Glottometrics 32
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, auf **CD-ROM** (in PDF Format) oder in **Buchform** 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**, obtained on **CD-ROM** (in PDF) or in form of **printed copies.**

Herausgeber – Editors

External academic peers for Glottometrics

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. 32 (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 32 (2015) **ISSN 2625-8226**

Contents

Glottometrics 32, 2015, 1-8

A quantitative investigation of English compounds in prose texts

Hanna Gnatchuk (*Alpen-Adria University, Austria*)[1]

Abstract. The given article deals with a quantitative analysis of English compounds in six novels of the first half of the twentieth century. The objective of the article is twofold: a) we are intended to reveal the most frequent structural patterns of the English compounds statistically; b) it is necessary to determine and measure the cohesion for the English compounds. The material of the research is represented by six novels. Each fifth page has been analyzed there. The results have been statistically processed.

Key words: stylistics, functional style, literary style, compounds, cohesion.

1. Introduction

Stylistics is a branch of linguistics which investigates the choices of lexical, grammatical, phonetic and linguistic means with the aim of transferring the ideas and emotions. The focus of our attention is on the *stylistics of the speech* which deals with separate texts by observing how they transfer the contents. Moreover, it is worth mentioning the fact that stylistics is usually divided into *functional stylistics* and *literal stylistics*. As far as functional stylistics is concerned, it studies all functional styles of the language. According to Galperin (1981), there are 5 functional styles: belles-lettres, publicistic, newspaper, scientific styles and the style of official documents. Literal stylistics focuses on the total combination of linguistic means which are characteristic of a certain author's work, literal direction or the whole epoch. There are a considerable number of the experts who are engaged with a stylistic study of the works by Shakespeare, Milton, Byron, Keats, etc.

Functional styles of the language are considered to be the basic categories of stylistics. They are formed in the process of a long-lasting language function and development. The notion "functional style" was firstly formulated by the representatives of the Prague linguistic school at the beginning of the 20th century. In their works they emphasized the fact that the natural language can be divided into a variety of styles according to the communicative function.

In this case, it is better to clarify one point. English researchers refer the notion of "style" to the literal texts. The notion "register" is referred to the other spheres of communication. To be more exact, the register includes the following components:

a) Situation conditions of the communication
b) Oral or written form of the communication
c) Role structure of the communication

For example, it is possible to draw a distinction between the register of the oral unofficial talk, the register of a scientific lecture, sermons, judicial documents, advertisements, commercial correspondences, telephone talks, etc. But the classification of the registers does not exist at all. They are predominantly determined according to the spheres, forms and the

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

relations of the communication participants. It is also possible to consider the registers as the *language variants*.

The literal style has a top position in the hierarchy of the styles according to the emotion interaction. Nevertheless, the problem about the literal style and its place among other styles remains quite debatable: one group of the researchers includes this style in the system of functional styles, the others are against it. The reasons for their objections can be summarized as follows: a) the literal language includes a variety of styles, it does not have specific features which can be available only in this language; b) the literary language has a quite peculiar aesthetic function which is realized in the specific usage of language means. In spite of the stylistic ambiguity and the author's individuality, it has a variety of specific features which help make a demarcation between the literary speech and other styles. Therefore, it is relevant to regard the literal style as one of the functional styles.

At this point we are intended to outline two features of the literal language:

a) Openness to all vocabulary means (both literal and non-literal).

In some cases, the language of the literary style may violate its norms. Here dialectal words, jargons, professional lexemes and other non-literal elements are to be found. In such a way, the prose literature uses a word-stock of all styles (i.e. scientific, official, publicistic, oral, etc). But they are represented in specific combinations and in a modified manner. With the help of all language levels, it is possible to trace the emotional and expressive nature of the literal style. It is the only style where the interaction of all stylistic means is available. This can also be explained by the fact that the literal style is quite rich in different themes. In particular, the other functional styles are aimed at describing one sphere of life or a human activity. The prose piece includes all spheres and phenomena of social life. Hence, the literal style is characterized by a variety of stylistic tints which are realized by the language means.

b) The imagery of the language units at all levels.

In particular, it is possible to find a wide range of the lexis in the metaphorical meanings, the usage of the synonyms of all types, polysemy, etc. In contrast to the other styles, the literary one has its "laws" which concern the perception of a word. In this case, the meaning of a word is determined by the author as well as the genre and compositional peculiarities of a literal piece.

The aim of the literary language is to give a possible interpretation of life events by showing the audience the author's points of view. The English literary style is divided into three substyles: the language of poetry, emotive prose and the language of drama. In its turn, the substyle of the emotive prose deals with a) the style of a novel; b) the style of a novella; c) the style of a story; d) the style of a satirical piece. On the whole, the following linguistic features can be found here:

a) An individual choice of vocabulary and syntax (lexical and syntactic idio-syncrasy);
b) The vocabulary which reflects the author's personal attitude towards the subjects or phenomena.

Our attention should be drawn to the novel. This style contains the features which are characteristic of the literary style in general. But the imagery is not as prolific here as in the poetry. Apart from the difference in size and rhythm, the literary variant of the speech is combined with the colloquial one at both syntactic and lexical levels. In other words, it is a combination of both oral and written languages (monologue – author's speech, dialogue – character's speech).

The writer's speech should correspond to the literary norms of a certain period of the English language. The main character's speech is chosen in order to give her/his appropriate characterization. Nevertheless, it is susceptible to a certain modification throughout the whole prose piece. In such a way, the colloquial speech is not the authentic realization of the human natural speech.

It is also possible to find here the elements of other styles. In particular, the elements of the newspaper style can be revealed in the work "It Can't Happen Here" by Sinclair Levis; the style of official documents in "The Man of Property" by Galsworthy; the scientific style in "The Citadel" by Cronin. All these styles are modified due to the influence of literal prose. Nevertheless, the excerpts written in other styles can be regarded as interpolation (not as a part of a style).

The literal prose appeared late in the history of the English literal language. It is a well-known fact that the literal prose did not exist in the earlier Anglo-Saxon literature in so far as it consisted only of poetry, religious and war songs. The translations of the Bible and the Saints' Lives are considered to belong to the first English literal prose.

The literature of the Middle Ages was quite didactic. It was represented by the translations from Latin of the literal works. In spite of it, the Norman Conquest (1066) had a negative influence upon the development of the Anglo-Saxon literature. The literal prose renovated its existence only in the second half of the 15th century. In that case it is possible to trace the chronicles describing the life and adventures of the legendary kings and knights.

The 16th century is famous for huge progress in all spheres of social life which led to the dynamic development of the English literal prose. A variety of the Latin and Greek translations played a key role in order to create stylistic norms of the literal prose at that period. The fundamental contribution into the creation of the typical features for the literal prose was made by Shakespeare. Nevertheless, the literal prose of the 16th century was not formed as a separate functional style. In this case, it is possible to admit the tendency that regarded the colloquial speech of the English language to be of a lower quality and unworthy of being represented in the prose literature. But the prose cannot exist without the character's direct speech. Hence, a considerable number of the prose works of that period were represented by biographies, the reports on travelling, essays on various philosophical and aesthetic problems. Finally, it is the 18th century that gave a rise to the intensive development of the prose as a whole. Historically speaking, this epoch was characterized by a political and religious struggle. As a result, many written pieces were of publicistic character. At the given period of time, it is possible to admit the ultimate formation of English belles-lettres (or literal) functional styles. This was contributed by numerous analyses of literature works both in the earlier epochs and the texts of the 18th-20th centuries where the development in all branches of the human life was observed.

2. A statistical analysis of English compounds in the prose texts

The purpose of the present research is to find the most frequent structural patterns of the English compounds in six novels statistically.

The data for the research consist of six novels by different authors (the first half of the twentieth century):

1) Theodore Dreiser "Jennie Gerhardt"
2) John Galsworthy "The Forsyte Saga"
3) Somerset Maugham "Theatre"
4) Jack London "The White Fang"
5) Aldoux Huxley "Along the road"

6) Aldoux Huxley “Over the river”

We have analyzed each fifth page of the above-mentioned novels. As a result, 968 compounds have been written out. In such a way, 18 structural patters of the compounds have been found and are given below:

1) Noun + Noun: *bedroom, railroad, coal merchant, cave door, guidebook*;
2) Adjective + Participle 2: *middle-aged, gold-headed, good-natured, heavy-laden, rose-lined;*
3) Verb(ing) + Noun: *dining-room, laughing-stock, frying-pan, sleeping-tablet, smoking-room*
4) Noun + Participle 2*: sun-lit, putty-faced, dust-colored, sherry-colored, lynx-eyed*;
5) Adjective + Noun: *old/shoe, large-mass, rapid-fire, blackbird, high-water*;
6) Adverb + Participle 2: *well-rounded, newly/married, well/bred, well-aimed, well-known;*
7) Noun + Adjective: *paper-thin, life-long, skull-deep, sea-green, wine-red*;
8) Adjective + Verb + ing: *better-looking, funny-looking, strange-seeming, hard-closing, prosperous-looking;*
9) Numeral + Noun: *first-rate, thirty-night, first-hand, third-class;*
10) Preposition + Noun: ou*tside, insight, upstairs, nearpoint, indoors*
11) Verb + Preposition: *make-up, lookout, knockout;*
12) Preposition + Participle 2: *up-lifted, overcrowded, overpopulated*
13) Adverb + Preposition: *hereafter*
14) Noun +Verb (ing): *bridge-building, tea-planting, blood-curdling, horse-breeding, toilet-training;*
15) Adjective + Adjective: *blue-hot, gold-glacious, velvet-black, pale-brown, grey-white;*
16) Noun + Preposition: *runner-up*
17) Participle +Verb +ing: *distinguished-looking*
18) Noun + Noun + Noun: *oak extension table, snowshoe rabbit*

Table 1 includes the rank-frequency distribution of the compounds, their patterns and frequencies.

Table 1
Rank frequency distribution of English compounds in the novels

Rank	**Pattern of compound**	**Frequency**
1	Noun+Noun	528
2	Adjective+Participle 2	96
3	Verb+ing+Noun	91
4	Noun+Participle 2	45
5	Adjective+Noun	43
6	Preposition+Noun	31
7	Adverb+Participle 2	27
8	Noun+adjective	26
9	Adjective+Verb+ing	21
10	Numeral+Noun	21
11	Adjective+Adjective	11

12	Noun+Noun+Noun	10
13	Verb+Preposition	6
14	Noun+Verb+ing	5
15	Preposition+Participle 2	3
16	Participle+Verb+ing	2
17	Adverb+Preposition	1
18	Noun+Preposition	1

In such a way, the minimal value of this range is 1, the maximal 18. In this case we have 18 the most frequent compound patterns. We may ask whether there is some regularity in the use of these types. To this end we rank them – as is usual in linguistics – and find either a discrete distribution expressing this order or a simple function (i.e. a not normalized sequence or function). Since ranking is usually associated with the power function, $y = ax^b$, introduced already by G.K. Zipf, we apply this function and obtain the fitting as shown in Table 2. The determination coefficient $R^2 = 0.98$ yields an excellent fit hence further functions can be omitted.

Table 2
Fitting the power function to the ranking of compound types in English

Rank	**Frequency**	**Power function**
1	528	522.26
2	96	131.10
3	91	58.40
4	45	32.91
5	43	21.09
6	31	14.66
7	27	10.78
8	26	8.26
9	21	6.53
10	21	5.29
11	11	4.38
12	10	3.70
13	6	3.14
14	5	2.71
15	3	2.36
16	2	2.07
17	1	1.84
18	1	1.64
$a = 522.2600$, $b = -1.9940$, $R^2 = 0.98$		

The problem of fitting data by a function or a distribution is rather philosophical. We simply use some mathematical models to show some order in the data (cf. Mačutek, Altmann 2007). Models are neither true nor false, they are adequate for the data or not. We could apply here also a discrete distribution, e.g. the negative hypergeometric but this is not necessary; we merely search for a kind of order.

3. Cohesive types of English compounds in the prose texts

Fan and Altmann (2007) admit a problematic character of such a linguistic phenomenon as "cohesion" in so far as it differs in all natural languages. Aiming to investigate the cohesion, we follow the procedures undertaken by Fan and Altmann (2007). In particular, we select a domain of our research (6 prose novels), perform the analysis of English compounds (write out all the compounds in these novels) and scale their cohesion (compute the results).

The aim of the analysis is to measure (or scale) the cohesion for English compounds statistically.

The material of the research consists of six prose novels of the first half of the twentieth century. 968 English compounds have been under analysis.

We have found the following cohesive types of English compounds: joining, joining with an inserted element, hyphenized and blank compounds:

Joining is a kind of cohesion in which two lexemes are written together: *honeymoon, kitbag, riverside, bedroom, scarecrow, staircase, eyebrow, landscape;*

Joining with an inserted element is a kind of cohesion in which two words are put orthographically together with the help of an inserted element: *sportsman, statesman, washerwoman, sportscar, beeswax.*

Hyphenization is a type of the compound whose parts are joined by means of a hyphen, their elements can be without a fugue: *inn-keeper, bed-rock, roof-line, snow-field, oil-men;*

Blank compounds. The elements of the compounds are written separately*: tea plantation, motor car, passenger station, law firm, oak extension table.*

The results are presented in Table 3 where the rank, the type of the cohesion, the frequency of the cohesive types are displayed.

Table 3
Rank frequency distribution

Rank	**The type of cohesion**	**The total number**
1	Hyphenized	543
2	Joining	302
3	Blank	79
4	Joining with an inserted element	5

Dwelling on the cohesion of the analyzed compounds, we have revealed 4 types of cohesion in English prose texts. The most frequent ones - hyphenized (58.4%) and joined compounds (32.5 %), the least frequent are blank (8.6%) and joining with an inserted element (0.5 %). Our analysis has confirmed the results obtained by Fan/Altmann (2007) on the basis of newspaper texts that three cohesive types, namely, compounds with joining cohesion (the elements are written together: *footnote, sunlight, bridegroom, sideboard, nutshell,* etc), hyphenized compounds (two elements are linked with the help of a hyphen: *smoking-room, laughing-stock, looking-glass, night-club, lunch-time,* etc) and compound blanks (the elements are written without a fugue: *law firm, carriage company, garden entrance*) dominate in English prose texts.

Moreover, it is worth mentioning that our analysis has revealed only one type of compounds consisting of three elements: Noun + Noun + Noun (1.2%). It shows that the compounds with more than two components are not characteristic of English prose texts. Nevertheless, the analysis of Hungarian and German newspapers has shown quite different results (Fan, Altmann, 2007). German and Hungarian languages tend to have more complex compounds in

newspaper style. Therefore, it remains quite actual to compare the cohesion in English compounds in different functional styles (scientific, newspaper styles) and make a comparison with the other languages.

Conclusions:

- It is possible to distinguish three more frequent types of compounds in English prose: joining, hyphenized compounds and blank ones.
- The highest frequency is observed in hyphenized compounds whereas the lowest one is characteristic of blank ones. This can be ascribed to the development of English.
- The results coincide with Fan/Altmann's outcomes (2007). In particular, the given three types of cohesions were found in English newspapers.

Since here we ranked the compounds according to their cohesion, it can be expected that this relationship can also be expressed by means of a function. However, this is a secondary classification, hence the power function alone is not sufficient: The difference between observed and expected frequency in the highest rank is too great. As is usual in this analysis, we add a second factor and obtain

$$(1) \qquad y = cx^{a + b \ln x}.$$

the so-called Zipf-Alekseev function. The logarithmic addition is proposed on the basis of psychological considerations. Fitting (1) to our data, we obtain the results presented in Table 4.

Table 4
Fitting the Zipf-Alekseev function to the ranks of compound types ordering

Rank	**The type of cohesion**	**Number of compounds**	**Zipf-Alekseeev function**
1	Hyphenized	543	542.91
2	Joining	302	302.90
3	Blank	79	73.36
4	Joining with an inserted element	5	6.57
	a = 0.8334, b = -2.4169, c = 542.9076, R^2 = 0.9991		

Evidently, the fitting is very satisfactory. In this way we obtained the first two models of compound behavior in English.

References

Arnold, I. V. (2002). *Stilistika. Sovremenn'j anglijskij jaz'k [Stylistics. The Modern English language]*. Moskva: Izd-vo "Flinta".

Baldick, C. (2008). *Oxford Concise Dictionary of Literary Terms*. Oxford: Oxford University Press.

Fan, F., Altmann, G. (2007). Some properties of English compounds. In: Kaliuščenko, V., Köhler, R., Levickij, V. (eds.*), Problems of typological and quantitative lexicology: 177-189*. Černovcy: Ruta.

Galperin, I. (1981). *Stylistics.* Moscow: Vysshaja shkola..

Mačutek, J., Altmann, G. (2007). Discrete and continuous modelling in quantitative linguistics. *Journal of Quantitative Linguistics 14(1), 81-94.*

Glottometrics 32, 2015, 9-20

A Quantitative Investigation of the Genre Development of Modern Chinese Novels

Cong Zhang, Haitao Liu[*]

Abstract: This study mainly investigates the genre development of modern Chinese novels since 1919 from a perspective of quantitative linguistics. We choose the *a-index* and *lambda* as our quantitative indicators. Firstly, we test their applicability to distinguish different genres of texts written in Chinese. The results show that both the indicators work, and *lambda* performs better than the *a-index*. Then we obtained the data of modern Chinese novels from 1919 to 2015 with regard to *lambda*. Based on the findings above and the diachronic data, we conclude that the change of the genre characteristics of modern Chinese novels is not significant since 1919.

Keywords: Modern Chinese novels; Genre; The a-index; Lambda;

1. Introduction

It is generally known that novel is an important form of literature and it is the most popular literary genre among people all over the world, which can be gathered from the fact that most former winners of the Nobel Prize in literature are novelists. Since there is a slight difference between the definition of "novel" in China and in English, we decide to take the Chinese definition of "novel" as our standard for modern Chinese novels. The sixth edition of *The Contemporary Chinese Dictionary* (2012, p.1435) defines "novel" as a narrative genre of literature, presenting specific social life via the characterization and the description of figures, plots, environments, etc. Starting from *the New Culture Movement*, the history of modern Chinese novels spans almost 100 years until now. And during this period, the Chinese society has changed dramatically. We went through *the Warlord Era*, *the Anti-Japanese War*, *the Chinese Civil War*, and finally *the reunification in 1949*. After the unification, we were isolated from other countries in the world for several decades, but the Chinese society was still in flux. Along with *the reform and opening up policy* in the end of 1978, we reconnected with the whole world. Now China is the world's second largest economy body, known as "*the workshop of the world*". With the development of the society, the Chinese language is also changing continuously. A lot of old words die out, while new words emerge or new meanings to existing words are added. From the definition above, we know that novel reflects the changes of our society, as well as language itself, which leaves the question below for us:

Question 1: With all these changes in the Chinese society and the Chinese language in the past ten decades, have the genre characteristics of modern Chinese novels also significantly changed?

Most literary studies in China use qualitative methods to analyze the form and content of literary works. Empirical data and statistical evidence (quantitative methods) are rarely used, which results in subjective conclusions about many issues drawn by researchers. In order to

[*] Address correspondence to: Haitao Liu, Department of Linguistics, Zhejiang University, 310058, Hangzhou, Zhejiang, China. Email address: lhtzju@gmail.com

obtain a relatively objective and fair conclusion, we will not consider the form or content of the novels in this study, but consider all modern Chinese novels as a whole from the perspective of linguistics, and all these novels constitute the genre that differs from the genre of prose, poem and government work report, etc. From the empirical point of view, we use quantitative methods to study the genre of modern Chinese novels based on the related indicators of word frequency, which leads to the second question we are to answer:

Question 2: *Are word frequencies genre indicators at all? Or to be more specific, is there an indicator of word frequency in quantitative linguistics which can measure the genre change in the Chinese language?*

2. The quantitative indicators, materials and methods

2.1. The quantitative indicators

On the one hand, when we set genre as our research target, we have to draw a conclusion in accordance with the whole text rather than a part of it. For example, one novel can contain a poem, a prose, a news report, a dialogue, and so on. If we only extract these parts of a novel to analyze the genre of it, we may obtain a wrong conclusion, so we must study the whole text in order to gain its genre type. In addition, all the modern Chinese novels investigated in this paper are randomly chosen from 90 novels that were published in the years 1919 to 2015.

On the other hand, generally modern Chinese novels are divided into full-length novels, medium-length novels and short stories in terms of their text length. As the length of each text (measured by the number of words in texts) in modern Chinese novels varies tremendously, we must choose indicators that are not affected or only slightly affected by the length of texts in our research. Finally, in this study, we choose the *a-index* and *lambda* as the quantitative indicators, and we will briefly introduce these two indicators in the following part:

2.1.1. The *a-index*

The *a-index* is derived from the h-point. Popescu et al. (2009a, p. 24) define the h-point as a fixed point in the rank-frequency distribution of words formed by word frequency statistics, and it represents the fuzzy boundary between the content words and function words in the rank-frequency distribution of word forms. Its mathematical definition is:

$$h = \begin{cases} r, & if\ there\ is\ an\ r = f(r) \\ \frac{f(i)r_j - f(j)r_i}{r_j - r_i + f(i) - f(j)}, & if\ there\ is\ no\ r = f(r) \end{cases} \quad (1)$$

r_i and r_j represent any rank numbers of two neighboring words in a rank-frequency distribution, while $f(i)$ and $f(j)$ represent the corresponding word frequencies of r_i and r_j respectively.

Since the value of the h-point is associated with text length *N*, Popescu et al. (2009b) define the *a-index* as:

$$a = N/h^2 \quad (2)$$

Popescu et al. (2009b, p.23) state that the *a-index* is a textual characteristics of a language, and it is not affected by text length.

2.1.2. *Lambda*

Lambda is derived from *arc length (L)*, which is also based on the rank-frequency distribution of words formed by word frequency statistics. Popescu et al. (2009a, p.49) define *arc length (L)* as the sum of all Euclidean distances between all adjacent frequencies in a rank-frequency distribution, and the mathematical definition is:

$$L = \sum_{i=1}^{V-1}[(f_i - f_{i+1})^2 + 1]^{1/2} \quad (3)$$

In order to normalize it, Popescu et al. (2011, p.2) define the indicator *lambda* as:

$$\Lambda = \frac{L(Log_{10}N)}{N} \quad (4)$$

Popescu et al. (2011) suppose that *lambda* is only slightly affected by text length.

2.2. Materials and methods

In order to answer the second question mentioned above, we randomly chose 10 novels spanning from 1984 to 2015 (see the third column of *Appendix 2*), 10 poems, 10 prosaic works and 10 government work reports as the materials of our study (see *Appendix 1*). Then we start to process the texts: to begin with, we use the automatic segmentation software *segtag*[①] to recognize all the single words text by text. Next we use *QUITA*[②] to gain the rank-frequency distribution statistics of each text. Finally we use the formulas mentioned above to obtain the value of the *a-index* and *lambda* of each text.

We apply the one-way analysis of variance (ANOVA) to test the null hypothesis that our four samples are drawn from population with the same mean value. In other words, that the change of text's genre by the means of *a-index* cannot be detected. If the result shows that the difference of each text's *a-index* is significant, this indicator is capable of detecting genre change of Chinese. Otherwise, the null hypothesis is true. And after we test the *a-index*, we will also test *lambda* in the same way.

Since it is a parametric test, let us see whether these parameters meet the conditions of applying the one-way ANOVA.

Table 1
Tests of normality of the *a-index* of each genre

	Genre	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
a	prose	.184	10	.200*	.964	10	.834
	poem	.159	10	.200*	.937	10	.521
	government work report	.216	10	.200*	.925	10	.397
	novel	.196	10	.200*	.911	10	.286
*. This is a lower bound of the true significance.							
a. Lilliefors Significance Correction							

① http://cloudtranslation.cc/segtag.html
② https://code.google.com/p/oltk/

Table 2
Tests of Normality of *lambda* of each genre

	Genre	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistic	df		Statistic	Statistic	df
Λ	prose	.297	10	.013	.783	10	.009
	poem	.196	10	.200*	.876	10	.117
	government work report	.139	10	.200*	.937	10	.522
	novel	.255	10	.064	.874	10	.111
*. This is a lower bound of the true significance							
a. Lilliefors Significance Correction							

Table 1 and Table 2 show that both *a-index* and *lambda* of each sample obey normal distribution ($p > 0.05$) except the lambda value of prose ($p = 0.009$), so we remove the lambda value of prose from our data. Second, as to homogeneity of variances, the result of *a-index* is $p = 0.032$, and *lambda* is $p = 0.367$. Since the homogeneity of variances of *a-index* is significant, we use the Brown-Forsythe test instead of the one-way ANOVA to examine *a-index* of each genre, and the Tamhane's T2 test to do the post hoc test.

As to the first question mentioned above, when we complete the tests of the *a-index* and *lambda*, we choose the indicator with better performance to conduct the following investigation.

In order to investigate whether there is a significant change of the genre characteristics of modern Chinese novels during the years from 1919 to 2015, we divide this period into three stages by means of the major social and historical events in China. Thus we have the first period from 1919 (*the May 4th Movement*) to 1949 (*the founding of the People's Republic of China*), the second period from 1949 to 1984 (the decision of the economic system reform has been passed), and the last part from 1984 till now. And these events are carefully chosen. *The May 4th Movement* in 1919 has influenced the Chinese culture, political development, socio-economic trends, education and so on, so it is considered as the beginning of China's new democracy. Although modern Chinese novels started from the *New Culture Movement*, it became flourishing only after the *May 4th Movement* in 1919, so we choose the *May 4th Movement* in 1919 as our first boundary. The second boundary is the founding of the People's Republic of China in 1949, which is the most important historical event in modern China, and certainly has a huge impact on modern Chinese novels. The last dividing point is in the year 1984. In the year 1978, China began to carry out *the reform and opening up policy*, but the policy was mainly conducted in rural areas of China rather than nationwide during 1978-1983, and the decision of the economic system reform, made by the central committee of the communist party of China, has not been passed until the year 1984, when the policy has been carried out throughout China. Economy has a tremendous influence on culture during the time of peace. For this reason, we choose 1984 as our third boundary.

We randomly choose 10 representative novels from each of these three periods respectively as the materials of our study. For the detailed reference of the texts under study see *Appendix 2*. The testing procedure is the same as above:

we present our null hypothesis first: *the genre of modern Chinese novels have not changed significantly in these three periods (the difference of each periods' indicator is not significant)*. and then we use SPSS to perform the one-way ANOVA. If the result shows that the difference of each periods' indicator is significant, we can conclude that the genre characteristics of modern Chinese novels do have changed significantly from 1919 till now, otherwise the genre of novels have not changed significantly during the past nearly 100 years.

3. Results and discussions

For each text under study we compute the values of both indicators which are summarized in *Appendix 3*. Table 3 shows the descriptive statistics of the *a-index* of each genre, while Table 4 shows the results of the Brown-Forsythe test for *a-index* of each separate genre group.

Table 3
The descriptive statistics of the *a-index* of each genre

	N	Mean of the a-index	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum of the a-index	Maximum of the a-index
					Lower Bound	Upper Bound		
prose	10	11.0395	2.8348	.8964	9.0116	13.0673	6.5439	16.6389
poem	10	8.5273	2.1031	.6651	7.0228	10.0317	5.4167	11.3719
government work report	10	8.6758	.9463	.2992	7.9989	9.3528	7.3700	10.1742
novel	10	7.5940	1.0513	.3325	6.8419	8.3460	5.9612	8.8806
Total	40	8.9591	2.2345	.3533	8.2445	9.6738	5.4167	16.6389

Table 4
The Brown-Forsythe test's results of the *a-index* of each genre

a				
	Statistic[a]	df1	df2	Sig.
Brown-Forsythe	5.955	3	21.839	.004
a. Asymptotically F distributed				

The Brown-Forsythe test results of each genre's *a-index* suggest that the difference of the four genres' *a-index* is significant ($p < 0.05$), which means that the *a-index* can detect the genre change of Chinese.

According to Table 3, we can see that the order of the mean *a-index* of the four selected genres is as follows: $\bar{a}_{prose} > \bar{a}_{government\ work\ report} > \bar{a}_{poem} > \bar{a}_{novel}$. Popescu et al. (2009b, p.23) consider that: "Smaller *a-index* is a symbol of analytism," namely, less word forms in the text, which means that the word forms in the text are more likely to be repeated. Thus, the order of analytism (the chance the word forms in the text to be repeated) of genres in Chinese is: prose > government work report > poem > novel.

Also according to Table 3, we find that the mean *a-index* of each genre's 95% confidence interval overlaps, which means that we still do not know which of the specific genre groups differ from each other. Therefore, it is necessary to do the post hoc test. We continue to use SPSS to perform the Tamhane's T2 test of the *a-index*, and get the following results:

Table 5
Multiple comparisons of the *a-index* of each genre

Dependent Variable: a Tamhane						
(I) genre	(J) genre	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
prose	poem	2.5122	1.1162	.209	-.8177	5.8421
	government work report	2.3636	.9451	.164	-.6581	5.3853
	novel	3.4455*	.9561	.023	.4135	6.4775
poem	prose	-2.5122	1.1162	.209	-5.8421	.8177
	government work report	-.1486	.7293	1.000	-2.4221	2.1250
	novel	.9333	.7435	.793	-1.3616	3.2282
government work report	prose	-2.3636	.9451	.164	-5.3853	.6581
	poem	.1486	.7293	1.000	-2.1250	2.4221
	novel	1.0819	.4473	.149	-.2407	2.4045
novel	prose	-3.4455*	.9561	.023	-6.4775	-.4135
	poem	-.9333	.7435	.793	-3.2282	1.3616
	government work report	-1.0819	.4473	.149	-2.4045	.2407
*. The mean difference is significant at the 0.05 level.						

The Tamhane's T2 test shows significant difference ($p < 0.05$) between the mean *a-index* values of proses and novels, and the other ones are not significant. This means that though the *a-index* can detect the genre change of Chinese, its distinction between individual genres is not very satisfactory.

Next, we use the same method to deal with the *lambda* value of each genre. We obtain the descriptive statistics (cf. Table 6) and the one-way ANOVA's results of each genre's *lambda* (cf. Table 7).

Table 6
The descriptive statistics of *lambda* of each genre

Lambda								
	N	Mean of lambda	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum of lambda	Maximum of lambda
					Lower Bound	Upper Bound		
poem	10	1.4755	.1247	.0394	1.3863	1.5647	1.3508	1.7037
government work report	10	1.0598	.0818	.0259	1.0013	1.1183	.9365	1.1763
novel	10	.7880	.1569	.0496	.6758	.9002	.5608	1.0610
Total	30	1.1078	.3118	.0569	.9914	1.2242	.5608	1.7037

Table 7
The one-way ANOVA's results of *lambda* of each genre

Lambda					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2.398	2	1.199	76.781	.000
Within Groups	.422	27	.016		
Total	2.820	29			

The one-way ANOVA shows that the difference between each genre is significant, $F(2, 27) = 76.781$, $p < 0.001$, $R^2 = 0.85$, which means *lambda* can also detect the genre change of Chinese. Popescu et al. (2011, p. 8-9) deem that text with a greater *lambda* value tends to have greater vocabulary richness. *Lambda* is also related to the frequency structure of word forms, so texts with greater *lambda* tend to have more complicated frequency structure of word forms. From Table 6, we obtain the order of each genre's mean *lambda*: $\Lambda\overline{\Lambda}_{\text{poem}} > \Lambda\overline{\Lambda}_{\text{government work report}} > \Lambda\,\overline{\Lambda}_{\text{novel}}$, namely the order of each genre's vocabulary richness and the complexity of word forms' frequency structure is poem > government work report > novel.

Table 8 shows that the mean *lambda* of each genre's 95% confidence interval rarely overlaps (except that the lower limit of prose and the upper limit of poem slightly overlap). In order to get the exact details, we conduct the LSD test of *lambda*, and get the following results:

Table 8
Multiple comparisons of *lambda* of each genre

Dependent Variable: Lambda LSD						
(I) genre	(J) genre	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
poem	government work report	.4157*	.0559	.000	.3011	.5304
	novel	.6875*	.0559	.000	.5729	.8022
government work report	poem	-.4157*	.0559	.000	-.5304	-.3014
	novel	.2718*	.0559	.000	.1571	.3864
novel	poem	-.6875*	.0559	.000	-.8022	-.5729
	government work report	-.2718*	.0559	.000	-.3864	-.1571
*. The mean difference is significant at the 0.05 level.						

The LSD test shows that the difference between the mean *lambda* values of any two genres is significant, which means that in contrast to the *a-index*, *lambda* yields a better classification on genres in Chinese. Therefore, we decide to use *lambda* to detect whether the change of the genre characteristics significantly happens in modern Chinese novels during the last nearly

100 years.

The *lambda*'s distribution of residuals of the 30 modern Chinese novels also obeys normal distribution, and their test of homogeneity of variances is $p = 0.149$. So we also use SPSS to compute the results, and we obtain the descriptive statistics (cf. Table 9) and the one-way ANOVA's results of *lambda* (cf. Table 10).

Table 9
The descriptive statistics of *lambda* of the three periods

Lambda								
	N	Mean of lambda	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum of lambda	Maximum of lambda
					Lower Bound	Upper Bound		
19-49	10	.8623	.2186	.0691	.7060	1.0187	.5700	1.1941
49-84	10	.7319	.2163	.0684	.5771	.8866	.4728	1.0977
84-	10	.7880	.15688	.0496	.6758	.9002	.5608	1.0610
Total	30	.7941	.1998	.0365	.7195	.8687	.4728	1.1941

Table 10
The one-way ANOVA's results of *lambda* of the three periods

Lambda					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.086	2	.043	1.078	.354
Within Groups	1.073	27	.040		
Total	1.158	29			

The one-way ANOVA shows that, the difference among the three periods is not significant, $F(2,27) = 1.078$, $p > 0.10$, $R^2 = 0.07$, which means that the genre characteristics of Modern Chinese novels has not significantly changed since 1919. As shown in Table 9, the order of the three periods' mean *lambda* is: $\overline{\Lambda}\Lambda_{1919\text{-}1949} > \overline{\Lambda}\Lambda_{1984\text{-}now} > \overline{\Lambda}\Lambda_{1949\text{-}1984}$ and it is oscillating - this means that the genre characteristics of modern Chinese novels have no clear trend in the past ten decades since 1919.

Why did the genre characteristics of modern Chinese novels remain unchanged with the transformation of the society? We believe that many factors are responsible for it and one of the most vital factor is that modern Chinese novels did not appear abruptly in the history of Chinese literature. It has been derived from novels in vernacular Chinese, and the latter has emerged as early as the *Tang Dynasty*. The emergence of modern Chinese novels is not a kind of mutation (e.g. Chinese new poem is a kind of mutation in the history of Chinese literature), but a kind of gradual change, which is unlikely to cause significant change of the genre characteristics.

3. Conclusion

According to the results of our tests, we have drawn the following conclusions:

1. Word frequencies are genre indicators. Both the *a-index* and *lambda* can detect the genre change of Chinese, and *lambda* is a better indicator than the *a-index* in the classification of genres of Chinese.

2. In Chinese, the order of analytism (the chance that word forms in the text to be repeated) of genres is: prose > government work report > poem > novel. The order of each genre's vocabulary richness and the complexity of frequency structure of word forms is: poem > government work report > novel.
3. The change of the genre characteristics of modern Chinese novels is not significant since 1919.

Our conclusion is merely based on the analysis of the existing data and their testing results. It may not be precise, and only the exhaustive research for all texts written during that period is the proper way to get more precise results, which will be the focus of our follow-up research on this issue. Further, frequency of words is only one of the uncountable characteristics a genre may display.

Acknowledgments

This work is partly supported by the National Social Science Foundation of China (Grant No. 11&ZD188).

References

Jiang, L., Tan, J., & Cheng, R. (eds.). (2012). *Xiàndài hànyǔ cídiǎn (The Contemporary Chinese Dictionary). (6th edition)*. Beijing: The Commercial Press.

Popescu, I., Mačutek, J., & Altmann, G. (2009a). *Aspects of Word Frequency*. Lüdenscheid: RAM-Verlag.

Popescu, I., Altmann, G., Grzybek, P., Jayaram, B. D., Köhler, R., Krupa, V., Mačutek, J., Pustet, R., Uhlířová, L., & Vidya, M. N. (2009b). *Word Frequency Studies*. Berlin/New York: de Gruyter.

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

Software

Kubát, M., Matlach, V. (2014). QUITA (Version 1.1.9.0) [Computer Software]. Olomouc: Palacký University. Available from: oltk.upol.cz/software.

Shi, X. (2005). Segtag [Computer Software]. Xiamen: Xiamen University. retrieved from http://cloudtranslation.cc/segtag.html.

SPSS (Version 20.0.0.) [Computer Software]. (2011). IBM Corporation.

Appendix 1

Randomly selected proses, poems and government work reports of the state council

	Prose	**Poem**	**Government work report of the state council**
1	Shijiewujiqiguan	Heikediguo	2000
2	Xixinglushangzuogongliu	Kezaiqiangshangdewuyixiang	2002
3	Jinci	Laci	2004
4	Tonglizhisi	Alaisiledecaifeng	2006
5	Renshanzhishui	Chenmodemianyang	2008
6	Zhengdingsanri	Haidetuan	2010
7	Luzhendeheiyeyubaitian	Woshiyigerenxingdehaizi	2011
8	Gancaisuiyue	Heshuiduiyin	2012
9	Guanyusiguanyushengdeduihua	Honghuacao	2013
10	Zhenxifennu	Yikedongridezhongziqidaizhexinsheng	2014

Appendix 2

Each 10 representative novels randomly chosen respectively from 1919-1949, 1949-1984 and 1984-2015

	1919-1949	**1949-1984**	**1984-2015**
1	AQzhengzhuan	Sanliwan	Honggaoliang
2	Chenlun	Gaoyubao	Huozhe
3	Mengke	Tiemuqianzhuan	Fengrufeitun
4	Chunchan	QIngchunzhige	Sanchongmen
5	Linjiapuzi	Shanxiangjubian	Menglihualuozhiduoshao
6	Jia	Dihouwugongdui	Shouji
7	Biancheng	Kudou	Yibanshihuoyanyibanshihaishui
8	Luotuoxiangzi	Jinguangdadao Part1	Yidijimao
9	Hulanhezhuan	Qiaochangzhangshangrenji	Luohun -80 houdexinjiehunshidai
10	Weicheng	Jingxindongpodeyimu	Lufanyanshi

Appendix 3

Text length, the *a-index* and *lambda* of all novels, proses, poems and government work reports

Text name	Genre	Text length (N)	a	Lambda
AQzhengzhuan	novel.1919-1949	13597	7.353705	0.993822
Chenlun	novel.1919-1949	12683	7.732358	1.194132
Mengke	novel.1919-1949	16471	7.300166	1.112318
Chunchan	novel.1919-1949	8924	7.081135	1.056063
Linjiapuzi	novel.1919-1949	14311	6.963614	0.906363
Jia	novel.1919-1949	146099	7.279673	0.570037
Biancheng	novel.1919-1949	32064	6.734720	0.76747
Luotuoxiangzi	novel.1919-1949	87626	6.862401	0.664124
Hulanhezhuan	novel.1919-1949	64655	7.317225	0.682686
Weicheng	novel.1919-1949	130943	7.515094	0.676401
Sanliwan	novel.1949-1984	90129	5.861668	0.536156
Gaoyubao	novel.1949-1984	70138	5.667002	0.472806
Tiemuqianzhuan	novel.1949-1984	21905	6.985013	0.968302
QIngchunzhige	novel.1949-1984	220369	7.807859	0.628638
Shanxiangjubian	novel.1949-1984	62186	7.443895	0.786126
Dihouwugongdui	novel.1949-1984	172603	7.277914	0.637015
Kudou	novel.1949-1984	160081	8.309231	0.643846
Jinguangdadao Part1	novel.1949-1984	217261	7.067288	0.559544
Qiaochangzhangshangrenji	novel.1949-1984	16785	7.762747	1.097749
Jingxindongpodeyimu	novel.1949-1984	29574	7.947863	0.98853
Honggaoliang	novel.1984-2015	30920	8.646436	1.049998
Huozhe	novel.1984-2015	57251	5.961162	0.560784
Fengrufeitun	novel.1984-2015	284542	8.880559	0.720781
Sanchongmen	novel.1984-2015	94251	7.991463	0.805668
Menglihualuozhiduoshao	novel.1984-2015	97483	6.549516	0.731240
Shouji	novel.1984-2015	67350	7.307943	0.659920
Yibanshihuoyanyibanshihaishui	novel.1984-2015	28569	7.848387	1.060956
Yidijimao	novel.1984-2015	18930	6.145605	0.761446
Luohun - 80 houdexinjiehunshidai	novel.1984-2015	94057	8.215303	0.749209
Lufanyanshi	novel.1984-2015	136444	8.393326	0.780189
Shijiewujiqiguan	prose	1662	11.54167	1.733085
Xixinglushangzuogongliu	prose	1962	13.62500	1.841681
Jinci	prose	1309	11.87302	1.90375
Tonglizhisi	prose	1507	11.39509	1.861414
Renshanzhishui	prose	928	11.45679	2.001712
Zhengdingsanri	prose	2300	9.782609	1.682255
Luzhendeheiyeyubaitian	prose	2201	8.084481	1.695182
Gancaisuiyue	prose	599	16.63889	2.029977
Guanyusiguanyushengdeduihua	prose	3072	6.543905	1.051059
Zhenxifennu	prose	605	9.453125	1.743147
Heikediguo	poem	208	5.777778	1.350792
Kezaiqiangshangdewuyixiang	poem	135	8.437500	1.47668

Laci	poem	94	10.44444	1.36949
Alaisiledecaifeng	poem	95	10.55556	1.560352
Chenmodemianyang	poem	86	11.37190	1.703681
Haidetuan	poem	410	7.623968	1.440796
Woshiyigerenxingdehaizi	poem	421	6.578125	1.387304
Heshuiduiyin	poem	115	9.387755	1.449529
Honghuacao	poem	196	9.679012	1.653341
Yikedongridezhongziqidaizhexinsheng	poem	195	5.416667	1.363377
2000	government work report	7628	9.391197	1.137795
2002	government work report	7417	10.17421	1.156105
2004	government work report	8206	9.117778	1.049317
2006	government work report	10124	8.033327	0.943234
2008	government work report	11792	7.37	0.936521
2010	government work report	9588	7.826939	1.053654
2011	government work report	9586	7.825306	1.017817
2012	government work report	8861	8.136823	1.047244
2013	government work report	7504	9.238535	1.08004
2014	government work report	8393	9.644355	1.176307

Glottometrics 32, 2015, 21-54

Statistical approach to measure stylistic centrality

Peter Zörnig, Brasilia[1]
Ioan-Iovitz Popescu, Bucharest
Gabriel Altmann, Lüdenscheid

Abstract. We first study the formal similarity of texts of the same author using the simplified lambda indicator. A specific indicator is proposed expressing the stylistic centrality of the author based on the rank-frequency distribution of words. After that After that we make use of a graph theoretical approach, the concept of entropy and study the number of similarities to quantify the centrality.

Keywords: lambda, style, centrality, entropy, graph characteristics

1. Introduction

There are many definitions of style but the descriptions given by qualitative linguists or literary scientists in a variety of reviews or surveys provide only names, classifications and examples. There is no operationalized definition which could be used mechanically. In http://literarydevices.net/style/ one can read: "The style in writing can be defined as the way a writer writes and it is the technique which an individual author uses in his writing. It varies from author to author and depends upon one's syntax, word choice, and tone. It can also be described as a voice that readers listen to when they read the work of a writer" (accessed 19.12.2014).

The majority of such definitions are merely tautologies. They do not enable us to measure and compare texts, they are not quantitative and there are no tests possible because there are no testable hypotheses. Nevertheless, these attempts at least try to identify the phenomena that may contribute to capture some aspects of style, e.g. poetic or rhetoric figures. A part of these phenomena has phonetic, grammatical, semantic or lexical character. The lower levels are sometimes captured quantitatively but the definitions of "tone", e.g. *sadness*, lead to a circulus vitiosus.

Style is not a unique property but rather a hierarchy of properties. At its lowest level there are more concrete properties but the way of quantification is, again, a stratum of possibilities. The term "stylistic centrality" is somewhat vague but it expresses some kind of balance in the work of an author, the trend towards a kind of unified *ductus* of the text in whatever aspect. Of course, this property can be defined and measured in many various ways depending on the view we pursue. One can restrict oneself to sentence patterns, choice between synonyms, metaphors, poetic figures, use of foreign words, rhythm, associations, etc. Every discovered property can be measured and if some patterns appear, they can be modelled mathematically. In this article stylistic centrality is concretized by the proportion of similar pairs among all possible pairs of texts of an author. We first calculate the ratio of the number of similar text pairs divided by the total number of pairs. Since this is only a single aspect, we also model the similarity structure more realistically by means of a graph. Finally, we study the distribution of the number of similarities.

[1] Address correspondence to: peter@unb.br

2. The simplified lambda indicator

In a previous article (Popescu, Altmann 2015) we tried to characterize quantitatively the similarity of texts based on the simplified lambda indicator. The procedure could be performed mechanically because this indicator is a simple number. We consider

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

approximating quite truly the arc length between the ranked frequencies of words. Here V is the extent of the vocabulary (practically the greatest rank), f_1 is the frequency of the word with rank 1, and h is the fixed h-point which can be computed in the usual way (cf. e.g. Popescu et al. 2009: 18 ff.). In order to make it independent of the text length, the simplified lambda indicator has been defined as

$$(2) \quad \Lambda^* = \frac{L^*(\log_{10} N)}{N}$$

Better normalizations were also proposed (c.f. Popescu, Zörnig, Altmann, 2013). Here we take the decadic logarithm. Instead of the complex computation of L and its variance, Popescu and Altmann (2015) proposed a very good approximation defined by the relationship (1) and the above transformation (2) of lambda. By computing these values for texts, one can determine the similarities using the asymptotic normal test that only needs the variance of (2) defined as

$$(3) \quad Var(\Lambda^*) = \frac{Var(f_1)(\log_{10} N)^2}{N^2} = \frac{f_1(N - f_1)(\log_{10} N)^2}{N^3},$$

because except for f_1 all other quantities (V and h) are constants (cf. Popescu, Altmann 2015). A corresponding relation can be obtained using the repeat rate. The inverse relationship arises by computing the entropy: the greater the entropy, the smaller is the concentration of the text. Hence applying the above indicator we can extend the research and compare texts. Using the asymptotic normal test

$$(4) \quad u = \frac{|\Lambda_1^* - \Lambda_2^*|}{\sqrt{Var(\Lambda_1^*) + Var(\Lambda_2^*)}}$$

for comparing individual texts, one obtains a matrix in which one can find significant dissimilarities ($|u| > 1.96$) and similarities ($|u| \leq 1.96$). The more texts are similar, the more a writer tends to a well-balanced, individual style. We may conjecture that (s)he has a subconscious pattern of text writing represented by word repetition. We may call it *style centrality* or in our case, more exactly *lexical style centrality.*

Lambda itself may have a double interpretation: Increasing the size of the text both V and f_1 can increase. However, when N is large (greater than about 5000), the increase of V is very small, while that of f_1 may increase constantly.

The (dis)similarity matrix can be visualized in form of a graph. Graphs as images are not very practical and lucid if they contain many vertices and edges, but their properties can be evaluated and authors or text sorts can be ordered according to these properties. In order to

exemplify the procedure, we present the relevant numbers and results concerning some works by Vergilius. The basic data are given in Table 1.

Table 1
Simplified lambda with Vergilius

Text	***N***	***Λ****	**Var(*Λ**)**
1.Vergilius, Georgicon Liber I	3306	2.4774	0.000145
2.Vergilius, Georgicon Liber II	3518	2.4149	0.000142
3.Vergilius, Georgicon Liber III	3698	2.3954	0.000132
4.Vergilius, Georgicon Liber IV	3658	2.4255	0.000127
5.Vergilius, Aeneid I	4880	2.2660	0.000092
6.Vergilius, Aeneid II	5172	2.2162	0.000099
7.Vergilius, Aeneid III	4533	2.3468	0.000118
8.Vergilius, Aeneid IV	4569	2.2805	0.000088
9.Vergilius, Aeneid V	5556	2.2074	0.000076

The simplified lambdas values do not differ significantly, but not all texts are structurally similar. Comparing pairs of texts using (4) we obtain the results presented in Table 2.

Table 2
u-tests for differences of simplified Lambda with Vergilius

	1	2	3	4	5	6	7	8	9
1	0.0000	3.6852	4.9266	3.1467	13.7529	16.7421	8.0550	12.9161	18.1699
2	3.6852	0.0000	1.1811	0.6440	9.7429	12.8054	4.2224	8.8681	14.0476
3	4.9266	1.1811	0.0000	1.8717	8.6491	11.7904	3.0682	7.7469	13.0233
4	3.1467	0.6440	1.8717	0.0000	10.7934	13.9366	5.0278	9.9005	15.3094
5	13.7529	9.7429	8.6491	10.7934	0.0000	3.6072	5.5891	1.0844	4.5244
6	16.7421	12.8054	11.7904	13.9366	3.6072	0.0000	8.8790	4.7088	0.6689
7	8.0550	4.2224	3.0682	5.0278	5.5891	8.8790	0.0000	4.6294	10.0149
8	12.9161	8.8681	7.7469	9.9005	1.0844	4.7088	4.6294	0.0000	5.7128
9	18.1699	14.0476	13.0233	15.3094	4.5244	0.6689	10.0149	5.7128	0.0000

Retaining only those cells in which $|u| < 1.96$ in the lower triangle we obtain the results in Table 3

Table 3
Similarities between texts by Vergilius

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

4		1	1						
5									
6									
7									
8					1				
9						1			

Table 3 is of course symmetrical with respect to the main diagonal, since a text A is similar to a text B, when B is similar to A. This matrix can be represented by the graph in Fig. 1, where the isolated vertices 1 and 7 have been omitted. The vertices of the graph represent the texts and two different vertices are connected by an edge, if the corresponding texts are similar ($|u| < 1.96$).

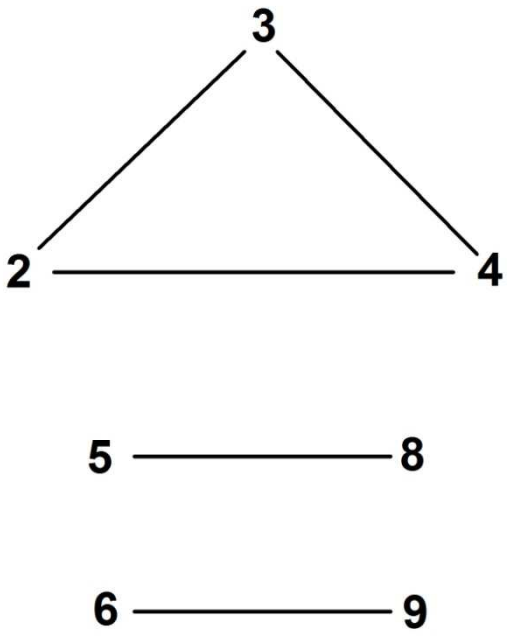

Figure 1. The graph related to Table 3
(number of texts $n = 9$, number of similarities $S = 5$)

3. Stylistic centrality as a proportion of similar pairs

The stylistic centrality concerning word frequency structures can be evaluated by computing

$$SI = \frac{S}{\binom{n}{2}} = \frac{2S}{n(n-1)} \quad (5)$$

where *S* is the number of cells with 1 in the lower (or upper) triangle (expressing similarity), and $\binom{n}{2}$ is the number of text pairs. This indicator also corresponds to the *(edge) density* of the respective graph, since it is the number of edges divided by the number of pairs of vertices. For Vergilius we obtain

$$SI = 2(5)/[9(8)] = 0.1389.$$

The graph representing the matrix in Table 3 (see Fig. 1) is quite simple and lucid but the more texts are analyzed, the complexity increases. Now, since the resulting *SI* is a simple proportion, one can easily compare the *SI*-values of different writers or text sorts. First, we derive the variances for *S* and *SI*. Assuming that each of the $\binom{n}{2}$ text pairs is similar with probability *p* and that similarities occur independently from each other, hence, *S* is binomially distributed with parameters *p* and $m = \binom{n}{2}$. The variance of *S* is therefore $p(1-p)m$, implying

$$(6) \qquad \mathrm{Var}(SI) = \mathrm{Var}(S/m) = \frac{1}{m^2}\,\mathrm{Var}(S) = \frac{p\,(1-p)}{m} = \frac{2\,p\,(1-p)}{n(n-1)} = \frac{2\,SI\,(1-SI)}{n(n-1)}$$

The latter equation holds approximately, since the proportion *SI* can be considered as an approximation for the unknown probability *p*.

For the above considered case we obtain $Var(SI) = 2(0.1389)(1-0.1389)/[9(8)] = 0.003322$.

The computation using also the upper triangle is analogous, the result does not change.

Let us present the computations for several authors. A small part of them has already been published earlier (cf. Popescu, Altmann 2015). Table 4 shows *n*, *S*, *SI* and *Var*(*SI*). The texts do not represent the full production of the author but merely a selection. The texts are ranked alphabetically by language (here *n* = number of texts, *S* = number of similarities).

Table 4
SI-values of individual writers, all texts
(Notes: 1. The lambda data were taken from the attached Appendix (German texts) or from the tables of the article by Popescu, Altmann (2015); 2. KZS means N. Ostrovskij' novel *Kak zakaljalas stal'*)

Language alphabetically	**Writer**	**Genre**	**Tab.**	***n***	***S***	***SI***	**Var(*SI*)**
Belorussian	translation, Ostrovskij's KZS	prose	20	10	4	0.0889	0.001800
Bulgarian	translation, Ostrovskij's KZS	prose	20	10	1	0.0222	0.000483
Croatian	translation, Ostrovskij's KZS	prose	20	10	5	0.1111	0.002195
Czech	Gottwald, New Year speeches	prose	5	5	2	0.2000	0.016000
Czech	Havel, New Year speeches	prose	5	13	23	0.2949	0.002666

Czech	Husák, New Year speeches	prose	5	15	31	0.2952	0.001982
Czech	Klaus, New Year speeches	prose	5	8	13	0.4643	0.008883
Czech	Novotný, New Year speeches	prose	5	11	11	0.2000	0.002909
Czech	Svoboda, New Year speeches	prose	5	6	3	0.2000	0.010667
Czech	translation, Ostrovskij's KZS	prose	20	10	4	0.0889	0.001800
Czech	Zápotocký, New Year speeches	prose	5	4	2	0.3333	0.037037
English	Byron	poetry	18	40	362	0.4641	0.000319
English	Joyce, Finnegans Wake	prose	6	17	24	0.1765	0.001069
German	Chamisso	prose	App.	11	13	0.2364	0.003282
German	Droste-Hülshoff	poetry	11	91	2164	0.5284	0.000061
German	Eichendorff	prose	App.	10	8	0.1778	0.003249
German	Goethe	poetry	9	7	12	0.5714	0.011662
German	Heine	poetry	8	20	78	0.4105	0.001274
German	Kafka	prose	App.	28	125	0.3307	0.000586
German	Keller	prose	App.	4	1	0.1667	0.023152
German	Lessing	prose	App.	10	20	0.4444	0.005487
German	Löns	prose	App.	13	12	0.1538	0.001669
German	Meyer	prose	App.	11	28	0.5091	0.004544
German	Novalis	prose	App.	13	9	0.1154	0.001309
German	Paul	prose	App.	55	563	0.3791	0.000159
German	Raabe	prose	App.	5	1	0.1000	0.009000
German	Rückert	prose	App.	5	7	0.7000	0.021000
German	Schiller	poetry	10	27	115	0.3276	0.000628
German	Schnitzler	prose	App.	14	10	0.1099	0.001075
German	Sealsfield	prose	App.	28	41	0.1085	0.000256
German	Tucholsky	prose	App.	5	1	0.1000	0.009000
German	Wedekind	prose	App.	8	2	0.0714	0.002368
Hawaiian	Laieikawai	prose	17	33	268	0.5076	0.000473
Hungarian	Ady Endre	poetry	13	23	98	0.3874	0.000938
Italian	Ciampi, End-of-Year speeches	prose	19	7	12	0.5714	0.011662
Italian	Cossiga, End-of-Year speeches	prose	19	7	6	0.2857	0.009718
Italian	Einaudi, End-of-Year speeches	prose	19	6	7	0.4667	0.016593
Italian	Gronchi, End-of-Year speeches	prose	19	7	19	0.9048	0.004103
Italian	Leone, End-of-Year speeches	prose	19	7	15	0.7143	0.009718
Italian	Napolitano, End-of-Year speeches	prose	19	8	17	0.6071	0.008519
Italian	Pertini, End-of-Year speeches	prose	19	7	5	0.2381	0.008638
Italian	Saragat, End-of-Year speeches	prose	19	7	11	0.5238	0.011878
Italian	Scalfaro, End-of-Year speeches	prose	19	7	2	0.0952	0.004103
Italian	Segni, End-of-Year speeches	prose	19	2	1	1.0000	0.000000
Latin	Apuleius, Metamorphoses	prose	7	11	14	0.2545	0.003450
Latin	Horatius	poetry	4	7	2	0.0952	0.004102

Latin	Vergilius	poetry	4	9	5	0.1389	0.003322
Macedonian	translation, Ostrovskij's KZS	prose	20	10	5	0.1111	0.002195
Polish	translation, Ostrovskij's KZS	prose	20	10	4	0.0889	0.001800
Romanian	Eminescu	poetry	14	146	3734	0.3528	0.000022
Russian	Lermontov	poetry	16	30	194	0.4460	0.000568
Russian	Ostrovskij, Kak zakalialas' stal'	prose	20	10	5	0.1111	0.002195
Russian	Pushkin	poetry	15	35	251	0.4218	0.000410
Serbian	translation, Ostrovskij's KZS	prose	20	10	5	0.1111	0.002195
Slovak	Bachletová	poetry	12	54	701	0.4899	0.000175
Slovak	Svoraková	prose	1	20	70	0.3684	0.001225
Slovak	translation, Ostrovskij's KZS	prose	20	10	7	0.1556	0.002919
Slovenian	translation, Ostrovskij's KZS	prose	20	10	9	0.2000	0.003556
Sorbian	translation, Ostrovskij's KZS	prose	20	10	8	0.1778	0.003248
Ukrainian	translation, Ostrovskij's KZS	prose	20	10	2	0.0444	0.000944

As can be seen, the degree of centrality or uniformity of style does not depend on the language or the text sort. It is a personal feature of a writer. Of course, here measured from a specific point of view. There may be common features for languages or text sorts but it will last a long time until more of them will be scrutinized.

If we order the authors and languages according to the size of *SI*, we obtain the results presented in Table 5

Table 5
Text collection ordered according to *SI*
(*N* = number of texts, *SI* = number of similarities)

Language	**Writer**	**Genre**	**Tab**	***n***	***S***	***SI* ranked**	**Var(*SI*)**
Bulgarian	translation, Ostrovskij's KZS	prose	20	10	1	0.0222	0.000483
Ukrainian	translation, Ostrovskij's KZS	prose	20	10	2	0.0444	0.000944
German	Wedekind	prose	App.	8	2	0.0714	0.002368
Belorussian	translation, Ostrovskij's KZS	prose	20	10	4	0.0889	0.001800
Czech	translation, Ostrovskij's KZS	prose	20	10	4	0.0889	0.001800
Polish	translation, Ostrovskij's KZS	prose	20	10	4	0.0889	0.001800
Latin	Horatius	poetry	4	7	2	0.0952	0.004102
Italian	Scalfaro, End-of-Year speeches	prose	19	7	2	0.0952	0.004103
German	Raabe	prose	App.	5	1	0.1000	0.009000
German	Tucholsky	prose	App.	5	1	0.1000	0.009000
German	Sealsfield	prose	App.	28	41	0.1085	0.000256
German	Schnitzler	prose	App.	14	10	0.1099	0.001075
Russian	Ostrovskij, Kak zakalialas' stal'	prose	20	10	5	0.1111	0.002195

Croatian	translation, Ostrovskij's KZS	prose	20	10	5	0.1111	0.002195
Macedonian	translation, Ostrovskij's KZS	prose	20	10	5	0.1111	0.002195
Serbian	translation, Ostrovskij's KZS	prose	20	10	5	0.1111	0.002195
German	Novalis	prose	App.	13	9	0.1154	0.001309
Latin	Vergilius	poetry	4	9	5	0.1389	0.003322
German	Löns	prose	App.	13	12	0.1538	0.001669
Slovak	translation, Ostrovskij's KZS	prose	20	10	7	0.1556	0.002919
German	Keller	prose	App.	4	1	0.1667	0.023152
English	Joyce, Finnegans Wake	prose	6	17	24	0.1765	0.001069
Sorbian	translation, Ostrovskij's KZS	prose	20	10	8	0.1778	0.003248
German	Eichendorff	prose	App.	10	8	0.1778	0.003249
Czech	Gottwald, New Year speeches	prose	5	5	2	0.2000	0.016000
Czech	Novotný, New Year speeches	prose	5	11	11	0.2000	0.002909
Czech	Svoboda, New Year speeches	prose	5	6	3	0.2000	0.010667
Slovenian	translation, Ostrovskij's KZS	prose	20	10	9	0.2000	0.003556
German	Chamisso	prose	App.	11	13	0.2364	0.003282
Italian	Pertini, End-of-Year speeches	prose	19	7	5	0.2381	0.008638
Latin	Apuleius, Metamorphoses	prose	7	11	14	0.2545	0.003450
Italian	Cossiga, End-of-Year speeches	prose	19	7	6	0.2857	0.009718
Czech	Havel, New Year speeches	prose	5	13	23	0.2949	0.002666
Czech	Husák, New Year speeches	prose	5	15	31	0.2952	0.001982
German	Schiller	poetry	10	27	115	0.3276	0.000628
German	Kafka	prose	App.	28	125	0.3307	0.000586
Czech	Zápotocký, New Year speeches	prose	5	4	2	0.3333	0.037037
Romanian	Eminescu	poetry	14	146	3734	0.3528	0.000022
Slovak	Svoraková	prose	1	20	70	0.3684	0.001225
German	Paul	prose	App.	55	563	0.3791	0.000159
Hungarian	Ady Endre	poetry	13	23	98	0.3874	0.000938
German	Heine	poetry	8	20	78	0.4105	0.001274
Russian	Pushkin	poetry	15	35	251	0.4218	0.000410
German	Lessing	prose	App.	10	20	0.4444	0.005487
Russian	Lermontov	poetry	16	30	194	0.4460	0.000568
English	Byron	poetry	18	40	362	0.4641	0.000319
Czech	Klaus, New Year speeches	prose	5	8	13	0.4643	0.008883
Italian	Einaudi, End-of-Year speeches	prose	19	6	7	0.4667	0.016593
Slovak	Bachletová	poetry	12	54	701	0.4899	0.000175
Hawaiian	Laieikawai	prose	17	33	268	0.5076	0.000473
German	Meyer	prose	App.	11	28	0.5091	0.004544
Italian	Saragat, End-of-Year speeches	prose	19	7	11	0.5238	0.011878
German	Droste-Hülshoff	poetry	11	91	2164	0.5284	0.000061
German	Goethe	poetry	9	7	12	0.5714	0.011662

Italian	Ciampi, End-of-Year speeches	prose	19	7	12	0.5714	0.011662
Italian	Napolitano, End-of-Year speeches	prose	19	8	17	0.6071	0.008519
German	Rückert	prose	App.	5	7	0.7000	0.021000
Italian	Leone, End-of-Year speeches	prose	19	7	15	0.7143	0.009718
Italian	Gronchi, End-of-Year speeches	prose	19	7	19	0.9048	0.004103
Italian	Segni, End-of-Year speeches	prose	19	2	1	1.0000	0.000000

The greater is the value of *SI*, the stronger is the stylistic centrality, i.e. the stronger is the tendency that the texts are similar. The End-of-Year speeches of Italian presidents are of special importance. Though Segni with 2 texts is not decisive, the other *SI*-values are very high. This may be caused by the fact that the text sort displays a certain stereotypy. However, this is not the case with Czech presidents but one never knows who wrote the speeches. It is rather a problem for historians.

Only a small part of texts satisfies $SI > 0.5$. Evidently one needs a great number of writers in order to venture a conjecture concerning the causes of stereotypy.

4. A graph theoretic approach

While *SI* displays an overall image, one can try to look at the similarities from another point of view. Individual texts display a certain tendency to be similar to other ones and this tendency can vary. In order to capture it, we consider the associated graph defined above. For each vertex i we denote by g_i the degree of the vertex i, i.e. the number of edges containing i. Thus, g_i also represents the number of texts similar to the text i. Then the *degree vector* is defined as $g = (g_1, g_2,\ldots,g_n)$. It is a well known fact of graph theory that

$$(7) \qquad \sum_{i=1}^{n} g_i = 2S$$

holds, i.e. the sum of degrees is $2S$, where S corresponds to the number of edges of the graph. One can imagine that by adding the degrees, the edges are counted, such that each edge is counted twice (one time for each end). We now consider the nonnegative numbers $p_i := \frac{g_i}{2S}$ which due to relation (7) sum up to 1. These numbers can therefore be interpreted as "probabilities" defined on the set of vertices. The entropy is now defined as

$$(8) \qquad H = -\sum_{i=1}^{n} p_i \log_2(p_i) = -\frac{1}{2S} \sum_{i=1}^{n} g(v_i) \log_2\left(\frac{g(v_i)}{2S}\right).$$

One can use also the natural logarithm but the dual logarithm is more usual here.

To illustrate the computation we use Table 3 in which we insert also the symmetric values (cf. Table 6). The sums of individual columns represent the vector of the writer

Table 6
Similarities of individual texts by Vergilius

Text	1	2	3	4	5	6	7	8	9
1									
2			1	1					
3		1		1					
4		1	1						
5								1	
6									1
7									
8					1				
9						1			
g_i	0	2	2	2	1	1	0	1	1

The vector of Vergilius is g(Vergilius) = (0,2,2,2,1,1,0,1,1), satisfying 2+2+…+ 1+1 =10 = 2*S*. Hence the entropy is

$$H = -3(2/10 \log_2 (2/10)) - 4(1/10 \log_2 (1/10)) = 1.393157 + 1.328771 = 2.721928$$

The entropy expresses the extent of centrality: the greater it is, the smaller is the concentration to a certain type; the smaller it is, the greater the probability that there is a subconscious pattern in the writer's mind. From the graph theoretic point of view, the entropy is large, when the graph is *regular*, i.e. when all vertex degrees are equal. In this case the degree vector is $(g_1,\ldots,g_n) = (k,\ldots,k)$, where the sum of components is nk. This implies $(p_1,\ldots,p_n) = (\frac{k}{nk},\ldots,\frac{k}{nk}) = (\frac{1}{n},\ldots,\frac{1}{n})$, and the entropy is $H = -n\frac{1}{n}\log_2\frac{1}{n} = \log_2 n$ representing a stable state. If there are large deviations between the degrees, the entropy is small, indicating a state of instability. A theoretical example for small entropy is the star shaped graph S_n having the $(n+1)$ vertices $0,1,\ldots,n$ and the n edges $(0,1)$, $(0,2),\ldots,(0,n)$. The degree vector is then $(g_1,\ldots,g_n,g_{n+1}) = (n,1,\ldots,1)$, implying $(p_0, p_1,\ldots,p_n) = (\frac{1}{2},\frac{1}{2n},\ldots,\frac{1}{2n})$. We obtain

$$H = -(\frac{1}{2}\log_2\left(\frac{1}{2}\right) + n\frac{1}{2n}\log_2\left(\frac{1}{2n}\right)$$

$$= \frac{1}{2}\log_2 2 + \frac{1}{2}\log_2 (2n) = \frac{1}{2}(\log_2 2 + \log_2 2 + \log_2 n) = 1 + \frac{1}{2}\log_2 n.$$

As numerical examples we consider the 3-regular graph in Fig. 2, where all vertices have degree 3 and the star shaped graph S_{19} (Fig. 3) having both 20 vertices. The corresponding entropies are $H(G) = \log_2 20 = \ln(20)/\ln(2) = 4.3219$ and $H(S_{19}) = 1 + \frac{1}{2}\log_2 19 = 3.1240$. It appears that (for a given number n of vertices) any regular graph with n vertices and the graph (S_{n-1}) maximize and minimize the entropy respectively, if isolated vertices, e.g. vertices with degree 0 do not occur. For $n = 20$ we get the above calculated lower and upper limits 3.1240 and 4.3219, respectively.

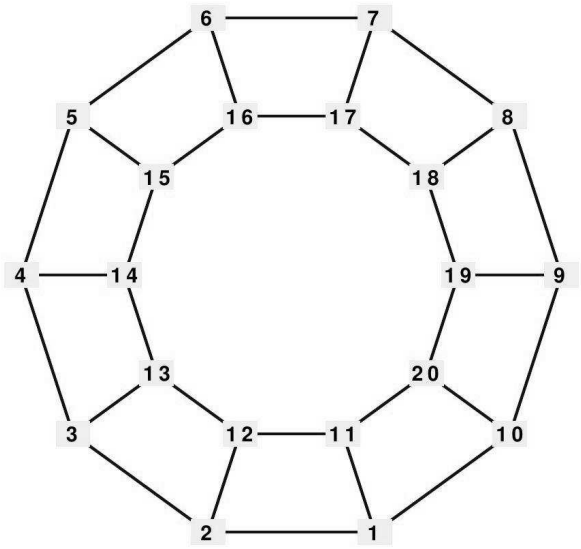

Figure 2. A 3-regular graph with 20 vertices

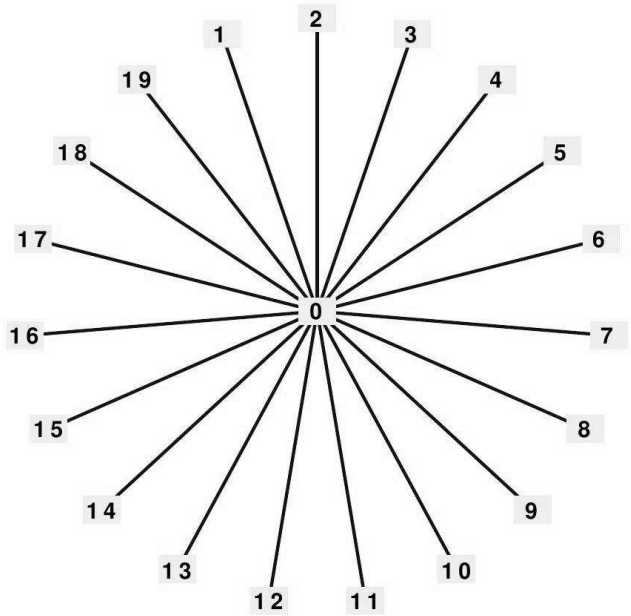

Figure 3. Star-shaped graph S_{19}.

In Table 7 the entropy of individual writers is displayed.

Table 7
Entropies of individual writers ranked *alphabetically* by language
(number of texts *n*, number of similarities *S*)

Language	**Writer**	**Genre**	**Table**	***n***	***S***	***H***
Belorussian	translation, Ostrovskij's KZS	prose	20	10	4	2.5000
Bulgarian	translation, Ostrovskij's KZS	prose	20	10	1	1.0000
Croatian	translation, Ostrovskij's KZS	prose	20	10	5	3.1219
Czech	Gottwald, New Year speeches	prose	5	5	2	2.0000
Czech	Havel, New Year speeches	prose	5	13	23	3.4405
Czech	Husák, New Year speeches	prose	5	15	31	3.7333
Czech	Klaus, New Year speeches	prose	5	8	13	2.7014
Czech	Novotný, New Year speeches	prose	5	11	11	3.2090
Czech	Svoboda, New Year speeches	prose	5	6	3	2.2516
Czech	translation, Ostrovskij's KZS	prose	20	10	4	2.7500
Czech	Zápotocký, New Year speeches	prose	5	4	2	1.5000
English	Byron	poetry	18	40	362	5.1891
English	Joyce, Finnegans Wake	prose	6	17	24	3.7207
German	Chamisso	prose	App.	11	13	3.2766
German	Droste-Hülshoff	poetry	11	91	2164	6.4248
German	Eichendorff	prose	App.	10	8	3.2500
German	Goethe	poetry	9	7	12	2.5342
German	Heine	poetry	8	20	78	4.2318
German	Kafka	prose	App.	28	125	4.6780
German	Keller	prose	App.	4	1	1.0000
German	Lessing	prose	App.	10	20	3.3219
German	Löns	prose	App.	13	12	3.0535
German	Meyer	prose	App.	11	28	3.3145
German	Novalis	prose	App.	13	9	3.4194
German	Paul	prose	App.	55	563	5.5337
German	Raabe	prose	App.	5	1	1.0000
German	Rückert	prose	App.	5	7	2.2170
German	Schiller	poetry	10	27	115	4.6180
German	Schnitzler	prose	App.	14	10	3.4464
German	Sealsfield	prose	App.	28	41	4.3581
German	Tucholsky	prose	App.	5	1	1.0000
German	Wedekind	prose	App.	8	2	2.0000
Hawaiian	Laieikawai	prose	17	33	268	4.8862
Hungarian	Ady Endre	poetry	13	23	98	4.3883

Italian	Ciampi, End-of-Year speeches	prose	19	7	12	2.6846
Italian	Cossiga, End-of-Year speeches	prose	19	7	6	2.6887
Italian	Einaudi, End-of-Year speeches	prose	19	6	7	2.2170
Italian	Gronchi, End-of-Year speeches	prose	19	7	19	2.7938
Italian	Leone, End-of-Year speeches	prose	19	7	15	2.7493
Italian	Napolitano, End-of-Year speeches	prose	19	8	17	3.1695
Italian	Pertini, End-of-Year speeches	prose	19	7	5	2.2464
Italian	Saragat, End-of-Year speeches	prose	19	7	11	2.6978
Italian	Scalfaro, End-of-Year speeches	prose	19	7	2	2.0000
Italian	Segni, End-of-Year speeches	prose	19	2	1	1.0000
Latin	Apuleius, Metamorphoses	prose	7	11	14	3.3249
Latin	Horatius	poetry	4	7	2	2.0000
Latin	Vergilius	poetry	4	9	5	2.7219
Macedonian	translation, Ostrovskij's KZS	prose	20	10	5	2.7219
Polish	translation, Ostrovskij's KZS	prose	20	10	4	3.0000
Romanian	Eminescu	poetry	14	146	3734	7.0683
Russian	Lermontov	poetry	16	30	194	4.7235
Russian	Ostrovskij, Kak zakalialas' stal'	prose	20	10	5	2.7219
Russian	Pushkin	poetry	15	35	251	4.9806
Serbian	translation, Ostrovskij's KZS	prose	20	10	5	3.1219
Slovak	Bachletová	poetry	12	54	701	5.6595
Slovak	Svoraková	prose	1	20	70	3.8929
Slovak	translation, Ostrovskij's KZS	prose	20	10	7	3.0391
Slovenian	translation, Ostrovskij's KZS	prose	20	10	9	3.2391
Sorbian	translation, Ostrovskij's KZS	prose	20	10	8	3.1250
Ukrainian	translation, Ostrovskij's KZS	prose	20	10	2	2.0000

The following figure illustrates the relation between entropy and number of similarities.

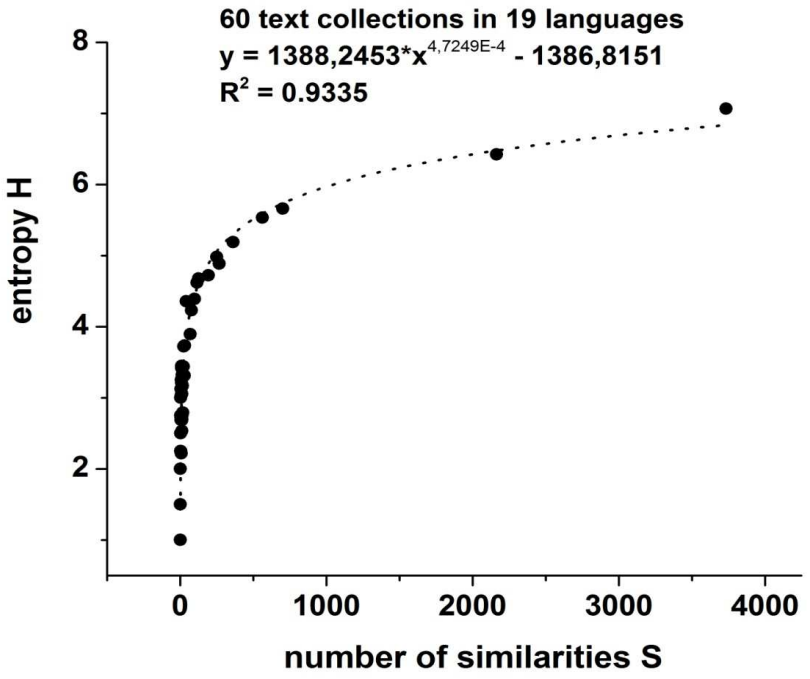

Figure 4. Increase of entropy with the number of similarities

Table 8
Entropy ordered by genre
(number of texts: *n*, number of similarities: *S*)

Language	Writer	Genre ranked	Table	*n*	*S*	*H*
English	Byron	poetry	18	40	362	5.1891
German	Droste-Hülshoff	poetry	11	91	2164	6.4248
German	Goethe	poetry	9	7	12	2.5342
German	Heine	poetry	8	20	78	4.2318
German	Schiller	poetry	10	27	115	4.6180
Hungarian	Ady	poetry	13	23	98	4.3883
Latin	Horatius	poetry	4	7	2	2.0000
Latin	Vergilius	poetry	4	9	5	2.7219
Romanian	Eminescu	poetry	14	146	3734	7.0683
Russian	Lermontov	poetry	16	30	194	4.7235
Russian	Pushkin	poetry	15	35	251	4.9806
Slovak	Bachletová	poetry	12	54	701	5.6595
Belorussian	translation, Ostrovskij's KZS	prose	20	10	4	2.5000
Bulgarian	translation, Ostrovskij's KZS	prose	20	10	1	1.0000
Croatian	translation, Ostrovskij's KZS	prose	20	10	5	3.1219
Czech	Gottwald, New Year speeches	prose	5	5	2	2.0000

Czech	Havel, New Year speeches	prose	5	13	23	3.4405
Czech	Husák, New Year speeches	prose	5	15	31	3.7333
Czech	Klaus, New Year speeches	prose	5	8	13	2.7014
Czech	Novotný, New Year speeches	prose	5	11	11	3.2090
Czech	Svoboda, New Year speeches	prose	5	6	3	2.2516
Czech	translation, Ostrovskij's KZS	prose	20	10	4	2.7500
Czech	Zápotocký, New Year speeches	prose	5	4	2	1.5000
English	Joyce, Finnegans Wake	prose	6	17	24	3.7207
German	Chamisso	prose	App.	11	13	3.2766
German	Eichendorff	prose	App.	10	8	3.2500
German	Kafka	prose	App.	28	125	4.6780
German	Keller	prose	App.	4	1	1.0000
German	Lessing	prose	App.	10	20	3.3219
German	Löns	prose	App.	13	12	3.0535
German	Meyer	prose	App.	11	28	3.3145
German	Novalis	prose	App.	13	9	3.4194
German	Paul	prose	App.	55	563	5.5337
German	Raabe	prose	App.	5	1	1.0000
German	Rückert	prose	App.	5	7	2.2170
German	Schnitzler	prose	App.	14	10	3.4464
German	Sealsfield	prose	App.	28	41	4.3581
German	Tucholsky	prose	App.	5	1	1.0000
German	Wedekind	prose	App.	8	2	2.0000
Hawaiian	Laieikawai	prose	17	33	268	4.8862
Italian	Ciampi, End-of-Year speeches	prose	19	7	12	2.6846
Italian	Cossiga, End-of-Year speeches	prose	19	7	6	2.6887
Italian	Einaudi, End-of-Year speeches	prose	19	6	7	2.2170
Italian	Gronchi, End-of-Year speeches	prose	19	7	19	2.7938
Italian	Leone, End-of-Year speeches	prose	19	7	15	2.7493
Italian	Napolitano, End-of-Year speeches	prose	19	8	17	3.1695
Italian	Pertini, End-of-Year speeches	prose	19	7	5	2.2464
Italian	Saragat, End-of-Year speeches	prose	19	7	11	2.6978
Italian	Scalfaro, End-of-Year speeches	prose	19	7	2	2.0000
Italian	Segni, End-of-Year speeches	prose	19	2	1	1.0000
Latin	Apuleius, Metamorphoses	prose	7	11	14	3.3249
Macedonian	translation, Ostrovskij's KZS	prose	20	10	5	2.7219
Polish	translation, Ostrovskij's KZS	prose	20	10	4	3.0000
Russian	Ostrovskij, Kak zakalialas' stal'	prose	20	10	5	2.7219
Serbian	translation, Ostrovskij's KZS	prose	20	10	5	3.1219

Slovak	Svoraková	prose	1	20	70	3.8929
Slovak	translation, Ostrovskij's KZS	prose	20	10	7	3.0391
Slovenian	translation, Ostrovskij's KZS	prose	20	10	9	3.2391
Sorbian	translation, Ostrovskij's KZS	prose	20	10	8	3.1250
Ukrainian	translation, Ostrovskij's KZS	prose	20	10	2	2.0000

As can be seen in the above table, the individual authors differ even in the same language. One may ask whether there is, in general, a difference also between prose and poetry regardless of language. To this end we consider the entropies H as some numbers and compare the means of prose and poetry. In poetry, we have n = 12 texts, the mean of poetry is $\bar{H}_{poetry}$ = 4.5450, the variance of $Var(H_{poetry})$ = 2.3410, and the variance of the mean is $Var(\bar{H}_{poetry})$ = 2.3410/12 = 0.195083. For prose, we have n = 48, the mean is $\bar{H}_{prose}$ = 2.8358, the variance is $Var(H_{prose})$ = 0.9776 and the variance of the mean is $Var(\bar{H}_{prose})$ = 0.9776/48 = 0.020367. Performing a normal test for difference we obtain

$$u = \frac{|4.5450 - 2.8358|}{\sqrt{0.195083 + 0.020367}} = 3.68$$

The test is asymptotic, the number of cases is not sufficient but we can preliminarily accept the conjecture that with regard to text sort, poetry is more concentrated than prose; a result that could be expected. One must, of course, increase the number of compared texts but one can also continue with other comparing other text sorts, or simply subdivide the "prose" in specified text-sorts. With regard to the number of languages and texts, this work has no end. Other indicators can be examined in the same way.

5. Vector of similarities

Looking at Table 9 below we see that the vectors of similarities whose components correspond to the vertex degrees of the corresponding graph, are not comparable in the given form. First, the numbers of texts are different; second, the order of similarities is not standard but freely ordered – just as the texts were analyzed. In order to make them comparable, one may proceed in several ways: (1) either one divides the similarities by their sum and sets up a distribution, or (2) one divides them by the number of texts/ chapters of the author and sets up a distribution, or (3) one sets up a new scale dividing the values by their maximum. Then one can compare the normalized distributions. This can be done either by considering them as vectors to be compared according to a certain metric, by comparing the ranks of the values, or by comparing the means, etc.

Here we shall first compute the criterion proposed by Ord (1972).

Table 9
Vectors of similarities

Author	Vector
Eminescu	66; 58; 39; 42; 58; 57; 69; 49; 57; 76; 71; 58; 37; 75; 65; 55; 69; 64; 57; 58; 50; 69; 33; 62; 4; 74; 72; 70; 81; 35; 57; 13; 53; 62;

	57; 35; 72; 63; 73; 39; 42; 67; 37; 30; 58; 83; 5; 19; 9; 15; 44; 34; 38; 64; 50; 55; 52; 15; 53; 75; 69; 49; 48; 18; 3; 58; 45; 40; 71; 27; 74; 72; 26; 65; 59; 59; 78; 62; 68; 44; 62; 58; 57; 39; 74; 50; 54; 70; 56; 43; 52; 55; 9; 57; 50; 70; 71; 64; 9; 64; 57; 22; 70; 55; 55; 58; 46; 62; 48; 18; 65; 77; 64; 58; 40; 45; 52; 22; 55; 34; 4; 54; 36; 62; 53; 44; 51; 37; 36; 52; 40; 71; 40; 53; 43; 33; 73; 71; 4; 63; 66; 20; 55; 73; 71; 62
Byron	4; 23; 22; 19; 21; 19; 29; 21; 21; 20; 21; 13; 22; 19; 29; 15; 18; 28; 6; 19; 3; 21; 21; 16; 16; 20; 3; 18; 20; 26; 7; 7; 20; 30; 19; 8; 30; 13; 19; 18
Joyce	0; 5; 4; 4; 5; 4; 2; 3; 4; 1; 4; 0; 2; 0; 3; 4; 3
Chamisso	3; 2; 2; 3; 3; 4; 2; 2; 3; 0; 2
Droste-Hülshoff	37; 44; 53; 32; 40; 19; 63; 51; 37; 61; 24; 52; 51; 20; 17; 61; 67; 54; 47; 43; 49; 49; 60; 65; 66; 40; 30; 39; 64; 59; 49; 40; 52; 32; 44; 57; 59; 42; 8; 57; 65; 30; 43; 27; 66; 38; 55; 66; 58; 47; 41; 50; 45; 62; 56; 66; 62; 66; 68; 49; 56; 48; 63; 59; 45; 56; 60; 59; 49; 37; 57; 51; 56; 3; 54; 19; 48; 23; 59; 20; 40; 39; 15; 56; 51; 57; 62; 39; 54; 34; 65
Eichendorff	2; 2; 1; 1; 2; 2; 2; 1; 2; 1
Goethe	4; 5; 4; 4; 5; 2; 0
Heine	6; 9; 2; 8; 11; 9; 9; 3; 9; 4; 9; 8; 11; 8; 7; 12; 8; 9; 9; 5
Kafka	0; 8; 11; 10; 9; 4; 14; 11; 8; 7; 13; 12; 11; 9; 8; 8; 17; 7; 6; 5; 8; 10; 12; 6; 5; 13; 11; 7
Keller	0; 1; 0; 1
Lessing	4; 4; 4; 4; 4; 4; 4; 4; 4; 4
Löns	0; 2; 3; 4; 4; 2; 2; 1; 0; 4; 0; 2; 0
Meyer	5; 6; 6; 1; 7; 7; 7; 7; 5; 3; 2
Novalis	1; 3; 1; 0; 2; 1; 2; 1; 1; 1; 1; 1; 3
Paul	18; 18; 27; 20; 26; 20; 26; 35; 20; 25; 23; 26; 5; 28; 24; 10; 23; 18; 18; 21; 4; 23; 24; 25; 20; 18; 25; 28; 25; 32; 15; 30; 26; 19; 21; 29; 27; 31; 12; 29; 9; 4; 24; 9; 6; 24; 25; 28; 33; 3; 19; 23; 4; 1; 20
Raabe	0; 0; 0; 1; 1
Rückert	3; 3; 3; 1; 4
Schiller	11; 4; 5; 9; 10; 8; 11; 6; 5; 11; 14; 11; 15; 10; 2; 9; 11; 4; 5; 13; 6; 11; 4; 6; 4; 13
Schnitzler	1; 3; 1; 2; 2; 3; 1; 1; 0; 1; 0; 1; 2; 2
Sealsfield	3; 1; 6; 6; 3; 6; 4; 0; 2; 2; 2; 3; 1; 1; 3; 6; 6; 5; 0; 6; 6; 0; 2; 1; 0; 2; 2; 3
Tucholsky	0; 1; 1; 0; 0
Wedekind	0; 0; 1; 1; 1; 0; 1; 0
Laieikawai	23; 5; 21; 3; 19; 10; 23; 22; 3; 17; 12; 20; 16; 18; 11; 15; 21; 19; 17; 21; 21; 20; 17; 22; 11; 22; 15; 6; 0; 23; 23; 21; 19
Ady	9; 13; 13; 5; 11; 10; 7; 5; 12; 12; 9; 1; 7; 7; 9; 2; 8; 11; 4; 7; 11; 10; 13
Ciampi	1; 4; 5; 4; 4; 4; 2
Cossiga	1; 3; 1; 2; 2; 2; 1
Einaudi	3; 0; 4; 1; 3; 3
Gronchi	4; 6; 5; 5; 6; 6; 6

Leone	6; 4; 2; 4; 5; 5; 4
Napolitano	5; 6; 5; 4; 5; 3; 1; 5
Pertini	0; 3; 0; 2; 2; 1; 2
Saragat	2; 4; 4; 4; 4; 3; 1
Scalfaro	1; 0; 1; 0; 0; 1; 1
Segni	1; 1
Apuleius	1; 2; 2; 4; 2; 4; 3; 4; 2; 3; 1
Horatius	0; 1; 1; 0; 1; 1; 0
Vergilius	0; 2; 2; 2; 1; 1; 0; 1; 1
Lermontov	13; 10; 1; 5; 17; 18; 20; 5; 17; 16; 5; 11; 16; 12; 7; 4; 5; 19; 17; 18; 19; 17; 17; 19; 17; 2; 20; 17; 12; 12
Ostrovskij	1; 2; 1; 0; 1; 1; 2; 0; 2; 0
Pushkin	6; 21; 12; 7; 17; 16; 20; 3; 5; 8; 7; 21; 5; 7; 13; 20; 14; 18; 22; 19; 20; 12; 17; 9; 16; 21; 18; 22; 17; 20; 20; 5; 18; 6; 20
Havel	5; 1; 6; 0; 5; 4; 5; 3; 3; 1; 3; 5; 5
Husák	2; 5; 1; 6; 5; 2; 4; 4; 5; 9; 5; 5; 2; 5; 2
Klaus	1; 4; 5; 3; 4; 0; 4; 5
Novotný	1; 1; 3; 2; 3; 2; 2; 2; 4; 0; 2
Svoboda	0; 2; 1; 1; 1; 1
Zápotocký	2; 1; 0; 1
Gottwald	0; 1; 1; 1; 1
Bachletová	31; 30; 32; 31; 11; 6; 20; 30; 34; 24; 25; 27; 28; 18; 20; 31; 35; 33; 15; 37; 25; 38; 7; 38; 12; 29; 23; 23; 29; 10; 31; 28; 9; 15; 33; 40; 32; 32; 23; 22; 26; 31; 12; 19; 38; 35; 28; 32; 11; 28; 28; 29; 36; 32
Svoráková	8; 6; 4; 8; 7; 5; 6; 2; 3; 2; 2; 1; 4; 6; 2; 2; 1; 1; 0
Slovak (Ostrovskij)	1; 2; 1; 0; 3; 1; 2; 1; 2; 1
Slovenian (Ostrovskij)	1; 2; 1; 2; 3; 1; 2; 2; 2; 2
Sorbian (Ostrovskij)	2; 2; 2; 0; 1; 2; 1; 2; 2; 2
Ukrainian (Ostrovskij)	1; 0; 0; 0; 0; 1; 0; 1; 1; 0
Serbian (Ostrovskij)	1; 2; 0; 1; 1; 1; 1; 1; 1; 1
Macedinian (Ostrovskij)	2; 1; 2; 0; 1; 2; 0; 1; 1; 0
Polish (Ostrovskij)	1; 1; 1; 0; 1; 1; 1; 1; 0; 1
Czech (Ostrovskij)	1; 2; 0; 0; 1; 1; 1; 1; 1; 0
Belorussian (Ostrovskij)	2; 1; 2; 0; 1; 1; 1; 0; 0; 0
Bulgarian (Ostrovskij)	1; 0; 0; 0; 0; 1; 0; 0; 0; 0
Croatian (Ostrovskij)	1; 2; 0; 1; 1; 1; 1; 1; 1; 1

For evaluating the vectors we proceed as follows. Consider, for example J. Joyce. His vector is (0, 5, 4, 4, 5, 4, 2, 3, 4, 1, 4, 0, 2, 0, 3, 4, 3). Setting up the distribution we obtain (number of texts = 17; sum of similarities = 48)

x	f
0	3
1	1
2	2
3	3
4	6
5	2

Since the number of text is too small, we cannot construct a reliable model. But one can compute the means and other characteristics of the distributions empirically. Here we shall characterize the distributions using Ord's criterion <I,S> where

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

are ratios of moments. For the individual texts we obtain the results presented in Table 10 and displayed in Figure 5. For the above vector of Joyce we obtain:

$$m_1' = \Sigma x_i f_i / N = [0(3) + 1(1) + \ldots + 5(2)]/17 = 48/17 = 2.8235$$
$$m_2 = \Sigma (x_i - m_1')^2 f_i / N = [(0 - 2.8235)^2(3) + (1 - 2.8235)^2(1) + \ldots + (5 - 28235)^2(5)]/17 = 2.7335$$
$$m_3 = \Sigma (x_i - m_1')^3 f_i / N = [(0 - 2.8235)^3(3) + (1 - 2.8235)^3(1) + \ldots + (5 - 2.8235)^3(5)]/17 = -2.6061$$

from which follows $I = 2.7335/2.8235 = 0.9681$ and $S = -2.6061/2.7335 = -0.9534$. The values for all writers are presented in Table 10.

Table 10
The <I,S> indicator for the analyzed texts

Text	m_1'	m_2	m_3	I	S
Eminescu	51.1507	352.6485	-5527.4610	6.8943	-15.6741
Byron	18.10000	50.5600	-192.5430	2.7978	-3.8022
Joyce	2.8235	2.7335	-2.6061	0.9681	-0.9534
Chamisso	2.3636	0.9587	-0.7303	0.4056	-0.7618
Droste-Hülshoff	47.5604	217.9167	-2843.1458	4.5819	-13.0469
Eichendorff	1.6000	0.2400	-0.0480	0.1500	-0.2000
Goethe	3.4286	2.8163	-4.9854	0.8214	-1.7702
Heine	7.8000	6.5600	-11.8560	0.8410	-1.8073
Kafka	8.9286	11.7806	-5.0466	1.3194	-0.4384
Keller	0.5000	0.2500	0.0000	0.5000	-
Lessing	4.0000	0.0000	0.0000	0.0000	-
Löns	1.8462	2.2840	0.4424	1.2371	0.1937
Meyer	5.0909	4.2644	-7.0729	0.8377	-1.6586
Novalis	1.3846	0.6982	0.4452	0.5043	0.6375
Paul	20.4727	70.2129	-451.5712	3.4296	-6.4314
Raabe	0.4000	0.2400	0.0480	0.6000	0.2000
Rückert	2.8000	0.9600	-0.8160	0.3429	-0.8500
Schiller	8.3846	12.8521	0.9595	1.5328	0.0747
Schnitzler	1.4286	0.8163	0.1574	0.5714	0.1929
Tucholsky	0.4000	0.2400	0.0480	0.6000	0.2000
Sealsfield	2.9286	4.4949	2.8207	1.5438	0.6275
Wedekind	0.5000	0.2500	0.0000	0.5000	0.0000
Laieikawai Anon.	16.2424	42.7897	-288.3462	2.6344	-6.7387
Ady	8.5217	11.3800	-21.3454	1.3354	-1.8757
Ciampi	3.4286	1.6735	-1.8017	0.4881	-1.0767

Cossiga	1.7143	0.4898	0.1574	0.2857	0.3214
Einaudi	2.3333	1.8889	-1.5926	0.8095	-0.8431
Gronchi	5.4286	0.5306	-0.3324	0.0977	-0.6264
Leone	4.2857	1.3147	-0.8921	0.3143	-0.6623
Napolitano	4.2500	2.1875	-3.6562	0.5147	-1.6714
Pertini	1.4286	1.1020	-0.2099	0.7714	-0.1905
Saragat	3.1429	1.2653	-1.2595	0.4026	-0.9954
Scalfaro	0.5714	0.2449	-0.0350	0.4286	-0.1429
Segni	1.0000	0.0000	0.0000	0.0000	-
Apuleius	2.5455	1.1570	0.1262	0.4545	0.1091
Horatius	0.5714	0.2449	-0.0350	0.4286	-0.1429
Vergilius	1.1111	0.5432	-0.0713	0.4889	-0.1313
Lermontov	12.9333	34.6622	-126.5339	2.6801	-3.6505
Ostrovskij	1.0000	0.6000	0.0000	0.6000	0.0000
Pushkin	14.3429	37.2539	-100.2443	2.5974	-2.6908
Havel	3.5385	3.3254	-3.6049	0.9398	-1.0840
Husák	4.1333	3.9822	3.7381	0.9634	0.9387
Klaus	3.2500	2.9375	-4.2188	0.9038	-1.4362
Novotný	2.0000	1.0909	0.0000	0.5454	0.0000
Svoboda	1.0000	0.3333	0.0000	0.3333	0.0000
Zápotocký	1.0000	0.5000	0.0000	0.5000	0.0000
Gottwald	0.8000	0.1600	-0.0960	0.2000	-0.6000
Bachletová	25.9630	77.4431	-460.6544	2.9828	-5.9483
Svoráková	3.6842	6.1108	5.5049	1.6586	0.9008
Slovak (Ostrovskij)	1.4000	0.6400	0.1680	0.4571	0.2625
Slovenian (Ostrovskij)	1.8000	0.3600	0.0240	0.2000	0.0667
Sorbian (Ostrovskij)	1.6000	0.4400	-0.4080	0.2750	-0.9273
Ukrainian (Ostrovskij)	0.4000	0.2400	0.0480	0.6000	0.2000
Serbian (Ostrovskij)	1.0000	0.2000	0.0000	0.2000	0.0000
Macedonian (Ostrovskij)	1.0000	0.6000	0.0000	0.6000	0.0000
Polish (Ostrovskij)	0.8000	0.16000	-0.0960	0.2000	-0.6000
Czech (Ostrovskij)	0.8000	0.3600	0.0240	0..4500	0.0667
Belorussian (Ostrovskij)	0.8000	0.5600	0.1440	0.7000	0.2571
Bulgarian (Ostrovskij)	0.2000	0.1600	0.0960	0.8000	0.6000
Croatian (Ostrovskij)	1.0000	0.2000	0.0000	0.2000	0.0000

The graphic representation of <I,S> yields a decreasing trend which can be captured by the concave power function $S = 0.1611 + 0.9876*I^{1.4889}$ as presented in Figure 5.

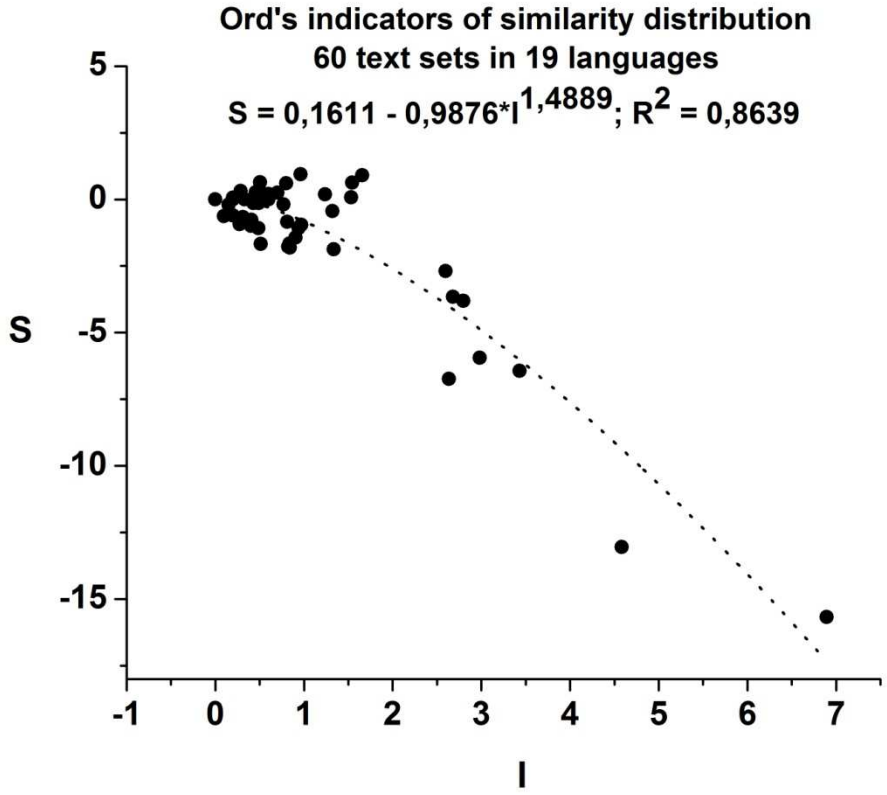

Figure 5. The relation of *I* and *S*

The <I,S>- relationship could be captured by other functions, too, the given result is preliminary but in any case it shows that there are certain mechanisms controlling the production of individual authors.

6. The ranking of similarities

If one ranks the texts of an author according to the number of similarities, one can see that there is some regularity which can be expressed by a function (see below). There are no great jumps distinguishing the centrality of texts, on the contrary, the centrality decreases continuously from the greatest centrality to the smallest. Adding new texts of the author would not significantly disturb this regularity. In order to model this course, we suppose that there are two forces controlling the centrality. The first is the subconscious pattern own to the writer, not causing any effort and realized in all his works. The second pattern represents his conscious striving for originality, differentiating every new text from the previous. The first effort can be expressed by the ratio c/(a + cx), the second, representing his originality by b/(1+bx). These expressions are in accordance with the unified theory (cf. Wimmer, Altmann 2005). The second component which modifies the first one must be subtracted. Hence the relative rate of change of centrality can be written as

$$\frac{dy}{y} = \left(\frac{c}{a+cx} - \frac{b}{1+bx} \right) dx \tag{10}$$

whose solution is

$$(11) \qquad y = \frac{a + cx}{1 + bx},$$

(omitting the integration constant).

Consider, for example the ranked centralities in the works by E. Bachletová as presented in Table 11. The fitting of (11) yields a result with $R^2 = 0.98$ and the computed values can be seen in the third column of the table. In Figure 6 the observed and computed values are displayed graphically.

Table 11
Fitting (11) to the ranked centralities of texts by E. Bachletová

Rank	Centrality	Computed		Rank	Centrality	Computed
1	40	36.8075		28	28	27.9528
2	38	36.5808		29	28	27.4643
3	38	36.3486		30	28	26.9584
4	38	36.1108		31	28	26.4343
5	37	35.8672		32	27	25.8907
6	36	35.6175		33	26	25.3268
7	35	35.3614		34	25	24.7412
8	35	35.0989		35	25	24.1328
9	34	34.8295		36	24	23.5002
10	33	34.5531		37	23	22.8418
11	33	34.2693		38	23	22.1561
12	32	33.9778		39	23	21.4414
13	32	33.6784		40	22	20.6957
14	32	33.3707		41	20	19.9170
15	32	33.0544		42	20	19.1032
16	32	32.7290		43	19	18.2516
17	31	32.3942		44	18	17.3597
18	31	32.0497		45	15	16.4245
19	31	31.6948		46	15	15.4428
20	31	31.3293		47	12	14.4109
21	31	30.9525		48	12	13.3251
22	30	30.5640		49	11	12.1810
23	30	30.1633		50	11	10.9736
24	29	29.7496		51	10	9.6976
25	29	29.3224		52	9	8.3470
26	29	28.8810		53	7	6.9151
27	28	28.4247		54	6	5.3943
a = 37,0289 , b = -0,0116 , c = -0,6484 , R^2 = 0,9838						

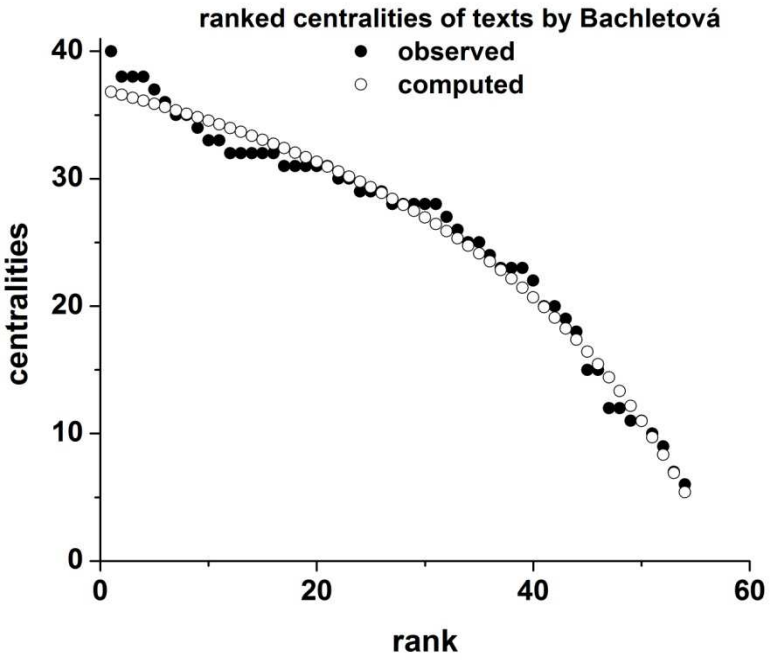

Figure 6. Fitting ranked centralities of texts by Bachletová

In Table 12, the results of fitting function (11) to the centralities of individual writers are presented. The same data are ordered by the number *S* of similarities in Table 13.

Table 12
The parameters and the determination coefficients of the fitting function
(the languages are ordered alphabetically)

Language	**Writer**	***S***	***a***	***b***	***c***	R^2
Belorussian	Transl. Ostrovskij's KZS	4	2.5270	0.0809	-0.2787	0.8531
Bulgarian	Transl. Ostrovskij's KZS	1	3.4263	1.6787	-0.5174	0.7065
Croatian	Transl. Ostrovskij's KZS	5	1.2084	-0.0952	-0.1208	0.4799
Czech	Gottwald, New Year speeches	2	1.0000	-0.2000	-0.2000	1.0000
Czech	Havel, New Year speeches	23	5.9346	-0.0452	-0.4604	0.9422
Czech	Husák, New Year speeches	31	7.8289	0.0161	-0.4115	0.8229
Czech	Klaus, New Year speeches	13	5.2864	-0.0879	-0.6670	0.9402
Czech	Novotný, New Year speeches	11	3.8546	0.0000	-0.3091	0.8447
Czech	Svoboda, New Year speeches	3	3.1209	0.5396	-0.1897	0.5300
Czech	Transl. Ostrovskij's KZS	4	1.6875	-0.0266	-0.1754	0.6926
Czech	Zápotocký, New Year speeches	2	2.5000	0.0000	-0.6000	0.7000
English	Byron	362	26.7450	-0.0159	-0.6496	0.9210
English	Joyce's Finnegans Wake	24	5.0441	-0.0324	-0.3103	0.9450
German	Chamisso	13	3.3894	-0.0659	-0.2980	0.7572

German	Droste-Hülshoff	2164	64.8195	-0.0077	-0.6803	0.9863
German	Eichendorff	8	2.2497	-0.0444	-0.1795	0.6881
German	Goethe	12	5.1542	-0.1143	-0.7375	0.9451
German	Heine	78	10.7463	-0.0352	-0.5097	0.9375
German	Kafka	125	14.7439	-0.0015	-0.4117	0.9121
German	Keller	1	1.5000	0.0000	-0.4000	0.4000
German	Lessing	20	-	-	-	-
German	Löns	12	4.8207	0.0185	-0.4053	0.9233
German	Meyer	28	7.6049	-0.0618	-0.6701	0.9638
German	Novalis	9	4.2728	0.2970	-0.1163	0.8316
German	Paul	563	30.4845	-0.0122	-0.5533	0.9671
German	Raabe	1	1.7589	0.2316	-0.4071	0.5721
German	Rückert	7	3.6088	-0.1772	-0.6986	0.7958
German	Schiller	115	14.6751	-0.0029	-0.4584	0.9645
German	Schnitzler	10	3.3439	0.0432	-0.2134	0.8773
German	Sealsfield	41	7.0900	0.0135	-0.2629	0.9434
German	Tucholsky	1	1.7589	0.2316	-0.4071	0.5721
German	Wedekind	2	1.3571	0.0000	-0.1905	0.6667
Hawaiian	Anonymous, Laieikawai	268	23.5281	-0.0226	-0.7187	0.9888
Hungarian	Ady Endre	98	13.1223	-0.0234	-0.5407	0.9788
Italian	Ciampi, End of Year speeches	12	4.8178	-0.1084	-0.6568	0.8602
Italian	Cossiga, End of Year speeches	6	3.7264	0.2156	-0.2026	0.8155
Italian	Einaudi, End of Year speeches	7	4.0619	-0.1097	-0.6818	0.8761
Italian	Gronchi, End of Year speeches	19	6.2143	-0.1087	-0.7512	0.8820
Italian	Leone,End of Year speeches	15	5.7432	-0.0893	-0.7003	0.8166
Italian	Napolitano, End of Year speeches	17	5.7003	-0.1028	-0.6907	0.9537
Italian	Pertini, End of Year speeches	5	3.1923	-0.0378	-0.4760	0.8700
Italian	Saragat, End of Year speeches	11	4.3898	-0.1077	-0.5952	0.9279
Italian	Scalfaro, End of Year speeches	2	1.2938	-0.0505	-0.1986	0.6494
Italian	Segni, End of Year speeches	1	-	-	-	-
Latin	Apuleius	14	4.6264	0.0141	-0.3186	0.9088
Latin	Horatius	2	1.2938	-0.0505	-0.1986	0.6494
Latin	Vergilius	5	2.2971	-0.0288	-0.2589	0.8368
Macedonian	Transl. Ostrovskij's KZS	5	2.4000	0.0000	-0.2546	0.8597
Polish	Transl. Ostrovskij's KZS	4	1.1164	-0.0862	-0.1128	0.7065
Romanian	Eminescu	3734	73.1356	-0.0048	-0.4949	0.9834
Russian	Lermontov	194	20.6301	-0.0193	-0.6853	0.9664
Russian	Ostrovskij's KZS	5	2.4000	0.0000	-0.2545	0.8597
Russian	Pushkin	251	23.0542	-0.0129	-0.6257	0.9671
Sebian	Transl. Ostrovskij's KZS	5	1.2084	-0.0952	-0.1208	0.4799

Slovak	Bachletová	701	37.0289	-0.0116	-0.6484	0.9838
Slovak	Svoraková	70	9.1905	0.0395	-0.4574	0.9698
Slovak	Transl. Ostrovskij's KZS	7	3.4826	0.1432	-0.2325	0.8245
Sloven	Transl. Ostrovskij's KZS	9	2.6875	-0.0266	-0.2020	0.6926
Sorbian	Transl. Ostrovskij's KZS	8	2.1567	-0.0872	-0.2159	0.8929
Ukrainian	Transl. Ostrovskij's KZS	2	1.4606	0.0869	-0.1772	0.6881

Table 13
The parameters and the determination coefficients of the fitting function
(number of similarities *S* descending)

Language	**Writer**	***S***	***a***	***b***	***c***	R^2
Romanian	Eminescu	3734	73.1356	-0.0048	-0.4949	0.9834
German	Droste-Hülshoff	2164	64.8195	-0.0077	-0.6803	0.9863
Slovak	Bachletová	701	37.0289	-0.0116	-0.6484	0.9838
German	Paul	563	30.4845	-0.0122	-0.5533	0.9671
English	Byron	362	26.7450	-0.0159	-0.6496	0.9210
Hawaiian	Anonymous, Laieikawai	268	23.5281	-0.0226	-0.7187	0.9888
Russian	Pushkin	251	23.0542	-0.0129	-0.6257	0.9671
Russian	Lermontov	194	20.6301	-0.0193	-0.6853	0.9664
German	Kafka	125	14.7439	-0.0015	-0.4117	0.9121
German	Schiller	115	14.6751	-0.0029	-0.4584	0.9645
Hungarian	Ady Endre	98	13.1223	-0.0234	-0.5407	0.9788
German	Heine	78	10.7463	-0.0352	-0.5097	0.9375
Slovak	Svoraková	70	9.1905	0.0395	-0.4574	0.9698
German	Sealsfield	41	7.0900	0.0135	-0.2629	0.9434
Czech	Husák, New Year speeches	31	7.8289	0.0161	-0.4115	0.8229
German	Meyer	28	7.6049	-0.0618	-0.6701	0.9638
English (Joyce)	Joyce's Finnegans Wake	24	5.0441	-0.0324	-0.3103	0.9450
Czech	Havel, New Year speeches	23	5.9346	-0.0452	-0.4604	0.9422
German	Lessing	20	-	-	-	-
Italian	Gronchi, End of Year speeches	19	6.2143	-0.1087	-0.7512	0.8820
Italian	Napolitano, End of Year speeches	17	5.7003	-0.1028	-0.6907	0.9537
Italian	Leone,End of Year speeches	15	5.7432	-0.0893	-0.7003	0.8166
Latin	Apuleius	14	4.6264	0.0141	-0.3186	0.9088
Czech	Klaus, New Year speeches	13	5.2864	-0.0879	-0.6670	0.9402
German	Chamisso	13	3.3894	-0.0659	-0.2980	0.7572
German	Goethe	12	5.1542	-0.1143	-0.7375	0.9451
German	Löns	12	4.8207	0.0185	-0.4053	0.9233
Italian	Ciampi, End of Year speeches	12	4.8178	-0.1084	-0.6568	0.8602

Italian	Saragat, End of Year speeches	11	4.3898	-0.1077	-0.5952	0.9279
Czech	Novotný, New Year speeches	11	3.8546	0.0000	-0.3091	0.8447
German	Schnitzler	10	3.3439	0.0432	-0.2134	0.8773
German	Novalis	9	4.2728	0.2970	-0.1163	0.8316
Sloven	Transl. Ostrovskij's KZS	9	2.6875	-0.0266	-0.2020	0.6926
Sorbian	Transl. Ostrovskij's KZS	8	2.1567	-0.0872	-0.2159	0.8929
German	Eichendorff	8	2.2497	-0.0444	-0.1795	0.6881
Italian	Einaudi, End of Year speeches	7	4.0619	-0.1097	-0.6818	0.8761
Slovak	Transl. Ostrovskij's KZS	7	3.4826	0.1432	-0.2325	0.8245
German	Rückert	7	3.6088	-0.1772	-0.6986	0.7958
Italian	Cossiga, End of Year speeches	6	3.7264	0.2156	-0.2026	0.8155
Italian	Pertini, End of Year speeches	5	3.1923	-0.0378	-0.4760	0.8700
Macedonian	Transl. Ostrovskij's KZS	5	2.4000	0.0000	-0.2546	0.8597
Russian	Ostrovskij's KZS	5	2.4000	0.0000	-0.2545	0.8597
Latin	Vergilius	5	2.2971	-0.0288	-0.2589	0.8368
Croatian	Transl. Ostrovskij's KZS	5	1.2084	-0.0952	-0.1208	0.4799
Sebian	Transl. Ostrovskij's KZS	5	1.2084	-0.0952	-0.1208	0.4799
Belorussian	Transl. Ostrovskij's KZS	4	2.5270	0.0809	-0.2787	0.8531
Polish	Transl. Ostrovskij's KZS	4	1.1164	-0.0862	-0.1128	0.7065
Czech	Transl. Ostrovskij's KZS	4	1.6875	-0.0266	-0.1754	0.6926
Czech	Svoboda, New Year speeches	3	3.1209	0.5396	-0.1897	0.5300
Czech	Gottwald, New Year speeches	2	1.0000	-0.2000	-0.2000	1.0000
Czech	Zápotocký, New Year speeches	2	2.5000	0.0000	-0.6000	0.7000
Ukrainian	Transl. Ostrovskij's KZS	2	1.4606	0.0869	-0.1772	0.6881
German	Wedekind	2	1.3571	0.0000	-0.1905	0.6667
Italian	Scalfaro, End of Year speeches	2	1.2938	-0.0505	-0.1986	0.6494
Latin	Horatius	2	1.2938	-0.0505	-0.1986	0.6494
Italian	Segni, End of Year speeches	1	-	-	-	-
Bulgarian	Transl. Ostrovskij's KZS	1	3.4263	1.6787	-0.5174	0.7065
German	Raabe	1	1.7589	0.2316	-0.4071	0.5721
German	Tucholsky	1	1.7589	0.2316	-0.4071	0.5721
German	Keller	1	1.5000	0.0000	-0.4000	0.4000

From Table 13 we notice a simple power law dependence of the fitting parameter *a* on the number *S* of similarities, as graphically presented in Figure 7 below.

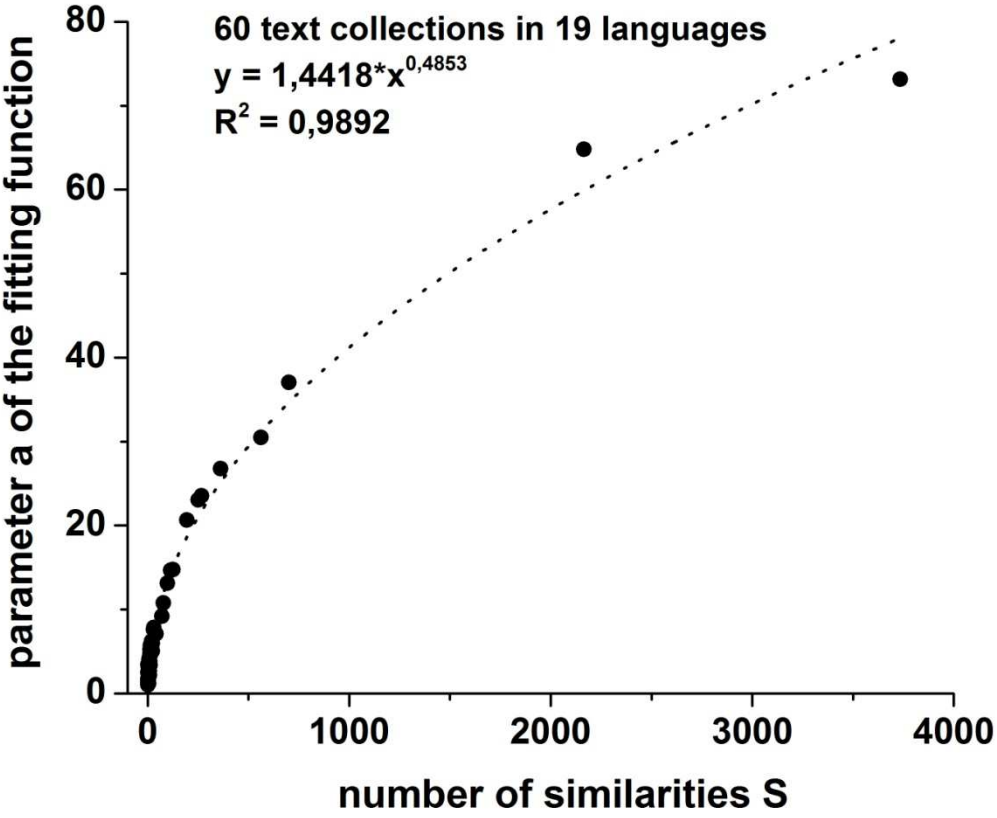

Figure 7. The dependence of the fitting parameter *a* on the number *S* of similarities

Preliminarily, it is not possible to draw conclusions about individual authors or languages. But a systematic study – especially of the origin of the texts - could reveal characteristic features of writers. In many cases we may suppose that the texts were written by the given author and a thorough study of the topical age of the author could show some new vistas. With some other texts, e.g. those written by the presidents, only historians could reveal who has written them.

The study of the complete work of a writer could reveal also the links between the individual parameters and other text properties and afford them better linguistic substantiation. In the present article we merely indicated some possible future research directions.

Acknowledgements
We thank Emmerich Kelih for the basic data of twelve Slavic languages, to Arjuna Tuzzi for the End-of-Year Speeches of Italian presidents, and to Radek Čech for the New Year Speeches of Czechoslovak or Czech presidents.

References

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

Popescu, I.-I., Altmann, G. (2015). A simplified lambda indicator in text analysis. *Glottometrics 30, 19-44.*

Popescu, I.-I., Zörnig, P., Altmann, G. (2013). Arc length, vocabulary richness and text size, *Glottometrics 25,43 – 53.*

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.

Appendix

Data of *N*, *Λ**, and Var(*Λ**) of German text collections used in the present article and not given in the reference article

ID	Writer	Text	*N*	*Λ**	Var(*Λ**)
1	Lessing 01	Der Besitzer des Bogens	114	1.4434	0.002139
2	Lessing 02	Die Erscheinung	208	1.6605	0.001514
3	Lessing 03	Der Esel mit dem Löwen	61	1.4048	0.003202
4	Lessing 04	Der Fuchs	47	1.4231	0.002424
5	Lessing 05	Die Furien	182	1.5026	0.001038
6	Lessing 06	Jupiter und das Schaf	362	1.6398	0.000626
7	Lessing 07	Der Knabe und die Schlange	231	1.6883	0.000906
8	Lessing 08	Minerva	74	1.6419	0.002414
9	Lessing 09	Der Rangstreit der Tiere	327	1.6148	0.001315
10	Lessing 10	Zeus und das Pferd	254	1.5054	0.001025
11	Novalis 01	Heinrich von Ofterdingen - Die Erwartung 1	2894	1.4903	0.000189
12	Novalis 02	Heinrich von Ofterdingen - Die Erwartung 2	3719	1.6052	0.000181
13	Novalis 03	Heinrich von Ofterdingen - Die Erwartung 3	5321	1.4187	0.000109
14	Novalis 04	Heinrich von Ofterdingen - Die Erwartung 4	2777	1.7274	0.000191
15	Novalis 05	Heinrich von Ofterdingen - Die Erwartung 5	8866	1.4275	0.000089
16	Novalis 06	Heinrich von Ofterdingen - Die Erwartung 6	4030	1.4502	0.000136
17	Novalis 07	Heinrich von Ofterdingen - Die Erwartung 7	1744	1.5836	0.000254
18	Novalis 08	Heinrich von Ofterdingen - Die Erwartung 8	2111	1.3748	0.000179
19	Novalis 09	Heinrich von Ofterdingen - Die Erwartung 9	8945	1.3651	0.000082
20	Novalis 10	Heinrich von Ofterdingen - Die Erfuellung	5367	1.4941	0.000110
21	Novalis 11	Hyazinth und Rosenblütchen	1358	1.6518	0.000415
22	Novalis 12	Neue Fragmente - Sophie	4430	1.5368	0.000126
23	Novalis 13	Neue Fragmente - Traktat vom Licht	1080	1.5701	0.000433
24	Goethe 01	Die neue Melusine	7554	1.2866	0.000080
25	Goethe 05	Der Gott und die Bajadere	559	1.7349	0.000686
26	Goethe 09	Elegie 19	653	1.7200	0.000532
27	Goethe 10	Elegie 13	480	1.7372	0.000541
28	Goethe 11	Elegie 15	468	1.7516	0.000563
29	Goethe 12	Elegie 2	251	1.6826	0.001208

30	Goethe 14	Elegie 5	184	1.6371	0.001433
31	Goethe 17	Der Erlkönig	225	1.3381	0.001143
32	Paul 01	Dr. Katzenbergers Badereise 1.	854	1.7609	0.000417
33	Paul 02	Dr. Katzenbergers Badereise 2. Reisezwecke	383	1.7671	0.000614
34	Paul 03	Dr. Katzenbergers Badereise 3. Ein Reisegefährte	520	1.7132	0.000674
35	Paul 04	Dr. Katzenbergers Badereise 4. Bona	580	1.7438	0.000459
36	Paul 05	Dr. Katzenbergers Badereise 5. Herr von Niess	1331	1.6618	0.000234
37	Paul 06	Dr. Katzenbergers Badereise 6. Fortsetzung der Abreise	526	1.6140	0.000415
38	Paul 07	Dr. Katzenbergers Badereise 7. Fortgesetzte Fortsetzung der Abreise	508	1.7205	0.000413
39	Paul 08	Dr. Katzenbergers Badereise 8. Beschluss der Abreise	402	1.7038	0.000873
40	Paul 09	Dr. Katzenbergers Badereise 9. Halbtagfahrt nach St. Wolfgang	1068	1.6249	0.000287
41	Paul 10	Dr. Katzenbergers Badereise 10. Mittags-Abenteuer	1558	1.6742	0.000215
42	Paul 11	Dr. Katzenbergers Badereise 11. Wagen-Sieste	2232	1.6428	0.000182
43	Paul 12	Dr. Katzenbergers Badereise 12. die Avantuere	620	1.7160	0.000487
44	Paul 13	Dr. Katzenbergers Badereise 13. Theodas ersten Tages Buch	1392	1.5312	0.000198
45	Paul 14	Dr. Katzenbergers Badereise 14. Missgeburten-Adel	1400	1.6809	0.000239
46	Paul 15	Dr. Katzenbergers Badereise 15. Hasenkrieg	1648	1.6436	0.000238
47	Paul 16	Dr. Katzenbergers Badereise 16. Ankunft-Sitzung	320	1.7927	0.000708
48	Paul 17	Dr. Katzenbergers Badereise I. Huldigungpredigt	1844	1.6896	0.000220
49	Paul 18	Dr. Katzenbergers Badereise II. Ueber Hebels alemannische Gedichte	870	1.7536	0.000456
50	Paul 19	Dr. Katzenbergers Badereise III. Rat zu urdeutschen Taufnamen	1236	1.7511	0.000230
51	Paul 20	Dr. Katzenbergers Badereise IIII. Dr. Fenks Leichenrede	2059	1.7252	0.000194
52	Paul 21	Dr. Katzenbergers Badereise V. Ueber den Tod nach dem Tode	3955	1.5098	0.000136
53	Paul 22	Dr. Katzenbergers Badereise 17. Blosse Station	478	1.7321	0.000457
54	Paul 23	Dr. Katzenbergers Badereise 18. Maennikes Seegefecht	656	1.7262	0.000460
55	Paul 24	Dr. Katzenbergers Badereise 19. Mondbelustigungen	1465	1.7201	0.000353
56	Paul 25	Dr. Katzenbergers Badereise 20. Zweiten Tages Buch	588	1.7426	0.000387
57	Paul 26	Dr. Katzenbergers Badereise 21. Hemmrad der Ankunft im Badeorte	1896	1.6113	0.000176
58	Paul 27	Dr. Katzenbergers Badereise 22. Niessiana	749	1.6349	0.000370
59	Paul 28	Dr. Katzenbergers Badereise 23. Ein Brief	241	1.7198	0.000756

60	Paul 29	Dr. Katzenbergers Badereise 24. Mittagtischreden	1825	1.6530	0.000209
61	Paul 30	Dr. Katzenbergers Badereise 25. Musikalisches Deklamatorium	388	1.6547	0.000724
62	Paul 31	Dr. Katzenbergers Badereise 26. Neuer Gastrollenspieler	1630	1.5962	0.000267
63	Paul 32	Dr. Katzenbergers Badereise 27. Nachtrag	163	1.6286	0.001064
64	Paul 33	Dr. Katzenbergers Badereise 28. Darum	596	1.7182	0.000479
65	Paul 35	Dr. Katzenbergers Badereise 30. Tischgebet und Suppe	1947	1.6236	0.000224
66	Paul 36	Dr. Katzenbergers Badereise 31. Aufdeckung und Sternbedeckung	425	1.6080	0.000553
67	Paul 37	Dr. Katzenbergers Badereise 32. Erkennszene	368	1.7013	0.000564
68	Paul 38	Dr. Katzenbergers Badereise 33. Abendtisch-Reden über Schauspiele	1218	1.6796	0.000248
69	Paul 39	Dr. Katzenbergers Badereise 34. Brunnen-Beaengstigungen	388	1.6881	0.000559
70	Paul 40	Dr. Katzenbergers Badereise 35. Theodas Brief an Bona	1370	1.5867	0.000267
71	Paul 41	Dr. Katzenbergers Badereise 36. Herzens-Interim	1032	1.6850	0.000351
72	Paul 42	Dr. Katzenbergers Badereise 37. Neue Mitarbeiter an allem	1546	1.5822	0.000206
73	Paul 43	Dr. Katzenbergers Badereise I. Die Kunst, einzuschlafen	4148	1.4967	0.000111
74	Paul 44	Dr. Katzenbergers Badereise II. Das Glueck	1881	1.6468	0.000193
75	Paul 45	Dr. Katzenbergers Badereise III. Die Vernichtung	2723	1.5617	0.000233
76	Paul 46	Dr. Katzenbergers Badereise 38. Wie Katzenberger …	3095	1.5260	0.000122
77	Paul 47	Dr. Katzenbergers Badereise 39. Doktors Hoehlen-Besuch	516	1.7296	0.000506
78	Paul 48	Dr. Katzenbergers Badereise 40. Theodas Hoehlen-Besuch	1200	1.6422	0.000315
79	Paul 49	Dr. Katzenbergers Badereise 41. Drei Abreisen	562	1.6929	0.000439
80	Paul 50	Dr. Katzenbergers Badereise 42. Theodas kuerzeste Nacht der Reise	430	1.6536	0.000817
81	Paul 51	Dr. Katzenbergers Badereise 43. Praeliminar-Frieden ...	3222	1.5439	0.000133
82	Paul 52	Dr. Katzenbergers Badereise 44. Die Stuben-Treffen	1731	1.6276	0.000238
83	Paul 53	Dr. Katzenbergers Badereise 45. Ende der Reisen und Noeten	1839	1.6403	0.000227
84	Paul 54	Dr. Katzenbergers Badereise I. Wuensche fuer Luthers Denkmal	6644	1.5137	0.000078
85	Paul 55	Dr. Katzenbergers Badereise II. Ueber Charlotte Corday	7854	1.4714	0.000076
86	Paul 56	Dr. Katzenbergers Badereise III. Polymeter	963	1.6049	0.000429
87	Chamisso 01	Peter Schlemihls wundersame Geschichte I	2210	1.4331	0.000181

88	Chamisso 02	Peter Schlemihls wundersame Geschichte II	1847	1.5475	0.000251
89	Chamisso 03	Peter Schlemihls wundersame Geschichte III	1428	1.5133	0.000325
90	Chamisso 04	Peter Schlemihls wundersame Geschichte IV	3205	1.4341	0.000142
91	Chamisso 05	Peter Schlemihls wundersame Geschichte V	2108	1.4396	0.000189
92	Chamisso 06	Peter Schlemihls wundersame Geschichte VI	1948	1.4489	0.000206
93	Chamisso 07	Peter Schlemihls wundersame Geschichte VII	1362	1.6108	0.000225
94	Chamisso 08	Peter Schlemihls wundersame Geschichte VIII	1870	1.4890	0.000234
95	Chamisso 09	Peter Schlemihls wundersame Geschichte IX	1320	1.5934	0.000498
96	Chamisso 10	Peter Schlemihls wundersame Geschichte X	1012	1.7105	0.000435
97	Chamisso 11	Peter Schlemihls wundersame Geschichte XI	1386	1.6026	0.000323
98	Droste 01	Die Judenbuche	16172	1.1813	0.000034
99	Droste 02	Der Tod des Erzbischofs Engelbert	884	1.7632	0.000504
100	Droste 03	Das Fegefeuer	700	1.8127	0.000489
101	Droste 04	Der Fundator	786	1.5840	0.000441
102	Droste 05	Die Schwestern	1274	1.6915	0.000291
103	Droste 08	Der Geierpfiff	965	1.6577	0.000358
104	Eichendorff 01	Aus dem Leben eines Taugenichts 1	3080	1.3977	0.000214
105	Eichendorff 02	Aus dem Leben eines Taugenichts 2	4100	1.2962	0.000155
106	Eichendorff 03	Aus dem Leben eines Taugenichts 3	4342	1.2458	0.000122
107	Eichendorff 04	Aus dem Leben eines Taugenichts 4	1781	1.4620	0.000252
108	Eichendorff 05	Aus dem Leben eines Taugenichts 5	1680	1.4437	0.000247
109	Eichendorff 06	Aus dem Leben eines Taugenichts 6	3223	1.2692	0.000148
110	Eichendorff 07	Aus dem Leben eines Taugenichts 7	2594	1.3582	0.000200
111	Eichendorff 08	Aus dem Leben eines Taugenichts 8	3987	1.3122	0.000125
112	Eichendorff 09	Aus dem Leben eines Taugenichts 9	3285	1.4098	0.000169
113	Eichendorff 10	Aus dem Leben eines Taugenichts 10	3052	1.3484	0.000163
114	Heine 01	Die Harzreise	19522	1.4637	0.000043
115	Heine 02	Die Heimkehr - Götterdämmerung	603	1.8489	0.000975
116	Heine 03	Die Heimkehr - Die Wallfahrt nach Kevlaar	394	1.4756	0.000863
117	Heine 04	Ideen. Das Buch Le Grand	20107	1.3276	0.000041
118	Heine 07	Belsazar	263	1.6562	0.001346
119	Rückert 01	Barbarossa	141	1.5548	0.002159
120	Rückert 02	Amor ein Besenbinder	327	1.5610	0.000518
121	Rückert 03	Der Frost	152	1.5790	0.001562
122	Rückert 04	Die goldne Hochzeit	721	1.6807	0.000335
123	Rückert 05	Erscheinung der Schnitterengel	212	1.6350	0.001147
124	Sealsfield 01	Das Cajuetenbuch - Die Praerie am Jacinto	1352	1.4613	0.000233
125	Sealsfield 02	Das Cajuetenbuch 1	4663	1.5255	0.000085
126	Sealsfield 03	Das Cajuetenbuch 2	3238	1.3974	0.000129
127	Sealsfield 04	Das Cajuetenbuch 3	3954	1.3964	0.000128
128	Sealsfield 05	Das Cajuetenbuch 4	3187	1.2664	0.000113

129	Sealsfield 06	Das Cajuetenbuch 5	2586	1.3936	0.000114
130	Sealsfield 07	Das Cajuetenbuch 6	2939	1.2851	0.000102
131	Sealsfield 08	Das Cajuetenbuch 7	4865	1.0936	0.000077
132	Sealsfield 09	Das Cajuetenbuch 8	7259	1.3435	0.000072
133	Sealsfield 10	Das Cajuetenbuch 9	4838	1.3183	0.000078
134	Sealsfield 11	Das Cajuetenbuch 10	3785	1.2630	0.000085
135	Sealsfield 12	Das Cajuetenbuch 11	3019	1.4581	0.000122
136	Sealsfield 13	Das Cajuetenbuch 12	2370	1.6261	0.000174
137	Sealsfield 14	Das Cajuetenbuch 13	2744	1.5788	0.000125
138	Sealsfield 15	Das Cajuetenbuch 14	4786	1.2925	0.000094
139	Sealsfield 16	Das Cajuetenbuch 15	4497	1.3907	0.000088
140	Sealsfield 17	Das Cajuetenbuch 16	6705	1.3890	0.000061
141	Sealsfield 18	Das Cajuetenbuch - Der Fluch Kishogues	4162	1.3148	0.000201
142	Sealsfield 19	Das Cajuetenbuch - Der Kapitaen	5626	1.1959	0.000074
143	Sealsfield 20	Das Cajuetenbuch - Callao 1825	8423	1.3851	0.000057
144	Sealsfield 21	Das Cajuetenbuch - Havanna 1816	6041	1.3958	0.000083
145	Sealsfield 22	Das Cajuetenbuch - Sehr Seltsam!	5748	1.1655	0.000065
146	Sealsfield 23	Das Cajuetenbuch - Ein Morgen im Paradiese	1752	1.5996	0.000262
147	Sealsfield 24	Das Cajuetenbuch - Selige Stunden	1696	1.5347	0.000237
148	Sealsfield 25	Das Cajuetenbuch - Das Diner	1368	1.6758	0.000204
149	Sealsfield 26	Das Cajuetenbuch - Der Abend	1517	1.4825	0.000188
150	Sealsfield 27	Das Cajuetenbuch - Die Fahrt und die Kajuete	4195	1.4422	0.000128
151	Sealsfield 28	Das Cajuetenbuch - Das Paradies der Liebe	1515	1.3435	0.000294
152	Keller 01	Romeo und Julia auf dem Dorfe	25625	1.1794	0.000039
153	Keller 02	Vom Fichtenbaum	301	1.7292	0.001266
154	Keller 03	Spiegel, das Kätzchen	13149	1.3131	0.000067
155	Keller 04	Das Tanzlegendchen	1896	1.7012	0.000291
156	Meyer 01	Der Schuss von der Kanzel 1	1523	1.7596	0.000236
157	Meyer 02	Der Schuss von der Kanzel 2	573	1.6751	0.000575
158	Meyer 03	Der Schuss von der Kanzel 3	1052	1.6805	0.000363
159	Meyer 04	Der Schuss von der Kanzel 4	2550	1.6057	0.000137
160	Meyer 05	Der Schuss von der Kanzel 5	1249	1.7156	0.000278
161	Meyer 06	Der Schuss von der Kanzel 6	833	1.7321	0.000401
162	Meyer 07	Der Schuss von der Kanzel 7	1229	1.7220	0.000286
163	Meyer 08	Der Schuss von der Kanzel 8	1028	1.7199	0.000354
164	Meyer 09	Der Schuss von der Kanzel 9	776	1.7540	0.000526
165	Meyer 10	Der Schuss von der Kanzel 10	940	1.6510	0.000392
166	Meyer 11	Der Schuss von der Kanzel 11	2398	1.6195	0.000168
167	Raabe 01	Im Siegeskranze	13045	1.1509	0.000065
168	Raabe 02	Eine Silvester-Stimmung	3173	1.1830	0.000156
169	Raabe 03	Ein Besuch	2690	1.3553	0.000208

170	Raabe 04	Deutscher Mondschein	6253	1.4333	0.000099
171	Raabe 05	Theklas Erbschaft	5087	1.4354	0.000100
172	Löns 01	Der Werwolf - 1. Die Haidbauern	1672	1.5133	0.000333
173	Löns 02	Der Werwolf - 2. Die Mansfelder	2988	1.2155	0.000182
174	Löns 03	Der Werwolf - 3. Die Braunschweiger	4063	1.1609	0.000130
175	Löns 04	Der Werwolf - 4. Die Weimaraner	3713	1.1758	0.000147
176	Löns 05	Der Werwolf - 5. Die Marodebruede	4676	1.1459	0.000148
177	Löns 06	Der Werwolf - 6. Die Bruchbauern	4833	1.2029	0.000135
178	Löns 07	Der Werwolf - 7. Die Wehrwoelfe	7743	1.1245	0.000099
179	Löns 08	Der Werwolf - 8. Die Schnitter	6093	1.2548	0.000120
180	Löns 09	Der Werwolf - 9. Die Kirchenleute	9252	1.0884	0.000079
181	Löns 10	Der Werwolf - 10. Die Hochzeiter	6546	1.1507	0.000089
182	Löns 11	Der Werwolf - 11. Die Kaiserlichen	4102	1.3062	0.000159
183	Löns 12	Der Werwolf - 12. Die Schweden	4432	1.2441	0.000142
184	Löns 13	Der Werwolf - 13. Die Haidbauern	1361	1.3839	0.000304
185	Wedekind 01	Mine-Haha I	4035	1.2788	0.000094
186	Wedekind 02	Mine-Haha II	6040	1.1756	0.000068
187	Wedekind 03	Mine-Haha III	7402	1.1370	0.000073
188	Wedekind 04	Mine-Haha IV	1297	1.6225	0.000245
189	Wedekind 05	Rabbi Esra	1935	1.1024	0.000245
190	Wedekind 06	Frühlingsstürme	5955	1.2063	0.000096
191	Wedekind 07	Silvester	605	1.6231	0.000448
192	Wedekind 08	Der Verführer	2033	1.5036	0.000221
193	Schnitzler 01	Der Sohn	2793	1.2943	0.000159
194	Schnitzler 02	Albine	1936	1.4703	0.000165
195	Schnitzler 03	Amerika	801	1.5443	0.000355
196	Schnitzler 04	Der Andere	2489	1.3412	0.000238
197	Schnitzler 05	Die Braut	2123	1.4308	0.000256
198	Schnitzler 06	Erbschaft	1539	1.4538	0.000207
199	Schnitzler 07	Die Frau des Weisen	5652	1.1140	0.000109
200	Schnitzler 08	Der Fürst ist im Hause	1711	1.3473	0.000217
201	Schnitzler 09	Das Schicksal	6552	1.2622	0.000068
202	Schnitzler 10	Welch eine Melodie	1349	1.5360	0.000254
203	Schnitzler 11	Frühlingsnacht im Seziersaal	1595	1.6144	0.000367
204	Schnitzler 12	Die Toten schweigen	6173	1.1323	0.000141
205	Schnitzler 13	Er wartet auf den vazierenden Gott	1184	1.4900	0.000285
206	Schnitzler 14	Mein Freund Ypsilon	3900	1.3084	0.000114
207	Kafka 01	In der Strafkolonie	10256	1.0665	0.000066
208	Kafka 02	Ein Bericht für eine Akademie	3181	1.4810	0.000183
209	Kafka 03	Betrachtung - Kinder auf der Landstraße	1072	1.5094	0.000263
210	Kafka 04	Betrachtung - Entlarvung eines Bauernfängers	625	1.4896	0.000443

211	Kafka 05	Betrachtung - Der plötzliche Spaziergang	247	1.6855	0.001239
212	Kafka 06	Betrachtung - Entschlüsse	178	1.7447	0.000927
213	Kafka 07	Betrachtung - Der Ausflug ins Gebirge	132	1.4940	0.002164
214	Kafka 08	Betrachtung - Das Unglück des Junggesellen	139	1.6342	0.002001
215	Kafka 09	Betrachtung - Der Kaufmann	596	1.6670	0.000519
216	Kafka 10	Betrachtung - Zerstreutes Hinausschaun	86	1.3721	0.001930
217	Kafka 11	Betrachtung - Der Nachhauseweg	151	1.5440	0.001762
218	Kafka 12	Betrachtung - Die Vorüberlaufenden	160	1.4327	0.001612
219	Kafka 13	Betrachtung - Der Fahrgast	232	1.5498	0.000899
220	Kafka 14	Betrachtung - Kleider	142	1.6824	0.002331
221	Kafka 15	Betrachtung - Die Abweisung	189	1.6501	0.000978
222	Kafka 16	Betrachtung - Zum Nachdenken für Herrenreiter	255	1.6987	0.000856
223	Kafka 17	Betrachtung - Das Gassenfenster	111	1.6031	0.003365
224	Kafka 18	Betrachtung - Wunsch, Indianer zu werden	61	1.3756	0.002443
225	Kafka 19	Betrachtung - Die Bäume	41	1.2981	0.004302
226	Kafka 20	Betrachtung - Unglücklichsein	1402	1.3399	0.000353
227	Kafka 21	Ein Brudermord	610	1.6940	0.000364
228	Kafka 22	Ein Landarzt	2129	1.4960	0.000208
229	Kafka 23	Der Geier	255	1.5005	0.001099
230	Kafka 24	Vor dem Gesetz	584	1.3785	0.000537
231	Kafka 25	Ein Hungerkünstler	3414	1.3392	0.000108
232	Kafka 26	Nachts	134	1.5874	0.001672
233	Kafka 27	Das Schweigen der Sirenen	428	1.5063	0.000512
234	Kafka 28	Die Sorge des Hausvaters	470	1.5691	0.000409
235	Tucholsky 01	Schloss Gripsholm 1	8544	1.2719	0.000071
236	Tucholsky 02	Schloss Gripsholm 2	7106	1.1415	0.000059
237	Tucholsky 03	Schloss Gripsholm 3	9699	1.1505	0.000055
238	Tucholsky 04	Schloss Gripsholm 4	7415	1.1201	0.000057
239	Tucholsky 05	Schloss Gripsholm 5	4823	1.1792	0.000098

Glottometrics 32, 2015, 55-62

Golden section in Chinese Contemporary Poetry

Xiaxing Pan
Hui Qiu
*Haitao Liu**[1]

Abstract. Golden section is one of the most famous aesthetic properties in the arts. The present study explores the golden section in Chinese contemporary poetry texts in terms of the 'h-point' on the word-frequency-distribution curves as well as their 'feet'. It demonstrates that the golden section of the selected Chinese poetry is not sitting on the 'h-point', but it is possible for us to investigate the 'feet' of them.

Keywords: *Chinese contemporary poetry, golden section, h-point, word-frequency distribution, Zipf-Alekseev model, syllable*

1. Introduction

'Golden section' is a notion representing an interesting aesthetic proportion. It is also named as 'golden ratio' or 'divine ratio'. Benjafield & Adam-Webber (1976: 11) claim that it has had a ubiquitous influence on Western thought. Lots of Western architects and artists often incorporate it in their works. For example, Median (1976) argues that the construction of artworks like the painting ─ the *Madonna Enthroned* by Duccio, graves ─ the *Dying Lioness*, architectures ─ the Parthenon, as well as music ─ piano sonatas of Mozart (Putz 1995), etc. all follow the principle of golden section. Obviously, this notion is widely accepted as a standard of beauty in aesthetics.

Beauty searching is one of the fundamental aesthetic functions of art, so as poetry. Poetry always pursues beauty through different ways, for instance, rhyme, rhythm, word forms, etc. Aristotle, in his famous work *On the Art of Poetry*,[2] considers that: 'Epic poetry and Tragedy, as also Comedy, Dithyrambic poetry, and most flute-playing and lyre-playing, are all, viewed as a whole, modes of imitation'. While golden section is the most astonishing number in natural world (Fett 2006: 173), it is the most interesting aesthetic properties poetry would like to imitate or create. However, compared with other forms of art, golden section in poetry is not obviously demonstrated, which invites a deep exploration.

The present paper tends to explore the phenomenon of golden section in Chinese contemporary poetry in the following steps: 1) The first step (section 2) is to define what the golden section is. 2) Being different from the music/painting-like aesthetic forms, poetic texts are art of language, so the second step (section 3) is to verify the golden section from the per

[1] Address correspondence to: Haitao Liu, Department of Linguistics, Zhejiang University, 310058, Hangzhou, Zhejiang, China. Email address: lhtzju@gmail.com
[2] the website of the book: http://www.authorama.com/the-poetics-1.html

spective of linguistics. 3) Data source of the paper is restricted to the contemporary poetic texts written later than 1916. The third part (section 4) is to interpret the reason why it prefers contemporary poetry to ancient ones. 4) It is not so easy to find out the golden section of Chinese contemporary poetry. The fourth part (section 5) mainly displays some attempts and experiments to explore the golden section of the poetry, and finally make a brief discussion.

2. Golden section

Mathematically, Mark (Fett 2006: 158) symbolized the golden section as *phi*, which is the first Greek letter in the name of Phidias. If we divide a line AB into two segments AC and BC, postulating AC = *x*, BC = 1, so the length of line AB is *x*+1. When the ratio of the larger segment AC is related to the smaller one BC exactly as the whole line AB is related to the larger part AC, we can get:

(1) $\frac{AC}{BC} = \frac{AB}{AC}$.

Inserting *x*, 1, *x*+1 into (1) respectively, we can obtain:

(2) $\frac{x}{1} = \frac{x+1}{x}$.

The solution of (2) is: $x = \frac{1\pm\sqrt{5}}{2}$, which is named ‘golden ratio’ or ‘golden section’.

3. Golden section in texts

It has been corroborated that golden section does exist in poetic texts. As mentioned in Bews (1970), Virgilian scholars believe that the mathematical approach to structure analysis has become a major aspect in Virgilian scholarship. Lots of aesthetic properties on golden section have been found out in Virgil’s epic poems, like *Eclogues*, *Georgics*, and *Aeneid*. Such kinds of research mainly focus on the counting of lines in speeches or passages. However, recent studies on golden section of texts concentrate on the so-called ‘h-point’ in the sequence of the word-frequency curves (Martináková et al. 2008; Popescu et al. 2009, 2012; Tuzzi et al. 2010a, 2010b).

Cited from Hirsch (2005), the concept of ‘h-point’ is introduced into linguistics by Popescu (2006), and soon the relative golden section is introduced by Popescu & Altmann (2006). Both concepts are related to the word rank-frequency distribution. According to Hirsch (2005), ‘h-point’ is considered as a simple and useful way to characterize the scientific output of a researcher. If the number of a researcher’s published papers is *N*, and the citation of every single paper is *P*, posting the papers in a descending order in terms of *P*, it is easy to find out a cross point on the descending curve, which is called ‘h-point’. Displaying similarly as the descending curve of citation, any word rank-frequency distribution curve has a ‘turning point’

which performs as an 'h-point' as well. As shown in Fig. 1, the 'h-point' on the word rank- frequency distribution curve divides it into two main areas: the one above the h-point is the synsemantic branch of a text, where most of the words are synsemantics, and the one below is the autosemantic branch, where most of the words are autosemantics.

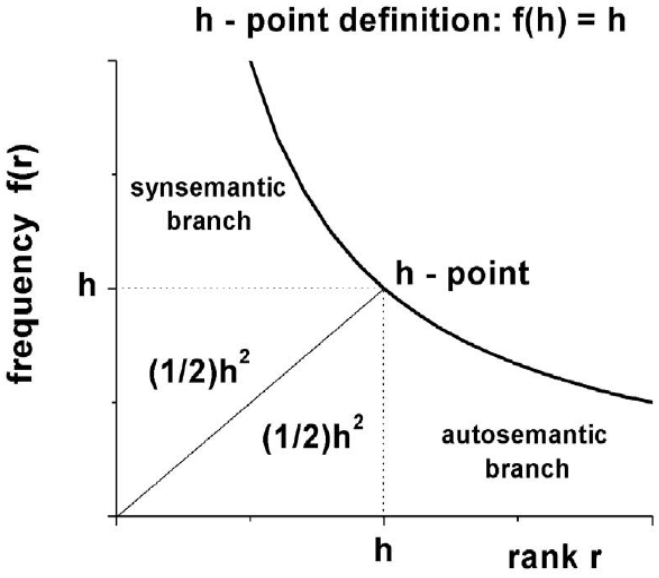

Figure 1. The definition of the 'h-point' (cf. Popescu & Altmann 2006: 25)

Three main points on any descending curve should be paid attention to. As can be seen on the word rank-frequency distribution curve of Lu Zhiwei's *Moonlight in the Cherry Tree*[3] in Fig. 2, point *A* is the word ranked 1st with the highest frequency 10, point *B* is the word ranked 68th with the lowest frequency 1, point *h* is the so-called 'h-point' whose frequency equals its rank order. Obviously, these *A, B,* h-like points on any descending curve can form a triangle with an angle α (∠*AhB* in Fig.2). This angle is metaphorically called 'writer's view', where the author 'sitting' and controlling the equilibrium between autosemantic and synsemantic (Popescu et al. 2007, 2009), with its value converges to the golden section.

[3] '*Moonlight in the Cherry Tree*' (lines are segmented into word units by spaces, and '//' represents lines of the poetry)
月光 在 樱树 , // 那 一 天 的 总 温习 // 早已 把 我 的 同年 朋友 // 一 个个 送 到 黑酣 乡里 。// 月光 在 樱树 , // 校钟 正 敲 过 十一 点 。 // 从 没有 见 过 这样 的 妙景 , // 樱树 里 浮 出 几 条 白线 ! // 月光 在 樱树 , // 我 的 心 像 天 一样 圆 , // 我 的 上帝 像 空气 一样 近 , // 我 见 他 在 樱树 下 生活 。// 月光 在 樱树 , // 那 一 天 我 亲自 看见 了 。// 我 的 祖宗 梦想 不 到 的 // 我 用 肉眼 同 他 会面 了 。// 月光 在 樱树 , // 那 是 何等 样 的 光 ! // 我 以后 不再 做 杜甫 的 奴隶 , // 我 亲自 见 了 宇宙 的 文章 。

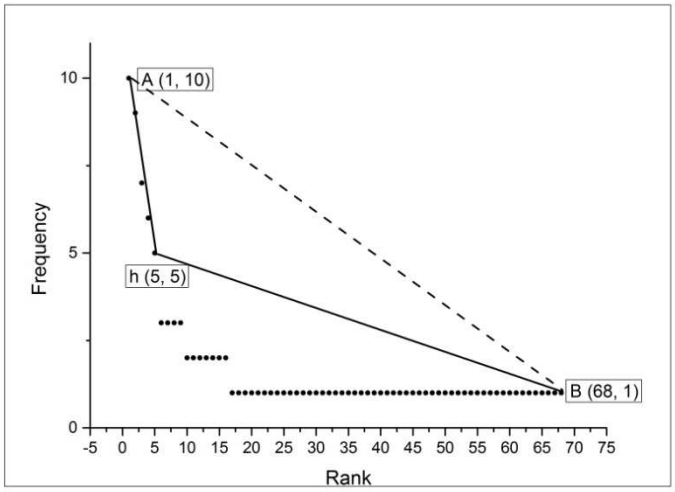

Figure 2. The word rank-frequency distribution curve of Lu Zhiwei's *Moonlight in the Cherry Tree*, as well as the h-point and the 'writer's view'

In order to calculate the value of *h* and the radian of ∠α, we need the following two functions: (3) describes the method approaching to the value of the 'h-point', and (4) for the cosine value of angle α.

$$(3)\quad h = \begin{cases} f(r), & if\ there\ is\ an\ r = f(r) \\ \frac{r_j * f(i) - r_i * f(j)}{r_j - r_i + f(i) - f(j)}, & if\ there\ is\ no\ r = f(r) \end{cases}$$

where $i > j$, $r_i < f(i)$, $r_j > f(j)$, r_n is the rank of words, and $f(r)$ is the relevant frequency.

$$(4)\quad cos\alpha = -\frac{(h-1)(f(1)-h)+(h-1)(V-h)}{\sqrt{(h-1)^2+(f(1)-h)^2} * \sqrt{(h-1)^2+(V-h)^2}}$$

where *V* is the number of word types, *f*(1) is the frequency of the word ranked 1. It can be found that, on a rank-frequency distribution curve, the rank order of the h-point doesn't always equal the corresponding frequency. For example, in Table 1, *f*(3) is 5, while *f*(4) equals 3. So we have to use the lower part of function (3) to calculate the correct value of *h* which is 3.67. Based on the value of *h*, the frequency of the first word, and the number of the word types of a text, it is easy to get the cosine of ∠α. Substitute the relevant values into function (4), the value of cosine α is -0.2432, and the radian is the arccosine of α which equals 1.8164.

Table 1
A rank-frequency distribution example

Rank	**Fr**	**Rank**	**Fr**	**Rank**	**Fr**	**Rank**	**Fr**	**Rank**	**Fr**	**Rank**	**Fr**
1	22	4	3	**7**	3	**10**	3	**13**	2	**16**	2
2	10	5	3	**8**	3	**11**	2	**14**	2	**17**	1
3	5	6	3	**9**	3	**12**	2	**15**	2	**18**	1

Tuzzi et al. (2010a: 31) propose that the golden section in texts is rather a matter of convergencence to the irrational golden number: 1.618... An investigation made by Tuzzi et al. (2010b: 95-106) on the trend of the radian of ∠α in 60 Italian presidential addresses texts proves that the values of the radians are relevant for the length of texts. The convergence curve in Figure 3 can be fitted by (5) well; the determination coefficient equals 0.833:

(5) $y = 1.618 + \frac{8.7094}{\sqrt{x}}$

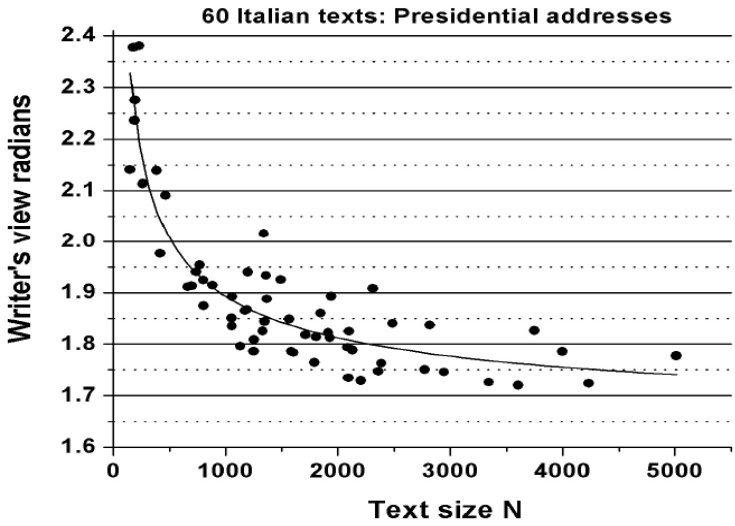

Figure 3. α radians in 60 Italian texts (cf. Tuzzi et al. 2010b: 101).

Accordingly, searching the golden section of Chinese contemporary poetry from the perspective of the h-point as shown in the previous studies seems like a reasonable attempt.

4. Data source

The two kinds of poetic texts, ancient poetry and contemporary poetry, are quite different. Fig. 4a is the rank-frequency curve of an ancient poetry *Zeng Wang Lun* (*To Wang Lun*) written by Li Bai in Tang Dynasty (A.D. 618-907), and Figure 4b is the rank-frequency curve of a contemporary work *YangChe Fu* (*Riskshaw Pullers*) written by Zang Kejia in 1930s. Segmenting the two texts into word units, reordering the word ranks into descending rankings, we can get the two rank-frequency curves. As can be seen, only the curve in Figure 4b fits the Zipf's distribution law well.

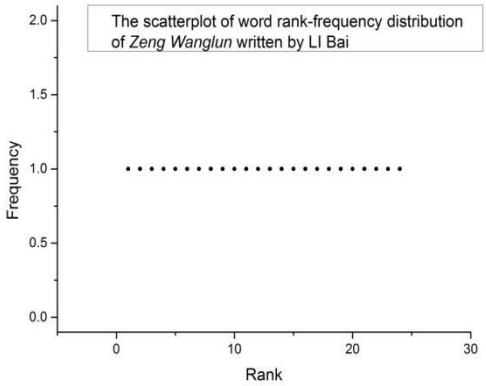

Figure 4a. The scatterplot of word rank-frequency distribution of *Zeng Wang Lun.*

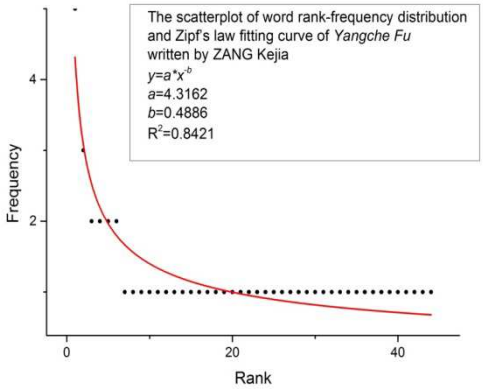

Figure 4b.The scatterplot of word rank-frequency distribution and Zipf's law fitting curve of *Yangche Fu.*

Like *Zeng Wang Lun* written by Li Bai, most of the ancient Chinese poetry are short and their TTRs (type-token ratio) are 1, which says that all the words in any poetry are different. Ancient poets in China kept their eyes on the arrangement of every single sentence of their works, even every single word. One of the most famous allusions is the selection between the two words TUI (push) and QIAO (knock) by Jia Dao.[4] The whole context of the poetry made an effort to create a quiet atmosphere. So the poet made a well-thought-out decision on choosing the word QIAO. This kind of poetic texts are under tight control of the poets.

It has been stated that texts which can follow Zipf's law are self-organized. Authors are always unconscious of the law. The contemporary poems are more normal and natural. The most convincing evidence is that most of the modern poetic texts abide by Zipf's law. Even there exist exceptions in contemporary poetry like *Shenghuo* (*Life*) written by Bei Dao (1949-) which is formed by a single word 'net'. Such kinds of texts are omitted from the study. The 'h-point' is sitting on a descending curve. So the precondition of the study is that the proceeding of a text has to be normal and natural. Thanks to the 'naturality' of contemporary poetry, we can search for their properties of beauty from the aspect of golden section.

We selected 297 plausible poetic texts arbitrarily from the website: http://www.shigeku.org/shiku/xs/index.htm.

The values of the radians of angle alpha are plotted in Figure 5.

[4] The poetry written by Jia Dao is *Ti Li Ning Youju (Inscription on the Tranquil House of Li Ning)*, and the famous verses are:

鸟宿池边树

僧敲月下门

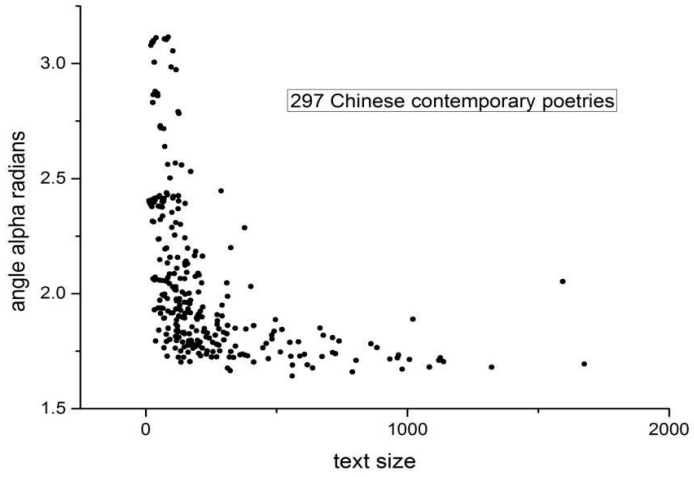

Figure 5. Angle alpha radians of the 297 selected contemporary poetry

5. Experiments and discussion

Fitting the function

(6) $y = 1.618 + \frac{a}{\sqrt{x}}$

to the relevant data we obtain a = 4.0934, and the determination coefficient R^2 = 0.3953, which means a bad fitting result. Then, we try to fit function (7) to the data:

(7) $y = a + \frac{b}{\sqrt{x}}$

we obtain a = 1.58916, b = 5.02352, R^2 = 0.40518, which still signalizes an unsatisfactory result. Accordingly, we may draw a conclusion that the radian of ∠α sitting on the 'h-point' cannot predict the golden section of Chinese contemporary poetic texts well. However, the data show that there are phenomena converging to the golden section.

We then turn to the study of rhythm and find that the word-frequency distributions of the poems act differently when we fit them by the Zipf-Alekseev model:

(8) $y = c * x^{a+b*lnx}$.

In the model fitting experiment, we picked up 24 out of the 297 pieces randomly and fitted their word frequency ranks by (8). Only 5 very short texts in the selected 24 poetic texts can be captured by this model. We then hypothesize that only the word frequency distribution of short and concise poetic texts can be fitted. In a further investigation of other 60 texts containing less than 20 lines, the result shows that 40 of them fit this model well, but the left 20 fail. With a comparison between the two groups of poetry, we conjecture that there is a boundary condition relevant to at least two factors: the number of poem lines and the number of word types -- when the lines of a poem are less than 20, and word types less than 76 at the same time, the model fits quite well, but fails if any of the conditions changes.

Meanwhile, the number of syllables plays an important role in the fitting process. According to our observation, we propose that, in Chinese contemporary poetry, lines, word types, and syllables are synthetically related. To measure the rhythms of the Chinese contemporary poems, the basic unit 'foot' is discussed. Mostly, one 'foot' is considered to be composed of two to three Chinese syllables (generally, a syllable coincides with a Chinese character), and one line of one poem is composed of three to five 'feet'. We suppose that the golden section of Chinese contemporary poetic texts may be hidden in the arrangement of the syllables, especially the proportion between the monosyllables and multi-syllables (including the disyllables), 'feet' in the verses, etc. We are looking forward to the testing in coming studies.

ACKNOWLEDGEMENTS

This work was supported by the National Social Science Foundation of China (11&ZD188).

References

Benjafield, J. & Adams-Webber, J. (1976). The golden section hypothesis. *British Journal of Psychology 67(1), 11-15.*

Bews, J. P. (1970). 'Aeneid' I and .618. *Phoenix 24(2), 130-143.*

Fett, B. (2006). An in-depth investigation of the divine ratio. *The Montana Mathematics Enthusiast 3(2), 157-175*.

Hirsch, J. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Science of the United States of America 102(46), 16569-16572.*

Martináková, Z., Mačutek, J., Popescu, I.-I. & Altmann, G. (2008). Some problems of musical texts. *Glottometrics 16, 80-110*.

Meian, H. (1976). The golden section and the artist. *Fibonacci Quarterly 14, 406-418*.

Popescu, I.-I. (2006). Text ranking by the weight of highly frequent words. *Exact methods in the study of language and text,* edited by Peter Grzybek and Reinhard Köhler: *557-567*. Berlin/New York: Mouton de Gruyter.

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

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

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

Popescu, I.-I., Čech, R. & Altmann, G. (2012). Some geometric properties of Slovak poetry. *Journal of Quantitative Linguistics 19(2), 121-131*.

Putz, J.F. (1995). The golden section and the piano sonatas of Mozart. *Mathematics Magazine 68(4), 275-282*.

Tuzzi, A., Popescu, I.-I. & Altmann, G. (2010a). The golden section in texts. *ETC - Empirical text and culture research 4: Dedicated to quantitative empirical studies of culture*, ed. by Andrew Wilson: *30-41*. Lüdenscheid: RAM-Verlag.

Tuzzi, A., Popescu, I.-I. & Altmann, G. (2010b). *Quantitative analysis of Italian texts*. Lüdenscheid: RAM-Verlag.

Glottometrics 32, 2015, 63-87

Probability distribution of interlingual lexical divergences in Chinese and English: 道 (*dao*) and *said* in *Hongloumeng*

Yu Fang & Haitao Liu[*]

Abstract: Previous studies have indicated that divergence exists in translation, which influences the quality of translation; such divergence follows some regularity which can be modeled by a probability distribution. The present article chose one verb *dao* from a Chinese literary classic *Hongloumeng*, and its English translation *said* to determine whether the frequencies of the two verbs develop according to a diversification process. Furthermore, we tend to find differences between the two English versions of Hawkes and Yang Xianyi regarding *dao*'s translation, and we also investigated the role of the three causes of divergence in the diversification process. The result indicates that both *dao*'s translations and the original text of *said* is in good agreement with the modified right-truncated Zipf-Alekseev distribution. Three major reasons were found for causing differences between the two translations: the nature of language, translators' subjectivity, and context. These are also the three major reasons for the diversification and differences of the two translated versions.

Keywords: probability distribution; interlingual lexical divergence; translation; dao; said

1. Introduction

A sentence, even a single word in a source language can be translated into different forms in a target language. Palmer & Wu (1995) investigated a specific lexical selection problem in translation — translating English *change-of-state* verbs into Chinese verbs. The result indicates that "break" in English can be translated as 打碎(*dasui*, hit into pieces), 击破 (hit into irregularly shaped pieces), 压断 (press into line shape), etc. in Chinese. This phenomenon can be attributed to the divergence of the two languages.

Since word choices have great impact on the quality of translation, causes of divergence have been widely studied, and one of the reasons lies in the nature of language. We all acknowledge that there are no two equivalent languages in the world, so divergence occurs when one language is translated into another. Saboor & Khan (2010) focused on lexical-semantic divergence for Urdu-to-English translation, and seven types of divergence have been discovered. Venkatapathy & Joshi (2007) proposed a generic discriminative re-ranking approach for the word alignment, which is able to make use of syntactic divergence features,

[*] Address correspondence to: Haitao Liu, Department of Linguistics, Zhejiang University, 310058, Hangzhou, Zhejiang, China. Email address: lhtzju@gmail.com

and to successfully decrease the alignment error rate by 2.3%. Similarly, Kulkarni et al. (2013) aimed to locate the structural and syntactic divergences in English-Marathi language pair through the translation pattern of English-Marathi constructions.

Those examinations have one commonality: all of them contribute to the divergence of the features of language, either in sentences or in words. Besides this reason, the translator's preference is another factor in causing such divergence. Lefevere (1992: p.1) defined translation as "a rewriting of an original text". This rewriting, of course, is conducted by translators with different backgrounds and can reflect that "a certain ideology and a poetics and as such manipulate literature to function in a given society in a given way" (p.1). Moreover, just like Shakespeare said there are one thousand Hamlets in one thousand people's eyes, translators may read the writers' intention in various ways. Crisafulli (1999), through analyzing H.F. Cary's translated version of Dante's 'Comedy', found that the translator not only conveys the original meaning but also works as a textual critic.

Context also cannot be ignored in causing such divergence, and it is central to translation to some extent (House, 2006). Malinowski (1935) argued that translation becomes ''rather the placing of linguistic symbols against the cultural background of a society than the rendering of words by their equivalents in another language'' (p. 18). In other words, the meaning of a linguistic unit cannot be fully understood and translated into another language unless one takes into consideration the interrelationship between linguistic units and the context of the situation.

Thus, it can be shown that previous studies focused on two aspects: (a) finding divergence to reduce the error of machine translation; (b) reasons for causing such divergence. Since reasons can be found to explain those divergences, we assume that those different translated versions of a word are predictable, in other words, those versions should follow a certain probability distribution; however, little research has been carried out in this dimension.

Studies on probability distribution could be found in word and sentence analysis. In word level, Rothe (1991) calculated all uses of *and* in text and examined their various denotations and functions. After getting a representative data base, a one-dimensional empirical curve was modeled to represent the distribution of the data. Voloshynovska (2011) applied the modification of Lavalette's function to scientific and belletristic literature and found that the fitting parameters of the function displayed characteristic values distinguishing between those two literature genres concerning a rather broad range of texts in English. And, at the sentence level, Köhler & Altmann (2000) found that the properties of syntactic constructions and categories are lawfully distributed according to a few probability distributions based on Susanne corpus and Negra-Korpus with constituency annotation. Liu (2009) investigated the probability distribution of the dependency relation extracted from a Chinese dependency treebank, and the results indicated that most of the investigated distributions could be excellently fitted using the modified right-truncated Zipf- Alekseev distribution (p. 256).

However, the probability distribution of interlingual lexical levels has rarely been studied. Liu (2009) found that the distributions of the active valency of a verb and the passive valency of a noun develop according to a diversification process (Altmann 2005). In this study, we hypothesize that a relatively simple probability distribution not only exists on the syntactic level and interlingual lexical level, but it makes its way also in the interlingual lexical level

during a translation process. “Diversification is a process of enlarging the number of forms or meanings of any linguistic entity” (Strauss & Altmann, 2006). It comprises a number of phenomena, among which, the increase in the meanings of a word (polysemy) is under research.

If a word diversifies and acquires several meanings, which are not used with the same frequency, in other words, the frequencies are heterogeneous (Altmann, 1996), and if we rank the frequencies of meanings of a word according to their magnitude, then we can conjecture that it follows a certain distribution, as “diversification is one of the lawlike processes operating on all levels of language.” (Altmann, 1996, p. 234).

Previous studies have proposed three sources for the emergence of divergence: the nature of languages, translators’ subjectivity and the context. In literature translation, divergence may evoke diversification for a word, which can acquire many translated versions due to these three reasons; thus, we can assume that the distribution of translations of one word develops according to a diversification process.

In this paper, we will select one verb 道(*dao*) from a Chinese literary work *Hongloumeng* and *said* from the two translated versions. In this paper, *dao* has a similar meaning with *say*, *talk* and *speak* in English. We will investigate the probability distribution of *dao* in all possible translations in the two English versions by David Hawkes and Yang Xianyi, with the probability distribution of the original text of *said* also being explored. Moreover, we will compare the differences of these two translations concerning the word *dao* and explain them in a specific context. This research is devoted to the following questions:

Question 1: Does the distribution of *dao*’s translations and the original text of *said* develop according to a diversification process? In other words, can the probability distribution of the translated Chinese word *dao* and the corresponding Chinese characters of *said* be smooth enough to be described by a relatively simple mathematical formula?

Question 2: If the answer to the first question is positive, is there any difference between the models of the two English translation versions?

Question 3: What is the function of the three causes of divergence, i.e. the nature of languages, translators’ subjectivity and the context, in this diversification process and in the differences between the versions?

This paper contains four sections. Section 2 describes the material and methods used. Section 3 presents the results of the distribution investigation, including the compatibility of translated *dao* and *said* to a distribution function, the analysis of the differences between the two translation versions, the causes of the diversification and those differences. Section 4 concludes this study.

2. Materials and method

To carry out this study smoothly, it is very important to choose an appropriate text: firstly, the literature work must have at least two translated versions; secondly, the chosen word in the original text must have a high frequency of usage, thus *Hongloumeng* and 道 *(dao)* is selected in this study.

Hongloumeng, as a masterpiece in Chinese traditional literature, written in the mideighteenth century, has nine complete or selective English translations (Chen & Jiang,

2003), and the two most popular versions are: *The Story of the Stone* translated by David Hawkes and John Minford and *A Dream of Red Mansions* translated by Yang Xianyi and his wife Gladys Yang. In a previous study, we have found that the standard frequency (a mean occurrence per 10000 words) of *said* in the selected 30 chapters is 1,596 in the selected Hawkes's version, and 530 in the selected Yang's version (Fang & Liu, 2015). It is easy for us to assume that *dao* in the original text also has a high frequency; thus, this word is suitable for this study.

The parallel Corpus of *A Dream of Red Mansions* (Ren, Sun & Yang, 2010), which was generated by Shaoxin University, was used to extract our material. In this corpus, Yang's translation is parallel with "Qixu version" and Hawkes' translation is parallel with "Chengyi version"[1] for the first 80 chapters; while for the remaining 40 chapters, both translations correspond to "Chengyi version". In this study, all 120 chapters are selected and *dao* is entered as the keyword. The Chinese character *dao* has several meanings such as "way, regularity and saying something". In this study, however, we only consider its meaning as "saying something", in other words, as an indicator of direct speech. Verb compounds with *dao* like 笑道(*said with a smile*) are beyond our consideration. Finally, we located 4737 concordance lines in Hawkes' translation and 3888 concordance lines in Yang's translation, after which we classified *dao*'s translations and calculated their frequencies. We found 89 translations of *dao* in Hawkes' version and 87 translations in Yang's version. The most frequent words is *said* in Hawkes' versions, with a standard frequency (a mean occurrence per 10000 words) of 4935 and without any indicator in Yang's version, with a standard frequency of 4105. There are only five words in Hawkes' version with a standard frequency over 100 times, but 13 words in Yang's version have a standard frequency over 100 times. See Appendix 1 and 2.

Corresponding with *dao*, we entered *said* as the keyword and selected those lines guiding the direct speech. Besides the single word *dao*, there were also some compound verbs containing 道, which were also translated into *said*. The result is shown in Table 1.

Table 1
Frequencies of *said* in Hawkes' and Yang's version

verbs in the original text	**Frequency in Hawkes**	**Frequency in Yang**
道(said)	2455	802
说(said)	361	317
笑道(said with a smile)	908	188
说道(spoke)	292	159
便说道(said)	98	74
因说道(so to speak)	46	33
因问道(so to ask)	50	20
叫(cried)	20	11

[1] There are twelve major versions of *Hongloumeng*, among which "Qixu version" and "Chengyi version" are included.

吩咐(ordered)	16	10
回道(replied)	18	8
劝道(persuaded)	10	8
忙道(said hurriedly)	51	7
回说(replied)	22	6
忙说(said hurriedly)	19	6
冷笑道(said coldly)	14	5
答道(answered)	11	5
云(said)	5	5
忙回(replied hurriedly)	7	3
答应道(answered)	7	2
命(ordered)	20	2
笑回道(answered…laughing)	5	2
啐道(said angrily)	8	0
骂道(scold)	7	0
问(ask)	65	0

As mentioned before, we hypothesize that the word's translation is a diversification process. This means that: "Every linguistic entity diversifies, i.e. it generates variants and secondary forms and acquires membership in different classes" (Strauss & Altmann, 2006). During the translation process, a word generates secondary meanings, and the ranked frequencies of a word's translations "abide by a rank-frequency distribution (or a rank-frequency series)" (Strauss & Altmann, 2006). More precisely, we assumed that the investigating distributions obey the Zipf-Alekseev model (Hřebíček, 1996, cited from Strauss & Altmann, 2006). Hřebíček used two assumptions:

(i) The logarithm of the ratio of the probabilities P_1 and P_x is proportional to the logarithm of the class size, i.e

$$\ln(P_1/P_x) \propto \ln x$$

(ii) The proportionality function is given by the logarithm of Menzerath´s law (Hierarchy), i.e.

$$\ln(P_1/P_x) = \ln(AX^b) \ln x$$

yielding the solution

$$P_x = P_1 x^{-(a+b \ln x)}, \quad x = 1,2,3, \ldots \qquad (1)$$

If (1) is considered a probability distribution, then P_1 is the normalizing constant, otherwise it is estimated as the size of the first class, $x = 1$. Very often, diversification distributions display a diverging frequency in the first class, while the rest of the distribution behaves regularly. In these cases, one usually ascribes the first class a special value α, modifying (1) as

$$P_x = \begin{cases} \alpha, & x = 1 \\ \frac{(1-\alpha)x^{(a+b\ln x)}}{T}, & x = 2,3,\ldots,(n) \end{cases} \quad (2)$$

where

$$T = \sum_{j=2}^{n} j^{-(a+b\ln j)}, 0 < \alpha < 1, \ \ a, b \ \in \Re$$

Distributions (1) or (2) are called Zipf-Alekseev distributions. If *n* is finite, (2) is called a modified right-truncated Zipf-Alekseev distribution.

3. Results and Discussion

3.1 Probability Distribution of translations of *dao* and *said*

We applied the Altmann-Fitter to the data shown in Appendix 1 and 2 and extracted the following information listed in Table 2, 3 and Figure 1, 2.

Table 2
Fitting the modified right-truncated Zipf-Alekseev distribution to different translations of *dao* in Hawkes' version

X[i]	F[i]	NP[i]		X[i]	F[i]	NP[i]
1	2352	2352.00		46	3	1.25
2	1293	890.80		47	3	1.19
3	318	435.09		48	3	1.13
4	175	255.89		49	3	1.07
5	83	167.38		50	3	1.02
6	47	117.34		51	3	0.97
7	47	86.40		52	3	0.93
8	25	64.35		53	2	0.88
9	24	50.55		54	2	0.84
10	23	40.46		55	2	0.81
11	22	33.08		56	2	0.77
12	19	27.47		57	2	0.74
13	19	23.11		58	2	0.71
14	15	19.68		59	2	0.68
15	14	16.92		60	2	0.65
16	13	14.68		61	2	0.62
17	13	12.83		62	2	0.60
18	12	11.29		63	2	0.57

19	12	10.00		64	2	0.55
20	11	8.91		65	1	0.53
21	9	7.98		66	1	0.51
22	9	7.18		67	1	0.49
23	8	6.48		68	1	0.47
24	8	5.88		69	1	0.46
25	7	5.35		70	1	0.44
26	7	4.89		71	1	0.42
27	7	4.47		72	1	0.41
28	6	4.11		73	1	0.39
29	6	3.79		74	1	0.38
30	6	3.50		75	1	0.37
31	5	3.24		76	1	0.36
32	5	3.00		77	1	0.34
33	5	2.79		78	1	0.33
34	4	2.60		79	1	0.32
35	4	2.43		80	1	0.31
36	4	2.27		81	1	0.30
37	4	2.12		82	1	0.29
38	4	1.99		83	1	0.28
39	4	1.87		84	1	0.27
40	3	1.76		85	1	0.27
41	3	1.66		86	1	0.26
42	3	1.56		87	1	0.25
43	3	1.48		88	1	0.24
44	3	1.40		89	1	0.24
45	3	1.32				
a = 1.5822, b = 0.1134, n = 89, α = 0.4965, DF = 59, R^2 = 0.9724						

In this and following similar tables: X[i] - the observed classes; F[i] - observed frequency; NP[i] - calculated frequency according to the modified right-truncated Zipf-Alekseev distribution; a, b, n and α- the parameters of the modified right-truncated Zipf-Alekseev distribution; DF - degrees of freedom; R^2 - the Coefficient of Determination

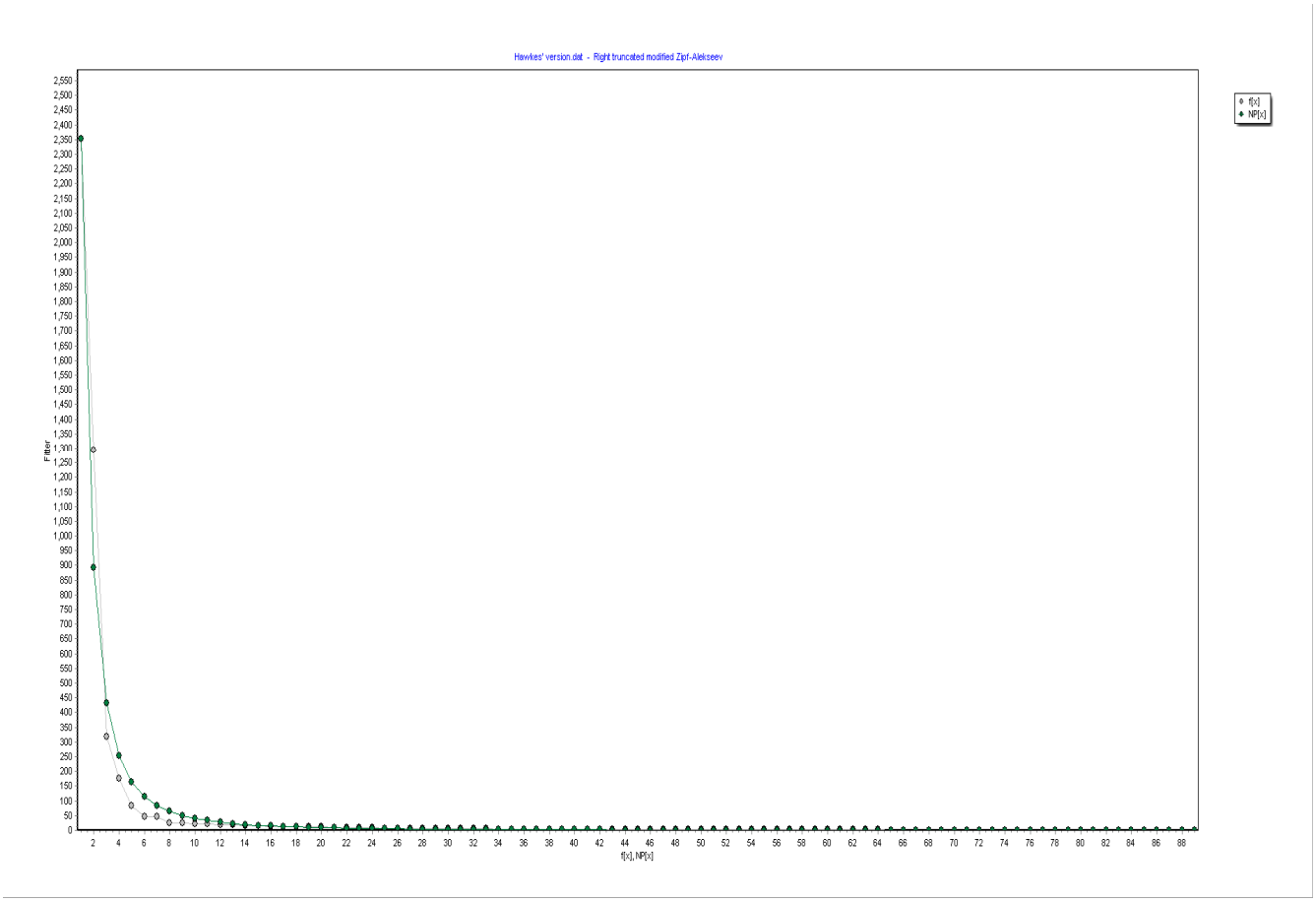

Figure 1 Fitting the modified right-truncated Zipf-Alekseev distribution to different translations of *dao* in Hawkes' version

Table 3
Fitting the modified right-truncated Zipf-Alekseev distribution to different translations of *dao* in Yang' s version

X[i]	F[i]	NP[i]		X[i]	F[i]	NP[i]
1	1596	1596.00		45	6	5.24
2	707	543.79		46	5	5.06
3	231	312.26		47	5	4.88
4	136	208.77		48	5	4.72
5	107	151.98		49	5	4.56
6	96	116.87		50	5	4.41
7	85	93.37		51	4	4.27
8	64	76.73		52	4	4.14
9	61	64.45		53	4	4.01
10	55	55.09		54	4	3.89
11	47	47.75		55	3	3.77
12	44	41.88		56	3	3.66
13	41	37.09		57	3	3.56
14	37	33.13		58	3	3.46

15	32	29.82		59	3	3.36
16	31	27.00		60	3	3.27
17	29	24.60		61	2	3.18
18	26	22.52		62	2	3.09
19	23	20.70		63	2	3.01
20	22	19.12		64	2	2.93
21	22	17.71		65	2	2.86
22	20	16.47		66	2	2.79
23	19	15.36		67	2	2.72
24	18	14.37		68	2	2.65
25	18	13.47		69	2	2.59
26	17	12.66		70	2	2.53
27	17	11.93		71	2	2.47
28	16	11.26		72	2	2.41
29	15	10.65		73	2	2.36
30	14	10.09		74	1	2.30
31	14	9.57		75	1	2.25
32	13	9.10		76	1	2.20
33	11	8.66		77	1	2.15
34	11	8.26		78	1	2.11
35	11	7.88		79	1	2.06
36	11	7.53		80	1	2.02
37	10	7.20		81	1	1.98
38	10	6.90		82	1	1.94
39	10	6.62		83	1	1.90
40	9	6.35		84	1	1.86
41	7	6.10		85	1	1.82
42	7	5.87		86	1	1.79
43	7	5.65		87	1	1.75
44	6	5.44				
$a = 1.2870, b = 0.0453, n = 87, \alpha = 0.4105, DF = 82, R^2 = 0.9861$						

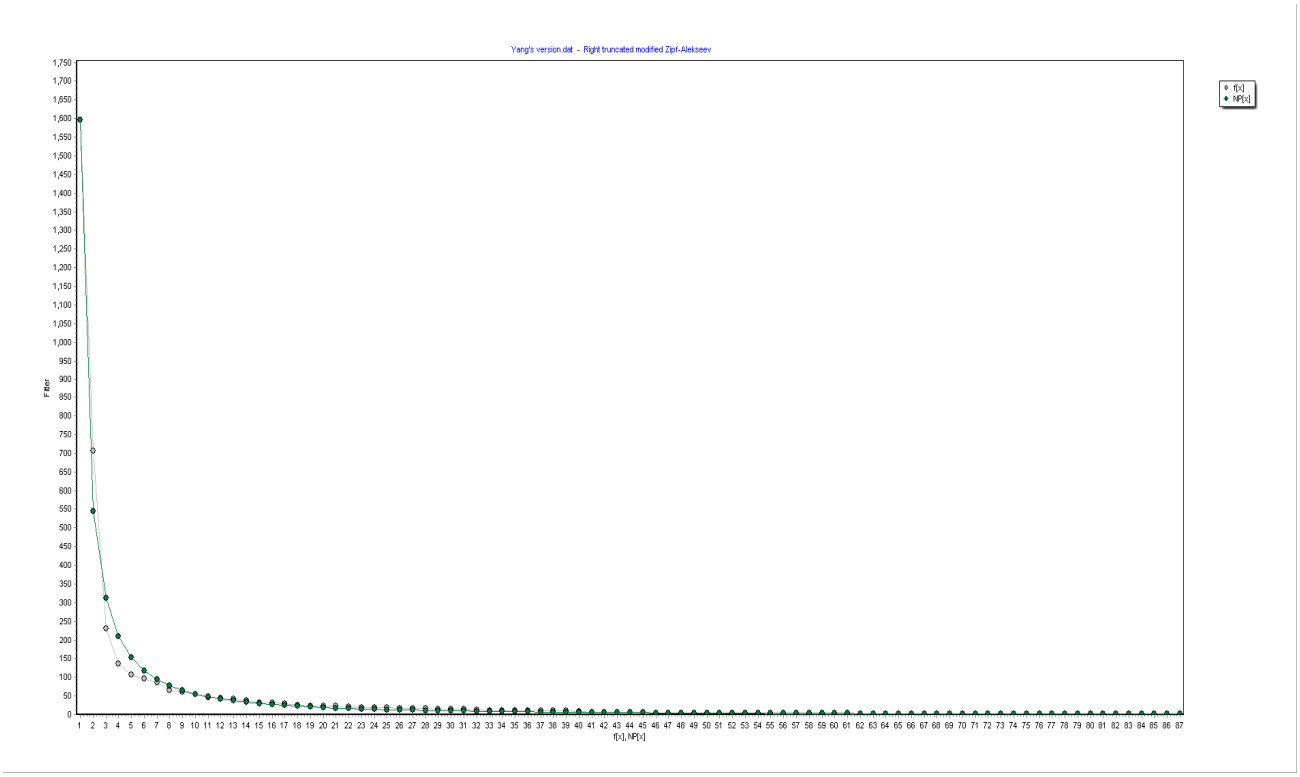

Figure 2 Fitting the modified right-truncated Zipf-Alekseev distribution to different translations of *dao* in Yang' s version

The Coefficient of Determination R^2, though defined for linear functions, "may be interesting in many cases and help to enlarge experience with this coefficient in connection with non-linear functions" (Altmann Fitter (3.1), User Guide, p. 10). The determination coefficients R^2 show that considering the data a simple function the two results are very good; that is to say, the distribution of *dao*'s translation in the two versions can be fitted well with the modified right-truncated Zipf-Alekseev function. Hence, if the frequencies of *dao*'s translations are in a descending order, they develop according to a diversification process.

To understand this diversification process better, we also applied the Altmann-Fitter to the data displayed in Table 1, and the results are listed in Table 4, 5 and Figure 3, 4.

Table 4
Fitting the modified right-truncated Zipf-Alekseev distribution to different expressions of *said* in Hawkes' version

X[i]	F[i]	NP[i]
1	2455	2455.00
2	908	835.21
3	361	408.35
4	292	234.69
5	98	148.75
6	65	100.74

7	51	71.60
8	50	52.79
9	46	40.08
10	22	31.15
11	20	24.69
12	20	19.90
13	19	16.26
14	18	13.46
15	16	11.26
16	14	9.51
17	11	8.10
18	10	6.95
19	8	6.01
20	7	5.22
21	7	4.57
22	7	4.01
23	5	3.55
24	5	3.15
a=1.3500, b=0.2315, n=24, α=0.5437, DF=19, R^2=0.9975		

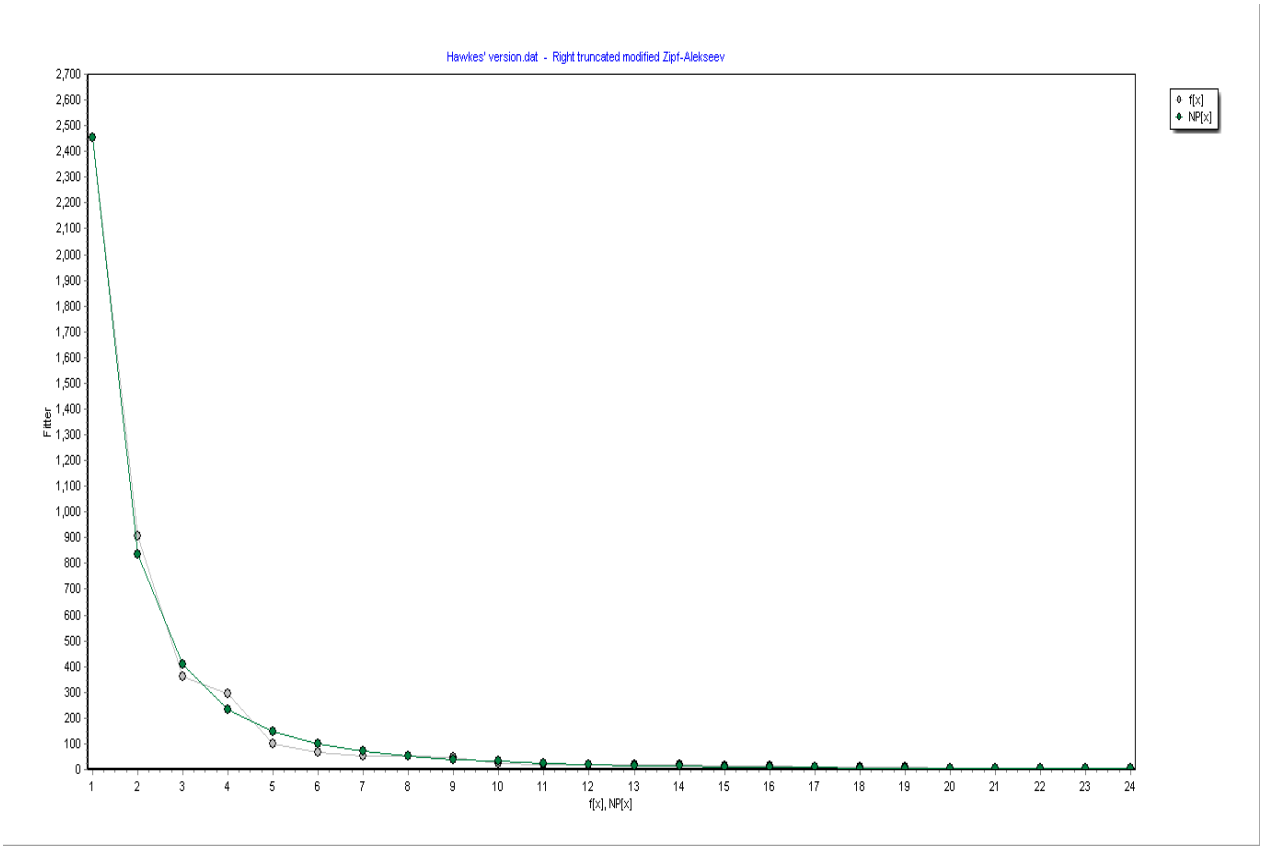

Figure 3. Fitting the modified right-truncated Zipf-Alekseev distribution to different expressions of *said* in Hawkes' version

Table 5
Fitting the modified right-truncated Zipf-Alekseev distribution to different expressions of *said* in Yang's version

X[i]	F[i]	NP[i]
1	802	802.00
2	317	320.70
3	188	188.94
4	159	114.90
5	74	72.85
6	33	47.98
7	20	32.66
8	11	22.86
9	10	16.39
10	8	11.99
11	8	8.94
12	7	6.77
13	6	5.20
14	6	4.04
15	5	3.18
16	5	2.53
17	3	2.03
18	2	1.64
19	2	1.34
20	2	1.10
a = 0.2078, b = 0.6122, n = 20, α = 0.4808, DF = 15, R^2 = 0.9962		

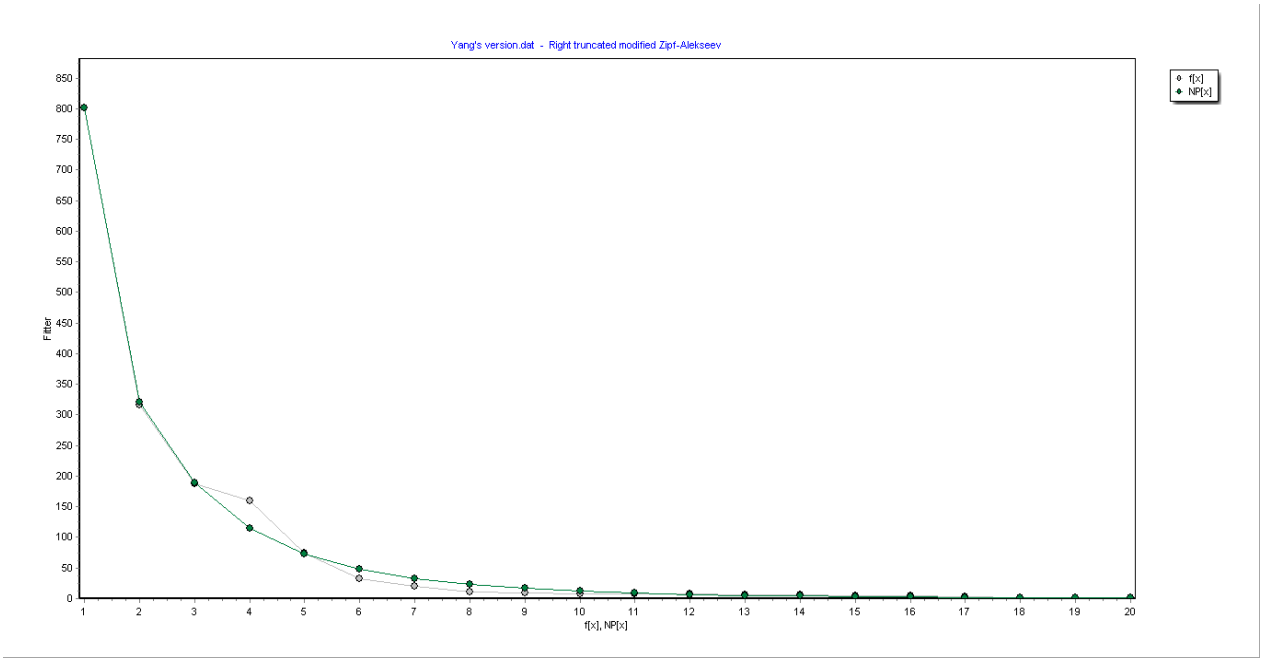

Figure 4. Fitting the modified right-truncated Zipf-Alekseev distribution to different expressions of *said* in Hawkes' version

Similar to the probability distribution of *dao*'s translations, the distribution of the original text of *said* can also be fitted well with the modified right-truncated Zipf-Alekseev distribution. That is to say, just like the Chinese word *dao*, the English word *said* is also polysemic and can convey the meanings of many words related to *dao* in Chinese.

3.2 Differences between the two versions

Though *dao*'s translation in the two versions follows the same distribution, there are still several differences between them.

Firstly, Hawkes tended to use *said* more often, while Yang preferred not to use any indicator. The most frequent translation in Hawkes' version is *said*, covering 49.35% of all extracted words. The second most frequent translation is without any indicator, covering 27.13%, and the third one *replied* covers 6.67%. In Yang's version, the most frequent translation is *without*, which covers 18.18%; the third one is *asked*, covering 5.94%.

Secondly, though both translators like to use certain words like *said, replied* and *asked*, Yang displays a wider range of diversification than Hawkes. Comparing the observed frequencies with the calculated frequencies according to the modified right-truncated Zipf-Alekseev distribution in these two versions, we can determine that the frequencies of 14 words (Rank 3 to rank 16) in Hawkes' version are below the calculated frequencies, while only nine words' frequencies (Rank 3 to Rank 11) are below the calculated frequencies in Yang's version. Moreover, if we consider those expressions with a standard frequency over 50 times as high frequency words, nine words (*said, without any indicator, replied, asked, exclaimed, cried, protested, retorted, continued*) belong to this category in Hawkes' version, covering about 10.11% of the total number. While there are 22 words (*without any indicator, said, asked, told, replied, exclaimed, answered, cried, remarked, urged, put in, retorted, agreed, protested, demanded, observed, objected, announced continued, called, explained, scolded*) in Yang's version meeting this standard. To find out whether Yang displays a larger variety of choices than Hawkes, we conduct a binomial test. The hypotheses are:

> H_0: High frequency words in Yang's version has the same proportion as that in Hawkes' version.
> H_1: High frequency words in Yang's version has a larger proportion than that in Hawkes' version.

The binomial test shows that high frequency words in Yang's version have a significantly larger proportion than that in Hawkes' version: $p < 0.001$.

Thirdly, most expressions are shared in the two translated versions, but each version also has some distinct words. In other words, some expressions can only be found in Hawkes' version, but not in Yang's version, whereas some expressions are peculiar to Yang's version, but not to Hawkes' version. They are listed in Table 6.

Table 6
Distinct verbs in Hawkes' and Yang's versions

verbs not in Hawkes	standard frequency in Yang	verbs not in Yang	standard frequency in Hawkes
assured	44	addressed	27
coaxed	5	barked	2
demurred	28	bragged	2
exhorted	3	babbled	2
fumed at	3	cried out	10
groaned	3	complimented	2
insisted	26	commanded	4
joked	5	conceded	2
joined in	3	echoed	6
prompted	5	expounded	2
reported	10	interceded	2
swore	36	mused out	6
sighed	33	moaned	2
summoned	10	pursued	2
scoffed	8	responsed	2
teased	13	reassured	2
wondered	18	rebuked	2
whined	3	restrained	6
		upbraided	6
		ventured	6

As we can see, nearly half of the words which are not in Hawkes' version, have a relatively high frequency (over 10 times) in Yang's version, but only two words, which are not in Yang's version, have a high frequency in Hawkes' version. From this point of view, we can infer that Yang tends to use uncommon words to describe the same meaning.

3.3 Reasons for differences

3.3.1 Differences of word structures between Chinese and English

Differences can be found in the verb system of Chinese and English. Chinese is typologically an isolating language, that is, a language with a lack of overt morphological processes, so verbs often concatenate to express meanings. "The meaning of the complete Chinese expression is usually composed from the meaning of the individual words" (Palmer & Wu, 1994, p. 63). We can see this phenomenon in the Chinese dictionary since many entries are compound words consisting of several distinct lexical items, for example, 干活 *(ganhuo, work)*, 打碎（*dasui, break*） and 笑说（*xiaoshuo, said with smile*）. Usually, there are two types of such Chinese verb compounds to make both the action and the details of the resulting state

explicit: one Verb-Verb (VV) compound (e.g. 赶跑 *ganpao, chase and run away*), and one Verb-Adjective (VA) compound (e.g. 吃饱 *chibao, eat and be full*).

English, however, has a significantly different system of verbs. Although English also makes explicit the result, it refers to the details of the resulting state through the use of a prepositional phrase (e.g. 笑道 *xiaodao, said with a smile* in the above example) or an adverb (e.g. 冷笑道 *lengxiaodao, said bitterly*) or only through one single word (e.g. 骂道 *madao, scold).*

This feature can be observed explicitly in the two translated versions. *Dao* and *said* are two major indicators of direct speech in Chinese and English; however, translations of *dao* are more varied than the source texts of *said*. See Table 7.

Table 7

dao's *p*ossible translations of and *said*'s corresponding verbs in the original text

verbs	**Hawkes**	**Yang**
dao	89	87
said	24	20

The single word *dao* does not convey any feelings of speakers, but from Appendix 1 and 2 we can observe many emotive verbs in the translation. For instance, *exclaimed* indicates the speaker's anxiousness, *retorted* shows the speaker's aggression, and *scolded* conveys the speaker's anger.

On the contrary, the source texts of *said* are mainly compound verbs, that is to say, two or more words work together to indicate the speech act and speakers' feeling. From Table 1, we can find that among the 24 Chinese (compound) verbs, 14 consists of *dao*, that is, say, *dao* is the nucleus-word.

Verb compounds containing *dao* can be classified into two groups: VV compound and VA compound. Their corresponding translations are also checked in the specific context. The result is listed in Table 8.

Table 8

verbs containing 道 translated into *said*

classification	verbs	Hawkes	Yang
VV compounds	说道	said	said
	便说道	said	said
	因说道	said	said
	因问道	said	said
	回道	said	said
	劝道	said	said
	答应道	said	said
	答道	said, said in answer to the question	said
VA compounds	忙道	said, said hurriedly, said in some surprise, etc.	said
	笑道	said with a smile, laughed and said,	said with a smile,

		said with a sneer, etc.	smiled (laughed) and said, etc.
	笑回道	said … laughing, etc.	said
	啐道	said angrily, etc.	
	冷笑道	said coldly, said bitterly	said with an ironic smile, said the apparition with a scornful laugh, said … tartly
	骂道	said bitterly, said scornfully, scold, etc.	

The result is consistent with the linguistic feature: Chinese action verbs can be either Verb-Verb (VV) compounds or Verb-Adjective (VA) compounds; their equivalents in English can be either one single word, or a verb together with a prepositional phrase or an adverb.

3.3.2 Translators' subjectivity

Translation is a rewriting process, so translators choose words according to their understanding and purpose.

David Hawkes is a famous British sinologist, who had studied in Oxford and Beijing University, so he is familiar with both English and Chinese. In translating *Hongloumeng*, Hawkes adopted a free and fluent translation strategy, trying to make it easier for English readers to understand. Guided by this principle, he chose words which are frequently used by ordinary people in their daily life.

Yang Xianyi, a famous Chinese translator, studied British literature at Oxford, and also displays a high proficiency in English. Yang is not a native speaker of English and did not comprehend the native English speakers' preference regarding word choice as well as Hawkes did, so he didn't show the tendency to use one or two words. However, Yang understood the Chinese culture better than Hawkes, and he followed a faithful and literal translation, so he placed more emphasis on the context when he selected the corresponding translation. In the previous study, we have discovered that the vocabulary richness of Yang's translation is higher than that of Hawkes' translation, which is reflected in *dao*'s translation.

3.3.3 Put verbs into specific context

Context is important in the study of meaning, because meaning is not abstract in the actual use of the language. It is especially true in literature for authors, who often wield the weapon of words, to achieve their desired intention and effect, so in order to understand the two translations, we need to put verbs into their context.

We had discovered that Yang did not use any corresponding verbs in translating *dao* at time, then how did Hawkes deal with them? We selected the following examples.

E.g. 1 (from Chapter 105)
The original text: 贾政**道**："究竟犯什么事？"
Hawkes' version: "But what are the charges?" **asked** Jia Zheng.

Yang's version: "What exactly are the charges against them?"

Yang did not indicate who asks this question here, but readers could infer it from the context. In previous text, Yang described *Xue Ke came running in* and had a conversation with Jia Zheng, so after Xue Ke finished his speech, Jia Zheng should give his response.

Hawkes, using *asked*, indicated that this was a question posed by Jia Zheng. Actually, during the whole dialogue, Hawkes gave indicators of direct speech explicitly: *Xue Ke came running into the courtyard and called out, Jia Zheng stepped out to greet him, said Xue Ke* and *asked Jia Zheng*.

E.g. 2 (from Chapter 120)

The original text: 王夫人听了，**道**："这个主意很是。不然，叫老爷冒冒失失的一办，我可不是又害了一个人了么！"

Hawkes' version: "That's a very good idea. You've thought it all out very well." **replied** Lady Wang. "If we do not take the initiative, Sir Zheng may go ahead himself and deal with her in a very tactless way, and then I will be responsible for yet another misfortune."

Yang's version: "That's an excellent idea. Otherwise, if I let the master dispose of her off-hand wouldn't that be the ruin of her?"

This is a reply from Lady Wang, but in the original text, the author still used *dao*. To avoid monotony, Hawkes translated it into *replied* to reveal Lady Wang's intention, and since the reply is relatively long, *replied Lady Wang* was put between two sentences to change the sentence structure, which could provide a smoother reading experience, as the second sentence is quite lengthy. Yang Xianyi, as always, did not use any indicator of direct speech.

There are still many sentences like the ones listed above. Looking deeper into those lines, we could find that Yang's preference could be attributed to the influence of Chinese. In Chinese literature, when a series of dialogues are carried out just between two speakers, authors tend to eliminate *said*-like words, while there is no such tendency in English literature.

Furthermore, since *said* had a standard frequency of 4935 in Hawkes' translation, which was much higher than 1818 in Yang's translation, the reasons for such a huge discrepancy need to be explored.

E. g. 3 (from Chapter 40)

The original text: 只见丰儿带了刘姥姥板儿进来，**道**："大奶奶倒忙的紧。"

Hawkes' version: You are very busy, Mrs. Zhu, **said** Grannie Liu.

Yang's version: "How busy you are, madam!" **remarked** Granny Liu.

In this dialogue, Hawkes translated the word into *said Grannie Liu*. Yang put it into a specific context and translated it as *remarked Granny Liu*. Though the two verbs have nearly the same function, *remark* convey more context and feelings than *said*, and if readers observe too many *said* in the text, they would be bored.

E.g. 4 (from Chapter 94)

The original text: 这里宝玉倒急了，**道**："都是这劳什子闹事！

Hawkes' version: All this only served to exasperate Bao-yu. "The amount of trouble that wretched thing has caused!" he **said**.

Yang's version: "All this trouble's due to that silly thing!" **burst out** Pao-yu.

The original text used *dao*, but translators often add their own understandings based on the context. Hawkes, again, chose *said* to lead the direct speech, but put *all this only served to exasperate Bao-yu* before the speech. Yang, instead, used *burst out* to imply Bao-yu's exasperation. In other words, readers could comprehend Bao-yu's feeling directly in Hawkes' version, while they need to infer his feelings in Yang's version.

We cannot deny that Hawkes also chose other verbs while Yang used *said* directly, but such occasions are much less than the former one. This can be explained through different translation tactics of the two translators: Hawkes wanted readers to understand the original text as easy as possible, so he described the speakers' feeling separately, and only used *said* to guide the direct speech; Yang, trying to convey Chinese culture to the West, combined the speakers' feeling with those verbs and let readers judge for themselves.

Some expressions do not appear in either of the two translations and they deserve our undivided attention. Here, we focus on those words, which are not in one version, but have a relatively high frequency in the other version. Four words (*demurred* and *insisted* from Yang's version; *addressed* and *cried out* from Hawkes) are chosen here for further analysis.

E.g. 5 demurred (from Chapter 119)

The original text: 平儿**道**："太太该叫他进来，他是姐儿的干妈，也得告诉他。"

Hawkes' version: "Perhaps you should ask her in, ma'am," **said** Patience. "After all she is Qiao-jie's godmother. We should tell her what is happening."

Yang's version: But Pinger **demurred**, "Better invite her in, madam. As Qiaojie's godmother she should be told about this."

To put this sentence back into the original text, we can infer that Pinger intended to change Lady Wang's (who has a higher position) opinion. Demur, "to disagree politely with another person's statement or suggestion" (Webster, p. 443), seems suitable in this situation, but *demur* is seldom used to guide direct speech.

E.g. 6 insisted (from Chapter 111)

The original text: 众上夜的人齐声说**道**："这不是贼，是强盗。"

Hawkes' version: "They were armed!" **cried** all the servants on night-duty.

Yang's version: The watchmen **insisted**, "They were brigands, not thieves."

"Cry" means "to shout or say something loudly" (Webster, p. 399); "insist" means "to say (something) in a way that is very forceful and does not allow disagreement" (Webster, p. 854) and it is usually followed by that-clause. This is a conversation between the watchmen and a constable. Obviously, the watchmen has no authority to demand a submissive agreement from them, so *insisted* is misused here by Yang Xianyi. Hawkes chose *cried* to indicate that the watchmen wanted to emphasize their statement.

E.g. 7 addressed (from Chapter 104)

The original text: 雨村便**道**："我是管理这里地方的，你们都是我的子民。

Hawkes' version: Jia Yu-cun **addressed** the offender directly: "This entire district, as you know, is in my charge, and every one of its residents falls under my jurisdiction."

Yang's version: "I am in charge of this district," Yucun **announced**. "All citizens here come under my jurisdiction."

"Address" means "to speak to (a person or group)" (Webster, p. 20); "announce" means "to say (something) in a loud and definite way" (Webster, p. 56). From the definition, we can conclude that *address* is a neutral word to guide direct speech, while *announce* displays a way of speech: loud and definite. In this sentence, both verbs are suitable.

E.g. 8 cried out (from Chapter 118)

The original text: 李纨、宝钗听了，诧异**道**："不好了！这人入了魔了。"

Hawkes' version: Li Wan and Bao-chai both **cried out** in alarm: "Lord save us! He's bewitched!"

Yang's version: Li Wan and Bao-chai **exclaimed**, "Oh dear! He's bewitched."

"Cried out" means "to speak in a loud voice" (Webster, p. 399); "exclaimed" means "to say (something) in an enthusiastic or forceful way" (Webster, p. 571). The speakers are in great surprise in this context, so these two verbs can express their feelings properly.

Though both translators use some verbs that the other translator do not, the word choice of Hawkes is obviously more proper than that of Yang.

4. Conclusion

Based on the above analyses, we came to the conclusions corresponding to the three research questions posed in the introduction section.

(1) The translated Chinese word *dao*, and the original text of *said* develop according to a diversification process, and is smooth enough to be modeled by a modified right-truncated Zipf-Alekseev distribution.

(2) Differences can be found between the two translated versions: Firstly, Hawkes tended to use *said* more often, while Yang preferred not to use any indicator. Secondly, though both translators prefer to use some words, Yang shows a wider range of word usage than Hawkes. Thirdly, most expressions are shared in the two translated versions, but they also have some distinct and unique words.

(3) The three causes of divergence play an important role in this diversification process: Firstly, Chinese and English are different in the nature of indicators of direct speech: Chinese often utilize concatenation to express meanings, that is, verb compounds are used to make both the action and the details of the resulting state explicit, while English refers to the details of the resulting state through the use of a prepositional phrase or only through one single word. Second, translators' subjectivity plays an important role in the divergence process, as they often choose words according to their understanding and purpose. Third, if we place those verbs that are not shared in the two translations into a specific context, we can discover

that Yang sometimes chooses some uncommon words in guiding direct speech, and Hawkes' choices seems more proper.

In this paper, we have tried to explore the probability distributions of *dao*'s translation and the original text of *said*, as well as the causes of divergence in this diversification process. The results are favorable. But there are still some limitations in this study: Firstly, due to the restrictions of time and space, we only chose *dao* and *said* as an example. For further research, we could investigate more verbs and find whether this result is suitable. Secondly, there are dozens of distinct verbs in Hawkes' and Yang' version, but we only chose two; if more words were taken into consideration, we could have certainly provided more interesting findings.

Acknowledgments

This work is partly supported by the National Social Science Foundation of China (Grant No. 11&ZD188).

References

Altmann, G. (1996). Diversification processes of the word. *Glottometrika,* 15, 234-240.

Altmann, G. (2005). Diversification processes. In R. Köhler, G. Altmann & R. G. Piotrowski (Eds), *Quantitative Linguistics: An International Handbook:* 646–658. Berlin: de Gruyter.

Chen, H. & Jiang, F. (2003). The translation of Hong Lou Meng into English: A descriptive study. *Chinese Translators Journal* 5, 46–52.

Crisafulli, E. (1999). The translator as textual critic and the potential of transparent discourse. *The Translator* 5(1), 83-107.

Fang, Y & Liu, H. (2015). Comparison of vocabulary richness in two translated *Hongloumeng. Glottometrics* 31, 54-75.

House, J. (2006). Text and context in translation. *Journal of Pragmatics* 38(3), 338-358.

Köhler, R. & Altmann, G. (2000). Probability distributions of syntactic units and properties. *Journal of Quantitative Linguistics* 7(3), 189-200.

Kulkarni, S.B., Deshmukh, P.D. & Kale, K.V. (2013). Syntactic and structural divergence in English-to-Marathi machine translation. In: *Computational and Business Intelligence (ISCBI), 2013 International Symposium on* (pp. 191-194). IEEE.

Lefevere A. (1992). *Translation, Rewriting, and the Manipulation of Literary Fame*. London and New York: Routledge.

Liu, H. (2009). Probability distribution of dependencies based on a Chinese Dependency Treebank. *Journal of Quantitative Linguistics* 16(3), 256-273.

Malinowski, B. (1935). *Coral gardens and their magic* (Vol. 2). London: G. Allen & Unwin.

Palmer, M. & Wu, Z. (1994). Verbs semantics and lexical selection. In: *Proceedings of the 32nd annual meeting on Association for Computational Linguistics* (pp. 133-138). Association for Computational Linguistics.

Palmer, M. & Wu, Z. (1995). Verb semantics for English-Chinese translation. *Machine translation* 10(1-2), 59-92.

Ren, L., Sun, H. & Yang, J. (2010). Chinese-English parallel corpus of *A Dream of Red Mansions*). Retrieved May, 20, 2015, from http://corpus.usx.edu.cn/.

Rothe, U. (1991). Diversification processes in grammar: An introduction. *Diversification Processes in Language: Grammar.* Hagen: Rottmann.

Saboor, A., & Khan, M. A. (2010). Lexical-semantic divergence in Urdu-to-English example based machine translation. In: *Emerging Technologies (ICET), 2010 6th International Conference on* (pp. 316-320). IEEE.

Strauss, U. & Altmann, G. (2006). *Diversification Laws in Quantitative Linguistics.* Retrieved June, 15, 2015, from http://lql.uni-trier.de/index.php/Diversification.

Venkatapathy, S. & Joshi, A.K. (2007). Discriminative word alignment by learning the alignment structure and syntactic divergence between a language pair. In: *Proceedings of the NAACL-HLT 2007/AMTA Workshop on Syntax and Structure in Statistical Translation* (pp. 49-56). Association for Computational Linguistics.

Voloshynovska, I.A. (2011). Characteristic features of rank-probability word distribution in scientific and belletristic literature. *Journal of Quantitative Linguistics* 18(3), 274-289.

Webster, M. (2010). *Merriam-Webster's Advanced Learner's English Dictionary.* Beijing: Encyclopedia of China Publishing House

Appendix 1

Frequencies of 道's translation in Hawkes' version

Rank	Translated words	Frequency in Hawkes	Standard Fre.*
1	said	2352	4935
2	Without any indicator	1293	2713
3	replied	318	667
4	asked	175	367
5	exclaimed	83	174
6	cried	47	99
7	protested	47	99
8	retorted	25	52
9	continued	24	50
10	told	23	48
11	spoke	22	46
12	explained	19	40
13	put in	19	40
14	called out	15	31
15	announced	14	29
16	addressed	13	27
17	called	13	27
18	interrupted	12	25
19	muttered	12	25

20	pleaded	11	23
21	answered	9	19
22	advised	9	19
23	inquired	8	17
24	went on	8	17
25	added	7	15
26	objected	7	15
27	whispered	7	15
28	agreed	6	13
29	remarked	6	13
30	suggested	6	13
31	cried out	5	10
32	grumbled	5	10
33	ordered	5	10
34	blurted out	4	8
35	begged	4	8
36	instructed	4	8
37	informed	4	8
38	snapped	4	8
39	scolded	4	8
40	burst out	3	6
41	declared	3	6
42	echoed	3	6
43	gasped	3	6
44	interposed	3	6
45	mused out	3	6
46	proceeded	3	6
47	rejoined	3	6
48	restrained	3	6
49	urged	3	6
50	upbraided	3	6
51	ventured	3	6
52	warned	3	6
53	chided	2	4
54	concluded	2	4
55	commanded	2	4
56	confessed	2	4
57	demanded	2	4
58	expostulated	2	4
59	implored	2	4
60	observed	2	4
61	persisted	2	4
62	reminded	2	4

63	sneered	2	4
64	yelled	2	4
65	barked	1	2
66	butted in	1	2
67	bragged	1	2
68	babbled	1	2
69	called back	1	2
70	complimented	1	2
71	countered	1	2
72	conceded	1	2
73	expounded	1	2
74	faltered	1	2
75	interceded	1	2
76	mumbled	1	2
77	moaned	1	2
78	panted	1	2
79	queried	1	2
80	quipped	1	2
81	prevaricated	1	2
82	pursued	1	2
83	remonstrated	1	2
84	recited	1	2
85	responsed	1	2
86	reassured	1	2
87	rebuked	1	2
88	stormed	1	2
89	volunteered	1	2
Total		4737	

*Standard Fre. = (frequency/token)*10000, all figures in the table are rounded to the nearest integer.

Appendix 2

Frequencies of 道's translation in Yang's version

Rank	Translated words	Frequency in Yang	Standard Fre.
1	Without any indicator	1596	4105
2	said	707	1818
3	asked	231	594
4	told	136	350
5	replied	107	275
6	exclaimed	96	247
7	answered	85	219
8	cried	64	165
9	remarked	61	157

10	urged	55	141
11	put in	47	121
12	retorted	44	113
13	agreed	41	105
14	protested	37	95
15	demanded	32	82
16	observed	31	80
17	objected	29	75
18	announced	26	67
19	continued	23	59
20	called	22	57
21	explained	22	57
22	scolded	20	51
23	warned	19	49
24	countered	18	46
25	interposed	18	46
26	assured	17	44
27	suggested	17	44
28	declared	16	41
29	ordered	15	39
30	rejoined	14	36
31	swore	14	36
32	sighed	13	33
33	added	11	28
34	begged	11	28
35	demurred	11	28
36	whispered	11	28
37	insisted	10	26
38	reminded	10	26
39	went on	10	26
40	pleaded	9	23
41	inquired	7	18
42	volunteered	7	18
43	wondered	7	18
44	blurted out	6	15
45	sneered	6	15
46	called out	5	13
47	expostulated	5	13
48	informed	5	13
49	snapped	5	13
50	teased	5	13
51	advised	4	10
52	gasped	4	10

53	reported	4	10
54	summoned	4	10
55	burst out	3	8
56	muttered	3	8
57	persisted	3	8
58	prevaricated	3	8
59	recited	3	8
60	scoffed	3	8
61	called back	2	5
62	coaxed	2	5
63	confessed	2	5
64	faltered	2	5
65	instructed	2	5
66	implored	2	5
67	joked	2	5
68	proceeded	2	5
69	prompted	2	5
70	panted	2	5
71	quipped	2	5
72	remonstrated	2	5
73	yelled	2	5
74	butted in	1	3
75	chided	1	3
76	concluded	1	3
77	exhorted	1	3
78	fumed at	1	3
79	grumbled	1	3
80	groaned	1	3
81	interrupted	1	3
82	joined in	1	3
83	mumbled	1	3
84	queried	1	3
85	spoke	1	3
86	stormed	1	3
87	whined	1	3
Total		3888	

*Standard Fre. = (frequency/token)*10000, all figures in the table are rounded to the nearest integer.

Glottometrics 32, 2015, 88-98

The relationship between word length and compounding activity in English

Christopher Michels[1]

Abstract. In the present article we test the hypothesis concerning the influence of word length on compounding activity and use extensive data contained in dictionaries.

1. Introduction

Whereas most linguistic models are based on variable descriptive rules and assumptions, as with all things, language as well as its behaviour is also governed by laws and law-like processes (Altmann and Schwibbe 1989, p. 2). These law-like processes "hold for all languages, they can be derived deductively, they can be embedded into a system of equivalent statements and the result can be tested statistically" (Altmann 1989, p. 101). With the help of an English dictionary, this study aims at testing one of Altmann's hypotheses about compounds, which is derived from Menzerath's law against the background of synergetic linguistics, namely the hypothesis that "the shorter a word, the more frequently it occurs in compounds" (1989, p. 104).

Following a brief description of the related grammatical concepts, the derivation of this hypothesis is outlined briefly. Then, the linguistic hypothesis is formalised and the involved linguistic properties *word length*, *compound*, and *compounding activity* are operationalised in order to test this formalised hypothesis empirically. Furthermore, the acquisition of test material from a dictionary is described as the essential prerequisite for testing. Finally, the estimated parameter values of the hypothetical equation and the test results for the adequacy of these estimates are presented.

2. Background

2.1. Word formation and compounds

According to the *Longman Grammar of Spoken and Written English (LGSWE)*, "complex word forms result from three main processes: inflection, derivation, and compounding" (Biber et al. 1999, p. 57). Whereas the former two processes involve the affixation of a base, the latter is defined as "independently existing bases combined to form new lexemes" (Biber et al. 1999, p. 58). English compounds occur in various types, their unity is evident "by their tendency to be pronounced with *unity stress* (i.e. stress on the first element) and written as one word or with a hyphen", and "they show limited possibilities of [the] substitution" of their

[1] Address correspondence to: christopher@cmich.eu

individual parts (Biber et al. 1999, p. 58). Furthermore, compounds are also described as a type of lexicalised "multi-word unit which tends to be written as a single word" (Biber et al. 1999, p. 59).

This description of compounding in English does not try to "form a sharp boundary between a 'pure compound' and a free combination of two words" (Altmann 1989, p. 100). Neither does the definition above restrict compounds to a maximum of two components, nor does it insist on a manifest distinctive feature of the bond between the components. Compounding is merely put in line with and distinguished from two other processes of word formation and the features of compounds are merely linked to various pieces of evidence which are mostly related to language in use, e.g. in terms of stress patterns or in terms of lexicalisation. Because of this close link to language in use resulting from the corpus-based approach of the *LGSWE* (Biber et al. 1999, p. 4), this grammatical description of compounds is suitable for the operationalisation of the linguistic properties involved in testing the hypothesis of this study.

2.2. Word length and compounds

"The basic motivation for forming compounds is" the "Bühlerian" need "to express oneself or a state of affairs", which is why compounding is basically "a specification, a narrowing of the (extension of the) meaning" (Altmann 1989, pp. 100, 103). Whereas the compounding process is highly arbitrary and chaotic from a semantic point of view, from a synergetic point of view, "*length*" is one of "Köhler's order parameters," namely "*polylexy*, *length*, *frequency* and *polytexty*", which are described as being responsible for law-like processes in compounding (Altmann 1989, p. 101).

3. Linguistic hypothesis

In the context of synergetic linguistics, the hypothesis that shorter words occur more frequently in compounds than longer words is the first concerning the order parameter *length* according to Altmann (1989, p. 104), which is considered a consequence of Menzerath's law (Altmann and Schwibbe 1989, p. 11). Essentially, it relies on two other existing hypotheses. Firstly, an "increase of polylexy leads to the shortening of words" (Altmann 1989, p. 105), and secondly, "the more meanings a word has, i.e. the greater its polylexy, the greater" is its compounding activity, i.e. "its chance of being used in a compound" (Altmann 1989, p. 103). Thus, the relationship between word length and compounding activity is indirect in nature and also involves polylexy as an important linguistic property.

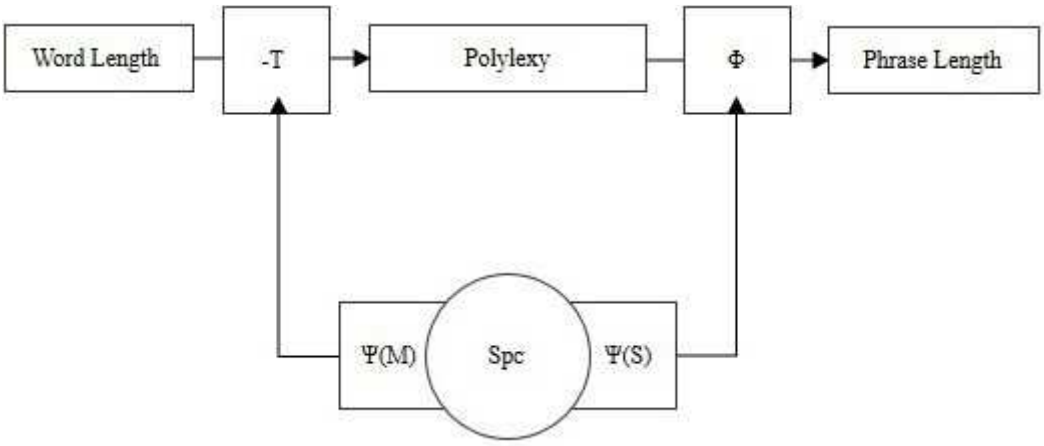

Figure 1. Menzerath's law on the phrasal level: A relationship between *Word Length*, *Polylexy*, and *Phrase Length* (adapted from: Köhler 1990, p. 12).

Concerning the first of the two underlying hypotheses, Figure 1 illustrates Menzerath's law on the phrasal level as a simplified relationship between *Word Length*, *Polylexy*, and *Phrase Length* (Köhler 1990, p. 12). *Polylexy* is inversely proportional to *Word Length*, whereas the coefficient of proportionality *T* is a function containing the need for specification *Spc* and the extent to which this need is met with the help of morphological ways of specifying meaning (Ψ(*M*)) (Köhler 1990, p. 4). Subsequently, *Phrase Length* is proportional to *Polylexy*, and the coefficient of proportionality Φ is a function of the need for specification *Spc* and the extent to which this need is fulfilled by syntactically reducing lexical ambiguity (Ψ(*S*)) (Köhler 1990, p. 10).

The illustration of this relationship is described as simplified because it does not include other factors which can be considered relevant to the property *Phrase Length*, for example, and it does not show the entire range of ways to reduce ambiguity or to encode meanings, namely lexically, morphologically, syntactically, or prosodically (Köhler 1990, pp. 4, 10). Similarly, the need for specification *Spc* is not the only need which can be considered relevant to these relationships (cf. Steiner 2002, pp. 220–221). The need to minimise the production effort also has an influence on the coefficients mentioned above, and these needs can be in competition with each other across the various levels of analysis existing beside the phrasal level. However, this illustration of the indirect, inversely proportional relationship between *Phrase Length* and *Word Length* clearly corresponds to the following specification of Menzerath's law from a linguistic point of view: "The bigger a linguistic construction [(i.e. a phrase)] is, the smaller are its components [(i.e. the words constituting that phrase)]." (Altmann and Schwibbe 1989, p. 5).

Figure 2 shows how the second underlying hypothesis combines with the first in order to illustrate the indirect relationship between *Word Length* and *Compounding Activity* which is essential to the central linguistic hypothesis of this study. The inversely proportional relationship between *Polylexy* and *Word Length* is linked with the proportional relationship between *Compounding Activity* and *Polylexy*, where the coefficient of proportionality *C* is a function of the need for specification *Spc* and the extent to which this need is met by morphologically reducing ambiguity (cf. Steiner 2002, p. 220).

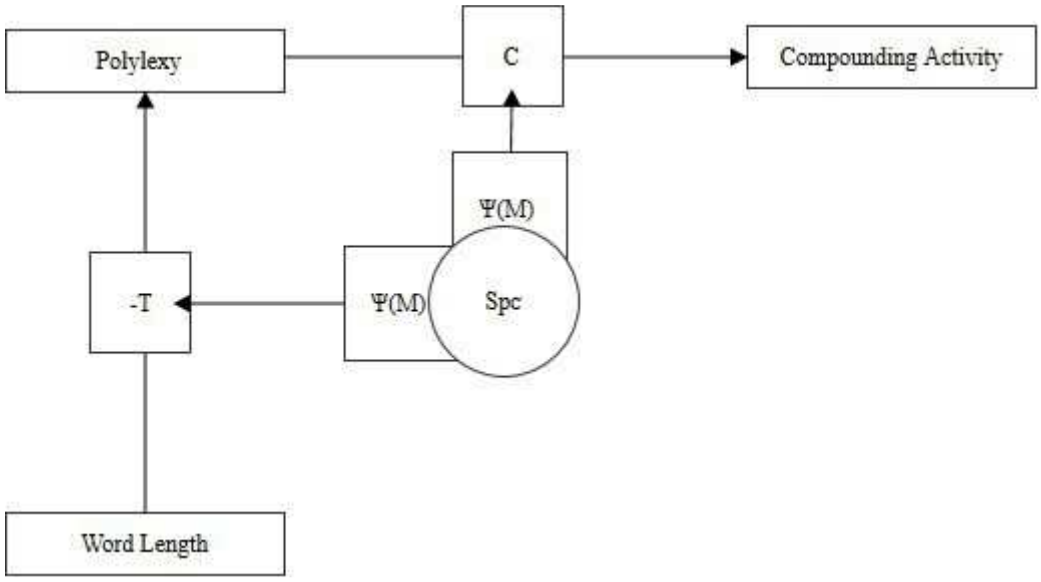

Figure 2. The indirect relationship between *Word Length* and *Compounding Activity* as the basis of Altmann's first hypothesis concerning *length* (cf. Steiner 2002, p. 220).

4. Statistical hypothesis

By analogy with the illustration of Menzerath's law on the phrasal level, the inversely proportional relationship between *Compounding Activity* and *Word Length* can be formalised with the general equation used by Altmann and Schwibbe (cf. 1989, p. 6), where *x* is the symbol of the independent variable *Word Length*, *y* is the symbol of the dependent variable *Compounding Activity*, and the parameters *a*, *b*, and *c* can be estimated with the method of least squares applied to the formula

$$y = ax^b e^{cx}$$

Leaning against the unified theory (cf. Wimmer, Altmann 2005), it can be conjectured that the relative rate of change of the dependent variable (*compounding activity*), *y'/y* is influenced by the language constant *k* and the relative rate of change of the independent variable (*length*), i.e.

$$\frac{dy}{y} = \left(k + \frac{b}{x} \right) dx$$

yielding the formula above (after reparameterisation).

The property *Word Length* is defined in terms of the number of syllables constituting a word. This can be considered a suitable definition for the purpose of this study because it employs a dictionary that includes stress patterns for its entries. In addition to the stress pattern symbols employed by the dictionary, white space and hyphens are also considered to be syllable separators. The number of substrings which are separated from each other by this set of characters determines the number of syllables, i.e. the word length

of a given lemma taken from the dictionary.

Following the description of compounding in the *LGSWE*, a lexeme is considered a compound if it consists of more than one base (Biber et al. 1999, p. 58) and if these bases are included in a set of elements which is compiled based on the entries in the dictionary (cf. Steiner 2002, p. 226). This set includes all lexemes of the dictionary which belong to entries with a lemma that do not contain any white space or hyphenation. For entries with complex lemmata, i.e. strings containing white space or hyphenation, the individual parts are added to the set of valid bases. Inflectional and derivational affixes have to be excluded from this set if the dictionary also contains entries for these elements. Inflectional genitive suffixes are ignored when valid bases are extracted from complex lemma strings.

Obviously, this cannot be considered a very precise definition of compounds because it also allows for "multi-word units" to be considered a compound. However, the *LGSWE* considers compounds "a type of multi-word unit which tends to be written as a single word", whereas it also mentions that the orthographic patterns linked with compounding are subject to variation to a large extent (Biber et al. 1999, pp. 58–59). Consequently, this way of operationalising compounds acknowledges the fuzziness of the boundary between compounds and multi-word units and also allows multi-word units to enter the set of compounds for a given lexeme.

According to Steiner, the compounding activity of a lexeme is defined as the number of compounds in which the lexeme occurs as a base (2002, p. 227).

5. Test material

The XHTML document *The Project Gutenberg E-Book of Webster's Unabridged Dictionary* (2009) serves as the basis for the acquisition of test data. The structure of the entries within this dictionary is illustrated in Listing 1 below. Using this structural information, 102,008 entries are extracted from approximately 286,000 relevant XHTML elements in the original file. This first step of data acquisition also filters the dictionary entries in order to exclude lemma strings with less than three characters in length. This restriction is related to the fact that the vast majority of these short lemmata were obsolete, rare, or did not contain any vowel, which is why no more elaborate filter was designed.

Listing 1: Example of the entry structure in *Webster's Unabridged Dictionary*

```
1  <?xml version='1.0' encoding='UTF-8'?>
2  <!DOCTYPE [...]>
3  <html xmlns="http://www.w3.org/1999/xhtml">
4    <head>[...]</head>
5    <body>[...]
6      <p id="id05920">AGREEABLE<br/> A*gree"a*ble, a. Etym:
          [...]<br/> </p>
7      <p id="id05921">1. Pleasing, either to the mind or senses;
          pleasant; grateful; [...]</p>
8      [...]
9      <h5 id="id29689">BOOK Book, n. Etym: [...]</h5>
10     <p id="id29690">1. A collection of sheets of paper[...]</p>
11     [...]
```

```
12     <p id="id114943">HEADMAN<br/> Head"man‘, n.; pl. Headmen.
          Etym: [...]<br/> </p>
13     <p id="id114944">Defn: A head or leading man, especially of a
          village community.</p>
14     [...]
15     <p id="id223311">SEA OWL<br/> Sea" owl‘. (Zoöl.)<br/> </p>
16     <p id="id223312">Defn: The lumpfish.</p>
17     [...]
18   </body>
19 </html>
```

The information extracted for the entries starting on the lines 6, 9, 12, and 15 in Listing 1 is shown in Table 1. Thus, each extracted entry comprises the lemma in capital letters as well as the same string in sentence case, enhanced with characters from the set containing the asterisk (*), the grave accent (`), and the double vertical quotation mark ("). These additional characters occurring in the sentence-cased strings indicate the stress pattern and also provide information about how a given lemma is divided into syllables. For a small set of 38 entries it was not possible to automatically extract the corresponding stress pattern, which is why these entries were excluded from the following steps. Additional elements such as the word class, plural forms, etymological information, or any meaning definitions are ignored entirely because they are not relevant to the following steps applied to the data.

Table 1
Extracted information for the examples in Listing 1

Lemma	Stress pattern	Ignored elements
AGREEABLE	A*gree"a*ble	a.
BOOK	Book	n.
HEADMAN	Head"man`	n.; pl. Headmen.
SEA OWL	Sea" owl`	(Zoöl.)

The 102,008 extracted entries constitute the set of candidates for compounds. The subset of 95,344 lemma strings not containing any white space or hyphen is considered the set of potential bases. Searching the entire candidate set for a given base string, a lemma string is accepted as a valid compound or multi-word unit only if it meets all of the following conditions. Firstly, the base string has to be a substring of the candidate string, requiring strict inequality between the compared strings. Secondly, the sum of the number of syllables in the current base string and the number of syllables in the substrings remaining after the base string is removed have to equal the number of syllables in the entire candidate string. As an illustration for the base string *OWL (Owl)*, *SEA OWL (Sea" owl`)* is a valid candidate string, the described equation holds, but *YOWLEY (Yow"ley)* is invalid. Lastly, these remaining substrings have to pass a validation process themselves.

This substring validation process requires the compilation of several sets of strings. One set is the result of splitting all extracted lemma strings with white space and hyphens as separators while ignoring any inflectional genitive suffixes. The members of this set are used as a positive list. A second positive list is necessary for substrings of up to four

characters in length because the *Webster's Unabridged Dictionary* includes several entries for words rarely occurring in Present Day English, such as *AGOG*, *FIR*, or *YAK*, causing errors in the search for candidate strings. Consequently, this complementary list only contains the members of the first positive list which fall into this range of string lengths and can be found on a list of the 10,000 most frequent words in English texts from *Project Gutenberg* (*Most Common Words in Project Gutenberg:* 2006). Furthermore, inflectional and derivational affixes have been removed from both of these lists in order to avoid the identification of derived forms as valid candidate strings. The list of these affixes is determined with the help of the dictionary, which identifies affix entries with leading or trailing hyphens. However, the dictionary does not consistently provide this information (cf. *ANTI*) in the case of derivational affixes, which is why the dictionary-based list of derivational affixes is supplemented with the most frequent derivational affixes for the four major lexical word classes (see Table 2) according to Biber et al. (1999, pp. 320–322, 400, 531, 540).

Table 2
Derivational affixes for the four major lexical word classes (Biber et al. 1999, p. 320 ff.)

Word class	Type	Derivational affixes
noun	prefix	anti- arch- auto- bi- bio- co- counter- dis- ex- fore- hyper- in- inter- kilo- mal- mega- mini- mis- mono- neo- non- out- poly- pseudo- re- semi- sub- super- sur- tele- tri- ultra- under- vice-
	suffix	-age -al -an -ian -ance -ence -ant -ent -cy -dom -ee -er -or -ery -ry -ese -ess -ette -fut -hood -ician -ie -y -ing -ism -ist -ite -ity -let -ment -ness -ship -tion -ure
verb	prefix	re- dis- over- un- mis- out- (most common) be- co- de- fore- inter- pre- sub- trans- under-
	suffix	-ize -ise -en -ate -ify -fy
adjective	suffix	-al -ent -ive -ous -ate -ful -less
adverb	suffix	-ly -ily -ally

The search for compounds or multi-word units for the 95,344 bases retrieved in the dictionary resulted in 5,920 bases with at least one validated candidate string. From this population, twenty samples are taken randomly for each of the occurring word lengths with a sufficient number of units in the population. Consequently, only the sets of bases with compounds for the word length values 1, 2, 3, 4, and 5 contributed to a total population size of 5,916 bases for which compounds or multi-word units exist in the dictionary. Since the validation of potential compound strings described above still fails in some cases, the random samples are checked manually according to the definition of compounds in section 4.2.2 above and replaced with new random samples if the set of identified valid compounds was faulty. By choosing a fixed sample size for every relevant word length value, homoscedasticity is provided for the methods of testing pursued in the following section (cf. Steiner 2002, p. 230; cf. Grotjahn 1992, pp. 126, 150). In addition to the distribution of positive search results across all word lengths, Table 3 also lists the number of base strings without any

validated compound candidates.

Table 3
Overview of the search results providing the basis for sample selection

Word Length	Bases with compounds	Bases without compounds	Total
1	2,629	4,183	6,812
2	2,319	25,425	27,744
3	701	27,503	28,204
4	211	19,916	20,127
5	56	9,364	9,420
6	4	2,580	2,584
7	0	417	417
8	0	34	34
9	0	2	2
Total	5,920	89,424	95,344

The base strings represented by the third column of Table 3 were ignored because of two problems. Firstly, this set of base strings is filtered only with the help of the criterion described above, namely the absence of any white space or hyphen in the lemma string of an extracted dictionary entry. One of the examples from Listing 1, *HEADMAN (Head"man`)*, obviously meets this condition but is identified as a compound of *HEAD* and *MAN* by the validation process outlined in the previous section. The indication of the "unity of compounds [...] is [subject to] a great deal of variation, however, both in phonological and orthographic patterns" (Biber et al. 1999, p. 58). This is only one of the reasons why the validation of compound candidates includes more complex methods than merely scanning for white space or hyphenation. By analogy, a validation process for the set of base strings would require a more elaborate solution. Secondly, considering both the base strings with positive results and those with negative results would have made the latter group the majority for each of the occurring word length values. Thus, the average number of compounds for a base of any given length would have been distorted towards zero. Inevitably, ignoring the latter group means a distortion in the opposite direction. However, exclusively considering the base strings with positive search results can be considered a simpler yet more appropriate alternative: It avoids the design of a more complex validation method for base strings and the poorly validated list of extracted base strings cannot influence the test results as an additional source of error.

Table 4
Examples of the randomly selected bases and their compound lists

Word Length	Base (stress pattern)	Compound list
1	Owl	barred owl, hornowl, jar-owl, owl-eyed, owllight, scops owl, sea owl
2	Na"tured	fair-natured, good-natured, ill-natured, well-natured
3	Wor"thi*ness	praiseworthiness, seaworthiness, thankworthiness
4	Im*pe"ri*al	crown-imperial
5	Dy‘na*mom"e*ter	split dynamometer, transmission dynamometer (accepted multi-word units)

6. Testing

In order to test whether the data extracted from *The Project Gutenberg E-Book of Webster's Unabridged Dictionary* comply with the hypothesis or not, the equation 4.1.1 above is linearised and the values for the parameters a, b, and c are estimated with the method of least squares. Furthermore, the adequacy of these estimations is examined with the help of two tests. The F-test checks whether the deviations of the theoretical values from the values originating from the data are sufficiently small (Köhler 1986, p. 99; Steiner 2002, p. 231). In addition, the coefficient of determination (R^2) is calculated in order to determine in how far the variance of the theoretical data explains the total variance based on the predicted variable (cf. Steiner 2002, p. 231). The following test results were obtained with the help of *NLREG*, Version 6.3.

Table 5
Parameter estimates based on mean values for each word length value

Parameter	Estimate
a	1773.3069
b	6.4086
c	-4.5439
R^2	0.9910

Table 6
Observed and computed values of the dependent variable *Compounding Activity* and F-test results

Word Length	*Compounding Activity*	
	Observed mean value	Computed mean value
1	18.85	18.8542
2	17.05	17.0296
3	2.25	2.4340
4	1.20	0.1635
5	1.35	0.0073
F = 109.96		Prob(F) = 0.00901

The results presented in Tables 5 and 6 confirm that the hypothetical equation complies with the sample data. The probability to obtain the F-value of 109.96 randomly is less than P = 0.01 and thus the F-test supports the adequacy of the estimated parameter values. Furthermore, the second test method also supports the hypothesis: According to the result for the coefficient of determination, the variance of the theoretical data explains 99.10% of the total variance. However, the result for the parameter *a* shows a high standard error. Finally, Figure 3 below shows the plot of the linear function based on the hypothesis and the estimated parameter values. The deviation of the coordinates resulting from the mean values of the data for the word length values 4 and 5 seems to be the most obvious. This might be due to the exclusion of bases without any compounds from the population discussed above, resulting in a distortion of the data.

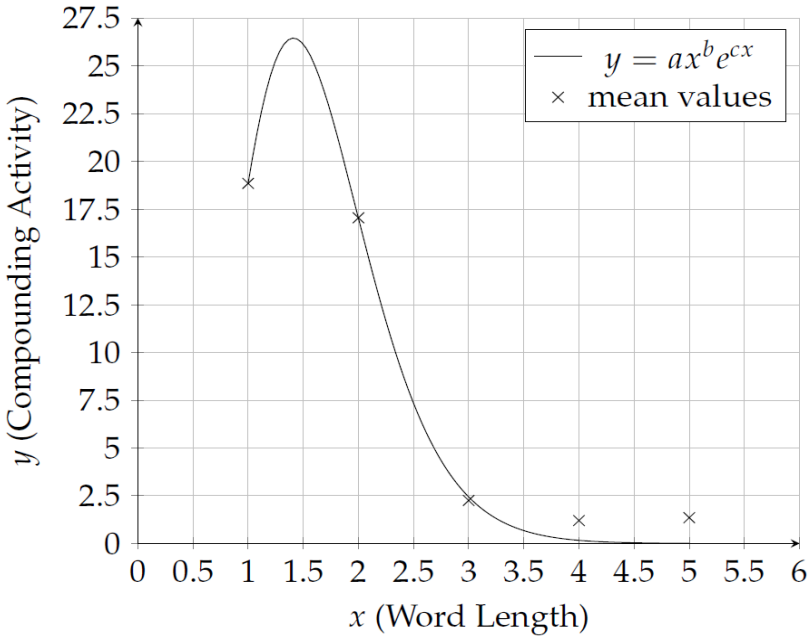

Figure 3. Plot of the linear function $y = ax^b e^{cx}$ with the estimated parameter values and the coordinates for the mean values from the sample

7. Conclusion

Despite several problems with the acquisition of data from *The Project Gutenberg E-Book of Webster's Unabridged Dictionary*, such as inconsistent information provided for stress patterns or derivational affixes, this study successfully confirmed the first hypothesis concerning the order parameter *length* according to Altmann (1989, p. 104) for English with the help of a dictionary as the main test material. Aside from the quality of the test material, this study also revealed other aspects that might benefit from further improvement or further investigation. The high standard error for the parameter *a* (cf. Table 5) might indicate that the significance of this parameter to the hypothetical equation needs to investigated, for example. In addition, according to Altmann and Schwibbe (1989, p. 7), the fact that the use of the general equation resulted in monotonically increasing parts occurring in Figure 3 above does not necessarily hint at unknown sources of error or random features, but it might just be a motivation to look for additional factors which might not have been included although they also influence properties which are relevant to the hypothesis of this study.

References

Altmann, Gabriel (1989). Hypotheses about Compounds. In: *Glottometrika 10: 100-107*. Ed. by Rolf Hammerl. Bochum: Brockmeyer.

Altmann, Gabriel; Schwibbe, Michael H. (1989). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*. Hildesheim: Olms.

Biber, Douglas et al. (1999). *Longman Grammar of Spoken and Written English*. Harlow, Essex: Longman.

Grotjahn, Rüdiger (1992). Evaluating the Adequacy of Regression Models: Some Potential Pitfalls. In: *Glottometrika 13: 121-172*. Ed. by Burghard Rieger. Bochum: Brockmeyer,

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

Köhler, Reinhard (1990). Linguistische Analyseebenen, Hierarchisierung und Erklärung im Modell der sprachlichen Selbstregulation. In: *Glottometrika 11: 1-18*. Ed. by Luděk Hřebíček. Bochum: Brockmeyer.

Most Common Words in Project Gutenberg: (2006). Project Gutenberg. URL: http://en.wiktionary.org/wiki/Wiktionary:Frequency_lists/PG/2006/04/1-10000 (visited on 04/03/2015).

Steiner, Petra (2002). *Polylexie und Kompositionsaktivität in Text und Lexik*. Ed. by Reinhard Köhler. Korpuslinguistische Untersuchungen zur quantitativen und systemtheoretischen Linguistik. URL: http://ubt.opus.hbz-nrw.de/volltexte/2004/279/pdf/07_steiner.pdf (visited on 04/11/2015).

The Project Gutenberg E-Book of Webster's Unabridged Dictionary (2009). Project Gutenberg. URL: http://www.gutenberg.org/cache/epub/29765/pg29765.html (visited on 04/17/2015).

Wimmer, Gejza; Altmann, Gabriel (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.

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.
19. K.-H. Best (Hrsg.), *Studies zur Geschichte der Quantitativen Linguistik. Band 1.* 2015, III + 159 pp.
20. P. Zörnig et al., *Descriptiveness, Activity and Nominality in Formalized Text Sequences.* IV+120 pp.