Can Zipf Analyses and Entropy Distinguish Between Artificial and Natural Language Texts?

A. Cohen[1], R. N. Mantegna[2], S. Havlin[1]

Jan 8 1996

[1] *Department of Physics, Bar-Ilan University, Ramat-Gan 52900, ISRAEL*

[2] *Dipartimento di Energetica ed Applicazioni di Fisica, Università di Palermo, Palermo, I-90128, ITALIA*

Abstract

We study statistical properties of natural texts written in English and of two types of artificial texts. As statistical tools we use the conventional and the inverse Zipf analyses, the Shannon entropy and a quantity which is a nonlinear function of the word frequencies , the frequency relative "entropy". Our results obtained by investigating eight complete books and sixteen related artificial texts suggest that the inverse Zipf analysis of the frequencies of distinct words may succeed better than the conventional Zipf analysis of the distinct words in distinguishing between natural and artificial texts. Our results also indicate that the frequency relative "entropy" is more useful than

the usual word entropy in distinguishing between natural and artificial texts. By studying the scaling behavior of both entropies as a function of the total number of words T of the investigated text, we find that the word entropy scales with the same functional form for both natural and artificial texts but with a different parameter while the frequencies relative "entropy" decreases monotonically with T for the artificial texts but has a minimum at $T \approx 10^4$ for the natural texts.

1 Introduction

Zipf analysis is a statistical tool used in several research fields. It was originally used by Zipf in statistical linguistics [1,2]. Early in the sixties was also used in the analysis of economic [3] and social [4] systems and recently it has been used in the study of systems such as chaotic dynamical systems [5], biological sequences [6,7] and economic systems [8,9].

Zipf found, for texts written in natural languages, a universal power-law behavior characterized by a power-law exponent close to 1. Several theoretical models [10,11,12,13] have been proposed to explain Zipf's law. Some of them [11,13] show, theoretically and empirically, that Zipf's law is also satisfied in randomly generated symbolic sequences, with an exponent ζ close to one. These theoretical models and empirical results suggest that Zipf's law and the value of its exponent, ζ, reflect little about the linguistic nature of a text. An experimental observation that a Zipf analysis of a symbolic sequence gives a Zipf plot is then not sufficient to prove that the symbolic sequence is a hierarchically structured language. On the other hand, the validity of the Zipf law implies that the observation of a Zipf behavior is

"necessary" in a natural text. Moreover, the analysis of the frequencies of the n-gram substrings of a text allows a language-independent categorization of topical similarity in unrestricted texts [14] so that a Zipf analysis may be useful for practical purposes. Another example of the usefulness of the Zipf's approach is given in [15] where it is suggested that the "distance" between two Zipf plots of two different texts is shorter when the texts are written by the same author than when written by two different authors.

In this paper, we analyze real and artificial texts by applying several statistical tools based on the analysis of occurrences of distinct words. They are the conventional and the inverse Zipf analysis and the study of the word entropy and of a new nonlinear function of the word frequencies which we address as frequency "entropy". We check if the tools we use are able to discriminate between real and artificial texts. We show that inverse Zipf analysis, word entropy and frequency entropy give different outcome in real and artificial texts while the conventional Zipf analysis gives approximately the same outcome when one investigates words with not too few occurrences. Our results suggest that the main difference between real and artificial texts are observed in the frequencies of the least frequent words.

The paper is organized as follow: In sect I, we discuss the statistical tools we use in our study and the technique we use to generate artificial symbolic sequences related to the investigated natural texts while in sect. III we present the results of our statistical analysis of 8 complete books and 16 related artificial texts. In sect. IV we discuss the scaling properties of the measured entropies and in sect. V we briefly draw our conclusion.

2 Conventional and new tools in statistical linguistic analysis

A conventional Zipf analysis is performed by counting the occurrences of distinct words in a given text. If one counts the occurrences of each distinct word in a long enough text and plots the number of occurrences as a function of the rank, r, i.e. of the position of the word in a table ordered with respect to the number of occurrences from the most to the least frequent word, one finds that $f(r)$, the number of occurrences of the word of rank r, and its rank are related by

$$f(r) \approx A_1 r^{-\zeta}, \tag{1}$$

where A_1 is constant and $\zeta \approx 1$.

Another kind of Zipf analysis [16] for natural texts, sometime called the inverse Zipf analysis, is the study of the words' frequencies distribution. Zipf found for natural texts a power law behavior that holds only for the low frequency words, and states that the number of distinct words, $I(f)$, that have the frequency f, is given by

$$I(f) \approx A_2 f^{-\alpha}, \tag{2}$$

where A_2 and α are constants (Zipf [16] estimated the constant α close to 2).

A different important tool used in the statistical analyses of symbolic sequences is the entropy or Shannon information [17,18]. For a text of R distinct words, the relative entropy is defined as

$$\tilde{H}_w = -\sum_{r=1}^{R} \frac{f(r)}{T} log_R \frac{f(r)}{T}. \tag{3}$$

where, T is the total number of words in the analyzed text and the label w in $\tilde{H}_w$ stands for the *words* entropy measured by using a base R logarithm. We choice this value for the base of the logarithm function because we wish to compare the results obtained by analyzing different texts with different R-s and the base R entropy, in contrast to the usual base 2 entropy, gives values in the range (0, 1) for any value of R.

Next, we introduce a new relative "entropy" which is calculated from the words' *frequencies* distribution instead of the words relative occurrences. We define the *frequencies "entropy"* as

$$\tilde{H}_f = -\sum_{f=1}^{F} \frac{I(f)}{R} log_F \frac{I(f)}{R}. \tag{4}$$

where $I(f)$ is the number of distinct words that have occurrence f, F is the total number of different occurrences, and the index f in $\tilde{H}_f$ stands for the *words' frequencies*. Here again, since the logarithmic base is F, $\tilde{H}_f$ is in the range $(0, 1)$. The advantage of $\tilde{H}_f$ compared with $\tilde{H}_w$ will be discussed in the next sections. It is probably worth noting that the calculation of an entropy value is range independent while the estimation of either the Zipf's law or the inverse Zipf's law exponents depends on the range where one performs the fitting procedure.

In this paper we compare several long natural texts written in English mainly with two types of artificial texts. These two types of artificial texts are:

1. An artificial text which is randomly generated with letters and space-mark probabilities equal to the distribution of letters and space-mark measured in a given natural text. We refer to this type of artificial text as "artificial 1".
2. An artificial text characterized by the same sequence of letters as the

natural one's but without the space and punctuation marks between the words. In the analyses of these texts an arbitrary letter plays the role of the space-mark as words delimiter. We refer to this type of artificial text as "artificial 2". We choose the letter "e" as words delimiter since it is the most frequent letter after the space-mark.

We wish to point out that wherever we use the phrase "artificial texts" we do not mean any type of artificial texts other than these two. In addition to these two types of artificila texts in Sect.IV we also investigate m-order Markovian texts , with m ranging from 1 to 4. A text of the kind we label as "artificial 1" is a symbolic Bernulli sequence in which the frequency of occurrence of the letters is the same as observed in the related natural text but the order of the letters is completely random. An "artificial 2" text mimics a symbolic sequence in which the letters are correctly ordered and carries information but the seperation into "words" is unknown. A study of these two types of artificial texts is important for obtaining preliminary information that may be used in a future analysis of symbolic sequences, other than literary texts, but which show evidence of a *possible* presence of an underlying hierarchical language. A noncoding DNA sequence, for example, is speculated to be such a case [7]. In such a sequence the possible seperation into elementary semantic units ("words") as well as the correct order the semantic unit is unknown.

3 Statistical properties of natural and artificial texts

In this study we investigate 8 complete books written in the English language. We ignore any symbol other than the "a-z" English letters and the space-mark. For each book, which is a natural text, we generate two artificial texts of the types "1" and "2" as defined above. We analyze the above three types of texts by performing the conventional Zipf and by the inverse Zipf analyses. We also measure, for each text, the words entropy, $\tilde{H}_w$, and the frequencies "entropy", $\tilde{H}_f$. Zipf's law exponent, ζ, is determined after a logarithmic binning of the rank. The inverse Zipf's law holds only for the low frequencies, and we arbitrarily choose to determine the inverse Zipf's law exponent, α, by analyzing only the first 20% of the frequencies, after a logarithmic binning of the frequency. Both binnings are performed with windows of the size 0.1 on the logarithmic scale.

As expected on the basis of previous theoretical and empirical studies [11,13] Zipf's exponent $\zeta \approx 1$ is not specific to hierarchical structured languages. $\zeta \approx 1$ is also observed when one analyzes "words" in an artificial random text (for not too rare "words"). We verify that the inverse Zipf's law is valid in natural languages ($< \alpha > \approx 1.6$). A representative example is shown in Fig. 1. We also find that the inverse Zipf's law is satisfied by artificial texts, with the exception of the first point in the inverse Zipf's plot (the number of distinct words that occur only once, not included in the calculation of α) . However, as it is evident from Fig. 1, the values of α are different in natural and artificial texts.

In Table 1 we give, for each book studied, the total number of words, T, the number of distinct words R, the Zipf's law exponent ζ, the inverse Zipf's law exponent α, the words entropy $\tilde{H}_w$, and the frequencies "entropy" $\tilde{H}_f$, for the three types of texts.

The most striking difference between the natural and the artificial texts is the ratio R/T, which is the relative amount of distinct words, or, in other words, the vocabulary size of a text. R/T is remarkably lower for the natural texts than for the artificial texts. This difference is expected since a natural text's vocabulary is constrained to be smaller than an artificial one's by the phonetic, by the grammar of the language, by the subject of discussion and by the author's style. This ratio has already been used, for example, by Trifonov in the study of DNA sequences [19]. It is also known in the linguistic literature [20] that the vocabulary of natural texts as a function of a text size follows a power law behavior,

$$R \sim T^{\beta} \tag{5}$$

where β slowly decreases with T but can be considered constant for a large range of T. Artifical texts have also be found to satisfy Eq. (5) but the vocabulary growth rate β assumes values significantly higher in artificial than in natural texts. This property seems to be a convinient tool for distinguishing between natural and artificial texts [21].

This remarkable difference in vocabulary size is not expressed through the value of the Zipf's law exponent ζ. One can see in Table 1 that for both natural and artificial texts ζ is very close to 1. The average difference in ζ between the natural texts and the "artificial 1" texts is 5% and and between the natural texts and the "artificial 2" texts is 14%. Our results suggest

that artificial and natural texts may be better discriminate by performing an inverse Zipf analysis. In fact, in the inverse Zipf analysis, the inverse Zipf exponent, α is systematically different in artificial and natural texts. The average differences in α between natural and "artificial 1" and "artificial 2" texts are 34% and 28%, respectively.

While the Zipf analysis is dominated by the high frequency region, the inverse Zipf analysis is dominated by the low frequencies. Indeed, by a careful observation of the conventional Zipf's plots one can see that for the artificial texts the plateaus at the low frequencies region are much broader than for the natural texts, which means that in an artificial text there are much more words that appear only once or twice. Due to the nature of a bilogarithmic plot, this fact only weakly affects the measured values of ζ. Conversely, in the inverse Zipf analysis the low frequency words mainly affect the value of the exponent α. The fact that the inverse Zipf analysis is found more useful for distinguishing between natural and artificial texts suggests that, in general, the analysis of the words' frequencies distribution should be more useful in this task than the analysis of words distribution.

One can see in Table 1 that for both relative entropies, $\tilde{H}_w$ and $\tilde{H}_f$, there is, consistently, a difference between the natural and the artificial texts. The average differences in $\tilde{H}_w$ between the natural and the artificial texts of types "1" and "2" are 17% and 20%, respectively, while the differences in $\tilde{H}_f$ are 410% and 224%. However, this results alone are not sufficient in order to conclude that the study of the words frequencies is more useful in distinguishing between natural and artificial texts. In fact, if one compares the differences between the natural and the artificial texts' related "redundancies", $R_w = 1 - \tilde{H}_w$ and $R_f = 1 - \tilde{H}_f$, one finds that the average differences

in R_w between the natural and the artificial texts of types "1" and "2", are 67% and 90%, respectively, while the average differences in R_f are 64% and 55%.

In spite of this last result, we find the frequencies "entropy", $\tilde{H}_f$, (or frequencies "redundancy", R_f), more useful than the words entropy, $\tilde{H}_w$, (or words "redundancy", R_w). There are two main reason for the above statement: (i) For almost all books studied the absolute difference in $\tilde{H}_f$ is more than twice the absolute difference in $\tilde{H}_w$; (ii) as will be shown in the next section, the scaling behavior of $\tilde{H}_f$ as a function of a text size has a different profile in the investigated range of T in natural and artificial texts while the scaling behavior of $\tilde{H}_w$ is similar for all three types of texts.

4 Scaling behavior of entropies

For each book's natural and artificial texts we determine the averages $\langle \tilde{H}_w \rangle$ and $\langle \tilde{H}_f \rangle$ as a function of T. The average is performed over all existing subtexts of length T in a given text. For example, in a text of 100000 words there exist 99001 subtexts of 1000 words. For each text, the smallest subtext is of 1000 words and the largest is the complete text. We find, as shown in Fig. 2, that, $\langle \tilde{H}_w(T) \rangle$ scales similarly for both natural and artificial texts and can be approximated by a decreasing linear function of $log(T)$. Thus,

$$\langle \tilde{H}_w(T) \rangle = A_w - B_w log_{10} T \tag{6}$$

where A_w and B_w are positive constants. It is worth noting that the constant B_w is systematically different in natural and artificial texts.

Converseley, as shown in Figs. 3a, 3b, we find a clear difference in the scaling behavior of $\langle \tilde{H}_f(T) \rangle$ in natural and artificial texts in the investigated range of T. For the artificial texts $\langle \tilde{H}_f \rangle$ decreases monotonically with T, while for the natural texts it reaches a minimum approximately at $T \approx 10^4$ (well inside the studied range of T). Both natural and artificial texts $\langle \tilde{H}_f \rangle$ can be well approximated by a parabolic function of $log(T)$,

$$\langle \tilde{H}_f(T) \rangle = A_f - B_f log_{10} T + C_f (log_{10} T)^2 \qquad (7)$$

where A_f, B_f and C_f are positive constants. By using the values of the constants A_f, B_f and C_f obtained by best fitting the data with Eq. (7), for the artificial texts of type "1" and "2" the minimum of the parabola is not predicted within the studied range of T and it is expected only at $T \approx 10^9$ and at $T \approx 10^6$, respectively. In fact C_f for the artificial texts is usually very small (see Table 2). The parameters of the best fits $\langle \tilde{H}_w(T) \rangle$ and $\langle \tilde{H}_f(T) \rangle$ are given in Table 2. In all cases there is a good agreement with Eq. (6) and in most cases with Eq. (7). Deviations from Eq. (7) were found for the natural texts of the books "Critique of Pure Reason" by Kant and "Descent of Man" by Darwin. This may be explained by the fact that these books include several subjects and therefore there are sharp changes in the vocabulary used in different sections. These changes do not have a strong effect on the behavior of $\langle \tilde{H}_w(T) \rangle$ but do strongly affect the behavior of $\langle \tilde{H}_f(T) \rangle$. For the other books, which are more uniform with respect to the use of the words, the behavior of $\langle \tilde{H}_f(T) \rangle$ for the natural texts is smoother and agrees well with Eq. (7).

The "artificial 1" texts are 0-order Markovian texts. We also perform our study on the scaling properties of $\langle \tilde{H}_f(T) \rangle$ and $\langle \tilde{H}_w(T) \rangle$ in Markovian texts of finite order $m = 1..4$. The transition probabilities used to generate

these Markovian texts are taken from the book "War & Peace". Again, only the "a-z" English letters and the space mark are considered and words are sequences of letters between two space marks. All Markovian texts are of the same number of words, 262677, as the natural text of the book "War & Peace". As m increases the Markovian text becomes more like the natural text. Hence, it is reasonable to expect that the statistical properties would also become more like the natural text's ones. We find, as shown in Fig. 4, that for all m studied $\langle \tilde{H}_w \rangle$ decreases with T. As m increases the decrease of $\langle \tilde{H}_w(T) \rangle$ is faster and gets closer to the natural text. For $\langle \tilde{H}_w(T) \rangle$ already the 3-order Markovian text overlaps the natural text. In contrast, as shown in Figs. 5a and 5b, for $\langle \tilde{H}_f \rangle$ even the 4-order Markovian text significantly differs from the natural text. For $m = 0..3$ $\langle \tilde{H}_f \rangle$ decreases with T in the complete range of T and only for the 4-order Markovian text it has a minimum within the studied range of T (at $T \cong 10^4$ as the natural text). Yet, $\langle \tilde{H}_f \rangle$ for the natural text is significantly higher than for the 4-order Markovian text and increases faster than the natural one with T.

Since $\tilde{H}_w$ and $\tilde{H}_f$ are limited, by definition, to be in the range $(0, 1)$, it is obvious that Eqs. (6) and (7) cannot hold for infinite texts. However a deviation from these equations is predicted only for very long texts, much longer than any human writing. For example, under the assumption of a logarithmic behavior with the constants given in Table 2 for the book War & Peace, the lower limit of $\tilde{H}_w$, $(\tilde{H}_w = 0)$, is achieved only at $T \sim 10^{14}$ for the natural text and at $T \sim 10^{20}$ for both artificial texts. For the same book, the upper limit of $\tilde{H}_f$ $(\tilde{H}_f = 1)$ is achieved only at $T \sim 10^{10}$ for the natural text, at $T \sim 10^{27}$ for the "artificial 1" text and at $T \sim 10^{16}$ for the "artificial 2" text.

5 Summary and Conclusions

We report a numerical investigation of natural texts and symbolic strings suggesting that artificial texts of two specific types are better distinguished from natural texts by the inverse Zipf analysis than by the conventional Zipf analysis. We also introduce the frequencies "entropy" and suggest it as a convenient and more useful statistical tool than the traditional words entropy for distinguishing between artificial texts and natural texts than the traditional words entropy. We empirically observe the advantage of the advantage of the use of the frequencies "entropy" with respect to the words entropy in distinguishing between artificial and natural texts. Our results may suggest that the linguistic constraints on a natural text are more strongly reflected in the behavior of the less frequent words.

Our numerical results suggest that a statistical analysis of the occurrences of different words in a symbolic text give information about the nature of the observed string especially when the behavior of the less frequent words is properly taken into account. We argue that that these linguistic and information theory tools may be relevant also for the analysis of symbolic sequences, other than human writing texts, which is speculated to have some of the characteristics of languages.

References

1. G. K. Zipf, *Human Behavior and the Principle of Least Effort*, Addison-Welsey Press, (1949).

2. G. K. Zipf, *The Psycho-Biology of Language, An Introduction to Dynamic Philology*, Cambridge MA: MIT Press, (1965).

3. B. Mandelbrot, *Oligopoly, Mergers, and Paretian Size Distribution of Firms*, Research note #NC-246 (Thomas J. Watson Research Center, Yorktown Heights, NY).

4. B. Mandelbrot, *A Class of Long-tailed Probability Distributions and the Empirical Distribution of City Sizes*, in Mathematical Explorations in Behavioral Science, Edited by F. Massarik and P.Ratoosh (R.D.Irwin, Inc and The Dorsey Press, Homewood, IL, 1965).

5. G. Nicolis, C. Nicolis, and J.S. Nicolis, *Chaotic Dynamics, Markov Partitions, and Zipf's Law*, Journal of Statistical Physics **54**, 915 (1989).

6. M. Yu. Borodovsky and S. M. Gusein-Zade, *A General Rule for Ranged Series of Codon Frequencies in Different Genomes*, J. Biomolecular Structure & Dynamics **6**, 1001 (1989).

7. R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C. K. Peng, M. Simons and H. E. Stanley, *Linguistic Features of Noncoding DNA Sequences*, Phys. Rev. Lett. **73**, 3169 (1994); *Systematic analysis of coding and noncoding DNA sequences using methods of statistical linguistics*, Phys. Rev. E**52**, 2939 (1995).

8. J.P.Bouchaud, *More Lévy Distributions in Physics*, in Lévy Flights and Related Topics in Physics, Edited by M.F.Shlesinger, G.M.Zaslavsky and U.Frish (Springer, Berlin, 1995) pp 239-250; H.E. Stanley, S. V. Buldyrev, A. L. Goldberger, S. Havlin, R.N. Mantegna, C. K. Peng, M. Simons and M. H. R. Stanley, *Long-Range Correlations and Generalized Lévy Walks in DNA Sequences*, in Lévy Flights and Related Topics in Physics, Edited by M.F.Shlesinger, G.M.Zaslavsky and U.Frish (Springer, Berlin, 1995) pp 331-347.

9. M. H. R. Stanley, S. V. Buldyrev, S. Havlin, R. N. Mantegna, M. A. Salinger and H. E. Stanley, *Zipf Plots and the Size Distribution of Firms*, Economic Lett. **49**, 453 (1995).

10. B. Mandelbrot, *An Informational Theory of the Statistical Structure of Language*, in Communication Theory, W. Jackson, Ed., Butterworths Scientific Publications, London (1953).

11. B. Mandelbrot, *Information Theory and Psycholinguistics: A Theory of Words Frequencies*, in *Readings in Mathematical Social Science*, P. F. ?feld and N. W. Henry, Eds., MIT Press (1966).

12. H. A. Simon, *On a Class of Skew Distribution Functions*, Biometrika **42**, 435-440 (1955).

13. W. Li, *Random Texts Exhibit Zipf's-Law-Like Word Frequency Distribution*, IEEE Trans. on Inf. Theory **38**, No. 6, (1992).

14. M. Damashek, *Gauging Similarity with n-Grams: Language-Independent Categorization of Text*, Science **267**, 843 (1995).

15. S. Havlin, *The Distance Between Zipf Plots*, Physica A **216**, 148-150 (1995).

16. Ref. [2], pp. 40-44.

17. C. E. Shannon, *A Mathematical Theory of Communication*, Bell System Technical Journal **27**, 379-423, (1948).

18. L. Brillouin, *Science and Information Theory*, Academic Press, New York (1956).

19. E. N. Trifonov, *Making Sense of the Human Genome*, Structure & Methods, Vol. 1: *Human Genome Initiative & DNA Recombination*, Adenine Press (1990), p. 69.

20. G. Herdan, *The Advanced Theory of Language as Choice and Chance*, Springer-Verlag. New York, Inc. 1966.

21. A. Cohen, to be published.

Figure captions

Fig. 1. Plot of the inverse Zipf's law for War & Peace. (a) The natural text, (b) "artificial 1" text, (c) "artificial 2" text. Measured distribution of frequencies (solid lines) and best fitting inverse Zipf's law (Eq. (2)) (dashed lines). For the natural text $\alpha = 1.61$, for the "artificial 1" text $\alpha = 2.18$ and for the "artificial 2" text $\alpha = 2.14$.

Fig. 2. The average words entropy $\langle \tilde{H}_w \rangle$ as a function of a text size T, for War & Peace. (a) The natural text, (b) "artificial 1" text, (c) "artificial 2" text. Symbols are the measured values and the dashed line is the best fit obtained with Eq. (6). $\langle \tilde{H}_w \rangle$ decreases linearly with $log(T)$ in the studied range of T. See Table 2 for the parameters of the best fit.

Fig. 3a. The average frequencies "entropy", $\langle \tilde{H}_f \rangle$ as function of text size T, for the artificial texts related to War & Peace. (a) "artificial 1" text, (b) "artificial 2" text. Symbols are the measured values and the dashed lines are the best fits obtained by using Eq. (7). For the artificial texts $\langle \tilde{H}_f \rangle$ decreases monotonically with T in the studied range of T. See Table 2 for the fitting parameters.

Fig. 3b. The average frequencies "entropy", $\langle \tilde{H}_f \rangle$ as a function of a text size T for the book War & Peace. Symbols are the measured values and the dashed line is the best fit obtained by using Eq. (7). In contrast with the artificial texts $\tilde{H}_f$ is nonmonotonic with T within the studied range of T.

Fig. 4. The words entropy, $\tilde{H}_w$ as a function of a text size T for the book War & Peace and artificially related m-order Markovian texts ($m = 0..4$). Symbols are the measured values and the dashed lines are the best fits obtained by using Eq. (5). Already for $m \geq 3$ the Markovian texts give values of $< \tilde{H}_w(T) >$ very close to the one observed for the natural text.

Fig. 5a. The average frequencies "entropy" $\langle \tilde{H}_f \rangle$ as a function of a text size T for m-order Markovian texts related to the book War & Peace ($m = 0..3$). $\langle \tilde{H}_f(T) \rangle$ decreases monotonically with T in the studied range of T.

Fig. 5b. The average frequencies "entropy" $\langle \tilde{H}_f \rangle$ of a 4-order Markovian text related to the book War & Peace. Symbols are measured values and the dashed line is the best fit obtained by using Eq. (7). Like the natural text (see Fig. 3b) Eq. (7) fits well the data of the 4-order Markovian text, however $\langle \tilde{H}_f(T) \rangle$ for the natural text is significantly higher and grows faster.

Table I

Book - Author	Type	T	R	ζ	α	$\tilde{H}_w$	$\tilde{H}_f$
Wealth of Nations - Smith	natural	365352	9598	1.11	1.46	.669	.513
" "	artificial 1	365352	176915	1.01	2.10	.810	.079
" "	artificial 2	213672	69565	0.93	2.03	.797	.164
War & Peace - Tolstoy	natural	262677	12925	1.05	1.61	.698	.451
" "	artificial 1	262677	124998	1.01	2.18	.808	.085
" "	artificial 2	138471	63656	0.92	2.14	.842	.126
Descent of Man - Darwin	natural	234210	12359	1.06	1.59	.687	.443
" "	artificial 1	234210	118802	1.00	2.18	.823	.081
" "	artificial 2	145512	52007	0.93	2.01	.806	.154
Moby Dick - Melville	natural	210966	18299	1.02	1.82	.706	.380
" "	artificial 1	210966	103014	1.01	2.24	.815	.086
" "	artificial 2	110310	55945	0.93	2.15	.853	.116
Critique of Pure Reason - Kant	natural	200274	6510	1.12	1.51	.670	.487
" "	artificial 1	200274	104308	1.00	2.19	.830	.081
" "	artificial 2	120723	43564	0.96	2.12	.797	.153
Anna Karenina - Tolstoy	natural	162907	9781	1.04	1.66	.705	.448
" "	artificial 1	162907	79727	1.00	2.22	.817	.090
" "	artificial 2	83082	40965	0.90	2.11	.859	.132
Don Quixote - Cervantes	natural	123873	8698	1.05	1.70	.700	.424
" "	artificial 1	123873	59432	1.00	2.25	.814	.096
" "	artificial 2	59573	30978	0.88	2.18	.872	.133
Faus - Goethe	natural	66696	8673	0.95	1.87	.760	.381
" "	artificial 1	66696	34885	0.98	2.22	.839	.102
" "	artificial 2	34505	20120	0.92	2.16	.879	.117

Table 1 - The Zipf's exponent ζ, the inverse Zipf's exponent α, the *words entropy* $\tilde{H}_w$, the *frequencies "entropy"* $\tilde{H}_f$ measured in the natural texts and in the related (see text) artificial texts of types "1" and "2". The quantities α and $\tilde{H}_f$ seem to be more useful for distinguishing between natural and artificial texts than the quantities ζ and $\tilde{H}_w$.

Table II

Book - Author	Type	A_w	B_w	A_f	B_f	C_f
Wealth of Nations - Smith	natural	1.08	.076	0.73	0.115	0.013
" "	artificial 1	1.11	.053	0.20	0.021	-
" "	artificial 2	1.13	.063	0.59	0.139	0.011
War & Peace - Tolstoy	natural	1.12	.079	0.59	0.096	0.013
" "	artificial 1	1.11	.056	0.29	0.054	0.003
" "	artificial 2	1.12	.053	0.42	0.096	0.008
Descent of Man - Darwin	natural	1.11	.081	0.59	0.086	0.012
" "	artificial 1	1.11	.054	0.27	0.050	0.003
" "	artificial 2	1.14	.065	0.42	0.072	0.004
Moby Dick - Melville	natural	1.13	.081	0.82	0.233	0.028
" "	artificial 1	1.11	.056	0.28	0.051	0.003
" "	artificial 2	1.11	.051	0.37	0.082	0.007
Critique of Pure Reason - Kant	natural	1.10	.083	0.51	0.016	0.002
" "	artificial 1	1.11	.053	0.29	0.060	0.004
" "	artificial 2	1.13	.067	0.54	0.135	0.012
Anna Karenina - Tolstoy	natural	1.13	.082	0.64	0.120	0.016
" "	artificial 1	1.12	.057	0.28	0.050	0.003
" "	artificial 2	1.11	.051	0.43	0.102	0.009
Don Quixote - Cervantes	natural	1.14	.087	0.77	0.194	0.025
" "	artificial 1	1.12	.060	0.27	0.039	0.001
" "	artificial 2	1.12	.051	0.33	0.055	0.003
Faust - Goethe	natural	1.16	.083	0.77	0.213	0.027
" "	artificial 1	1.12	.058	0.25	0.034	0.001
" "	artificial 2	1.10	.048	0.29	0.051	0.002

Table 2 - The parameters of the best fits of the natural and related artificial texts obtained by using the logarithmic scaling laws $\langle \tilde{H}_w \rangle = A_w - B_w log_{10} T$ and $\langle \tilde{H}_f \rangle = A_f - B_f log_{10} T + C_f (log_{10} T)^2$ (Eqs. (6) and (7)).