

Abstract. The aim of the article is to show that the quantitative indicators already applied to many texts are also useful for characterizing a special text containing many artificial components created by J. Joyce.

Keywords: James Joyce, Finnegans Wake, English, quantitative properties

1. Introduction

James Joyce (1882-1941) began his writing career in 1914, and ended it with the publication of *Finnegans Wake* in 1939, after he had worked for 17 years on his last book. Throughout his career, Joyce experimented with poetry, plays and prose and his writings were influenced by a variety of factors. These included, but were not limited to, the political instability of Ireland at the time, the Irish literary and cultural revival of the late 19th century, and the European shift towards a more experimental style of literature (Spinks, 2009: 1-14). Indeed, his contributions to this new experimentalism have led some literary critics to praise him very highly, for example describing him as "the greatest and most enigmatic literary figure of the twentieth century" (Spinks, 2009: 1).

Joyce achieved arguably the most formidable concentration of this experimentation with his book *Finnegans Wake.* Considering the lexis alone, the book mixes standard English lexical items with neologisms, portmanteaus and polyglot puns. Furthermore, many different languages are represented (see Christiani, 1966; O'Hehir, 1967). However there are also other aspects that can present difficulties for a reader; for example Joyce writes simultaneously on different narrative planes and draws upon private experiences. Due to its idioscyncrasy, when *Finnegans Wake* was first published, the response it received was largely bemused or un-favourable; however, it is now viewed by some as postmodern triumph (c.f. Levin, 1944: 124; MacCabe, 1979: 133). Despite this, it remains one of the most controversial literary texts of our times.

The large majority of previous literary criticism of *Finnegans Wake* has taken a qualit-ative approach and focused on specific stylistic aspects of the work (see Campbell and Robinson, 1947; Benstock 1969; DiBernard, 1980). Some works could be considered to have taken a slightly more quantitative approach, by systematically considering the text and attempting to capture the size of it. For example, Glasheen (1956) created a census of biographical information of the characters in *Finnegans Wake* and Hart (1962) created a primary index of the 63,924 words in the vocabulary, an alphabetical list of syllables in the compound words and also listed some 10,000 English words suggested by Joyce's puns and distortions. However such analyses are still heavily qualitative in their methodology. This paper, the first in a series of articles, will offer a new perspective to the study of *Finnegans*

Wake through taking a quantitative approach in order to consider the relationship between the author's creativity and language laws.

Whilst writing is a creative process, there is evidence to suggest it is constrained by language laws (see Zipf, 1935). These language laws can be seen as comparable to those in physics; however, whilst there are thousands of physicists trying to find laws in their field, there are a small number of linguists attempting to do the same for language laws. Fortunately, there are already several steps made by Köhler (2012) into the depth of syntax, and statistical evaluations from different domains (cf. Bybee, Hopper 2001, cf. also Janda 2013). In this study, our main aim is to examine whether, in a text of this sort, linguistic laws are strong enough to soften the exuberant self-organization in the vocabulary, to establish whether the usual mathematical models used to analyse texts are still valid.

2. Methodology

The Joycean texts and word frequencies used in the present article are provided by Sandulescu and Vianu (James Joyce: Finnegans Wake. Full Text. Contemporary Literature Press, posted on Internet at the addresses given in References).Most word frequency data in the present article were obtained with http://www.writewords.org.uk/word_count.asp, after removing apostrophes, hyphens, and accents from the text. We shall call these words "mechanical words".

To explore stratification (see sections 2.3 and 3.3) it was necessary to consider the proportion of standard English words in the text. Therefore, for episode one, "original words" were used and classified as "standard English" or "Joycean word". This classification was agreed, out of context, with the joint judgements of two native speakers with backgrounds in English linguistics.

Through this paper, we analyse some of the quantitative properties of *Finnegans Wake*, using methods that have been used in similar studies previously. Through this, we enable the reader to perform comparisons of these texts. Below, we give a theoretical description of the steps of our analysis. Please note, this is not intended to be an exhaustive analysis; it is a beginning of a complete quantitative description of Joyce's work.

2.1 Rank-frequency distribution

There are several laws that attempt to capture the regularities that seem to exist in the frequency structure of texts, by expressing the relationship between frequency and rank of words in a text. Zipf (1935) carried out a systematic investigation of several languages and found a stable relationship between rank and frequency, which he expressed through a power law function. Researchers have since built on Zipf's work (see Popescu, Altmann and Köhler, 2010), attempting to explain it further and find an equation that better expresses the relationship. It is now common practice for the rank-frequency distribution of a text to be modeled by the Zipf-Mandelbrot distribution, which is a normalized extended Zipf-distribution (cf. Wimmer, Altmann 1999a: 666). We will therefore use this to present the rank-frequency distributions of words in the 17 episodes of *Finnegans Wake*.

2.2 The Lambda indicator

The Lambda indicator is derived from the sum of Euclidean distances between the neighboring frequencies of the rank-frequency distribution, i.e. as

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

where L is the arc length of the word frequency distribution, V is the vocabulary (= highest rank) and f_r are the individual frequencies. Since this indicator increases with increasing text size N, it can be standardized by taking the ratio

(2) $$\Lambda = \frac{L}{N} Log_{10}(N)$$

yielding a relatively stable value independent of N.

Unfortunately, the variance of the Euclidian distance is a very lengthy expression containing the covariances, and it requires complex computing especially for text comparisons (cf. Popescu, Mačutek, Altmann 2010). In order to alleviate the use of Lambda, one found a simple approximation which minimally deviates from the Euclidean arc length and called it *simplified arc length* (Popescu, Altmann 2014)

(3) $L^* = V + f_1 - (h + 1)$

where h is the currently used h-point defined as

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

This point can be found and computed easily. Hence the standard *simplified Lambda* is defined as

(5) $$\Lambda^* = \frac{L^*}{N} Log_{10}(N) = \frac{(V + f_1 - (h+1))Log_{10}(N)}{N}.$$

Since in (5) the only variable is f_1 (V is given for the text and h is a fixed point), the variance of the simplified Lambda can easily be derived by expansion as

(6) $$Var(\Lambda^*) = \frac{f_1(N - f_1)(Log_{10}N)^2}{N^3}$$

For comparing two texts, one can use the asymptotic normal test defined as

(7) $$u = \frac{|\Lambda_1^* - \Lambda_2^*|}{\sqrt{Var(\Lambda_1^*) + Var(\Lambda_2^*)}}$$

The formulas are sufficient for characterizing the vocabulary richness in individual episodes of *Finnegans Wake*, identifying stylistic change within a text and performing comparisons between different texts. Needless to say, a work like the studied one does not arise spontaneously, so to say, in one go, but is steadily corrected, improved, parts are added or omitted, etc. Thus we obtain merely only a *grosso modo* image of the development, nevertheless, the whole is a true image of the vocabulary.

2.3 Stratification

Texts, partly due to characteristics of individual languages and partly due to language variability, are composed of a number of components. It is possible to confirm the existence of this stratification in a text through calculating the number of strata present at the word form level. Usually, this is done using the stratification formula (cf. Popescu, Altmann, Köhler 2010) defined as

$$(8) \quad y = 1 + A_1\exp\left(-x/r_1\right) + A_2\exp\left(-x/r_2\right) + ...$$

in which the number of exponential components signals the number of strata. If two coefficients are equal, or if a coefficient presents a nonsense number, or if the determination coefficient R^2 attains a value greater than 0.9, the last component may be eliminated as redundant.

However, the stating of the number of strata does not mean the recognition and identification of strata, but merely their existence and number (Knight 2013, p.36). However we will still carry out this analysis with *Finnegans Wake* as, firstly, the findings can still be compared with previous attempts and, secondly, the more texts that are analysed in this way, the more likely it is that we will be able to recognise and identify specific strata.

2.4 Ord's criterion

The aim of Ord's criterion (cf. Ord 1972) is to show that there is a unique structure if the values lie in a certain domain. The criterion has the form

$$(9) \quad I = \frac{m_2}{m_1'}, \qquad S = \frac{m_3}{m_2},$$

where m'_1 is the mean and m_r are the central moments of r-th order.

2.5 Pearson's excess

Pearson's excess is used as the indicator of excess of the distribution. Using simply

$$(10) \quad \beta_2 = \frac{m_4}{m_2^2},$$

without -3 which compares it with the normal distribution (cf. Kapur, Saxena 1970: 38).

2.6 Entropy and Repeat Rate

There are many definitions of entropy (cf. Esteban, Morales 1995). In our analysis, we use the best known measure, proposed by C. Shannon and applied currently in linguistics to show the diversity/uncertainty and the concentration of the distribution. This is defined as

(11) $$H = -\sum_{i=1}^{V} p_i \log_2 p_i$$

Here $p_i = f_i/N$, i.e. the relative frequencies of each word in the text. The variance of entropy can be obtained by expansion as

(12) $$Var(H) = \frac{1}{N}\left(\sum_{i=1}^{V} p_i \log_2^2 p_i - H^2\right)$$

It is possible to also use the natural logarithm. The entropy can be relativized dividing the value of H by its maximum which is simply $H_0 = log_2,V$, hence

(13) $H_{rel} = H/H_0$

and its variance is

(14) $$Var(H_{rel}) = \frac{Var(H)}{(\log_2 V)^2}.$$

Now, the greater is the diversity, the greater is vocabulary richness.

The Repeat Rate says asymptotically the same as the Entropy, but it is interpreted in reverse sense. If all frequencies are concentrated to one word, then the text is maximally concentrated. The smallest concentration is given if all words have the same frequency. The Repeat Rate is defined as

(15) $$RR = \sum_{i=1}^{V} p_i^2 = \frac{1}{N^2}\sum_{i=1}^{V} f_i^2 \quad .$$

The maximum is 1, the minimum is 1/V, the relative Repeat Rate is

(16) $$RR_{rel} = \frac{1-RR}{1-1/V},$$

and the variance is

(17) $$Var(RR) = \frac{4}{N}\left(\sum_{i=1}^{V} p_i^3 - RR^2\right).$$

2.7 Writer's view

Other aspects of this methodology section have highlighted that authors shape their texts both consciously and sub-consciously. Some aspects of the writing process are subconscious because they take their course according to laws (not rules). Laws cannot be learned but they can be captured conceptually. One of such laws is the abiding by the "golden section" which can be defined as

(18) $$\varphi = \frac{1+\sqrt{5}}{2} = 1.6180...$$

and in frequency analysis of texts it is represented by the so-called "writer's view" (cf. Popescu, Altmann 2007). One can imagine the writer sitting at a fixed point of the rank-frequency distribution and looking at the same time at the most frequent word (f_1) and at his vocabulary (V), i.e. the last word of the distribution. That means, his view encompasses an angle between his position - let us call it $P(h,h)$ - and the extreme points $P(1,f_1)$ and $P(V,1)$. The situation is visualized in Figure 1.

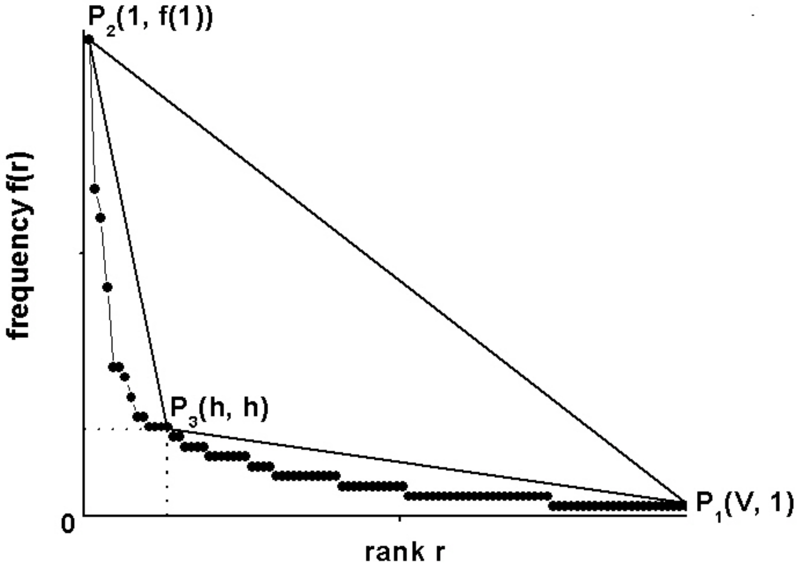

Figure 1. The writer's view angle ($P_2P_3P_1$)

The fixed point is defined as that point at which the rank and the frequency of that rank are equal. It is called h-point (cf. Popescu 2001). If there is no such point, it can be obtained by interpolation as shown in (6).

The cosine of the angle of the h-point can be computed classically as

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

and the radian of this angle is given as α *rad* = *arcos*(*cos* α). And this is exactly the value we call writer's view.

2.8 Vocabulary richness

In section 2.2, we outlined how we intend to analyse *Finnegans Wake* using the Lambda indicator. This will give us an indication of the vocabulary richness of the novel; however we wish to also use other methods to analyse this in more depth.

The number of indicators characterizing vocabulary richness is enormous. The concept itself can be interpreted in different ways, as can be seen in the history of its application (cf. e.g. Wimmer, Altmann 1999). Vocabulary richness may be considered as a function of any of the following: the number of different lemmas in text; the number of hapax legomena and the number of different tokemes (word form types). Alternatively, it is possible to study its evolution in text and perform several transformations. Regardless, text size N is always involved and this circumstance caused problems in the developing of indicators of richness (cf. Wimmer, Altmann 1999).

Popescu and Altmann (2006) introduced Gini's coefficient as a method of measuring vocabulary richness, as it takes into account all frequencies. However, frequencies play different roles. Fortunately, it is not necessary to revert and cumulate the distribution and the compute the sum of trapezoids to obtain the area above the Lorenz curve. Instead, one simply computes

$$(20) \qquad G = \frac{1}{V}\left(V + 1 - \frac{2}{N}\sum_{r=1}^{V} r f_r\right)$$

where V is the vocabulary (= highest rank), N is the text size, r is the rank and f_f the frequency of rank r. The authors defined a richness indicator as the complement to G, i.e.

$$(21) \qquad R_4 = 1 - G.$$

Since in (20) there are some constants (V and 2) and the mean, it is easy to define the variance as

$$(22) \quad Var(G) = Var(R_4) = \frac{4\sigma^2}{V^2 N}$$

where σ^2 is the variance of the distribution.

A quite different approach to vocabulary richness is considering the h-point. Words with ranks smaller than h are mostly auxiliaries, synsemantics and those (thematic) words which occur quite frequently but do not contribute to the richness. Richness is produced rather by words that seldom occur; in the history of this research one separated hapax legomena and considered them as unique indicators of richness. This is, of course, a slightly restricted view. But one can add also dis legomena or even tris legomena, but which of the approaches leads to "better" results? Where is the boundary?

Popescu et al. (2009: 29ff.) took into account the fixed point h and considered all words whose frequency is smaller than h (that is, the tail of the distribution) as contributors to richness. In order to obtain a comparable indicator we first define the cumulative probabilities up to h as

$$(23) \qquad F([h]) = F(r \le h) = \frac{1}{N}\sum_{r=1}^{[h]} f_r$$

That is, $F([h])$ is the sum of relative frequencies of words whose ranks are smaller or equal to h. A slight correction to $F([h])$ is the subtraction of the quantity $h^2/(2N)$, the half of the square of the h-point (cf. Popescu et al. 2009: 17). Using these conditions, one can define the indicator

(24) $$R_1 = 1 - \left(F([h]) - \frac{h^2}{2N} \right)$$

Since in (24) the only variable is $F([h])$ which can be considered a probability, one easily obtains the variance of R_1 as

(25) $$Var(R_1) = F([h])[1 - F([h])]/N.$$

This study will consider both of these approaches to vocabulary richness.

3. Results and analysis

3.1 Rank-frequency distribution

Unfortunately, the results of fitting the Zipf-Mandelbrot distribution are not satisfactory statistically. This may be due to some boundary conditions that have not been taken into account, and to the fact that the chi-square fitting has different weak points. However, considering the resulting formula as a simple function, we obtain a good result yielding R^2 = 0.9964.

Alternatively, it is possible to perform the fitting by means of a function known as Zipf-Alekseev function. One can obtain it from the differential equation

(26) $$\frac{dy}{y} = \frac{A + B \ln x}{Dx} dx$$

Which, when solved and reparametrized, yields the function

(27) $$y = cx^{a + b \ln x}.$$

In (26), *A* is the language/text-sort/style/,… constant, *B* is the force of the speaker/ writer and *D* is the equilibrating force of the community (cf. Wimmer, Altmann 2005). The check of sufficiency can be done again with the determination coefficient R^2.

Applying (27) to all episodes separately, we obtain the results presented in Table 1.

Table 1
Zipf-Alekseev Fitting (mechanical words)

Text	***a***	***b***	***c***	R^2
FW Episode 01	-0.6487	-0.0605	657.9873	0.9939
FW Episode 02	-0.5609	-0.0878	385.0283	0.9841
FW Episode 03	-0.5791	-0.0711	577.5572	0.9886
FW Episode 04	-0.6179	-0.0685	671.2932	0.9905
FW Episode 05	-0.6424	-0.0524	499.2077	0.9906
FW Episode 06	-0.4927	-0.0879	909.1371	0.9945
FW Episode 07	-0.5171	-0.0862	543.3030	0.9880

FW Episode 08	-0.3843	-0.1132	438.6174	0.9880
FW Episode 09	-0.4304	-0.0976	710.6777	0.9903
FW Episode 10	-0.5039	-0.0851	801.7924	0.9918
FW Episode 11	-0.6105	-0.0716	1674.9200	0.9945
FW Episode 12	-0.6983	-0.0575	487.0949	0.9595
FW Episode 13	-0.4000	-0.1034	490.0503	0.9876
FW Episode 14	-0.4322	-0.0902	902.7356	0.9959
FW Episode 15	-0.3987	-0.1032	1317.1361	0.9905
FW Episode 16	-0.4376	-0.0851	595.9386	0.9895
FW Episode 17	-0.5676	-0.0594	696.8380	0.9912

As can be seen, the parameters a and b are smaller than 0, and parameter b linearly depends on parameter a, namely $b = – 0.1683 – 0.1659a$ with $R^2 = 0.85$. This shows that even in a non-standard text such as *Finnegans Wake*, the background law is followed subconsciously by the writer. It may be possible to insert the parameter a and its relation to parameter b in a more general theory encompassing language levels. However, it must be further scrutinized whether the negative values of a are characteristic only to the given text or are a general feature of rank-frequency distributions of words. Since this is possible only with a great number of other texts, we must, for now, renounce this task.

The results show that, in the example of this unusual text, the Zipf-Alekseev function yields a better fit than Zipf-Mandelbrot. The text, due to its use of non-standard words, has a large number of hapax legomena (words that occur only one time). The result suggests that modeling a rank-frequency distribution, especially in cases having very long tail, may be done more adequately with a simple function.

3.2 The Lambda indicator

In Table 2, the computed values are presented.

Table 2
Simplified Lambdas for individual episodes of *Finnegans Wake* (mechanical words)
(Note: the difference between the actual Λ and the simplified Λ^* is a few per-mille)

Text	***N***	***V***	***f*(1)**	***h***	***L****	***Λ****	**Var (*Λ**)**
FW Episode 01	9850	4107	642	32.0000	4716.0000	1.9120	0.00009865
FW Episode 02	6025	2798	375	24.0000	3148.0000	1.9750	0.00013841
FW Episode 03	9830	4363	580	32.5000	4909.5000	1.9940	0.00009003
FW Episode 04	10389	4443	659	31.0000	5070.0000	1.9602	0.00009225
FW Episode 05	8150	3419	491	28.6000	3880.4000	1.8622	0.00010627
FW Episode 06	16137	6243	898	42.0000	7098.0000	1.8508	0.00005766
FW Episode 07	9524	4153	535	29.8571	4657.1429	1.9456	0.00008813
FW Episode 08	8044	3477	419	28.5000	3866.5000	1.8772	0.00009362
FW Episode 09	14348	6166	692	39.6667	6817.3333	1.9751	0.00005528
FW Episode 10	15309	6619	777	41.2500	7353.7500	2.0103	0.00005512
FW Episode 11	25952	9986	1672	51.0000	11606.0000	1.9741	0.00004526
FW Episode 12	6176	2402	452	27.5000	2825.5000	1.7342	0.00015782

FW Episode 13	9551	3961	474	33.8000	4400.2000	1.8336	0.00007823
FW Episode 14	17658	6237	898	44.2500	7089.7500	1.7052	0.00004930
FW Episode 15	26921	9986	1262	52.0000	11195.0000	1.8422	0.00003257
FW Episode 16	12870	5307	577	39.5000	5843.5000	1.8659	0.00005619
FW Episode 17	12994	5271	709	39.0000	5940.0000	1.8805	0.00006718

For the sake of illustration we show the computation for Episode 1 and compare it with Episode 2. We obtain

$$\Lambda^*_{E1} = \frac{[4107 + 642 - (32.00 + 1)]\log_{10}(9850)}{9850} = 1.9120,$$

and

$$u = \frac{|1.9120 - 1.9759|}{\sqrt{0.00009865 + 0.0001381}} = 4.15,$$

a highly significant value, which suggests there is a stylistic difference between the two episodes. This could be the effect of multiple factors, for example a long pause in writing.

Comparing all episodes with one another, we obtain the results presented in Table 3 below. Instead of presenting all numbers, we mark (**X**) those pairs of texts whose u is smaller than 1.96, as this indicates that there is no significant difference of Lambdas and that the texts share similarity.

Table 3

Similarities of simplified Lambdas in 17 episodes of *Finnegans Wake*

Episode	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1																	
2																	
3		X															
4		X															
5																	
6					X												
7		X		X													
8					X												
9		X	X	X													
10			X														
11		X	X	X					X								
12																	
13						X											
14																	
15					X	X							X				
16					X	X		X									
17					X			X								X	

Table 4 expresses this information in a different form, highlighting, for each episode, the number of other episodes it shares similarity with.

Table 4
Number of Lambda-similarities found for each episode of *Finnegans Wake*

Episode	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Number of similarities	0	5	4	4	5	4	2	3	4	1	4	0	2	0	3	4	3

The centrality (the stylistic gravitation of an episode) is the greater the more episodes are similar to it. Hence the sets of episodes according to decreasing centrality are

{2,5}, {3,4,6,9,11,16}, {8,15,17}, {7,13}, {10}, {1,12,14}.

It is clear that the episodes with the greatest centrality are 2 and 5, whereas the most divergent are episodes 1, 12 and 14. These results provide a new insight into the stylistic patterns found within *Finnegans Wake* and offer increased focus for a future qualitative study of the text.

Tables 5 and 6 show the mean and maximum lambdas calculated in previous studies for a range of text types.

Table 5
Mean lambdas of the rank-frequency distributions of some English writers
(taken from Popescu, Čech, Altmann 2011, Appendix, pp. 120 – 127)

Text sort	**# texts**	**mean Λ**
Table 6a: English poetry	18	1.4450
Table 6b: English prose	56	1.2922
Table 6c: English Nobel lectures	21	1.3079
Table 6d: English scientific texts	10	1.0528
Table 6e. English stories told by children	39	1.2651

Table 6
Maximal Lambdas in some works by English writers
(taken from Popescu, Čech, Altmann 2011, Appendix, pp. 120 – 127)

Text sort	**Genre**	**Text containing maximum Λ**	**Text author**	**maximum Λ**
Table 6a	Poetry	Howl (1956)	Ginsberg, A.	1.7905
Table 6b	Prose	Rosinante to the road again. XIV	Dos Passos, J	1.7679
Table 6c	Nobel	Literature (banquet speech) (1953)	Churchill, W.	1.6126
Table 6d	Science	Rorty's Inspirational Liberalism (2003)	Bernstein, R.J.	1.2412
Table 6e	Children	The Rift	Toni, boy, 11 years	1.5024

If we consider the maximum Lambdas for other texts, we see that the values seem to differ for different genres. Poetry has the highest value, followed by prose. Nobel and science have lower values. It seems reasonable to question whether the more a text deviates from realism in its content and the stronger is its creative component the greater its Lambda is. Our analysis of *Finnegans Wake* seems to fit with this hypothesis. Due to its play with words it is arguably the most creative text so far analyzed, and it has the highest scoring mean of Λ* **(**1.8940) and highest scoring maximum of Λ* (2.0103). Of course, a number of different texts in different languages would be necessary to test this further. The interested reader can perform further analyses concerning languages, text sorts, styles, development, etc. in order to obtain an overall image of this indicator (cf. Popescu, Čech, Altmann 2011).

Finally, Table 2 and Table 7 allow a comparison between Joyce's novels *Finnegans Wake* (1939) and *Ulysses* (1922), the latter written in standard English. The difference is enormous when one compares the Λ* columns, the corresponding lambda averages being 1.8940 for *Finnegans Wake* versus 1,3671 for *Ulysses*.

Table 7
Simplified Lambdas for individual episodes of *Ulysses* (mechanical words)
(Note: the difference between the actual *Λ* and the simplified *Λ** is small per-mille)

Text	***N***	***V***	***f*(1)**	***h***	***L****	***Λ****	**Var (*Λ**)**
Ulysses Episode 01	7189	2043	399	30.3333	2410.6667	1.2932	0.00010846
Ulysses Episode 02	4394	1508	265	24.0000	1748.0000	1.4492	0.00017116
Ulysses Episode 03	5697	2320	284	25.0000	2578.0000	1.6995	0.00011727
Ulysses Episode 04	5874	2026	395	25.4000	2394.6000	1.5364	0.00015168
Ulysses Episode 05	6390	2026	353	27.7500	2350.2500	1.3997	0.00011828
Ulysses Episode 06	10903	2817	630	37.5000	3408.5000	1.2622	0.00008140
Ulysses Episode 07	10151	2840	638	34.0000	3443.0000	1.3589	0.00009314
Ulysses Episode 08	12903	3529	565	40.5000	4052.5000	1.2911	0.00005483
Ulysses Episode 09	11968	3491	626	39.0000	4077.0000	1.3892	0.00006888
Ulysses Episode 10	12442	3429	626	36.0000	4018.0000	1.3224	0.00006440
Ulysses Episode 11	12153	3205	432	38.0000	3598.0000	1.2093	0.00004707
Ulysses Episode 12	21274	5660	1608	49.5000	7217.5000	1.4683	0.00006152
Ulysses Episode 13	16755	3571	811	48.4000	4332.6000	1.0923	0.00004905

In order to state the significance of the difference we compute the asymptotic normal test between the means of the two simplified lambdas in the two tests according to

$$u = \frac{\bar{\Lambda}_1 - \bar{\Lambda}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

and obtain

$$u = \frac{1.8940 - 1.3671}{\sqrt{\frac{0.00763}{17} + \frac{0.02353}{13}}} = 11.0863$$

which is highly significant. Hence, *Finnegans Wake* strongly differs from a "normal" text.

3.3 Stratification

The results of the computation of strata in *Finnegans Wake* are presented in Table 8.

Table 8
The two-strata structure of rank-frequency distributions of words in all episodes (mechanical words)

Text	N	A_1	r_1	A_2	r_2	R^2
FW Episode 01	9850	800.5245	2.4216	105.2927	31.3232	0.9956
FW Episode 02	6025	438.3131	2.9998	51.4478	33.2732	0.9910
FW Episode 03	9830	620.5005	3.0397	90.4213	33.5005	0.9848
FW Episode 04	10389	800.7309	2.3973	122.1213	27.2785	0.9906
FW Episode 05	8150	566.8180	2.8675	67.8039	39.9975	0.9897
FW Episode 06	16137	975.8178	3.0202	169.7285	32.5279	0.9920
FW Episode 07	9524	589.3728	3.2088	82.4540	35.7731	0.9900
FW Episode 08	8044	457.4715	3.1030	99.9073	28.2512	0.9911
FW Episode 09	14348	741.8399	3.3278	134.5352	35.0325	0.9917
FW Episode 10	15309	889.3433	2.9443	142.0732	34.7241	0.9951
FW Episode 11	25952	1973.5895	2.4524	297.9667	29.1142	0.9894
FW Episode 12	6176	664.7541	2.1508	67.7475	31.9517	0.9774
FW Episode 13	9551	503.3348	3.2776	105.1176	31.3593	0.9895
FW Episode 14	17658	903.1733	3.1081	211.5357	30.9411	0.9888
FW Episode 15	26921	1380.8318	3.1462	287.4493	32.8846	0.9900
FW Episode 16	12870	619.4422	3.2579	120.6342	37.8397	0.9931
FW Episode 17	12994	772.4971	2.4798	152.8376	31.4530	0.9846

As can be seen, the second coefficient r_2 is always greater than r_1, signaling the weak expression of the second stratum. The fitting is very adequate in all cases. Hence we can conjecture that there are two word strata in all texts.

To explore this further, we shall consider strata of original words (as defined in section 2). If we consider separately the frequencies of English words (eliminating all the others), we obtain again a two strata relation

$$y = 1 + 803.6911\exp(-x/2.4385) + 102.3272\exp(-x/30.6489)$$

with R^2 = 0.9960. Since the parameters are quite different, we have again two strata and may continue the procedure. But here, there are as many possibilities as we are able to define. Separating autosemantics and synsemantics would not finish the work. From the linguistic point of view, this would be a fertile way into the depth but from the textological view its relevance is not yet known.

Consider the non-English words, such as the most frequent ones: *willingdone, jinnies, lipoleums, prankquean, hoother,…* it is not easy to find a linguistic or textological criterion which would enable us to perform a classification. If we fit the stratification formula to this data, we obtain again two strata

$$y = 1 + 36.2053\exp(-x/1.6548) + 3.4349(-x/39.7718)$$

with R^2 = 0.9783. Even a tri-stratal function yields non-equal parameters. Therefore much philological work would still be necessary to find the exact nature of the strata.

Since the difference of parameters may be caused also by the different size of data, we compute the lambda indicator for both and compare them. We obtain the results presented in Table 9.

Table 9
Simplified lambda for the three variants of Episode 1
(words separated by blanks)

All words (standard English and invented)						
N	***V***	***f*(1)**	***h***	***L****	***Λ****	**Var (*Λ**)**
9767	4146	642	31.6667	4755.3333	1.9425	0.00010009
Standard English words						
N	***V***	***f*(1)**	***h***	***L****	***Λ****	**Var (*Λ**)**
7562	2116	642	31.6667	2725.3333	1.3979	0.00015456
Joyce's invented words						
N	***V***	***f*(1)**	***h***	***L****	***Λ****	**Var (*Λ**)**
2205	2030	25	6.0000	2048.0000	3.1054	0.00005683

One can see that the frequency distribution of Joyce`s invented words has a much greater simplified lambda than the one of standard English words only. Performing the asymptotic normal test between the latter two distributions, we obtain

$$u = |1.3979 - 3.1054|/[0.00015456 + 0.00005683]^{1/2} = 117.44.$$

an extremely significant value whose probability is very small.

The above example supports the findings of section 3.2, suggesting that lambda can be drastically increased by enriching the vocabulary with enough *x* unique words (actual or invented). The general formula results directly from the definition (5), namely

$$\Lambda^*(x) = \frac{L^*+x}{N+x} Log_{10}(N+x) \tag{28}$$

To explore this further, we will draw on the example of the poem *Jabberwocky* by Lewis Carroll. Like *Finnegans Wake*, this text contains many words originally made up by the author. We used the values of *N* and *L**, given below in Table 10.

Table 10
Lambda for *Jabberwocky*

Lewis Carroll, Jabberwocky (1871)						
N	***V***	***f*(1)**	***h***	***L****	***Λ****	**Var (*Λ**)**
168	92	19	4.5000	105.5000	1.3974	0.00295660

We get

$$\Lambda*(x) = \frac{105{,}5 + x}{168 + x} Log_{10}(168 + x)$$

in terms of *x* additional unique words as shown in Figure 2.

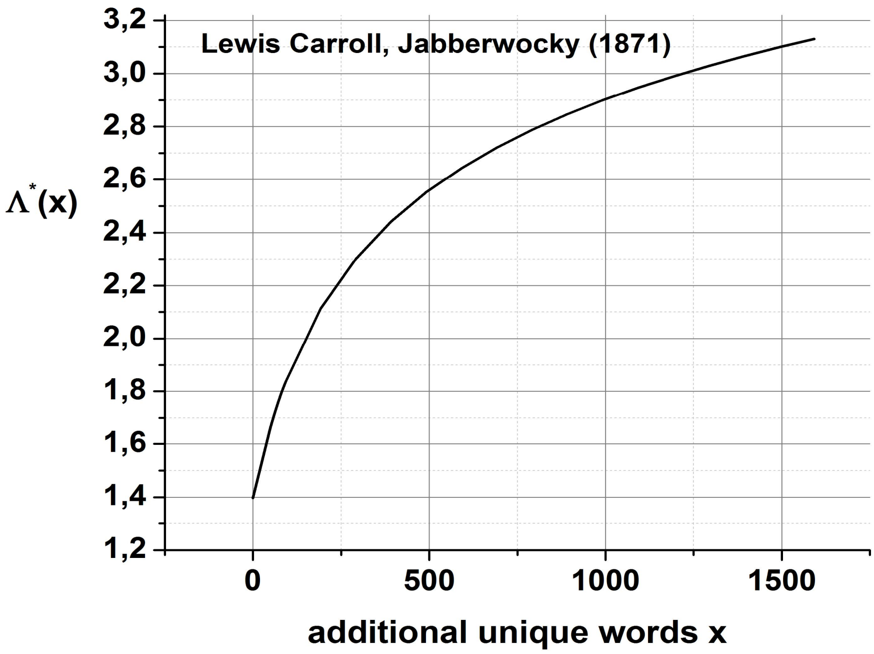

Figure 2. Lambda amplification by additional unique words

As it can be seen, a middle lambda text of about Λ* = 1.4 can be increased to a lambda of about 3.1 by inserting about 1500 new unique words (hapax legomena). However, this freedom is given only to the text author, not to the researcher who must adhere to the state of affairs.

3.4 Ord's criterion

In Table 11 the values of Ord's criterion for each individual episode of *Finnegans Wake* are shown.

Table 11
Ord's criterion for individual episodes of
Finnegans Wake (mechanical words)

Episode	*N*	*V*	m_1'	m_2	m_3	*I*	*S*
1	9850	4107	18.3284	1403	142294	76.5266	101.4493
2	6025	2798	17.8944	1445	152210	80.7499	105.3374
3	9830	4363	17.4356	1358	140841	77.9017	103.6918
4	10389	4443	17.6995	1365	139515	77.1060	102.2289
5	8150	3419	20.1931	1586	158093	78.5312	99.6933
6	16137	6243	18.5976	1417	143401	76.1719	101.2280
7	9524	4153	18.4012	1450	148444	78.7856	102.3927
8	8044	3477	18.3802	1348	134480	73.3131	99.7993
9	14348	6166	17.6029	1334	135979	75.8000	101.9106
10	15309	6619	16.9289	1282	130904	75.7198	102.1209
11	26642	10676	16.0859	1193	121971	74.1423	102.2692
12	6176	2402	20.3339	1580	159757	77.6954	101.1219
13	9551	3961	18.9060	1429	144060	75.5798	100.8178
14	17658	6237	20.1035	1515	149985	75.3757	98.9796
15	27373	10438	17.6353	1320	133823	74.8546	101.3749
16	12870	5307	18.8625	1411	140567	74.7842	99.6493
17	12994	5271	19.7454	1482	145483	75.0404	98.1860

The relationship between *I* and *S* is visualized in Figure 3.

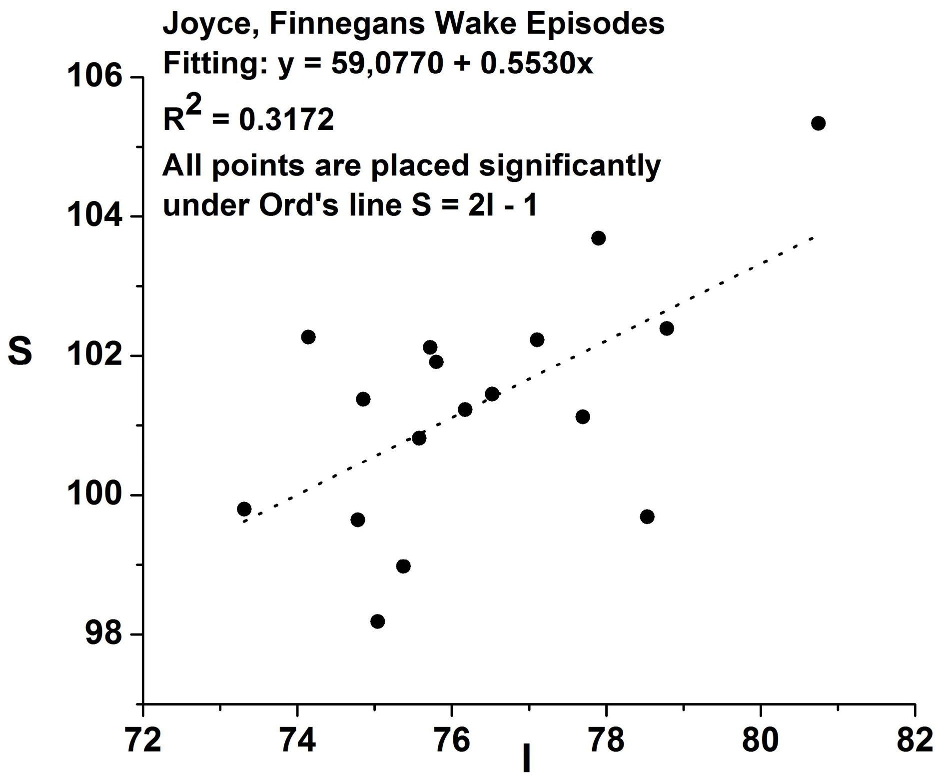

Figure 3. Ord's criterion <I,S> for the individual episodes

Ord's criterion displays a certain tendency but this tendency cannot be captured by a straight line. As can be seen in Figure 3, a very weak tendency exists.

The aim of Ord's criterion is to show that there is a unique structure if the values lie in a certain domain. The separator of the domains is the line $I = 2S - 1$, separating the negative hypergeometric domain under the line from several other ones. Since the <I,S> points are under the line, it would be interesting to substantiate linguistically its position. This is surely a task for the future; if one joined the neighboring points, one would obtain a strong oscillation which could be captured merely using some polynomials.

The aim of any indicator in text analysis is to identify some property of the given text, show its location in the two dimensional space, find its links to other indicators and show the inner mechanism controlling the self-regulation. Here, we must dispense with this aim because we analyze only one text.

3.5 Pearson's excess

We obtained the results presented in Table 12.

Table 12
Pearson's excess

Episode	***N***	***V***	m_2	m_4	β_2
1	9850	4107	1403	19979792	10.1558
2	6025	2798	1445	21787444	10.4348
3	9830	4363	1358	19952925	10.8153
4	10389	4443	1365	19586475	10.5162
5	8150	3419	1586	22281134	8.8602
6	16137	6243	1417	20189479	10.0606
7	9524	4153	1450	20913274	9.9503
8	8044	3477	1348	18761611	10.3326
9	14348	6166	1334	19122332	10.7408
10	15309	6619	1282	18367567	11.1783
11	26642	10676	1193	17101986	12.0233
12	6176	2402	1580	22624458	9.0646
13	9551	3961	1429	20271044	9.9281
14	17658	6237	1515	21035004	9.1608
15	27373	10438	1320	18773335	10.7731
16	12870	5307	1411	19705541	9.9030
17	12994	5271	1482	20287021	9.2405

As can be seen, β_2 is almost constant. It does not bring any possibility of classification or modeling a development trend. A thorough comparison with other texts would show whether this property is constant also for "normal" texts.

3.6 Entropy and Repeat Rate

All values necessary for evaluation and comparison of Entropy and Repeat Rate for all individual episodes of *Finnegans Wake* are presented in Table 13 below.

Table 13
Entropy and Repeat Rate of individual episodes of *Finnegans Wake*

Text	*N*	*V*	*H*	Var(*H*)	*RR*	Var(*RR*)
FW Episode 01	9850	4107	9.7437	0.001166	0.010183	1.362E-07
FW Episode 02	6025	2798	9.5711	0.001619	0.009937	2.077E-07
FW Episode 03	9830	4363	9.9722	0.001123	0.008632	1.005E-07
FW Episode 04	10389	4443	9.8648	0.001124	0.009796	1.206E-07
FW Episode 05	8150	3419	9.7025	0.001236	0.008983	1.302E-07
FW Episode 06	16137	6243	10.0712	0.000793	0.008725	5.710E-08
FW Episode 07	9524	4153	9.9052	0.001138	0.008628	9.940E-08
FW Episode 08	8044	3477	9.5949	0.001324	0.009236	1.152E-07
FW Episode 09	14348	6166	10.2781	0.000837	0.007399	4.790E-08
FW Episode 10	15309	6619	10.3844	0.000801	0.007482	4.930E-08
FW Episode 11	26642	10676	10.5383	0.000585	0.009250	4.380E-08
FW Episode 12	6176	2402	9.0835	0.001678	0.013649	3.645E-07
FW Episode 13	9551	3961	9.7812	0.001114	0.008287	8.140E-08
FW Episode 14	17658	6237	9.9978	0.000706	0.008113	4.180E-08
FW Episode 15	27373	10438	10.5862	0.000526	0.007297	2.410E-08
FW Episode 16	12870	5307	10.1697	0.000851	0.006801	4.430E-08
FW Episode 17	12994	5271	10.0400	0.000882	0.007762	6.000E-08

As can be seen in Table 13, the richness of all episodes is relatively stable. That means, Entropy and Repeat Rate are effects of some laws working in the background; the writer abides by them unconsciously and creates them in spite of his originality. Though, in theory, there is a clear relationship between Entropy and Repeat Rate (cf. e.g. Altmann 1988: 45), in practice we obtain at least a power relationship as visualized in Figure 4.

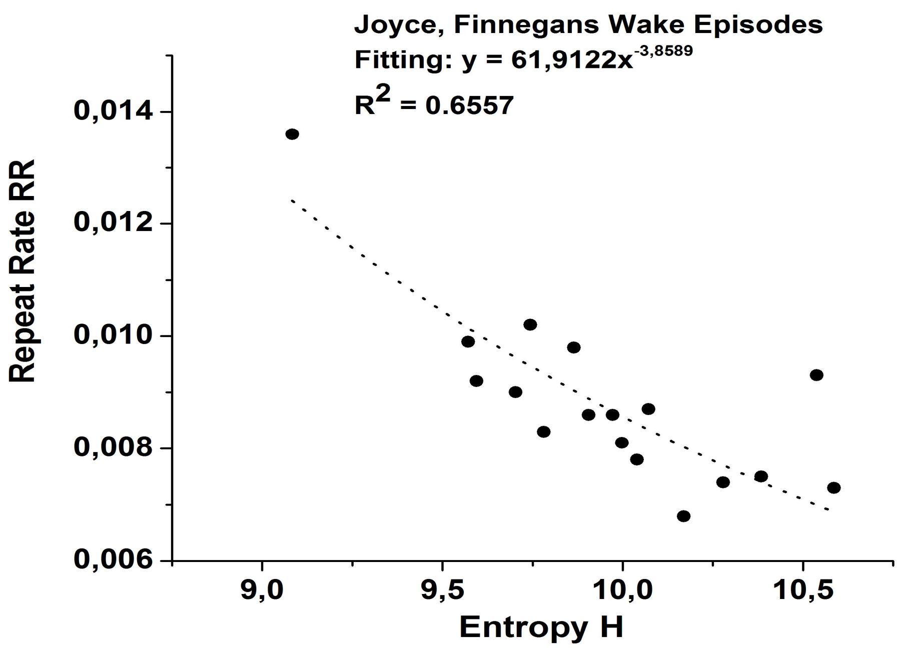

Figure 4. Entropy and Repeat Rate for *Finnegans Wake* episodes

This analysis will allow the mean Entropies or Repeat Rates of other works to be compared with *Finnegans Wake* using the variances, enabling new insights into these texts.

3.7 Writer's view

The computation of this value for the individual episodes of *Finnegans Wake* yielded values presented in Table 14.

Table 14
Writer's view of individual episodes of Finnegans Wake

Text	***N***	***V***	***f*(1)**	***h***	***cos α***	***α* rad**
FW Episode 01	9850	4107	642	32.0000	-0.0584	1.6292
FW Episode 02	6025	2798	375	24.0000	-0.0737	1.6445
FW Episode 03	9830	4363	580	32.5000	-0.0647	1.6355
FW Episode 04	10389	4443	659	31.0000	-0.0545	1.6253
FW Episode 05	8150	3419	491	28.6000	-0.0677	1.6386
FW Episode 06	16137	6243	898	42.0000	-0.0544	1.6253
FW Episode 07	9524	4153	535	29.8571	-0.0640	1.6349
FW Episode 08	8044	3477	419	28.5000	-0.0782	1.6491
FW Episode 09	14348	6166	692	39.6667	-0.0655	1.6363
FW Episode 10	15309	6619	777	41.2500	-0.0607	1.6316
FW Episode 11	25952	9986	1672	51.0000	-0.0359	1.6067
FW Episode 12	6176	2402	452	27.5000	-0.0734	1.6443
FW Episode 13	9551	3961	474	33.8000	-0.0826	1.6535
FW Episode 14	17658	6237	898	44.2500	-0.0576	1.6284
FW Episode 15	26921	9986	1262	52.0000	-0.0472	1.6181
FW Episode 16	12870	5307	577	39.5000	-0.0787	1.6496
FW Episode 17	12994	5271	709	39.0000	-0.0639	1.6347

Ordering the episodes according to increasing *N*, we obtain the course visualized in Figure 5.

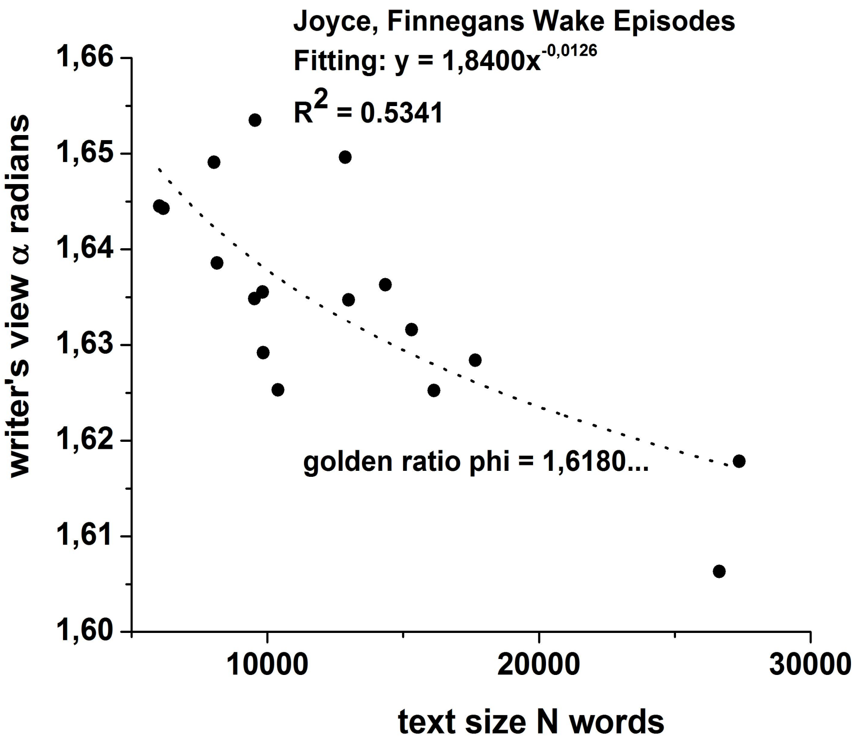

Figure 5. Writer's view for *Finnegans Wake* episodes

It has been shown in 20 languages and 176 texts that with increase of text size α *rad* converges to the value φ = 1.6180… that is, to the golden section (cf. Popescu, Altmann 2007). In all of the examined texts, α *rad* was situated in the neighborhood of this value. One cannot consider it a random event but rather a law concealed in some human senses and thinking.

The power function fitted to the data displays irregular oscillation but the direction is unmistakable. In the longest text (episode 15) α rad is almost identical with the golden section. Since the golden section exists also in other domains of human activity, it is not a purely linguistic phenomenon. Its origin should be sought somewhere in our evolution or in our physical and mental constitution. Nevertheless, comparisons of texts are possible because the parts of a text display different α rad, hence a textual whole has a mean and the individual parts have a spread which can be captured e.g. by the variance. The theoretical golden section is a constant having no spread.

When comparing *Finnegans Wake* with other texts, we may consider *Finnegans Wake* as expected values and use them for comparison in an asymptotic normal test. The mean "writer's view" of *Finnegans Wake* is $\overline{WW}$ (FW) = 1.6344 and the variance is Var (WW) = 0.00014401, hence Var ($\overline{WW}$) = 0.0001441/17 = 0.000008476. Comparing *Finnegans Wake* with *Ulysse*s, also by Joyce, we obtained α rad = 1.5880, we obtain *u* = 15.94 which is highly significant in spite of the small optical difference. However, Ulysses has been evaluated as a whole, not in parts.

3.8 Vocabulary richness

When considering vocabulary richness of each individual episode of *Finnegans Wake* using Gini's coefficient, we obtained the results presented in Table 15.

Table 15
Vocabulary richness of individual episodes
of *Finnegans Wake* using Gini's coefficient

Text	***N***	***V***	***G***	***R*₄**	**Var(*G*)**
FW Episode 01	9850	4107	0.5643	0.4357	0.000034
FW Episode 02	6025	2798	0.5153	0.4847	0.000055
FW Episode 03	9830	4363	0.5383	0.4617	0.000034
FW Episode 04	10389	4443	0.5546	0.4454	0.000032
FW Episode 05	8150	3419	0.5575	0.4425	0.000041
FW Episode 06	16137	6243	0.5940	0.4060	0.000021
FW Episode 07	9524	4153	0.5453	0.4547	0.000035
FW Episode 08	8044	3477	0.5522	0.4478	0.000041
FW Episode 09	14348	6166	0.5544	0.4456	0.000023
FW Episode 10	15309	6619	0.5504	0.4496	0.000022
FW Episode 11	26642	10676	0.5850	0.4150	0.000013
FW Episode 12	6176	2402	0.5841	0.4159	0.000054
FW Episode 13	9551	3961	0.5653	0.4347	0.000035
FW Episode 14	17658	6237	0.6240	0.3760	0.000019
FW Episode 15	27373	10438	0.6009	0.3991	0.000012
FW Episode 16	12870	5307	0.5666	0.4334	0.000026
FW Episode 17	12994	5271	0.5764	0.4236	0.000026

Though one may see the slow linear decrease of R_4 and the F-test yields a significant result, fitting a straight line to the number in column R_4 yields merely $R^2 = 0.36$ and ordering according to increasing *N* improves slightly the linear tendency.

Popescu et al. (2009) analyzed and evaluated 173 texts in 20 languages using the same method. In other English texts, all Nobel lectures, R_4 was in the interval of 0.2640 and 0.4605. The mean of the Nobel lectures was 0.3478. In comparison, the mean of *Finnegans Wake* is 0.4336. The difference seems to be quite great, but we shall not perform any further test here until it can be compared to a wider range of English texts.

Moving on, when we analyse vocabulary richness using formula (25) we achieve the results shown below in Table 16.

Table 16
Vocabulary richness in individual episodes of *Finnegans Wake*

Text	***N***	***V***	***h***	***F*([*h*])**	***R*₁**	**Var(*R*₁)**
FW Episode 01	9850	4107	32.0000	0.3709	0.6811	2.3689E-05
FW Episode 02	6025	2798	24.0000	0.3349	0.7129	3.6970E-05
FW Episode 03	9830	4363	32.5000	0.3517	0.7020	2.3195E-05
FW Episode 04	10389	4443	31.0000	0.3646	0.6817	2.2299E-05

FW Episode 05	8150	3419	28.6000	0.3401	0.7101	2.7538E-05
FW Episode 06	16137	6243	42.0000	0.3956	0.6591	1.4817E-05
FW Episode 07	9524	4153	29.8571	0.3464	0.7004	2.3772E-05
FW Episode 08	8044	3477	28.5000	0.3717	0.6788	2.9033E-05
FW Episode 09	14348	6166	39.6667	0.3671	0.6877	1.6193E-05
FW Episode 10	15309	6619	42.0000	0.3624	0.6952	1.5093E-05
FW Episode 11	25952	9986	51.0000	0.4054	0.6447	9.2883E-06
FW Episode 12	6176	2402	27.5000	0.3873	0.6739	3.8423E-05
FW Episode 13	9551	3961	33.8000	0.3729	0.6869	2.4484E-05
FW Episode 14	17658	6237	44.2500	0.4055	0.6499	1.3652E-05
FW Episode 15	26921	9986	52.0000	0.4004	0.6498	8.9179E-06
FW Episode 16	12870	5307	39.5000	0.3625	0.6981	1.7956E-05
FW Episode 17	12994	5271	39.0000	0.3773	0.6812	1.8081E-05

This method has previously been applied to 176 texts in 20 languages and yielded values for R_1 in the interval of 0.4308 and 0.9369 (cf. Popescu et al. 2009: Table 3.6). If we consider only the texts in English, they were in the interval of 0.6290 and 0.7545 with a mean of 0.6767. All of the episodes of *Finnegans Wake* are within the interval previously found for texts of English, yet have a little higher mean of 0,6829. This is to be expected since Joyce created many new words which were used only once, thus leading to a slight increase of the vocabulary richness R_1. This effect appears much more visible when the vocabulary richness is measured by lambda, as it results from the comparison of Table 2 for *Finnegans Wake* with Tables 5 and 6 for other English texts. Nevertheless, the almost infinite task to analyze all English texts remains an enterprise for the future.

Though the differences between R_1 of individual chapters are visually very small, it can be shown that some neighbouring episodes are significantly different. In Table 17 the R_1 of the neighbouring episodes are compared. The resulting value is the asymptotic u of the normal distribution.

Table 17
Normal tests for the differences of R_1 of the neighbouring episodes

Episodes	**u**
1-2	4.08
2-3	1.40
3-4	3.01
4-5	4.02
5-6	7.84
6-7	6.65
7-8	2.97
8-9	1.32
9-10	1.34

10-11	10.20
11-12	4.23
12-13	1.64
13-14	5.99
14-15	0.00986
15-16	9.31
16-17	2.81

All values greater than 1.96 signal a significant difference. As we saw in section 3.2, there is a significant different between episodes 1 and 2. However, if one draws a figure of R_1 for the episodes, one can observe a very strong oscillation, hence significant differences are not exceptional in this case.

If we compare all episodes with all other ones, we obtain a matrix displaying the similarities as shown in Table 18.

Table 18
Similarities of vocabulary richness as expressed by R_1

Id #	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**
1																	
2																	
3		X															
4	X																
5		X	X														
6																	
7		X	X		X												
8	X			X													
9	X			X				X									
10			X				X		X								
11																	
12	X			X				X	X								
13	X			X			X	X	X	X		X					
14						X					X						
15						X					X			X			
16			X		X		X		X	X			X				
17	X			X				X	X			X	X				

Table 19 expresses this information in a different form, highlighting, for each episode, the number of other episodes it shares similarity with.

Table 19
Number of similarities found for each episode of *Finnegans Wake*

Episode	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Number of similarities	6	3	5	6	4	2	6	6	8	5	2	6	9	3	3	6	6

As can be seen, there is quite a difference in the number of similarities shown by individual episodes. Episode 13 shares similarities with 9 other episodes, the highest scoring example, and is therefore the episode with the highest centrality in this instance. As can be seen, there is a great difference between the similarity in vocabulary richness computed in this way and using other indicators /cf. section 3.2).

A logical continuation of this study of centrality would be the comparison of concrete entities of Episode 13 with those of other ones. Unfortunately, the number of entities that could be compared is infinite and one would never know whether one found the pertinent ones.

The fact that R_1 and R_4 express the same property can be documented by their power relationship as visualized in Figure 6 below. It is worth noting that the Lorenz-curve is based on cumulative probabilities, too, but computed by an equivalent procedure. One can, of course, propose other different indicators (e.g. omitting synsemantics) but all must at least positively correlate with the above ones.

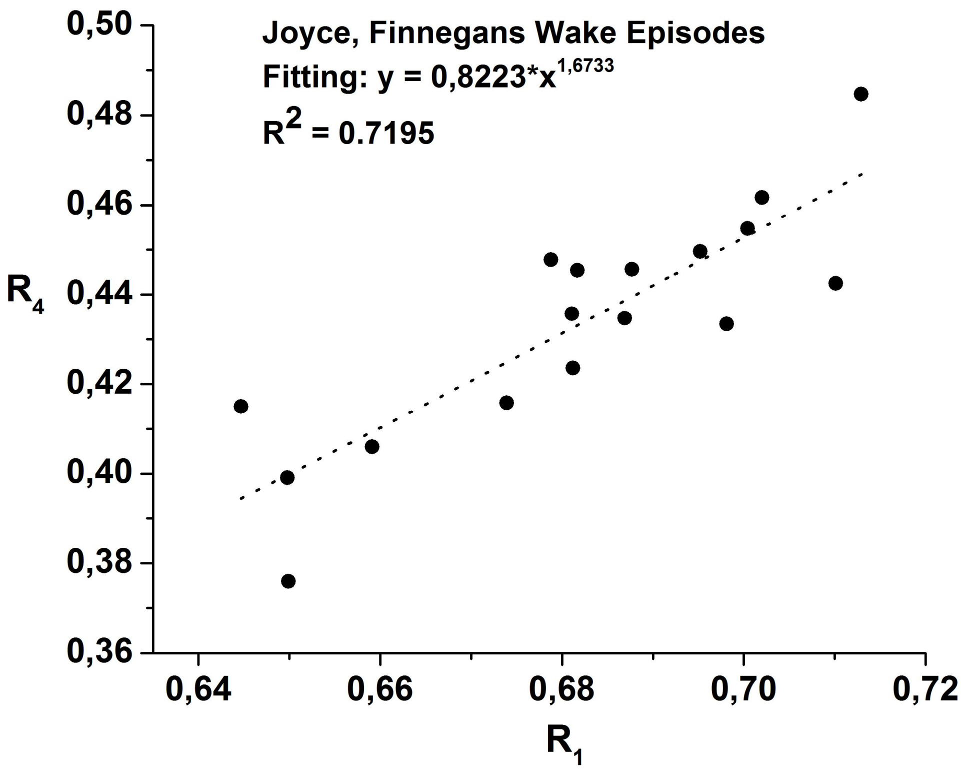

Figure 6. The relationship between R_1 and R_4

If there is at least a positive correlation between two indicators, one of them is sufficient for characterizing the text. But in that case one can show that the indicators merely show various aspects of the text and one can incorporate both in a synergetic control cycle. In special texts like FW, the dependence may be expressed by the difference between the parameters.

In order to obtain a wider perspective, we will also consider the link between R_1 and R_4 based on the data of Popescu et al. (2009), where 176 texts in 20 languages[1] were considered. The results are shown in figure 7.

[1] The 20 languages included were Bulgarian, Czech, English, German, Hungarian, Hawaii, Italian, Indonesian, Kannada, Lakota, Latin, Maori, Marathi, Marquesan, Rarotongan, Romanian, Russian, Samoan, Slovene and Tagalog.

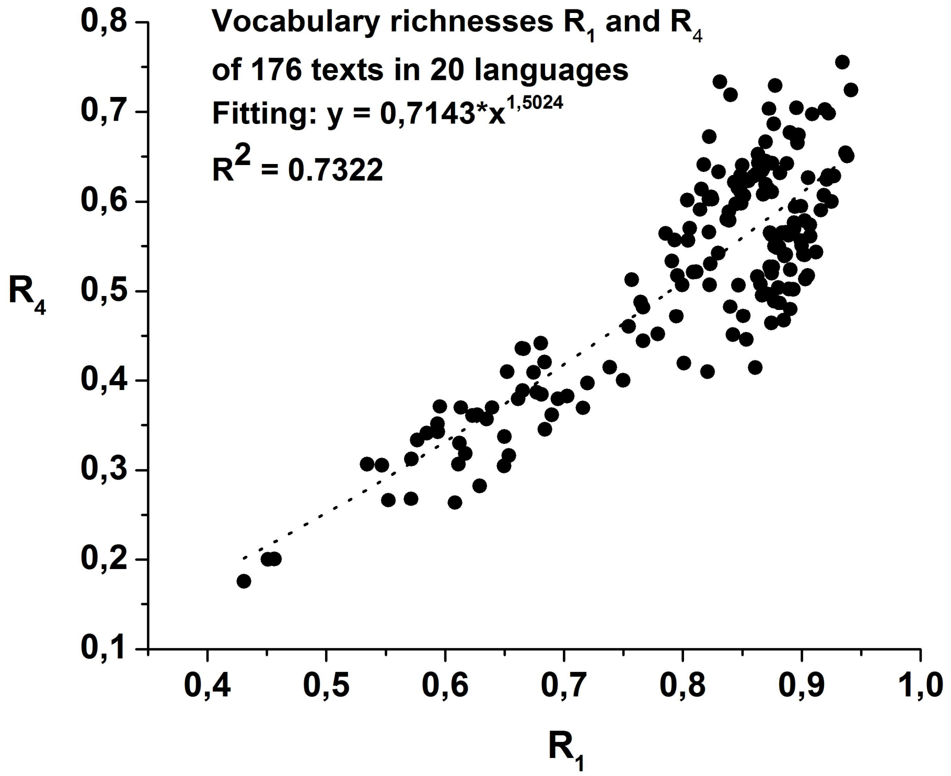

Figure 7. The link between R_1 and R_4 in 176 texts in 20 languages.

Richness cannot come into existence without influencing other properties. Finding those which are related with it may lead to a discovery of a law. To this end, a synthesis of all the computed above indicators of individual episodes of *Finnegans Wake* is presented in Table 20.

Table 20
Synthesis of all the above indicators
of individual episodes of *Finnegans Wake*

Text	N	V	Λ^*	I	S	H	RR	R_1	R_4	**α rad**	β_2
FW 01	9850	4107	1.9120	76.5266	101.4493	9.7437	0.0102	0.6811	0.4357	1.6292	10.1558
FW 02	6025	2798	1.9750	80.7499	105.3374	9.5711	0.0099	0.7129	0.4847	1.6445	10.4348
FW 03	9830	4363	1.9940	77.9017	103.6918	9.9722	0.0086	0.7020	0.4617	1.6355	10.8153
FW 04	10389	4443	1.9602	77.1060	102.2289	9.8648	0.0098	0.6817	0.4454	1.6253	10.5162
FW 05	8150	3419	1.8622	78.5312	99.6933	9.7025	0.0090	0.7101	0.4425	1.6386	8.8602
FW 06	16137	6243	1.8508	76.1719	101.2280	10.0712	0.0087	0.6591	0.4060	1.6253	10.0606
FW 07	9524	4153	1.9456	78.7856	102.3927	9.9052	0.0086	0.7004	0.4547	1.6349	9.9503
FW 08	8044	3477	1.8772	73.3131	99.7993	9.5949	0.0092	0.6788	0.4478	1.6491	10.3326
FW 09	14348	6166	1.9751	75.8000	101.9106	10.2781	0.0074	0.6877	0.4456	1.6363	10.7408
FW 10	15309	6619	2.0103	75.7198	102.1209	10.3844	0.0075	0.6952	0.4496	1.6316	11.1783

FW 11	26642	10676	1.9741	74.1423	102.2692	10.5383	0.0093	0.6447	0.4150	1.6067	12.0233
FW 12	6176	2402	1.7342	77.6954	101.1219	9.0835	0.0136	0.6739	0.4159	1.6443	9.0646
FW 13	9551	3961	1.8336	75.5798	100.8178	9.7812	0.0083	0.6869	0.4347	1.6535	9.9281
FW 14	17658	6237	1.7052	75.3757	98.9796	9.9978	0.0081	0.6499	0.3760	1.6284	9.1608
FW 15	27373	10438	1.8422	74.8546	101.3749	10.5862	0.0073	0.6498	0.3991	1.6181	10.7731
FW 16	12870	5307	1.8659	74.7842	99.6493	10.1697	0.0068	0.6981	0.4334	1.6496	9.9030
FW 17	12994	5271	1.8805	75.0404	98.1860	10.0400	0.0078	0.6812	0.4236	1.6347	9.2405

4. Conclusion

In this study, our main aim was to state whether, in a text of this sort, linguistic laws are strong enough to soften the exuberant self-organization in the vocabulary, to establish whether the usual mathematical models used to analyse texts are still valid. Our analysis shows that clearly even extraordinary texts, where the writer tries to deviate from the standard, follow some subconscious laws. We showed that it is possible to trace these laws by computing different indicators representing the degrees of some properties and searching for their links to other properties. In some cases, for example sections 3.2, 3.3, 3.4, 3.6 and 3.7, standard mathematical models could be used to achieve this. In such instances, it was possible to characterize the text as a whole, compare episodes and perform comparisons between different texts. This provided new insights into the structure and vocabulary of *Finnegans Wake* and presents opportunities for further analysis to be carried out. In others, the mathematical models needed to be adjusted or did not provide results consistent with any previously found data, limiting further analysis. This point shows that the interpretation of all of our findings is limited by the amount of comparable data and, as summarised in section 1, few linguists are perusing the study of language laws. In every language there are some boundaries that cannot be surpassed; *Finnegans Wake* may represent such a boundary, but this can be overcome once we can compare the results with thousands of texts in English and other languages.

References

Altmann, G. (1987). Zur Anwendung der Quotiente in der Textanalyse. *Glottometrika 1, 91-106.*

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

Bakhtin, M. (1986). *Speech genres and other late essays.* Austin: University of Texas Press.

Benstock, B. (1969). Every telling has a taling: A reading of the narrative of Finnegans Wake. *Modern Fiction Studies, 15, 3-25.*

Busemann, A. (1925). *Die Sprache der Jugend als Ausdruck der Entwicklungsrhythmik*. Jena: Fischer.

Bybee, J., Hopper, P. (eds.) (2001). *Frequency and the emergence of linguistic structure*. Amsterdam-Philadelphia: Benjamins.

Campbell, J., Robinson, H.M. (1947). *A Skeleton Key to Finnegans Wake.* London: Faber and Faber.

Christiani, D.B. (1966) Scandinavian elements of Finnegans Wake. Evanston: Northwestern University Press

DiBernard, B. (1980). *Alchemy and Finnegans Wake*. Albany: State University of New York Press.

Esteban, M.D., Morales, D. (1995). A summary of entropy statistics. *Kybernetica 31(4), 337-346.*

Glasheen, A. (1956). *A Census of Finnegans Wake: An Index of the Characters and Their Roles.* Evanston: Northwestern University Press.

Hart, C. (1962), *Structure and Motif in Finnegans Wake* (London: Faber and Faber).

Janda, L.A. (ed.) (2013). *Cognitive linguistics: The quantitative turn.* Berlin-Boston: de Gruyter-Mouton.

Kapur, J.N., Saxena, H.C. (1970). *Mathematical Statistics.*Delhi: Chand.

Knight, R. (2013). Laws governing rank frequency and stratification in English texts. *Glottometrics 25(1), 30-42.*

Köhler, R. (2012). *Quantitative syntax analysis*. Berlin-Boston: de Gruyter-Mouton.

Levin, H. (1944). *James Joyce: A critical introduction*. London: Faber and Faber.

MacCabe, C. (1979). *James Joyce and the revolution of the world.* London : Macmillan.

O'Hehir, B. (1967), *A Gaelic lexicon for Finnegans Wake.* Berkeley, Los Angeles: University of California Press.

Ord, J.K. (1972). *Families of Frequency Distributions*. London: Griffin.

Popescu, I.-I., Altmann, G. (2014). A simplified lambda indicator in text analysis. *Glottometrics* (pending).

Popescu, I.-I. (2001). Cited papers ranked by descending citation frequency. http://www.iipopescu.com/CITSH.htm

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

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

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

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

Popescu, I.-I., Čech, R., Altmann, G. (2013). Descriptivity in Slovak lyrics. *Glottotheory 4(1), 92-104.*

Popescu, I.-I., Mačutek, J., Altmann, G. (2010). Word forms, style and typology. *Glottotheory 3(1), 89-96*

Popescu, I.-I., Mačutek, J., Altmann, G. (2010). *Aspects of word frequencies*. Lüdenscheid: RAM-Verlag.

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

Sandulescu, G., Vianu, L. (2010). *A Manual for the Advanced Study of James Joyce's Finnegans Wake in One Hundred and Five Volumes*. Contemporary Literature Press http://editura.mttlc.ro/Joyce%20Lexicography.html http://sandulescu.perso.monaco.mc/

Spinks, L. (2009). *James Joyce: A critical guide*. Edinburg: Edinburgh University Press.

Wimmer, G., Altmann, G. (1999). On vocabulary richness. *Journal of Quantitative Linguistics 6(1), 1-9.*

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

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

Ziegler, A., Best, K.-H., Altmann, G. (2002). Nominalstil. *ETC – Empirical Text and Culture Research 2, 75-85.*
Zipf, G.K. (1935). *The Psycho-Biology of Language: An introduction to dynamic philology*. Boston: Houghton Mifflin

Chronology of Joyce's works
http://www.ricorso.net/rx/az-data/authors/j/Joyce_JA/apx/schema/Wks_chron.htm

Glottometrics 30, 2015, 73-85

Conceptual inertia in texts

Ruina Chen
Gabriel Altmann

Abstract. Conceptual unity of the text can be captured in different ways. Here, we use the Belza-chains and their presence in texts to perform a kind of measurement, comparisons, tests and modeling.

Keywords: Belza-chains, conceptual continuity, synergetics, text linguistics

1. Introduction

Every "normal" text has a theme; it speaks about something. To this end some concepts are usually repeated, though not always in the same form. In order to measure the degree of this kind of inertia, Belza (1971) introduced the concept of sentence chains which are called today Belza chains (cf. Skorochod'ko 1981; Altmann 2014). The Belza chain is an uninterrupted sequence of sentences containing the same concept. The concept is an autosemantic (explicitly presented or merely referred to). Synsemantics are taken into account only if they refer to or replace the respective autosemantic.

A concept need not be represented by the same word. One can take into account also synonyms, metaphors, hypernyms, hyponyms, pronouns, any kind of reference, and occurrence in other word classes. The last criterion means that a concept may be in the first sentence e.g. a noun, in the next one an adjective (e.g. German: *Gött, göttlich;* English: *dead, deadly, death*), etc. However, there are no prescriptions; every researcher can state his own criteria which are adequate for the given language and for his problem. There will be surely differences between the criteria for strongly analytic and strongly synthetic languages, the latter having a number of redundant forms. For example, in German "ich spreche" either "ich" or the affix "-e" is redundant. In Hungarian one may use merely the verb "beszélek" (I speak) with personal ending; in Indonesian, the verb does not have a personal ending: "saya bicara", just as in English, which has redundancy only in the present tense for the singular third person, e.g., "I speak" but "he speaks". But one must begin somewhere and improve the conceptual background step by step. It must be emphasized that if the same concept occurs in a non-immediate subsequent sentence, then the given sentence does not belong to the same chain. For thorough descriptions of cohesion and coherence types see the books on text linguistics (e.g. Linke, Nussbauer, Portmann 1994) which is, unfortunately, still qualitative.

A sentence may belong to a chain or be conceptually isolated. The chain length is measured by the number of sentences belonging to it. Here the frequency of a concept is not important but its presence in chains is. There may be a sentence whose predecessor and follower do not contain a common concept with it hence there are chains of length 1. If a set or subset of sentences contains more common concepts, then one counts as many chains as necessary and measures their length separately. On the other hand, the repetition of a concept in the same sentence need not be taken into account. The length and the number of chains in a text express its conceptual inertia.

Instead of concepts one can consider also speech acts in a stage play. In stage plays we expect many interruptions of concept chaining because sentences may simply be some reactions to preceding sentences but need not contain the same concept. An answer "Yes!" to any question is, of course, in some semantic connection with the preceding question but not a conceptual one. Hence stage plays may differ strongly from e.g. scientific texts. But even here, one can consider ellipsis a mute repetition of the concept.

In order to obtain a comparable indicator, one can use the mean length of chains. If there are no chains, then each sentence represents a chain of length 1, hence the minimum inertia is 1. The maximum cannot be given. In any case, one can test whether the inertia significantly differs from its minimum (representing the null hypothesis), and one can test the difference of two texts, because the variances of lengths can be computed in the usual way. Another comparable indicator is the proportion of chains of length 1 which indicate the interruption of the conceptual chaining.

In order to perform the analysis in a unified way, we consider sentence a unit separated from other units by a full stop, colon, semicolon, question mark, and exclamation mark, but this must be stated just at the beginning because different treatment of these signs may lead to different results. In poetry, the boundaries are unequivocal: each verse is a separate unit.

2. Measurement

Conceptual inertia can be measured by means of the properties of the distribution of lengths, for example in terms of the mean length of chains, i.e. by

$$\overline{CI} = \frac{1}{L}\sum_{k=1}^{L} l_k \tag{1}$$

where l_k are the individual chain lengths and L is the number of chains. If one finds a theoretical distribution/function capturing the observed distribution, then one of the parameters of the function can be used as an indicator, e.g. Ord's criterion. Since we operate with lengths, the distribution of chain lengths may be, perhaps, captured by the Zipf-Alekseev distribution or replaced by a respective (not normalized) function (cf. Popescu, Best, Altmann 20014). Having a theoretical function, one can construct a number of different indicators.

Now, one can perform the analysis stepwise, for each chapter of a novel separately, or one can take the complete text (simply by adding all results) which is automatically given with short texts. Then a number of various hypotheses can be tested. In the sequel, we mention some of them.

(a) If one considers separate parts of the text, then the evolution of inertia can be studied. Either one compares some empirical indicators or, if one has a theoretical function one studies the change of a parameter of the function.

(b) Since the first result of the analysis is a vector of lengths (lengths written as they occur/begin in text), one can study the properties of the vector, test the hypothesis that the more distant are two (whole) parts of the text, the greater will be the difference of the vectors (or the given functions). This is an analogue to the Skinner hypothesis (1957). This hypothesis is usually applied to test the phonetic similarity of verses with increasing distance. It can be used, of course, also in semantics or other domains of linguistics.

(c) The conceptual inertia of a given text can as a whole be compared with other texts. In this way one could trace down one of the properties of text sorts, development of the writer, and tendencies in the culture represented by individual languages. A scientific text is surely written with different conceptual inertia than a poetic or a didactical text.

(d) Belza chains are not an isolated phenomenon. They may display relations to other texts/language properties which would open an infinite domain because the number of text properties depends on the development of science. This way leads to the construction of control cycles (cf. Köhler 2005), setting up parts of a theory, searching for laws, etc. One can, for example, mention mean sentence length, entropy of different kinds, text stratification, or any property of the Köhlerian control cycle.

Investigations of this kind are few and far between because they cannot be performed by the computer which cannot discover synonymy or metaphor, etc.; they must be performed by hand. Homonymy is not taken into account, e.g. in German “der Leiter” (*leader*) and “die Leiter” (*ladder*) do not represent the same concept. Still more complex is the situation in written Chinese. Here we shall present only some examples from some languages in order to stimulate this kind of research.

For German we use the poem *Der Erlkönig* by J.W.v. Goethe. The chain is marked in the line on the topical concept. The unit is the verse. Verses not belonging to any chain are weighted by 1. It is to be noted that if the same concept occurs twice in a line it is taken into account only once. The words in the second column of the table are only representatives of a concept in the given chain. In order to make a chain more lucid we insert one of the typographical symbols ▀ ▄ █ ■ ▬ ▲ ► ▼ ◄ ● ◘ ☺ behind the respective concept

Table 1
Inertia in a German text (Goethe, Der Erlkönig)

Wer▀ reitet so spät durch Nacht und Wind?	6 ▀ (wer,Vater, er, er, mein, Vater)
Es ist der *Vater*▀ mit seinem *Kind*▬;	4 ▬ (Kind, Knaben, ihn, Sohn)
Er▀ hat den *Knaben*▬ wohl in dem Arm.	
Er▀ fasst *ihn*▬ sicher, er hält ihn warm.	
Mein▀ *Sohn*▬, was birgst du so bang dein Gesicht?	
Siehst, *Vater*▀, du den *Erlkönig*▲ nicht?	2 ▲ (Erlkönig, Erlkönig)
Den *Erlkönig*▲ mit Kron und Schweif?	
Mein *Sohn*▼, es ist ein Nebelstreif.	3 ▼ (Sohn, du, dir)
Du▼, liebes Kind, komm, geh mit *mir*! ●	2 ● (mir, ich)
Gar schöne Spiele spiel *ich*● mit *dir*▼;	
Manch bunte Blumen sind an dem Strand,	1
Meine Mutter hat manch gülden Gewand.	1
Mein► Vater, mein Vater, und hörest du nicht,	3 ► (mein, mir, Kind)
Was Erlenkönig *mir*► leise verspricht?	
Sei ruhig, bleibe ruhig, mein *Kind*►:	
In dürren Blättern säuselt der Wind	1
Willst, feiner *Knabe*☺, du mit *mir*▬ gehn?	2 ☺ (Knabe, dich);
Meine ▬*Töchter*◘ sollen *dich*☺ warten schön;	3 ▬ (mir, meine, meine)
Meine▬ *Töchter*◘ führen den nächtlichen Reihn	3 ◘ (Töchter, Töchter, wiegen)
Und *wiegen*◘ und tanzen und singen dich ein.	
Mein Vater, mein Vater, und siehst du nicht dort	1
Erlkönigs Töchter am düstern Ort?	1
Mein Sohn, mein Sohn, ich sehe es genau:	1
Es scheinen die alten Weiden so grau.	1
Ich▼ liebe *dich*◄, mich reizt deine schöne Gestalt;	4 ▼ (ich, ich, er, Erlkönig);
Und bist *du*◄ nicht willig, so brauch *ich*▼ Gewalt.	4 ◄ (dich, du, mich, mir)

Mein Vater, mein Vater, jetzt faßt *er*▼ *mich*◄ an! *Erlkönig*▼ hat *mir*◄ ein Leids getan! Dem *Vater*► grausets, er reitet geschwind, *Er*► hält in Armen das ächzende Kind, *Erreicht*► den Hof mit Mühe und Not: In *seinen*► Armen das Kind war tot.	 4 ► (Vater, er, erreicht, seinen)

It can be seen that the verbal affix in the third person (*erreicht*) and the pronouns identify the respective person.

Setting up the vector of chain lengths we obtain [6, 4, 2, 3, 2, 1, 1, 3, 1, 2, 3, 3, 1, 1, 1, 1, 4, 4]. The mean is 43/18 = 2.3889 (= sum of lengths divided by their number). The variance of the length is $s^2 = 2.1340$. Since the smallest Belza length of a text is 1 – occurring when there are no conceptual chains – one can express the weight of chaining using the normalization by the u-criterion showing the weight of deviation of the mean from 1. Using the above numbers we obtain

$$IW = (2.3889 - 1)/(2.1340/18)^{1/2} = 4.0624.$$

This indicator is adequate for simple classifications but not for comparisons.

However, the strengths of inertia can be estimated rather by the number of conceptual interruptions in the text. The text may be semantically coherent – as is usual – but for expressing something, the author may use different concepts. Hence, the number of isolated sentences, i.e. $f(1)$, is an image of continuity. In order to characterize a text, one can take the relative frequency of the isolated sentences, i.e. $P = f(1)/N$ where N is number of chains, which can easily be used for comparisons. For example, there are seven isolated lines in Goethe and N = 18 chains, hence $P(1)$ = 7/18 = 0.3889. The variance of a proportion is $V(P) = PQ/N$, here $V(P_{Gethe})$ = 0.3889(1-0.3889)/18 = 0.0132. Below, we perform all comparisons of texts using this indicator.

In order to show another example, we analyze explicitly a Slovak text, a piece of prose by E. Bachletová written in a very poetic vein (cf. Table 2).

Table 2
A Slovak prose text by E. Bachletová

Leto v nás	
Rozpálené cesty, levandulové záhony, tisíce vôní v povetrí.	1
A mierne ospalé, pomalé popoludnie na terase kaviarne.	1
Sedíme, hodnotíme svoj život a jednoducho – sme.	1
Leto je výbornou kulisou k rozprávaniu o bytí.	1
Akoby sme▼ boli náhle posunutí do inej dimenzie, kde sa konečne nikam nenáhlime. Nie sme▼ zasýpavaní mailami ani správami na mobile, svet má jednoducho inú príchuť. A tak sa možno vraciame▼ do detstva, do čias mladosti, akosi nechtiac porovnávame, či hľadáme isté spojenia. A možno zisťujeme,▼ že roky nezvratne posunuli život a my sa nemáme o čo oprieť, alebo v komsi nájsť tichého spojenca.	5▼ (sme, sme, vraciame, zistujeme, starneme)

Starneme▼ v čase.	
Chvíľa sentimentu je tu.	1
No vzápätí si uvedomíme, že sme■ sa ocitli na vnútornej križovatke a nie je isté, či zvolíme správne.	2■ (sme, nás)
Leto► v nás■.	2► (leto, horúce)
Horúce►, dráždivé, znepokojivé.	
S nábojom výziev, ktoré priniesla doba, spoločenský tlak, okolnosti v súkromí či v profesii.	1
Ako sledujem životy mojich priateľov, je zrejmé, že vari každý z nich prežíva akýsi zlom či prerod.	1
Nové zamestnanie, zdravotné problémy, syndróm vyhorenia, namáhavá opatera rodičov, rozvod, strata domova či narušená komunikácia s deťmi.	1
Mnohé zmeny sa však v našich životoch dejú buď prirýchlo alebo priveľmi pomaly.	1
O to ťažšie je nájsť vnútornú rovnováhu, alebo aspoň dočasnú spokojnosť so stavom, ktorý nie je optimálny.	1
Byť trpezlivým, rozvážnym, pokojným v čase neistoty a obáv o finančné zabezpečenie nie je jednoduché.	1
Avšak aj nad touto situáciou má moc Boh. ●	5● (Boh, Boh, Božích, Pán, Boží)
Práve v okamihoch nášho najhlbšieho vnútorného temna má Boh● s nami svoje plány.	
A takmer vždy ide naozaj o trpezlivé, no zároveň odvážne odovzdanie sa do Božích● rúk.	
Pán● má totiž pripravené svoje riešenie, no v inom chápaní času,◘ ako si my▼ predstavujeme.	2 ◘ (času, čas)
Boží● čas◘ zahŕňa totiž priestor▬ pre naše ▼ duchovné prijatie novej situácie.	2 ▼(predstavujeme, naše) 2 ▬ (priestor, ten)
A ten▬ sa jednoducho nedá – odmerať.	
Je leto.	1
Prázdniny otvorili svoju náruč, deti sa rozbehli do táborov a kolóna áut sa nedočkavo posúva po diaľnici k moru.	1
Sme ■vytrhnutí z každodennosti a možno sa cítime neisto v novej úlohe, ● ktorú máme.	8 ■ (sme, máme, sme, naša, nedokážeme, nemáme, naša, stratíme)
Áno, máme ■totiž novú úlohu ●– oddychovať. ☺	
Hoci sme ■mnohí už v strednom veku, oddychovať ☺ akosi nedokážeme.	2 ● (úlohe, úlohu) 2 ☺(oddychovať, oddychovať)
Naša ■myseľ je zaneprádznená starosťami, úzkosťami rôzneho druhu.	
Nedokážeme ■jednoducho vypnúť a prežiť radosť zo slobody a oddychu.	
A možno nemáme ■ani tie správne podmienky na relax.	
No jedno je isté, že naša ■preťažená duša potrebuje načerpať novú energiu, novú nádej a novú vášeň pre život.	
Inak úplne stratíme ■nadhľad nad vlastnou či vnútenou realitou.	
Želám vám všetkým, aby ste si dovolili splniť úlohu letného oddychu bez výčitiek svojej mysle či okolia.	1
Boh nás totiž potrebuje silných, aby sme Mu opäť mohli slúžiť s radosťou a úsmevom na perách a v duši.	1

The results for this and other texts in other languages are presented in Table 3.

Table 3
Survey of conceptual inertias in some texts

Text	**Vector**	**Length**	**Chains**	**Mean**	**Variance**	**P**	**V(P)**
German Goethe	[6,4,2,3,2,1,1,3,1,2,3, 3,1,1,1,1,4,4]	43	18	2.3889	2.1340	0.3389	0.0132
Slovak: Bachletová	[1,1,1,1,5,1,2,2,1,1,1, 1,1,1,5,2,2,2,1,1,8,2, 2,1,1]	47	25	1.8800	2.8600	0.6000	0.0096
Slovak: Svoráková	[6,2,6,2,3,2,2,3,2,3,1, 1,1,7,1,2,1,1,2,1,2]	51	21	2.4286	3.1571	0.3333	0.0106
Indonesian Rosidi	[4,2,5,2,2,1,1,2,2,2,1, 3,3,1,2,2,3,1,2,2,3,2, 1]	49	23	2.1304	1.0277	0.2609	0.0084
Hungarian: Petöfi	[1,1,1,1,2,2,1,5,4,1,2, 3,2,4,3,2]	35	16	2.1875	1.6292	0.3750	0.0146
Italian Napolitano	[2,2,2,2,1,1,1,6,2,2,1, 5,2,1,2,1,2,2,1,3,1,3, 7,2,2,1,3,2,3,1,1,1,1, 1,1,1,3,1,1,1,2,1,1,4, 2,1,1,3,2,8,1,1,3,2]	110	54	2.0370	2.3005	0.4630	0.0046
Czech Havel 1990	[2,1,1,1,1,1,1,1,1,2,2, 4,1,1,5,3,2,2,8,2,5,2, 1,2,3,1,1,2,1,1,2,1,2, 2,1,2,2,2,2,1,2,2,1, 1,1,1,1,4,1,7,2,2,1,1, 5,1,2,2,2,2,3,1,1,1,1, 1,1,1,1,1,1,1,1,1,1,1, 1,2,2,2,2,2,2,3,1,2,1]	157	87	1.8046	1.5474	0.5172	0.0029
Czech Havel 1991	[2,4,2,1,5,4,2,1,1,6,2, 16,2,2,1,1,6,3,4,3,1, 4,4,2,4,2,1,4,1,1,2,3, 1,1,2,1,1,6,2,1,2,3,5, 3,3,2,2,4,1,1,1,2,8,3, 4,2,3,2,3,1,2,3,1,1,1]	175	65	2.6923	5.2163	0.3231	0.0036
French St.-Exupé-ry (Ch. 1)	[3,1,2,2,4,1,2,2,2,2,2, 1,4,2,7,2,3,1,3,4]	50	20	2.5000	2.0526	0.2000	0.0080
Chinese 1[1]	[4,3,4,1,2,6,2,2,1,4,4, 4,2,2,2,4,5,3,3,6,2,2, 3,3,2,2,2,3,6,2,2,2,2, 2,3,2,2,2,2,3,3,2,1,3, 2,2,4,3,3,2,2,2,2,2]	146	54	2.7038	1.3823	0.0556	0.0010
Chinese 2[2]	[5,2,8,9,14,3,2,2,2,5, 3,3,2,2,2,4,2,4,2,3,4,	159	47	3.3830	4.9371	0	0

[1] The consumption tax of refined oil will be adjusted from today on, while its price in the domestic market remain unchanged (From *People's Daily*, Nov 29, 2014)

[2] The three pillars consolidate the harvest base (policy interpretation)(From *People's Daily*, Dec 5, 2014)

	3,2,2,5,3,5,3,3,3,5,2, 3,2,3,2,2,5,2,2,2,2,2, 3,5,2,3]						
English Press[3] 1	[2,3,2,7,2,2,2,2,2,3,2, 2,2,2,2,1,2,3,4,2,3,2, 2,1,1,3,2,2,2,1,2,3,2, 2,6,1,1]	85	37	2.2973	1.4925	0.1081	0.0026
English Press[4] 2	[3,2,2,4,2,2,1,3,7,2,5, 1,3,3,3,4,2,2,4,2,3,2, 2,3,2,3,2,2, 8,2,2,2, 2,3,2,2, 2]	101	37	2.7297	2.0360	0.0541	0.0014

The order of texts in Table 3 does not follow any principle. If one orders the texts according to the mean length of chains, one obtains: Havel 90 – Bachletová – Napolitano – Rosidi – Petöfi – English 1 – Goethe – Svoráková – St.-Exupéry – Havel 91 – Chinese 1 – English 2 – Chinese 2 (Cz, Sk, It, Ind, Hu, E, G, Sk, Fr, Cz, Ch, E, Ch); if one orders them according to P, one obtains: Chinese 2 - Chinese 1 – English 2 -English 1 – St. Exupéry – Rosidi – Havel 91 – Svoráková – Goethe – Petöfi – Napolitano – Havel 90 – Bachletová (Ch, Ch, E, E, Fr, Ind, Cz, Sk, G, Hu, It, Cz, Sk). The orders are "almost" symmetric but there is no linguistic principle. Further texts are necessary in order to discover the background.

3. Comparisons

The direct comparison of the two texts can be performed using the u-test for testing the difference of two means. For the first two analyzed texts we obtain

$$u = \frac{\overline{CI}_{Goethe} - \overline{CI}_{Bachletová}}{\sqrt{Var(\overline{CI}_G) + Var(\overline{CI}_B)}} = \frac{2.3889 - 1.8800}{\sqrt{\frac{2.1340}{18} + \frac{2.8600}{25}}} = 1.0544$$

hence the mean chaining inertia of the two texts is not significantly different.

Now, since we are interested in the inertia which is interrupted by isolated sentences, we may compare the proportions of isolated sentences in two texts, i.e. the interruptions of the conceptual stream. One can perform the exact binomial test, Fisher's test, or one can use the asymptotic normal test.

Though the values of mean chain lengths do not differ visually, we can state that none of the means differs significantly from the other ones. The resulting *u* are not significant. Using the proportions of isolated sentences, we apply the asymptotic normal test and compute

$$u = \frac{|P_1 - P_2|}{\sqrt{P(1-P)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

[3] The Cuban embargo. If not now, when? (From *The Economist*, April 5th 2014)
[4] Wooing Mrs Merkel (From *The Economist,* March 1st 2014)

where P can be estimated from $P = (f_{1,1} + f_{1,2})/(N_1 + N_2)$, where $f_{1,1}$ is the frequency of chains of length 1 in the first text, $f_{1,2}$ that in the second text. For the individual authors we obtain the results presented in Table 3. An example: comparing Goethe and Bachletová we obtain P = (7 + 15) /(18 + 25) = 0.5116, hence

$$u(Goethe, Bachletová) = \frac{|0.3889 - 0.6000|}{\sqrt{0.5116(1-0.5116)\left(\frac{1}{18}+\frac{1}{25}\right)}} = 1.37\,.$$

This difference is not significant. All the other comparisons are presented in Table 4.

Table 4
Comparison of inertia interruptions in texts

	Goethe	Bachletová	Svoráková	Petöfi	Rosidi	Napolitano	Havel 90
Goethe	-						
Bachletová	1.37	-					
Svoráková	0.36	1.80	-				
Petöfi	0.08	3.43*	0.26	-			
Rosidi	0.87	2.37*	0.53	0.76	-		
Napolitano	0.55	3.68*	1.02	0.62	1.66	-	
Havel 90	0.99	0.73	1.51	1.04	2.19*	0.63	-
Havel 91	0.52	2.40*	0.09	0.39	0.56	1.56	2.39*

	St.-Exupéry	Chinese 1	English 1	Chinese 2	English 2
Goethe	1.28	3.54*	2.44*	3.94*	2.68*
Bachletová	2.70*	5.37*	4.12*	5.97*	4.73*
Svoráková	0.96	3.18*	2.10*	4.18*	2.88*
Petöfi	1.16	3.35*	2.28*	4.41*	3.00*
Rosidi	0.47	2.56*	1.54	3.66*	2.29*
Napolitano	2.06*	4.83*	3.37*	5.38*	4.19*
Havel 90	2.57*	5.62*	4.26*	6.05*	4.86*
Havel 91	1.06	3.62*	2.43*	4.32*	3.13*
St.-Exupéry		1.89	0.95	3.16*	1.71
Chinese 1			0.92	1.64	0.03
English 1				1.91	0.75
Chinese 2					1.61

As can be seen, the Chinese texts differ significantly from almost all other texts. This is caused, perhaps by the language (probably also by the genre, since the two chosen texts were taken from the press. Information in this genre type is usually densely concentrated, which may contribute to the consecutively connected concept or thematic chains within, as reflected in the low percentages of value-“1”-chain), but this conjecture must be further scrutinized. Quite peculiar is the difference between the two texts of the Czech president. However, we conjecture that his texts have some extreme links with other properties which must be studied separately. Each text displays some differences but one needs a thorough investigation to find the causes.

Considering the number of significant differences between texts and languages we may set up the following order: Chinese 2 (9), Slovak: Bachletová (9), English 2 (8), Chinese 1 (8), Chinese 2 (8), English 1 (7), Czech: Havel 90 (7), Czech: Havel 91 (6), Italian (6), Hungarian (5), Indonesian (5), German (4), Slovak: Svoráková (4), French (4). One cannot recognize any system.

Of course, one could measure also the radians between the vectors but the length of the compared texts plays here an important role. In order to apply this method, one would be forced to take text parts consisting of the same number of sentences. This is possible – without violating the structure of the texts – e.g. in sonnets which have the same length in all languages.

4. Fitting

The lengths of the Belza chains follow a probability distribution which can be modeled. But since we have to do with lengths, we prefer a simple function whose adequacy has already be shown for any type of lengths in language (cf. Popescu, Best, Altmann 2014). It is the Zipf-Alekseev function obtained as a special case of the unified theory (cf. Wimmer, Altmann 2005), this time considering *A* as the situation in the language and substituting the function *B ln x* for the influence of the speaker/writer. The logarithmic influence of the speaker is known also from psychology. The differential equation

(2) $$\frac{dy}{y} = \frac{A + B\ln x}{Dx}dx$$

after reparametrization yields the function

(3) $$y = cx^{a+b\ln x},$$

where the independent variable *x* is the length and the dependent variable *y* is the frequency of the given length.

Results of fitting (3) to the individual poems are presented in Table 5.

Table 5
Fitting the Zipf-Alekseev function to Belza chains

	German (Goethe)		Slovak (Bachletová)		Slovak (Svoráková)		Hungarian (Petöfi)		Indonesian (Rosidi)	
1	7	6.8512	15	15.0011	7	7.1626	6	6.0424	6	6.0152
2	3	4.0253	7	6.9937	8	7.3043	5	4.7147	1	10.9623
3	4	2.9684	-	-	3	4.1913	2	2.7095	1	4.0959
4	3	2.3985	-	-	-	-	2	1.5231	4	1.1203
5	-	-	2	2.0274	-	-	1	0.8776	1	0.2912
6	1	1.7834	-	-	2	0.6153				
7			-	-	1	0.3401				
8			1	0.9707	-	-				

	a = -0.7774	a = -0.9921	a = 0.9104	a = 0.2781	a = 2.9441
	b = 0.0147	b = -0.1559	b = -1.2727	b = -0.9177	b = -2.9983
	c = 6.8512	c = 15.0011	c = 7.1626	c = 6.0424	c = 6.0152
	R^2 = 0.84	R^2 = 1.00	R^2 = 0.89	R^2 = 0.96	R^2 = 0.99

	Italian: Napolitano		Czech: Havel 1990		Czech: Havel 1991		French: St-Exupéry		Chinese 1	
1	25	24.9923	45	45.0073	21	21.0135	4	4.1617	3	3.9574
2	17	17.0961	31	30.9451	18	17.8089	9	8.5872	28	27.2697
3	7	6.4049	4	4.5211	10	11.0475	3	4.2260	12	14.0525
4	1	2.2714	2	0.5385	9	6.6393	3	1.5390	7	3.9361
5	1	0.8369	3	0.0668	2	4.0577	-	-	1	0.9273
6	1	0.3263	-	-	3	2.5475	-	-	3	0.2114
7	1	0.1347	1	0.0014	-	-	1	0.0646		
8	1	0.0586	1	0.0002	1	1.0888				
16					1	0.0785				
	a = 0.6343		a = 2.1116		a = 0.3537		a = 2.8076		a = 5.5732	
	b = -1.7054		b = -3.8261		b = -0.8546		b = -2.5429		b = -4.0230	
	c = 24.9923		c = 45.0073		c = 21.0134		c = 4.1617		c = 3.9574	
	R^2 = 0.99		R^2 = 0.99		R^2 = 0.97		R^2 = 0.87		R^2 = 0.96	

	English: Press 1		Chinese 2		English: Press 2	
1	6	6.0056	-	-	2	2.2691
2	22	21.9929	21	21.2035	20	19.8401
3	6	6.0305	13	11.3731	9	9.4718
4	1	0.9605	3	6.7457	3	2.2817
5	-	-	7	4.2966	1	0.4525
6	1	0.0197	-	-	-	-
7	1	0.0030	-	-	1	0.0169
8			1	1.4570	1	0.0035
9			1	1.0805		
14			1	0.3187		
	a = 5.0676		a = -0.8142		a = 6.2524	
	b = -4.6093		b = -0.4030		b = -4.5073	
	c = 6.0056		c = 45.2475		c = 2.2691	
	R^2 = 0.99		R^2 = 0.9305		R^2 = 0.99	

All fittings are satisfactory. The result indicates that even length of this kind has a law-like background. Of course, many texts more must be analyzed in order to accept it definitively.

5. Control

Since we work with a function whose parameters are interpreted (*A* = state of the language, *B* influence of the speaker, *D* = control by the community), a part of the structuring is concealed in the relationship between the parameters. In the resulting formula, the forces have been reparametrized and *c* is merely the integration constant depending on the frequency of length 1. But there may be a link between the parameters *a* and *b*. It would be of course better to analyze a great number of texts in many languages but this is a task for a future team.

Using the results repeated in Table 6, we compare simply the parameters *a* and *b* in form of a graph because any computation would be premature, and obtain Figure 1. Here, the values of *a* are simply ordered increasingly; the result is evident but preliminarily, therefore we cannot propose an appropriate function. It looks quite linear but one cannot generalize with only 11 texts. In any case, this relation is a sign of self-regulation, a kind of equilibrating the influencing forces.

Table 6
Results of computations

Text	a	b	c	R^2
Slovak: Bachletová	-0.9921	-0.1559	15.0011	1.00
German: Goethe	-0.7774	0.0147	6.8512	0.84
Hungarian: Petöfi	0.2781	-0.9177	6.0424	0.96
Czech: Havel 1991	0.3537	-0.8546	21.0134	0.97
Italian: Napolitano 2013	0.6343	-1.7054	24.9923	0.99
Slovak: Svoráková	0.9104	-1.2727	7.1626	0.89
Czech: Havel 1990	2.1116	-3.8261	45.0073	0.99
French; St.-Exupéry	2.8076	-2.5429	4.1617	0.87
Indonesian: Rosidi	2.9422	-2.9993	6.0152	0.99
English: Press text (1)	5.0676	-4.6093	6.0056	0.99
English: Press text (2)	6.2524	-4.5073	2.2691	0.99
Chinese: Press text (1)	5.5734	-4.0230	3.9574	0.96
Chinese: Press text (2)	-0.8142	-0.4030	45.2475	0.93

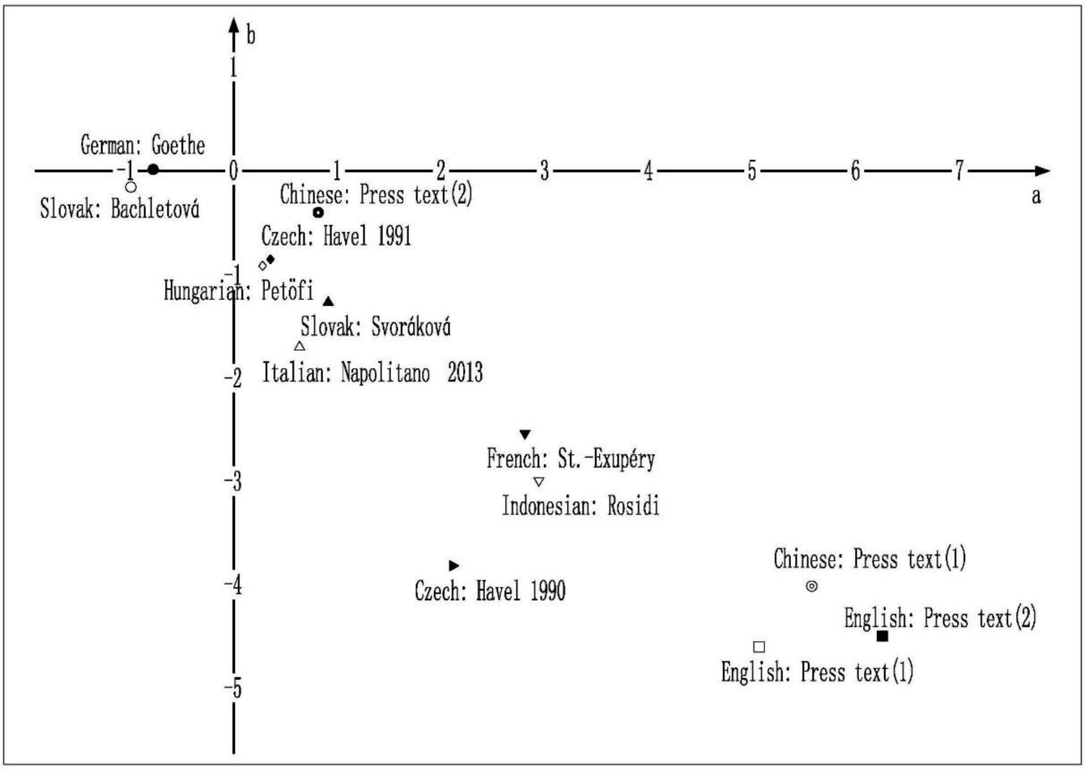

Figure 1. Relation between parameters *a* and *b* of the Zipf-Alekseev function

It can be conjectured that the longer a chain, the more different words or parts of words represent the given concept. This, however, strongly depends also on the language, its analytism or synthetism. In order to find an adequate expression of this dependence, one has to analyze several texts of the same author or many texts in the given language. We simply conjecture that a link of this kind could be captured by the same formula but with different parameters. Its placing in Köhler's (2005) control cycle would be the next step. Mixing languages leads to a preliminary Lorentzian function but one could be satisfied also with a straight line.

6. Conclusions

The above results represent only one of the many possible approaches to the measurement of the conceptual unity of texts. Here, two types of direct continuation of this research can be sketched. (1) One may count all occurrences of a given concept with its text-linguistic representatives. In this way one obtains a different distribution which may be called concept distribution. The possibilities to derive the distribution theoretically and evaluate its properties analogously to the word distribution are sufficiently known. A great number of indicators can be used for the characterization of texts. (2) The representatives of a concept do not have the same weight. The main concept may be represented by all text-linguistic categories, and this representation may be weighted. There are many possibilities, one must decide for one of them. To show merely an example: Personal pronouns in singular refer exactly but those in plural concern several concepts. For example in Indonesian, "kami" (we) concerns "I" and some other persons, but "kita" means "I" and "you". Hence the weights of representation differ. A personal ending has a different weight than direct naming, etc. Up to now, there is no trial to evaluate the categories of text-linguistics in this way. Nevertheless, it will be necessary in the future.

References

Altmann, G. (2014). Supra-sentence levels. *Glottotheory 5(1), 25-39*.

Bachletová, E. (2014). Leto v nás. In: *Riadky bytia*. Bratislava. (poetry in prose)

Belza, M.I. (1971). K voprosu o nekotorych osobennostjach semantičeskoj struktury svjaznych textov. In: *Semantičeskie problemy avtomatizacii i informacionnogo potoka: 58-73*. Kiev.

Havel, V. (1990, 1991). *End-of Year Speeches*.

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

Köhler, R., Altmann, G. (2009). *Problems in Quantitative Linguistics Vol. 2: 57-58.* Lüdenscheid: RAM

Linke, A., Nussbaumer, M., Portmann, P.R. (1994^2). *Studienbuch Linguistik*. Tübingen: Niemeyer.

Napolitano, G. (2013). *End-of-Year speech.*

Petöfi, S. (1847). *Szeptember végén* (poem).

Popescu, I.-I., Best, K.-H., Altmann, G. (2014). *Unified modeling of length in language.* Lüdenscheid: RAM.

Rosidi, Ajip (1968). *Tjerita pendek Indonesia*. Djakarta: Gunung Agung. (Chapter "Kawan bergelut", p. 35-37).

Skinner, B.F. (1957). *Verbal behavior.* Skinner Foundation.

Skorochod´ko, E.F. (1981). *Semantische Relationen in der Lexik und in Texten*. Bochum: Brockmeyer.

Svoráková, S.(2003). Čakanie na Štraussa. Recenzia knihy : Tomáš, Štrauss : Metamorfózy umenia XX. storočia. Bratislava : Kalligram, 2001. *D art -- Revue súčasného výtvarného umenia* 10, 37.

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

BIBLIOGRAPHY

A bibliography of quantitative studies on sound symbolism

Hanna Gnatchuk

Aaker, Jennifer L. (1997). Dimensions of brand personality. *Journal of Marketing Research* 34.3. 347-356. http://www.haas.berkeley.edu/groups/finance/Papers/Dimensions%20of%20BP%20JMR%201997.pdf

Abelin, A. (2012). Relative frequency and semantic relations as organizing principles for the psychological reality of phonaesthemes. *Selected papers from UK-CLA Meetings* 1, 128-145. http://www.uk-cla.org.uk/files/proceedings/Abelin.pdf

Abelin, A. (1999). *Studies in sound symbolism*. Göteborg University.

Athaide, Gerard A., Klink, Richard R. (2012). Creating global brand names: The use of sound symbolism. *Journal of Global Marketing* 25.4. 202-212.

Baxter, S., Lowrey, T. (2011). Phonetic symbolism and children's brand name preferences. *Journal of Consumer Marketing* 28(2), 516-523. http://faculty.business.utsa.edu/tlowrey/Baxster%202011.pdf

Bentley, M., Varon, E. (1933). An accessory study of "phonetic symbolism". *American Journal of Psychology 25*, 76-86.

Bergen, B. (2004**).** The psychological reality of phonaesthemes. *Language* 80, 290 -314.

Birch, D., Erickson, M. (1958). Phonetic symbolism with respect to three dimensions of the semantic differential. *The Journal of Generative Psychology* 58(2), 291-297.

Brown, Roger, Nuttall, Ronald. (1959). Method in phonetic symbolism experiments. *Journal of Abnormal and Social Psychology* 54. 441-445

Bystrova, L.V., Levickij, V.V. (1973). Fonetičeskoe shodstvo semantičeski svjazannyh slov, *Zeitschrift fur Phonetik, Sprachwissenschaft und Kommunikationsforschung*, 26.(6), 44-87.

Bystrova, L.V., Levytskyj V.V. (1976). Shche raz pro symvolichne znachenna dejakuh holosnuh ta pryholosnuh ["Again about symbolic meaning of some vowels and consonants"]. *Inozemna filologija*, 81, 75-80 (in Ukrainian).

Chisnall, Peter M. (1974). Aluminium household foil in the common market: Research for an effective brand name. *Journal of Management Studies* 11, 246-255.

Coulter, K. (2009). The effects of phonetic symbolism on comparative price perceptions. In: Ann L. McGill and Sharon Shavitt (eds.), *NA - Advances in Consumer Research Volume 36, 986-987.* Duluth, MN: Association for Consumer Research.

Doyle, John R., Bottomley, Paul, A. (2011). Mixed messages in brand names: Separating the impacts of letter shape from sound symbolism. *Psychology & Marketing* 28(7) 749-762.

Drellishak, S. (2006) Statistical Techniques for Detecting and Validating Phonesthemes. http://depts.washington.edu/uwcl/matrix/sfd/Drellishak%20-%20Phonesthemes.pdf

Fischer Jorgensen, E. (1967). Perceptual dimensions of vowels. *Language Typology and Universals: 667-671*. The Hague: Mouton.

Hall, Kenneth Roland, Oldfield, Richard Charles. (1950). An experimental study on the fitness of signs to words. *Quarterly Journal of Experimental Psychology* 2(2), 60-70.

Heaton, Eugene E. (1967). Testing a new corporate name. *Journal of Marketing Research* 4(3), 279-285.

Hennessey, Judith E., Bell, Theodore S., Kwortnik, Robert J. (2005). Lexical interference in semantic processing of simple words: Implications for brand names. *Psychology & Marketing* 22(1), 51-69.

Horelov, I.I., Sedov, K.F. (1998). Osnov' psiholingvistiki ["Fundamentals of psycholinguistics"]. Moscow: Labirint (in Russian).

Hurdzueva, A. (1973). Zvukovoj simvolizm i factor', vlijajushchije na nego. ["Sound symbolism and factors that influence it"]. *Zbornik nauchnih trudov Moskovskogo IIJA*, 72, 242-255 (in Russian).

Imai, Mutsumi, Kita, Sotaro, Nagumo, Miho, Okada, Hiroyuki. (2008). Sound symbolism facilitates early verb learning. *Cognition* 109(1), 54-65. http://web.sfc.keio.ac.jp/~imai/pdf/soundsymbolism_Cognition.pdf

Irwin, Francis W., Newland, Elizabeth. (1940). A genetic study of the naming of visual figures. *The Journal of Psychology* 9. 3-16.

Keller, Kevin Lane, Heckler, Susan E., Houston, Michael J. (1998). The effects of brand name suggestiveness on advertising recall. *Journal of Marketing* 62, 48-57.

Klank, Linda J. K., Huang, Yau-Huang, Johnson, Ronald C. (1971). Determinants of success in matching word pairs in tests of phonetic symbolism. *Journal of Verbal Learning and Verbal Behavior* 10. 140-148.

Klink, R.R. (2000). Creating brand names with meaning: The use of sound symbolism. *Marketing Letters* 11(1), 5 - 20.

Klink, R.R. (2001). Creating meaningful brand names: A study of semantics and sound symbolism. J*ournal of Marketing Theory and Practice* 9(2), 27-34.

Klink, Richard R. (2009). Gender differences in brand name response. *Marketing Letters* 20(3), 313-326.

Klink, Richard R., Wu, Lan. (2014). The role of position, type, and combination of sound symbolism imbeds in brand names. *Marketing Letters* 25, 13-24. http://download.springer.com/static/pdf/318/art%253A10.1007%252Fs11002-013-9236-3.pdf?auth66=1419086173_099a2daab3dc2effc7c54d87b6368c03&ext=.pdf

Kushneryk, V.I. (1996). Do problem vzaemozvazku phonetichnogo znachenna iz lexuchnum znachennam slova ["To the problem of the connection of phonetic meaning and lexical meaning of the word"]. *Naukovuj visnuk. Germanska Philologija* 2, 14-17. Chernivtsi: Vud-vo Chernivtsi universitet (In Ukrainian).

Lowrey, M.T., Shrum, L.J. (2007). Phonetic symbolism and brand name. *Journal of Consumer Research* 34, 406 - 414. http://faculty.business.utsa.edu/tlowrey/JCR2007.Final.pdf

Lvova, N. (2005). Semantic functions of English initial clusters. *Glottometrics* 9, 21-27.

Lvova, N. (2011). Determining phonetic symbolism of the text. In: Kelih, E., Levickij, V., Matskulyak, Y. (eds.), *Issues in Quantitative Linguistics* 2: 94-104. Lüdenscheid: RAM-Verlag.

Levitskii V., Naidesh O. (2011). The phonosemantic properties of a poetic text. In: B.P. Sherr, J. Bailey, E.V. Kazartsev (eds.), *Formal Methods in Poetics. A Collection of Scholary Works Dedicated to the Memory of Professor M.A. Krasnoperova: 227-235.* Lüdenscheid: RAM-Verlag.

Levickij, V.V. (2013). Phonetic symbolism in natural languages. *Glottotheory* 4(1), 72-91.

Levickij, V. V. (1973). *Semantika i fonetika. Posobije, podgotovlennoje na mateiale eksperimentalnch issledovanij* ["Semantics and phonetics. The book is based on the material of the experimental research"]. Chernovtsi: Izd-vo Chernov. University (in Russian).

Levickij, V.V., Sternin, I.A. (1989). Eksperimental'nje metod' v semasiologii ["Experimental methods in semasiology"]. Voronezh: Izd-vo Voronezh. Universiteta (in Russian).
Levickij, V. V. (2008). *Zvukovoj simvolizm: Mify i real'nost'* ["Sound symbolism: myths and reality"]. Chernivtsi National University (in Russian).
Levitskij, V.V. (1973). Symvolichni znachenna ukrajinskuh holosnyh ta pryholosnyh, *Movoznavstvo 2*, 36-49 (in Ukrainian).
Levickij, V. V. (1989). *Statisticheskoje izuchenie lexicheskoj semantiki* ["Statitical study of lexical semantics"]. Kiev: UMK VO (in Russian).
Levitskij, V.V. (1986). Semantychni i stulistychni funktsiji pochatkovuh spoluchen fonem u nimetski movi ["Semantic and stylistic functions of initial combinations of phonemes in the German language"]. *Inozemna philologija* 81, 75-80 (in Ukrainian).
Leont'ev A. A. (1976). Psiholingvisticheskij aspect jaz'kovogo znachenja ["Psycholinguistic aspect of language meaning"]. In: *Principles and methods of semantic research:* 46-73. Moscow: Nauka (in Russian).
McMurray, Gordon A. (1958). A study of "fittingness" of signs to words by means of the semantic differential. *Journal of Experimental Psychology* 56.4. 310-312.
Monaghan, P., Mattock, K., Walker, P. (2012). The role of sound symbolism in language learning". *Journal of Experimental Psychology. Learning, Memory, and Cognition* 38(5), 1152-1164.
Newman, S. (1933). Further experiments in phonetic symbolism. *American Journal of Psychology* 45, 53–75.
Nielsen, A., Rendall, D. (2012). The source and magnitude of sound-symbolic biases in processing artificial word material and their implications for language learning and transmission. *Language and Cognition* 4(2), 115-125.
http://people.uleth.ca/~d.rendall/Drew_Rendall/Publications_files/Nielsen-Rendall-2012.pdf
Osgood, Charles E., Suci, George J. (1955). Factor analysis of meaning. *Journal of Experimental Psychology* 50(5), 325-338.
Ohala John J., Hinton L., Nichols J. (1994). *Sound Symbolism*. Cambridge: Cambridge University Press.
Orlova, E.V. (1966). O vosprijatii zvykov ["About the perception of the sounds"]. *Razvitije fonetiki russkogo jaz'ka:* 144-154. Moscow.
Otis, K., Sagi, E. (2008). Phonaesthemes: A Corpus-Based Analysis. Conference paper in: *Proceeding of the 30th Annual Meeting of the Cognitive Society:* 65-70. http://csjarchive.cogsci.rpi.edu/Proceedings/2008/pdfs/p65.pdf
Nygaard, L.C., Cook, A.E., Namy, L.L. (2009). Sound to meaning correspondences facilitate word learning. *Cognition* 11(2), 181-186.
Parault, J.S., Schwanenflugel, P.J. (2006). Sound symbolism: A piece in the puzzle of word learning. *Journal of Psycholinguistic Research* 35, 329-351.
Park, Tschang-Zin. (1966). Experimentelle Untersuchungen über Sinnzusammenhang, Lautgestalt und Wortbedeutung. *Psychologische Forschung* 29. 52-88.
Parise, V.C., Pavani, F. (2011). Evidence of sound symbolism in simple vocalizations. *Experimental Brain Research* 214, 373-380.
Pavia, Teresa M., Costa, Janeen Arnold (1993). The winning number: Consumer perceptions of alpha-numeric brand names. *Journal of Marketing* 57. 85-98.
Roper, Carolann W., Dixon, Paul W., Ahren, Elsie H., Gibson, Verner L. (1976). Effect of language and sex on universal phonetic symbolism. *Language and Speech* 19.4. 388-396.
Sapir, E. (1929). A study in phonetic symbolism, *Journal of Experimental Psychology*12(3), 225-239.

Scheerer, Martin, Lyons, Joseph (1957). Line drawings and matching responses to words. *Journal of Personality* 25(3), 251-273.

Shrum, L.J., Lowrey, T.M, Luna, D., Liu, M. (2012). Sound symbolism affects across languages: Implications for global brand names. *International Journal of Research in Marketing* 29, 275-279. http://isiarticles.com/bundles/Article/pre/pdf/1984.pdf

Schmitt, B. H., Pan, Y., Tavassdi, N. (1994). Language and consumer memory. The impact of linguistic differences between Chinese and English. *Journal of Consumer Research* 21(3), 419-431. http://www.jstor.org/stable/2489683?seq=1#page_scan_tab_contents

Slobin, Dan I. (1968). Antonymic phonetic symbolism in three natural languages. *Journal of Personality and Social Psychology* 10(3), 301-305.

Smith, L.B., Sera, D.M. (1992). A developmental analysis of the polar structure of dimensions. *Cognitive Psychology* 24, 99-142.
http://www.iub.edu/~cogdev/labwork/smith_sera1992.pdf

Tamaritz, M. (2008). Exploring systematicity between phonological and context-cooccurrence representations of the mental lexicon. *The Mental Lexicon* 3(2), 259-278.

Tarte, R., Barrit, L. (1971). Phonetic symbolism in adult native speakers of English. Three Studies. *Language and Speech* 76(2), 231-239.

Tarte, Robert D., Boyle, Michael W. (1982). Semantic judgements of compressed monosyllables: Evidence for phonetic symbolism. *Journal of Psycholinguistic Research* 11(3), 183-196.

Thompson, Patrick D., Estes, Zachary. (2011).Sound symbolic naming of novel objects is a graded function. *The Quarterly Journal of Experimental Psychology* 64(12), 2392-2404.

Thorndike, E. (1945). On Orr's hypotheses concerning the front and back vowels. *British Journal of Psychology* 36(1), 10-13.

Weiss, J. (1968). Phonetic symbolism and perception of connotative meaning. *Journal of Verbal Learning and Verbal Behavior* 7, 574-576.

Weiss, J. H. (1963). Role of ‚meaningfulness' versus meaning dimensions in guessing the meaning of foreign words. *Journal of Abnormal and Social Psychology* 66(6), 541-546.

Weiss, J.H. (1963). Further study of the relation between the sound of a word and its meaning. *The American Journal of Psychology* 76(4), 624- 630.

Wertheimer, M. (1958). The relation between the sound of a word and its meaning. *American Journal of Psychology* 71, 412- 415.

Wicker, F.W. (1968). Scaling studies of phonetic symbolism. *Journal of Personality and Social Psychology* 10(2), 175-182.

Westbury, C. (2005). Implicit sound symbolism in lexical access. *Brain and Language* 93(1), 10-19. http://www.sciencedirect.com/science/article/pii/S0093934X04002202

Wichmann, S., Holman, E.W, Brown, C.H. (2010). Sound symbolism in Basic Vocabulary. *Entropy* 12, 844-858.

Wrembel, M. (2010). Sound symbolism in foreign language acquisition. *Research in Language* 8, 175-188.

Wu, Lan, Klink, Richard R., Guo, Jiansheng. (2013). Creating gender brand personality with brand names: The effects of phonetic symbolism. *Journal of Marketing Theory and Practice* 21(3), 319-329.

Yorkston, E. Menon, G. (2004). A sound idea: Phonetic effects of brand names on Consumer Judgements", *Journal of Consumer Research* 31, 43-51.
http://web.stanford.edu/class/linguist62n/yorkston.pdf

Zhernovej, A.N., Levitskij V.V. (1988). Nachal'nje sochetnija fonem v nemetskom jaz'ke. Psycholingvisticheskie issledovanija znachenij slova i ponimanija texta [„Initial combinations of phonemes in German]. In Psiholingvisticheskie issledovanija znachenija slova I ponimanija texta [*Psycholinguistic research of the meanings of words and the comprehension of the text]*, 124-132. Kalinin (in Russian).

Zhuravlov, A. P. (1974). *Foneticheskoje znachenije* ["Phonetic meaning"]. Leningrad: University Press (in Russian).

Zhuravlov, A.P. (1981). *Zvuk i smysl* ["Sound and sense"]. Leningrad: Izd-vo Leningrad university (in Russian).

Other linguistic publications of RAM-Verlag:

Studies in Quantitative Linguistics

Up to now, the following volumes appeared:

1. U. Strauss, F. Fan, G. Altmann, *Problems in Quantitative Linguistics 1*. 2008, VIII + 134 pp.
2. V. Altmann, G. Altmann, *Anleitung zu quantitativen Textanalysen. Methoden und Anwendungen.* 2008, IV+193 pp.
3. I.-I. Popescu, J. Mačutek, G. Altmann, *Aspects of word frequencies*. 2009, IV +198 pp.
4. R. Köhler, G. Altmann, *Problems in Quantitative Linguistics 2.* 2009, VII + 142 pp.
5. R. Köhler (ed.), *Issues in Quantitative Linguistics.* 2009, VI + 205 pp.
6. A. Tuzzi, I.-I. Popescu, G. Altmann, *Quantitative aspects of Italian texts*. 2010, IV+161 pp.
7. F. Fan, Y. Deng, *Quantitative linguistic computing with Perl.* 2010, VIII + 205 pp.
8. I.-I. Popescu et al., *Vectors and codes of text*. 2010, III + 162 pp.
9. F. Fan, *Data processing and management for quantitative linguistics with Foxpro***.** 2010, V + 233 pp.
10. I.-I. Popescu, R. Čech, G. Altmann, *The lambda-structure of texts*. 2011, II + 181 pp
11. E. Kelih et al. (eds.), *Issues in Quantitative Linguistics Vol. 2*. 2011, IV + 188 pp.
12. R. Čech, G. Altmann, *Problems in Quantitative linguistics 3*. 2011, VI + 168 pp.
13. R. Köhler, G. Altmann (eds.), *Issues in Quantitative Linguistics Vol 3.* 2013, IV + 403 pp.
14. R. Köhler, G. Altmann, *Problems in Quantitative Linguistics Vol. 4.* 2014, VI + 148 pp.
15. K.-H. Best, E. Kelih (Hrsg.), *Entlehnungen und Fremdwörter: Quantitative Aspekte.* 2014, IV + 163 pp.
16. I.-I. Popescu, K.-H. Best, G. Altmann, *Unified modeling of length in language.* 2014. III + 123 pp.
17. G. Altmann, R. Čech, J. Mačutek, L. Uhlířová (eds.), *Empirical approaches to text and language analysis.* 2014, IV + 230 pp.
18. M. Kubát, V. Matlach, R. Čech, *QUITA. Quantitative Index Text Analyzer*. 2014, IV + 106 pp.