Quantitative analysis of translation texts

A Kiv*, L Bodnar, E Sedov, O Britavska, N Yaremchuk, M Yakovleva

South Ukrainian National Pedagogical University named after K D Ushinsky

Ukraine, Odessa, Staroportofrankovskaya 26, 65020

Received 1 October 2014, www. cmnt. lv

Abstract

A new, stylistic-mathematical approach for analysis of literary works, particularly for analysis of translation works, is developed. The important requirement to the translation is its compliance with the structure of language in which the translation is done. We showed that this analysis can be carried out using Zipf's laws and information characteristics of literary work.

Keywords: analysis of translations, Zipf's laws, information methods

I checked the harmony with algebra…
Mozart and Saliery
S. Pushkin

1 Introduction: Zipf's distribution

In Ref [1] Zipf argued that in the development of a language, a certain vocabulary balance should eventually be reached as a result of action of two opposing forces: the force of unification and the force of diversification. The first force tends to reduce the vocabulary and corresponds to a principle of the least effort seen from the point of view of the speaker. The second force has the opposite effect and is connected with the people desire to understand the meaning of speech. Though Zipf does not transform his ideas into a mathematical model, we consider his basic consideration as a two-person game, however without a precise definition of the cost-functions involved.

Zipf discovered empirically for word tokens in an English corpus that if f is the frequency of a word in the corpus and r is the rank, then

$$f = k / r \ , \tag{1}$$

where *k* is Zipf's constant for the corpus. When log f is drawn against log r in a graph (which is called a Zipf's curve), a straight line is obtained with a slope of –1. Zipf discovered his law by analysing manually the frequencies of words in the novel "Ulysses" by James Joyce. It contains a vocabulary of 29,899 different word types associated with 260,430 word tokens.

Following Zipf's discovery many experiments aided by the appearance of computers confirmed that the law is correct. The slope of the curve was found to vary slightly from –1 for some cases. Also the frequencies for the highest ranked words sometimes deviated from the straight line, which suggested several modifications of the law, and in particular one derived theoretically by Mandelbrot [2] with the form:

$$f = k / (k + \alpha)^{\beta} \ , \tag{2}$$

where α and β are constants for the corpus being analysed. However, generally the constants α and β were found to be only small varying deviations from the original law by Zipf.

In his book "The Psycho-Biology of Language" publicshed in 1935, Zipf called attention for the first time to the phenomenon that has come to bear his name. This book contains Zipf's first diagram of the log (frequency)-vs. -log (rank) relationship, a Zipf's curve for his count of words in the Latin writings [3].

The laws of Zipf that give correlations "Rank – Frequency" and "Quantity – Frequency" allowed solving actual problems of linguistics and literary theory. The important conclusion from these laws is an existence of universal quantitative characteristics of all human texts independently on the language groups. Over the last time the mathematical linguistics has intensively developed [4, 5]. Hyperbolic word frequency distributions were analysed in application to different texts and finally were admitted as adequate mathematical expressions reflecting regularities of language structure.

2 Information characteristics of text

In addition to Zipf's constant, we introduce another quantitative parameter that characterizes any text. In all languages where the frequencies of words are described by a hyperbolic distribution small children use a few words they know with relative frequencies very different from the probabilities given for these words in the native language. They form only simple sentences, and at this stage the

* ***Corresponding author*** e-mail: kiv@bgu.ac.il

number of bits per word is small. The parents talk to their children at a lower bit rate than they normally use, but with a higher bit rate than their children. Thereby new words and grammatical structures can be presented to the child. When the child knows the basic words and structures of the language its bit rate increases. From this example one can see that a bit rate is an important characteristic for any text and its manipulation leads to significant changes in the expression of human thoughts and the transfer of information. In this connection an account of the information quantity and a bit rate in the text under study is undoubtedly useful. This statement can be illustrated also by comparison of poetical and prosaic works.

It's possible to define the bit rate in the text by different ways: for example, it may be a quantity of bits per chapter. In the literature works one can see large differences in the bits per word for poetry and prose. These differences can be seen by comparison of poetical and prosaic works of the same author.

We performed such comparison for Alexander Pushkin's works **[**6**]**. For Zipf's constant we obtained the value ~ 0. 06 – 0. 07 in the case of prosaic works and ~ 0. 04 – 0. 05 in the case of poetical works. It means that the number of words in the middle stripe of ranks that contains the most important words for understanding the content and the sense of the text is smaller for poetical works. Zipf showed that the approximation of his law is much better for the middle ranks than for the very lowest and the very highest ranks. Thus we are able to estimate the number of bits in the middle stripes of ranks with good accuracy. The quantity of information (*I*) can be estimated using Hartley formula **[**7**]:**

$$I = \log_2 N \,, \tag{3}$$

where N is a total number of words in the text. The obtained results show the necessity to distinguish between quantity and quality of information in the analysis of literary works. In the poetry work the same thought is expressed by fewer words, and every word carries more semantic meaning. It is a problem for another study.

Below we apply quantitative approaches to evaluate the quality of translations.

3 Evaluation of transition quality

The translating is essentially a skill and needs for study the use of a series of disciplines, such as linguistics, cultural anthropology, philology, psychology, and theories of communication. In contrast with the other sciences, translation is an activity that all bilingual people can engage in without special studies of technical procedures. As efficient bilinguals they quickly sense the degrees of equivalence in comparable texts.

Translators try first of all to transmit the style and genre of the original text. In each case a translator, proceeding from the original text, tries to find the subtle details of genre to reflect them in the translation version. For example, the basic features of speech genres, as a rule, coincide in English and Russian languages. Therefore for these texts it is reasonable to provide in the translated text a simple adequate transmission. But in most cases, the translation requires a versatile (semantic, grammatical and quantitative) analysis.

We introduce new stylistic-mathematical approach (SMA) for analysis of literary works, particularly for analysis of translation works, and proceed from the idea that the important requirement to translation is its compliance with the language in which the translation was done from the *point of view of Zipf's laws and information characteristics.* According to SMA any translation should satisfy at least to the following requirements:

- ✓ The sense of the translation version must exactly reflect the ideas of the original text.
- ✓ It is necessary to find the equivalent constructions in the translation version for idioms and other specific expressions in the original text.
- ✓ The translator should provide the appropriate difference between Zipf's constants and the appropriate compliance of the information characteristics for the original text and for the translation.

3. 1 THE QUANTITY OF INFORMATION IN THE TEXT

Analyzing some literary work on the base of Zipf's Laws, we determine the borders for the middle stripe of ranks and thus the number N of the most important words for this text. Thus we can estimate the quantity of information by formula (3). Then we use the same borders for the main stripe of ranks when evaluate the quantity of information for the individual parts of the whole text (chapters, sections etc.).

A computer program based on using Zipf's Laws was worked out for analysis of English and Russian literary texts. This program uses the algorithms of texts data processing from the Microsoft Company's products, such as Microsoft C#, Microsoft SQL 2008. The Microsoft SQL 2008 was chosen because now it is one of the most powerful full-text modules, *realized on more than 10 languages*, and at the same time is the most accessible for our study. The algorithm realized in the developed program allows processing any texts in order to present them as tables of database with necessary parameters. As a result of uniting capabilities of these products, we've obtained the client-server structure of the program, where the program is a client and Microsoft SQL 2008 is a server. The user enables to specify a set of search criteria. The program gets the answers and outputs from the server in the acceptable to user form.

3. 2 ANALYSIS OF TRANSLATIONS BASED ON ZIPF'S LAWS

This study was aimed to compare different translations of the famous play of William Shakespeare "Hamlet. Prince of Denmark" **[**8**]**. Various characteristics of this work given by critics were accounted. They analyzed these translations from the point of view of the exact reflection of Shakespeare's ideas, preservations of original thoughts, and the quality of the translation language. Then we estimated Zipf's constants for the original text (Edition of 1828) and translations taken from **[**8**]**. In Table 1 one can see the obtained results.

In Table 1 Zipf's constants are varied from 0. 0954 (that is close to English language) to 0. 0684 (that is close to Russian language). On the basis of these results and the stylistic analysis performed by other researchers we come to conclusion that the translation of Pasternak satisfied the conditions of the high level translation described above. His translation is the most closely to the native Russian language. At the same time in this translation Pasternak reproduced the music and the spirit of Shakespeare's masterpiece **[**8**]**. The opposite translation approach we see in the Radlova's work. She tried do not omit any word in the original text. As a result she did not reproduce in Russian version the sense of Shakespeare's work and at the same time her text is closer to the structure of English language.

TABLE 1 Evaluation of Zipf's constants

Author and translators	Year	Zipf's constant	Comments
Shakespeare	1603	0,1191	Original
Pasternak	1940	0,0684	Translation
Romanov	1899	0,0822	Translation
Averkiev	1895	0,0827	Translation
Kroneberg	1925	0,0837	Translation
Lozinski	1933	0,0877	Translation
Radlova	1937	0,0954	Translation

3. 3 ANALYSIS OF TRANSLATIONS BASED ON INFORMATION CHARACTERISTICS OF TEXT

We have added the mathematical analysis of text based on Zipf's laws by entering the information characteristics of literary work. It was introduced the characteristic for any text that is defined as a *number of bits per word* (BPW). The number of bits can be calculated using (3) if to find the full number of words in main stripe of ranks of Zipf's curve. For this we integrate the formula (1):

$$N = \int_{r1}^{r2} f(r)\,dr \quad , \tag{4}$$

where $r1$ and $r2$ are borders for the main stripe of ranks of Zipf's curve. The result of integration (4) we substituted into the formula (3). Thus the expression for BPW is:

$$\text{BPW} = I / N. \tag{5}$$

Parameter BPW was calculated for each Scene of three Acts of "Hamlet – Prince of Denmark". In Figure 1 we show how BPW changes from one Scene to another in the first Act of William Shakespeare's play. We see that BPW changes through the text. Moreover these changes are quasi-periodical. One can assume that this peculiarity of the information transmission in the literary work activates the reader's perception. Of course, the author does not do it purposely. It is a component of the creative process.

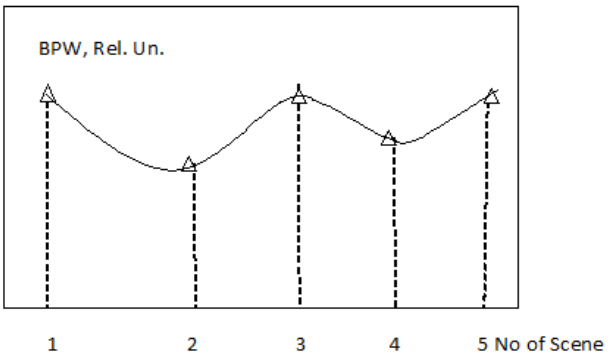


FIGURE 1 Illustration of BPW changes in the first act of the William Shakespeare play "Hamlet. Prince of Denmark"

4 Conclusions

Our study shows that the mathematical analysis of texts based on the use of Zipf's laws and the calculation of information characteristics leads to new possibilities to evaluate the quality of literary translations. It is possible to determine more precisely the conformity of the translation to the structure of language in which the translation is done. A new opportunity to explore the features of the creative process is opened.

References

[1] Zipf G K 1949 Human Behaviour and the Principle of Least Effort: Addison-Wesley, Cambridge
[2] Mandelbrot B 2012 *The Fractals: Memoir of a Scientific Maverick* Pantheon Books, NY
[3] Zipf George K 1935 *The Psycho-Biology of Language* Boston: Houghton Mifflin Co
[4] Lem Stanislaw 1999 The Extraordinary Hotel, or the Thousand and First Journey of Ion the Quiet, In Imaginary Numbers *An Anthology of Marvelous Mathematical Stories, Diversions, Poems, and Musings*, Ed William Frucht NY: Wiley 185-190
[5] Baayen H 2001 *Word Frequency Distributions* N-Y: Kluwer Academic Publishers
[6] Kiv A, Goncharenko D, Sedov Y, Bodnar L, Yaremchuk N 2008 Mathematical study of evolution of Russian Language, *Computer Modelling&New Technologies* 12(1), 55-58
[7] Hartley R V L 1928 *Bell System Technical Journal* 130-138
[8] William Shakespeare Hamlet, Prince of Denmark *Selected translations* 1985 Moscow: Raduga Publishers

Authors

Arnold Kiv

Current position, grades: Professor-researcher of Department of materials engineering, Ben-Gurion University of the Negev, Head of Department of physical and mathematical modelling, South-Ukrainian national pedagogical university after K. D. Ushinskij.
Scientific interest: Nanomaterials and nanoelectronics;
Experience: Expert in the theory of real structure of solids and physical processes in electronic devices. In last decade he obtained significant results in the field of nanomaterials, track nanostructures and track electronics.

Lilia Bodnar

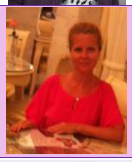
Current position, grades: PhD, Associated Professor, Department of physical and mathematical modelling, South-Ukrainian national pedagogical university after K. D. Ushinskij. She obtained Master Degree from this university.
Scientific interest: Application of information technologies in humanities.
Publications: About 30 publications.
Experience: Expert in computer modelling of processes and phenomena in sociology, philology, musical art and education.

Eugen Sedov

Current position, grades: PhD, Vice-rector on information technologies of South-Ukrainian national pedagogical university after K. D. Ushinskij, Associated Professor, Department of physical and mathematical modelling.
Scientific interest: Application of information-communication technologies (ICT) in education, cloud technologies in education, new applications of the Internet of things.
Publications: About 100 publications.
Experience: Expert in the field of ICT, in the practical application of new products of Microsoft and Intel.

Ellen Britavska

Current position, grades: PhD, Associated Professor, Department of physical and mathematical modelling, South-Ukrainian national pedagogical university after K. D. Ushinskij. She obtained Master Degree from this university.
Scientific interest: Penetration of fast ions through the crystals, ion-induced creation of the surface relief for electronic devices, methodology of physics teaching, computer modelling in humanities.
Publications: About 45 publications.
Experience: Expert in computer modelling of the ion-induced atomic processes in the surface layers of crystals and computer modelling in literary sciences.



Natalia Yaremchuk

Current position, grades: PhD, Senior Lecturer, Department of Ukrainian and Foreign literature, South-Ukrainian national pedagogical university after K. D. Ushinskij. The main results are obtained in the field of comparative literary criticism.
Scientific interest: Comparative analysis of literary works of various nations. Application of information technologies in philology,
Publications: About 25 publications.
Experience: Expert in the analysis of the translations quality.

Marina Yakovleva

Current position, grades: PhD, Senior Lecturer, Department of differential psychology, South-Ukrainian national pedagogical university after K. D. Ushinskij, Head of Odessa Science Centre on Euro-Atlantic Cooperation. The main results are obtained in the field of the motivation of foreign languages learning.
Scientific interest: Psychological aspects of foreign languages learning. Application of information technologies in linguistics.
Publications: About 25 publications.
Experience: Expert in the information and psychological analysis of the translations quality.