AVERAGE WORD LENGTH AND TEXT REDUNDANCY VARIABILITY: FRENCH TEXTS CASE STUDY

Malvina Marinashvili
PhD, Associate Professor, Odessa I. I. Mechnikov National University, Ukraine
e-mail: malvimari@gmail.com, orcid.org/0000-0002-3041-7064

Summary

The redundancy and average word length correlation in French texts have been researched. This correlation has been evaluated on the basis of analysis of entropy, redundancy and average word length for literary, scientific, and publicistic texts. It has been revealed that the variability of text redundancy correlates well with the variability of average word length, if calculating the average word length of an individual text we exclude the length of words belonging to the exponential tail of entropy curve. In this regard it is proposed to distinguish between two average word lengths of text: the average length of a word belonging to the exponentially decaying tail of entropy and the average length of a word not belonging to the exponential tail of entropy.

Keywords: text entropy, text redundancy, word length, information capacity, quantitative linguistics

DOI: https://doi.org/10.23856/3849

1. Introduction

Natural language is a complex system with a hierarchical structure, number of set rules and internal connections. To solve the present day problems of linguistics, in particular, quantitative ones, it appears to be important the study of regularities reflecting the inner properties or structure of a natural language.

In great number of linguistic researches *(Zipf, 1949; Miller et al., 1958; Mikros et al., 2005; Köhler, 2005; Strauss et al., 2007; Popescu et al., 2013 and others)* the word frequency, its length (or their correlation: Zipf's law) have been investigated and some regularities of words distribution in the texts of different functional styles, genres and various language case study revealed. In these researches it has been proved that text symbols distribution in terms of frequency is stated to be a stable characteristic neither that of the author nor the subject area of a text, but of a language.

Redundancy and word length (or the correlation of frequency and word length) as separate objects of research have been under analysis in many scientific works *(Shannon, 1948; Zipf, 1949; Miller, 1958; Newmann, 1960; Arapov, 1988; Grzybek et al., 2005; Guerrero, 2005; Köhler, 2005; Strauss et al., 2007; Grudeva, 2010; Kalimeri et al., 2012; Kalimeri et al., 2015; Alontseva, Ermoshin, 2019)*. However, the relation between these characteristics hasn't been studied enough yet.

The present paper intends to investigate the average word length and text redundancy correlation regularities based on French case study.

We suggest the average word length of a text consists of two lengths and in assessing entropy and text redundancy it is important to take account of not one average word length, but two. To test this hypothesis, we have studied the variability of average word length and redundancy versus maximum entropy of text, which in information theory is understood as information capacity of a message.

It should be noted that the researchers M. Kalimeri et al. *(Kalimeri et al., 2012)* comparing texts of different genres and in different languages, also differentiate words related and not related to the exponential tail of n-grams (words) relative frequency, taking into account their size (number of letters in a word). We will refer to this work in details while discussing the results of our research.

To study the regularities in frequency and word length distribution, many researchers refer to methods of information theory, which primarily was created to solve diverse practical tasks, in particular to calculate the system effectiveness for rendering information and increase the amount of information. Herewith, researchers made their attempts to apply math theory of information to literary, scientific and publicistic texts.

In present work informational entropy, redundancy and word length regarded as basic notions in information theory are also considered to be text characteristics.

Using linguistic redundancy C. Shannon measured the volume of information contained in different messages. Redundancy has many interpretations and in information theory is considered as the excessive information (in other words repeated or unnecessary information), defined as percentage content of excessive information in the texts of a given language. Shannon defined redundancy as the difference between the entropy of the messages actually transmitted and the maximum entropy that the channel could transmit. The simplest cause of this difference is probability distribution of message elements (e.g. letters, words, etc.).

Redundancy means that information may be discarded from the text without the harm to its meaning and easily restored as it is determined by the structure of the language itself. In connection with this fact redundancy can't be considered as the phenomenon of language imperfection or incompleteness of a message structure. Any text can have redundancy in any natural language and depending upon the type of a message the degree of redundancy can vary as well. Redundancy is in existence on all levels of a language *(Dubois et al., 1970; Martinet, 1991; Gillette, Wit, 1999; Grudeva, 2008)*, beginning with letters, and words up to a text and can be used as a measure of knowledge of a language and its culture by a person *(Raatz, Klein-Braley, 1981)*. Namely language redundancy assists to text easily recreation, even if it's not complete or contains a great number of errors. In this connection a lot of researchers consider language redundancy to be one of the factors increasing the reliability of received information. It is worth noting that despite numerous definitions of redundancy, linguistic redundancy is mainly defined from the point of view of information communication.

The correlation of redundancy and average word length is important to consider when transmitting information over communication channels for which messages are a coherence of letters that form words and phrases having a certain meaning. In this regard, we define the word size as a number of letters. It is also important to mention that in this case the message source is completely subordinated to the statistical structure of a language conveying the message. By statistical structure we understand the relation between such text characteristics as average word length, the probabilities of one-, two-, three- and multi-letter combinations and others which specify the structure of a language.

2. Materials and Methods

French texts of different functional styles have been used as a source of materials for our research analysis: literary *(Clavel, 1974; Gavalda, 2013)*, publicistic *(Fulda, 2017; Laine, Feldman, 2018)* and scientific *(Barthes, 1972; Derrida, 1996)*. Entropy, redundancy and average word length of these texts have been studied when changing word size (measured as number of letters per word).

As many scientists do *(Baker, 1951; Miller, 1958; Kalimeri et al., 2012 among others)* we use the letters of an alphabet as a basic element for measuring word length. To estimate word frequency in text we have used the absolute frequency although some researchers as is, for instance, M.V. Arapov *(Arapov, 1988)*, use a word rank.

We analyzed the texts based on the entropy of Claude Shannon *H(p) (Shannon, 1951)*. It is a statistical parameter that measures the average amount of information per one letter of a language text:

$$H(p) = -\sum_{i=1}^{N} p_i \log_2(p_i)\,, \tag{1}$$

where p_i is the probability of appearance of the *i*-th word, that is the relative frequency defined as:

$$p_i = \frac{n_i}{M}\,, \tag{2}$$

here n_i is the absolute frequency of appearance of the *i*-th; *M* is the total number of words in a text.

Meanwhile, informational entropy is defined as a measure of uncertainty or unpredictability of information content. In equation (1), *H(p)* is measured in bits per letter.

The average word length Lm is defined as:

$$L_m = \sum_{i=1}^{N} L_{mi} = \sum_{i=1}^{N} p_i l_i \tag{3}$$

where l_i is the length of *i*-th word (the number of letters in the word) and p_i is its relative probability determined by the formula (2).

Redundancy is determined using classic formula, which C. Shannon called "redundancy of a language" *(Shannon, 1948)*:

$$R = 1 - \frac{H}{H_0}\,, \tag{4}$$

In equation (4) *H* refers to entropy determined by formula (1), whereas H_0 indicates maximum entropy (information capacity of the message) and is defined as $H_0 = log_2(N)$.

To ensure the reliability of research findings, all punctuation marks and bibliographical references have been removed from the texts. Besides, we consider apostrophes as letters. We transformed hyphenated text occurrence such as "*finit-elle*" into separate words "*finit*" and "*elle*".

The text processing technique includes sequential stages, the first three of which are presented in table 1, using the example of literary text "Pirates du Rhône" *(Clavel, 1974)*. To change the average word length, we successively removed from the text the words beginning with the shortest one (i.e., one-letter words), then two-, three-, four-letter words, etc. In the interests of concision we give in this paper only the steps for removing one-letter (columns 4, 5, 6) and two-letter words (columns 7, 8, 9).

Thus, we first estimated H_0, *H* and L_m for the primary series (columns 1-3 in Table 1). At the next stage first row (i.e. all one-letter words) was removed and for the new series (columns 4–6 in Table 1) H_0, *H* and L_m were calculated again. Then next length words, i.e. two-letter words, were removed and the same calculations for H_0, *H and* L_m (columns 7–9) were made. After that three-letter words got the same method and so on.

The lengths of words after each stage are as follows: $L_{17} = \sum_{i=1}^{17} p_i l_i$, $L_{16} = \sum_{i=2}^{17} p_i l_i$, $L_{15} = \sum_{i=3}^{17} p_i l_i$,..., $L_5 = \sum_{i=13}^{17} p_i l_i$. The word removal process is completed before the exponential tail

Table 1

Statistical characteristics and some text processing stages on the sample of a literary text "Pirates du Rhône" *(Clavel, 1974)*

Word length (number of letters)	**Entropy, H_i**	**Average word length, L_{mi}**	**Word length (number of letters)**	**Entropy, H_i**	**Average word length, L_{mi}**	**Word length (number of letters)**	**Entropy, H_i**	**Average word length, L_{mi}**
1	2	3	4	5	6	7	8	9
1	0.187389	0.040512	-	-	-	-	-	-
2	0.497877	0.492527	2	0.497877	0.492527	-	-	-
3	0.406407	0.440522	3	0.406407	0.440522	3	0.406407	0.440522
4	0.438934	0.695853	4	0.438934	0.695853	4	0.438934	0.695853
5	0.351783	0.554310	5	0.351783	0.554310	5	0.351783	0.554310
6	0.347780	0.651414	6	0.347780	0.651414	6	0.347780	0.651414
7	0.297525	0.579351	7	0.297525	0.579351	7	0.297525	0.579351
8	0.191553	0.334642	8	0.191553	0.334642	8	0.191553	0.334642
9	0.136352	0.232715	9	0.136352	0.232715	9	0.136352	0.232715
10	0.082690	0.132581	10	0.082690	0.132581	10	0.082690	0.132581
11	0.038432	0.055378	11	0.038432	0.055378	11	0.038432	0.055378
12	0.022770	0.031946	12	0.022770	0.031946	12	0.022770	0.031946
13	0.009301	0.011992	13	0.009301	0.011992	13	0.009301	0.011992
14	0.004984	0.006273	14	0.004984	0.006273	14	0.004984	0.006273
15	0.001996	0.002372	15	0.001996	0.002372	15	0.001996	0.002372
16	0.000400	0.000421	16	0.000400	0.000421	16	0.000400	0.000421
17	0.000400	0.000448	17	0.000400	0.000448	17	0.000400	0.000448
$H_0 = 4.087463$	$H = \sum_{i=1}^{17} H_i = 3.016583$	$L_m = \sum_{i=1}^{17} L_{mi} = 4.263265$	$H_0 = 4.000000$	$H = \sum_{i=2}^{17} H_i = 2.888987$	$L_m = \sum_{i=2}^{17} L_{mi} = 4.401049$	$H_0 = 3.906891$	$H = \sum_{i=3}^{17} H_i = 2.888987$	$L_m = \sum_{i=3}^{17} L_{mi} = 4.401049$

of the entropy (or relative frequency), in this case, before $L_5 = \sum_{i=13}^{17} p_i l_i$. As seen from table 1 with deleting of words different in length, i.e. words with different amount of letters change both average length and text entropy. This correlation is well illustrated on Figure 1, characterizing the relative frequency (Figure 1*a*) and entropy (Figure 1*b*) distribution for different *N*. For *N* = 17 the curves of relative frequency and entropy characterize the original text (unchanged), *N* = 16 – the text without one-letter words, *N* = 15 – the text without two-letter words, *N* = 14 illustrate a text without three-letter words, etc.

From Figure 1 it can be seen that the change of the relative frequency transforms the slope of the curve (i.e. decay rate) and the text entropy, despite the fact that their distribution character remains unchanged. This also changes the average word length. Moreover, after the consecutive removal of words, the average length calculation was performed for points that are not placed on the exponential tail of the entropy curve. In Figure 1 these are the points which abscissas are equal to: 14, 15, 16, 17. These points stand out well on the entropy curve.

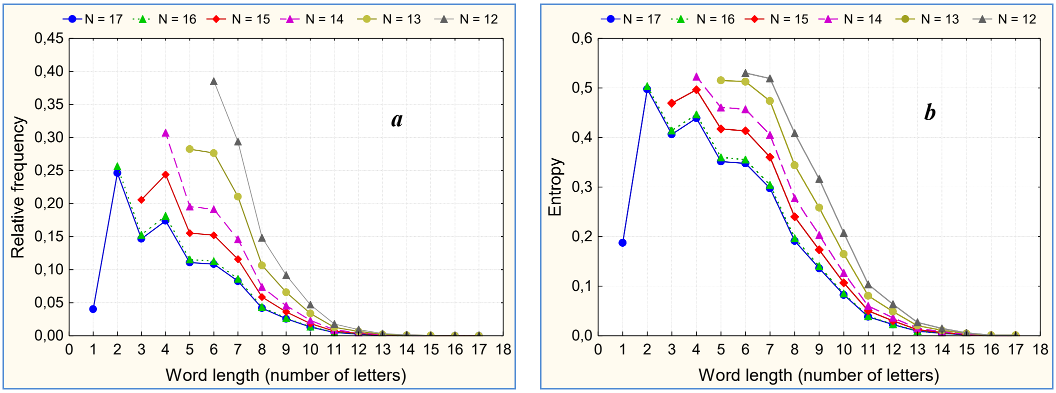

Figure 1. Relative frequency (a) and entropy (b) distribution versus word length for different values of N *(for literary text "Pirates du Rhône", B. Clavel, 1974)*

The researches done by M. Kalimeri and others *(Kalimeri et al., 2012)* state the text entropy in different languages and functional styles differs if only account 5–10-letter words. These differences are not observed on more than 10-letter words (i.e. related to the exponential decay of the relative frequency). In fact, these authors' studies have also resulted in distinguishing between the word lengths related and not related to the exponential tail of the relative frequency (or entropy).

3. Results and Discussion

The idea of differentiating text words to the two types related and not related to the exponential tail of entropy brought to the necessity the study of redundancy *R* and average word length L_m variability depending on the maximum entropy H_0. The choice of H_0 as a parameter referring to which there are the changes of *R* and L_m are explained by the fact that in such a way all words have equiprobable distribution. The results of our research are shown in Figure 2 and Figure 3, which reflect variation of redundancy *R* and L_m depending on H_0.

Variability of R and L_m has been studied taking into account (Figure 2) and not taking into account (Figure 3) the average word length relating to the exponential decay of entropy. On the figures R and L_m are presented in normalized relative units.

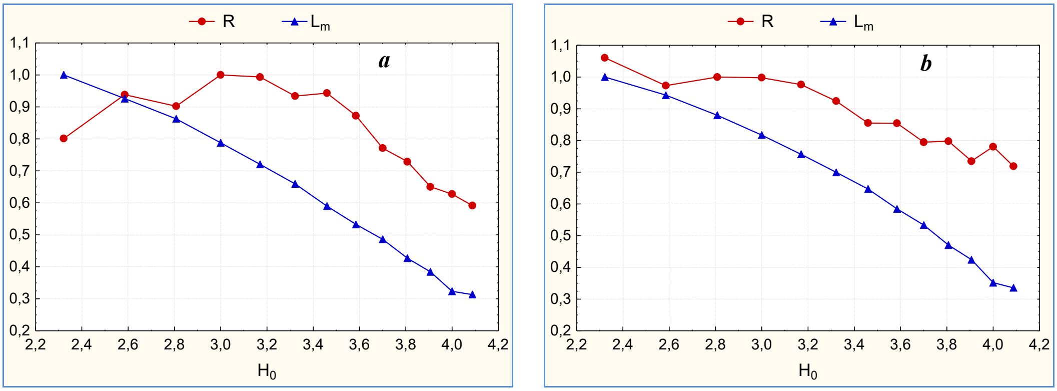

Figure 2. Redundancy *R* and average word length L_m variation versus H_0, for different texts: a) literary *(Clavel, 1974)*; **b) scientific** *(Derrida, 1996)*

Figure 2 demonstrates the dependences $R(H_0)$ and $L_m(H_0)$ taking account all average word lengths including words related to the exponential tail of entropy. These dependences are presented for literary (Figure 2*a*) and scientific (Figure 2*b*) texts. Figure 2 shows that in this case the dependences $R(H_0)$ and $L_m(H_0)$ demonstrate a different character of average word length and text redundancy distribution.

The dependences $R(H_0)$ and $L_m(H_0)$ for the case without taking into account average word lengths related to the exponential tail of entropy are presented in Figure 3. On this figure $R(H_0)$ and $L_m(H_0)$ have almost the same non-monotonous character and maximums. Herewith, the rise and the fall of the average length and redundancy occur in approximately the same part of H_0.

Besides arranging the texts of different functional styles in the sequence: 1) literary (Figure 3*a*, 3*b*); 2) scientific (Figure 3*c*, 3*d*); 3) publicistic (Figure 3*e*, 3*f*) reveals the fact that the maximums of the curves for R and L_m shift towards larger values of H_0. That is clearly expressed for redundancy.

Thus, comparison of Figure 2 and Figure 3 shows that the dependences $R(H_0)$ and $L_m(H_0)$ demonstrate the same character excluding words related to the entropy exponential tail from the calculation of average word length. Thereat, it seems appropriate to distinguish between two average word lengths of the text: the average word length related and not related to the exponential tail of entropy.

4. Conclusions and Suggestions

In the present article the relation of redundancy and average word length in literary, scientific and publicistic French texts has been studied. Variability of the text redundancy R correlates well with the variability of the average word length L_m of a individual text, if not taking into account the word lengths related to the exponential tail of entropy. Moreover, the dependences of redundancy and average length on the maximum entropy have almost the same

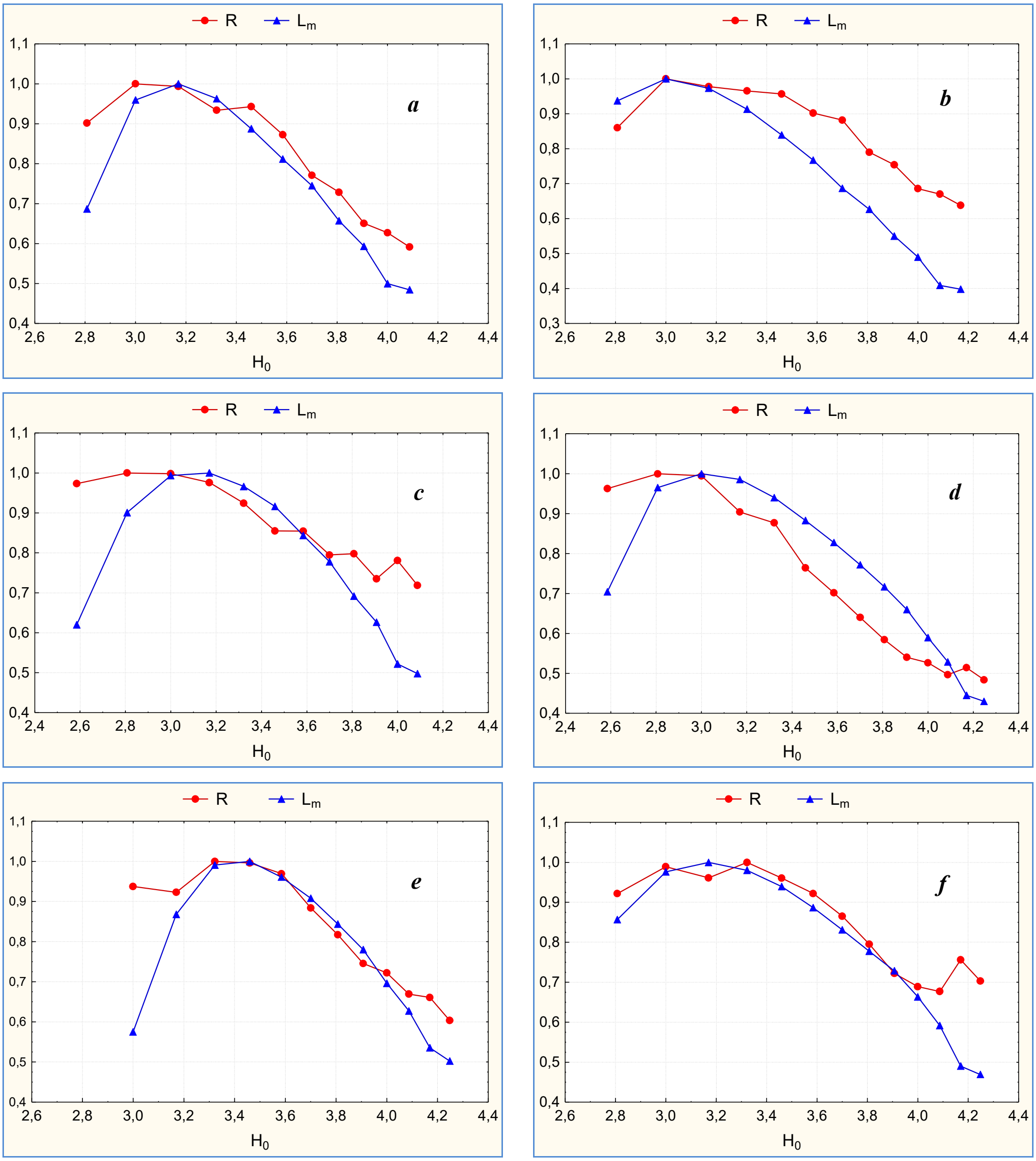

Figure 3. Texts redundancy *R* and average word length L_m variation without considering words related to the entropy exponential tail in dependence of the H_0: *a, b* – literary; *c, d* – scientific; *e, f* – publicistic

non-monotonous character and maximums. On this occasion it is preferable to distinguish average word lengths related and not related to the exponential tail of entropy.

Taking into account the identified patterns can be useful when assessing the text redundancy, transferring information (text) over the communication channel, as well as modeling of informational entropy.

In our opinion the coincidence of redundancy and average word length variability character makes possible to determine the range of word lengths (words consisting of letters of different numbers) that can be removed from the message (text) with the minimum damage to the meaning of the original text. Verification of this hypothesis comprises the prospect of further researches.

Acknowledgements

The author is grateful to Andrey Olinchuk for proving the useful program on C# for determination of French text statistical characteristics. Also special thanks to Dr Dzhema V. Melkonyan for her precious assistance in mathematical data analysis.

References

Alontseva, N. V., Ermoshin, Y. A. (2019). Problem of language redundancy on the example of a scientific text. RUDN Journal of Language Studies, Semiotics and Semantics, 10 (1), 129–140. DOI: 10.22363/2313-2299-2019-10-1-129-140. [in English].

Arapov, M. V. (1988). Kvantitativnaya lingvistika [Quantitative linguistics]. Moscow: Nauka. [in Russian].

Baker, S. J. (1951). A linguistic law of constancy: II. The Journal of General Psychology, 44, 113–120. [in English].

Barthes, R. (1972). Le degré zéro de l'écriture [Writing Degree Zero]. Paris: Seuil. [in French].

Clavel, B. (1974). Pirates du Rhône [Fishermen of the Rhône]. Paris: Robert Laffont. [in French].

Derrida, J. (1996). Le monolinguisme de l'autre où la prothèse de l'origine [Monolingualism of the Other or the Prosthesis of Origin]. Paris: Galilée. [in French].

Dubois, J., Edeline, F. Klinkenberg, J.M., Minguet, P., Pire, F., Trinon, H. (1970). Rhétorique générale [A General Rhetoric]. Paris: Larousse. [in French].

Fulda, A. (2017). Emmanuel Macron, un jeune homme si parfait [Emmanuel Macron, a young man so perfect]. Paris: Plon. [in French].

Gavalda, A. (2013). Billie [Billie]. Paris: Le Dilettante. [in French].

Gillette, M., Wit, E.J.C. (1999). What is Linguistic Redundancy? A Technical Report. University of Chicago, U.S.A. Retrieved from: http://www.math.rug.nl/~ernst/linguistics/redundancy3.pdf. [in English].

Grudeva, E.V. (2008). Izbytochnost teksta: istoriya voprosa i metodika issledovaniya [Redundancy of the text: the history of the issue and the methodology of the research]. Izvestiya Rossijskogo gosudarstvennogo pedagogicheskogo universiteta imeni A.I. Gercena [News of the Russian A.I. Herzen State Pedagogical University], 59, 106–114. [in Russian].

Grudeva, E.V. (2010). Izbytochnost yazyka i izbytochnost teksta: nekotorye razmyshleniya [Redundancy of the language and redundancy of the text: some reflexions]. Acta linguistica Petropolitana. Trudy Instituta lingvisticheskih issledovanij [J. of the Institute for Linguistic Studies], 6 (2), 73–89. [in Russian].

Grzybek, P., Standlober, E., Kelih, E., Antic, G. (2005). Quantitative Text Typology: The Impact of Word Length. C. Weihs and W. Gaul (Eds.). Classification – The Ubiquitous Challenge. Heidelberg: Springer, 53–64. [in English].

Guerrero, F.G. (2005). A new look at the classical entropy of written English. IEEE Transactions of Information Theory. preprint arXiv:0901.4784. Retrieved from: https://www.researchgate.net/

publication/45883885_A_New_Look_at_the_Classical_Entropy_of_Written_English. [in English].

Kalimeri, M., Constantoudis, V., Papadimitriou, C., Karamanos, K., Diakonos, F.K., and Papageorgiou, H. (2012). Entropy analysis of word-length series of natural language texts: Effects of text language and genre. International Journal of Bifurcation and Chaos, 22(9). DOI:10.1142/S0218127412502239. [in English].

Kalimeri, M., Constantoudis, V., Papadimitriou, C., Karamanos, K., Diakonos, F.K., and Papageorgiou, H. (2015). Word-length entropies and correlations of natural language written texts. Journal of Quantitative Linguistics, 22 (2), 101–118. [in English].

Köhler, R. (2005). Synergetic linguistics. Quantitative Linguistics. Köhler, R., Altmann, G., Piotrowski, R.G.(eds.). An International Handbook. Berlin/New York: de Gruyter. 760–774. [in English].

Laine, M., Feldman J.-Ph. (2018). Transformer la France [To transform France]. Paris: Plon. [in French].

Martinet, A. (1991). Éléments de linguistique générale [Elements of General Linguistics]. Paris: Armand Colin. [in French].

Mikros, G. K., Hatzigeorgiu, N., and Carayannis, G. (2005). Basic quantitative characteristics of the modern greek language using the hellenic national corpus. Journal of Quantitative Linguistics, 12 (2–3), 167–184. DOI: 10.1080/09296170500172478. [in English].

Miller, G.A., Newman, E.B., Friedman, E.A. (1958). Length-frequency statistics for written English. Information and Control, 1, 370–389. [in English].

Newman, E. B., Waugh, N. C. (1960). The redundancy of texts in three languages. Information and Control, 3, 141–153. https://doi.org/10.1016/S0019-9958(60)90731-2. [in English].

Popescu, I.-I., Naumann, S., Kelih E., Rovenchak, A. et al. (2013). Word length: aspects and languages. Issues in quantitative linguistics. Köhler, R., Altmann, G. (eds), 3, 224–281. [in English].

Raatz, U., Kelein-Braley, C. (2002). Introduction to the language and the C-Test. University Language Testing and the C-Test. J. A. Coleman, R. Grotjahn, & U. Raatz (Eds.). Bochum: AKS-Verlag, 75–86. [in English].

Shannon, C. E. (1948) A mathematical theory of communication. The Bell System Technical Journal, 27 (3), 379–423. [in English].

Shannon, C. E. (1951). Prediction and entropy of printed English. Bell System Technical Journal (BSTI), 30, № 1, 50–64. [in English].

Strauss, U., Grzybek, P., Altmann, G. (2007). Word Length and Word Frequency. Contributions to the Science of Text and Language. Text, Speech and Language Technology. Grzybek, P. (eds), 31. Dordrecht: Springer, 277–294. [in English].

Zipf, G. K. (1949). Human behaviour and the principle of least effort. Cambridge: Addison-Wesley Press. [in English].