

Entropy and Long-Range Correlations in Literary English

To cite this article: W. Ebeling and T. Pöschel 1994 *EPL* **26** 241

View the [article online](#) for updates and enhancements.

Related content

- [A New Method to Calculate Higher-Order Entropies from Finite Samples](#)
A. O. Schmitt, H. Herzel and W. Ebeling
- [Entropy of Symbolic Sequences: The Role of Correlations](#)
W. Ebeling and G. Nicolis
- [Nonlinear dynamical model for texts](#)
K. Rateitschak, W. Ebeling and J. Freund

Recent citations

- [A new universality class in corpus of texts: A statistical physics study](#)
Elham Najafi and Amir H. Darooneh
- [Quantifying Evolution of Short and Long-Range Correlations in Chinese Narrative Texts across 2000 Years](#)
Heng Chen and Haitao Liu
- [Is Natural Language a Perigraphic Process? The Theorem about Facts and Words Revisited](#)
ukasz Dbowski

Entropy and Long-Range Correlations in Literary English.

W. EBELING and T. PÖSCHEL(*)

*Institut für Theoretische Physik, Humboldt-Universität zu Berlin
Invalidenstraße 110, D-10099 Berlin, Germany*

(received 13 September 1993; accepted in final form 10 March 1994)

PACS. 05.40 – Fluctuation phenomena, random processes, and Brownian motion.

PACS. 87.10 – General, theoretical, and mathematical biophysics (inc. logic of biosystems, quantum biology and relevant aspects of thermodynamics, information theory, cybernetics, and bionics).

Abstract. – We investigated long-range correlations in two literary texts, «Moby Dyck» by H. Melville and Grimm's tales. The analysis is based on the calculation of entropylike quantities as the mutual information for pairs of letters and the entropy, the mean uncertainty, per letter. We further estimate the number of different subwords of a given length n . Filtering out the contributions due to the effects of the finite length of the texts, we find correlations ranging to a few hundred letters. Scaling laws for the mutual information (decay with a power law), for the entropy per letter (decay with the inverse square root of n) and for the word numbers (stretched exponential growth with n and with a power law of the text length) were found.

From a formal point of view a book may be considered as a linear string of letters. In this respect there exists a similarity to other linear structures [1]. Usually the strings generated by dynamical systems show only short-range correlations, except under critical conditions where, in analogy to equilibrium phase transitions [2], correlations on all scales may be observed [3]. Recently several studies on long-range correlations in DNA sequences [4] and in human writings [5] have been published. The intrinsic difficulties connected with the analysis of long-range correlations in DNA led to a controversial discussion about the authentic character of long-range structures in DNA [6].

This work is devoted to the investigation of long-range correlations in texts. We use the methods of entropy analysis, which were first applied to texts by Shannon in 1951 [7]. For several reasons we expect the existence of long-range structures in these sequences. Since a book is written in a unique style and according to a general plan of the author, we expect correlations which are ranging from the beginning of a text up to the end [8].

Another strong argument for long correlations is based on the combinatorial explosion. Uncorrelated sequences generated on an alphabet of λ letters have a manifold of λ^n different subwords of length n . A subword (block) is here any combination of letters including the space, punctuation marks and numbers. For $n > 100$ the number $N(n)$ is extremely large.

(*) Also at HLRZ, Research Center Jülich, D-52425 Jülich, Germany.

Hence we must expect that only a very small subset $N^*(n)$ of the possible words appears in a text. Bounds of this kind are given by the rules of writing texts, *i.e.* by the rules of syntax as well as by semantic relations, which do not allow for an arbitrary concatenation of letters to words and of words to sentences. The problem we address here is whether the function $N^*(n)$ follows a simple scaling law.

In earlier papers the conjecture has been made that the number of allowed subwords scales according to a stretched exponential law [3, 9]:

$$N^*(n) \sim \exp[cn^\alpha] \quad \text{with} \quad \alpha < 1, \quad c = \text{const.} \quad (1)$$

The scaling rule (1) reduces the number of the allowed subwords drastically ($N^*(n) \ll \lambda^n$) for large n . In order to describe a given string of length L using an alphabet of λ letters we introduce the following notations [3]: let $A_1 A_2 \dots A_n$ be the letters of a given substring of length $n \leq L$. Let further $p^{(n)}(A_1 \dots A_n)$ be the probability for this substring (block). A special case is the probability of finding a pair with $(n-2)$ arbitrary letters in between $p^{(n)}(A_1, A_n)$. Then we may introduce the mutual information for two letters in distance n [10, 11]:

$$I(n) = \sum_{A_i A_j} p^{(n)}(A_i, A_j) \log \left[\frac{p^{(n)}(A_i, A_j)}{p^{(1)}(A_i) p^{(1)}(A_j)} \right], \quad (2)$$

which is closely related to the autocorrelation function [4, 10-12]. Further we define the entropy per block of length n [13]:

$$H_n = - \sum p^{(n)}(A_1 \dots A_n) \log p^{(n)}(A_1 \dots A_n). \quad (3)$$

The block entropy is related to the mean number of words [3] by

$$N^*(n) \sim \lambda^{H_n}. \quad (4)$$

As shown already by Shannon, H_n/n is an important quantity expressing the structure of sequences. In [3] we assumed the scaling

$$H_n/n = h + gn^{\mu_0-1} + e/n, \quad 0 \leq \mu_0 < 1, \quad n \rightarrow \infty. \quad (5)$$

Here h , the limit of the mean uncertainty, is called the entropy of the source. This quantity is positive for stochastic as well as for chaotic processes, g and e are constants; if $h, e > 0$ and $g = 0$ the correlations in the string are short range corresponding to a Markov process with a finite memory [13]. For periodic strings one finds $h = g = 0, e > 0$. The existence of a long-range order in strings may be characterized by the condition $g > 0$ describing a slowly decaying contribution to the asymptotics of the entropy per letter for large n . Of special interest for the further consideration of texts is the case $h \ll 1, g > 0$ corresponding to a power law tail of the entropy decaying slower than $1/n$. It would be interesting and important to estimate the limit entropy h for «homogeneous texts», however, there are not enough data to do it with sufficient reliability. In the following we assume that $h \approx 0.01 \dots 0.1$. Therefore it may be neglected in our investigations which are restricted to $n < 30$.

The mutual information is not a monotonic function of n . We define long-range effects by power law tails of the averaged mutual information $I(n)$. Here the averaging is carried out over a window comprising several of the typical oscillations (fluctuations). Several authors have demonstrated that DNA sequences show a slowly decaying fluctuation at large scales [10-12, 14, 15].

We will apply the methods of entropy analysis to literary English represented by the

books «Moby Dick» by Melville ($L \approx 1170200$) and Grimm's Tales ($L \approx 1435800$). Pieces of music may be treated in a similar way [8]. For simplification we use an alphabet consisting of $\lambda = 32$ symbols: the small letters $a \dots z$, the marks , . () # and the space; # stands for any number. In order to get a better statistics we have used for the entropy calculations also a restricted alphabet consisting of only $\lambda = 3$ letters O, M, L . O codes for vowels, M for consonants and L for spaces and marks.

To estimate the mutual information we count frequencies of pairs of letters at distance n . Figure 1 shows the mutual information calculated for «Moby Dick» and for Grimm's Tales ($\lambda = 32$). The results show well-expressed correlations in the range $n = 1 \dots 25$ which are followed by a long slowly decaying tail. The obtained values for the transinformation $I(k)$ become meaningless if they are smaller than the level of the fluctuations $\delta I(k)$ due to the finite length L of the text [10, 14]:

$$\delta I(k) = \frac{\lambda^2 - 2\lambda}{2 \ln \lambda L} . \tag{6}$$

For our rather long texts with $L > 10^6$ the fluctuation level has a value of about 10^{-4} . The smoothed values for the mutual information for the range $n = 25 \dots 1000$ may be fitted by the scaling law $I(k) = c_1 n^{-0.37} + c_2$ with $c_1 = 1.5 \cdot 10^{-4}$, $c_2 = 1.1 \cdot 10^{-4}$. The constant c_2 corresponds here to the level of fluctuations. Our results show that long texts show pair correlations which decay, at least up to distances of several hundred letters, according to a power law; however, in the range $n = 10 \dots 1000$ the fluctuations are rather strong and the mean-square deviation reaches 20 ... 40%.

For the calculation of entropies we count the frequencies of subwords, where a subword of length n is defined as any combination of n letters. The results for $n = 4, 9, 16, 25$ letters in Grimm's Tales are shown in fig. 2 in a rank-ordered representation. The structure of the rank-ordered distributions is for both texts rather similar, however the list of words is of course very different.

The form of the subword distributions is distinctly not Zipf-like, it does not follow a power law. On the contrary, with increasing n there is a tendency to form a plateau [8]. The effects

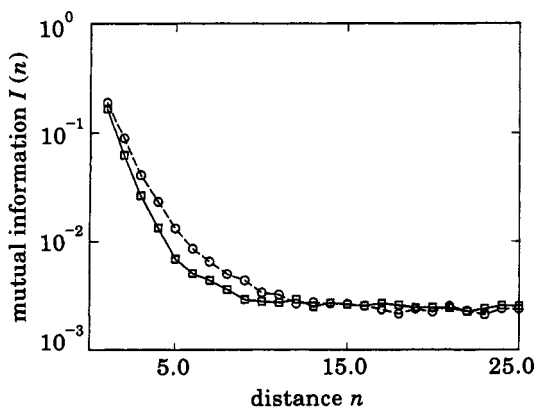


Fig. 1.

Fig. 1. - The mutual information calculated for Melville's «Moby Dick» (○) and for Grimm's Tales (□) ($\lambda = 32$).

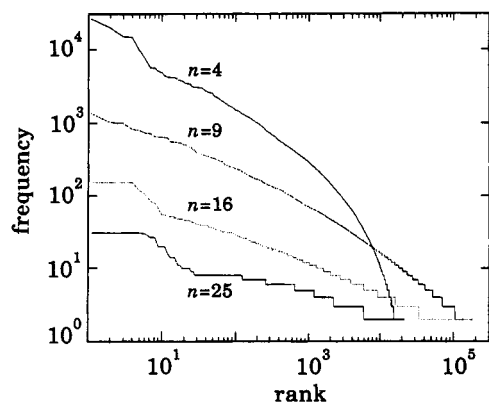


Fig. 2.

Fig. 2. - The observed rank-ordered distribution of words of length $n = 4, 9, 16, 25$ for Grimm's Tales.

due to finite n and the effects of finite length L tend to smooth the edges of the distribution [14]. The importance of length corrections for estimating the frequencies of subwords was considered by several authors [10,13]. For a deeper analysis of this problem we refer to recent articles [14,16,17]. Our method for the entropy analysis uses an extrapolation of the entropy to infinite text length [3].

The probabilities which we need for the calculation of entropies are unknown and can only be estimated from the frequencies $N_i(n)$ of the subwords of length n in a text of length L containing $N = L + 1 - n$ subwords. Introducing the observed subword frequencies into the entropy definition leads to the observed entropies

$$H_n^{\text{obs}} = \log(N) - \frac{1}{N} \sum_i N_i(n) \log(N_i(n)). \quad (7)$$

This is a random variable with the expectation value

$$H_n^{\text{exp}} = \langle H_n^{\text{obs}} \rangle = \log(N) - \frac{1}{N} \sum_i \langle N_i(n) \log(N_i(n)) \rangle. \quad (8)$$

Assuming a Bernoulli distribution for the letter combinations, the mean values can be calculated explicitly [10,14]. The result is

$$H_n^{\text{exp}} = \begin{cases} H_n - \frac{N^*(n)}{2N}, & \text{if } N^*(n) \ll N, \\ \log(N) - \log(2) \frac{N}{N^*(n)}, & \text{if } N^*(n) \gg N. \end{cases} \quad (9)$$

The latter case corresponds to the situation that only a minor part of the possible subwords of length n have a chance to appear in the text.

The relation between the effective number of words $N^*(n)$ and the block entropy H_n is given by eq. (4). Hence the expected block entropy may be represented as a function of $\log N$ with one free parameter H_n which is found by fitting the curves. In this way the block entropies for both books were calculated up to $n = 26$. For small word length, *i.e.* $n \leq 16$ for $\lambda = 3$ and $n \leq 5$ for $\lambda = 32$, we used the approximation (9) for $N^*(n) \ll N$. For larger n , *i.e.* $n \geq 20$ for $\lambda = 3$ and $n \geq 10$ for $\lambda = 32$, we applied the approximation (9) valid for $N^*(n) \gg N$. More concretely, we measured the deviation between the observed entropy and $\log(N)$. Then the theoretical entropy H_n was estimated from $N^*(n)$ using eq. (4). In the intermediate region we applied a smooth Padé approximation between both formulae. In a procedure of successive approximations the entropy H_n was considered as a free parameter which was fitted in a way that $H_n^{\text{exp}}(\log N)$ came as close as possible to the measured (observed) entropy values. In practice this method breaks down for $n \geq 30$ if $\lambda = 3$ and for $n \geq 25$ if $\lambda = 32$. Longer subwords do not have a chance to appear several times in the text, which leads to large statistical errors.

The calculations for $n \leq 26$ show that the square-root law yields a reasonable approximation for the scaling of the entropy per letter with the word length n

$$\begin{cases} H_n/(n \log(\lambda)) \approx (4.84/\sqrt{n}) - (7.57/n) & (\lambda = 3), \\ H_n/(n \log(\lambda)) \approx (0.9/\sqrt{n}) + (1.7/n) & (\lambda = 32). \end{cases} \quad (10)$$

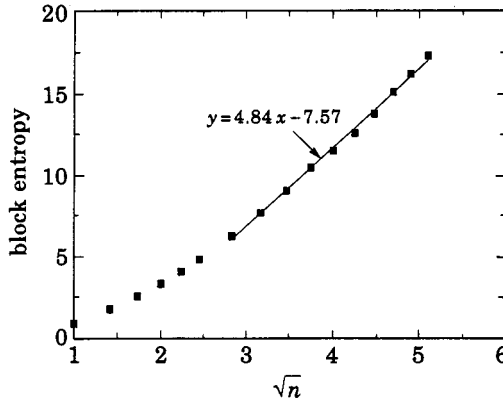


Fig. 3. – The scaling behaviour of the block entropy H_n with the square root of the word length n for «Moby Dick» encoded by the alphabet $\lambda = 3$.

Figure 3 shows the fit for the alphabet $\lambda = 3$. The scaling law of the square-root type was first found by Hilberg [9] by fitting Shannon’s original data. For $n = 100$ and $\lambda = 32$ our scaling formula yields $H_{100} \approx 10 \log(\lambda)$, which is not far from Shannon’s estimation $H_{100} \approx 40$ bits.

The number of subwords increases according to a stressed exponential law. For the growth we found the approximation

$$N_n^* \approx 2^{7.45\sqrt{n}-11.3} \quad (\lambda = 3), \tag{11}$$

$$N_n^* \approx 2^{4.5\sqrt{n}+8.5} \quad (\lambda = 32). \tag{12}$$

We summarize now the results obtained for the two books: the scaling of the mutual information and the entropy per letter show, in agreement with earlier work [3], that long texts are neither periodic nor chaotic but somehow in between. Taking into account length corrections we calculated block entropies up to $n = 26$ and mutual-information values up to distances of a few hundred letters. Based on these data we formulated a hypothesis about the long-range scaling. For the range $n \geq 100$, the pair correlations contained in the transformation of long texts, $L > 10^6$, decay according to a power law, however the level of fluctuations is rather high. A reliable estimation of the block entropies for $n > 30$ is still an open question. The results for the entropy of the two books suggest in agreement with Shannon’s data and Hilberg’s findings that the mean entropy per letter decays to its limit according to a square-root law. As a consequence, the number of different subwords in texts increases with the number of letters n according to a stretched exponential law. Our estimations for the growth yield for $n = 100$ a total number of about 2^{53} different subwords. Most of the subwords which would be possible from the combinatorial point of view are actually forbidden and do not appear in real texts. We also investigated how the number of genuine English words $N(L)$ (formally defined here as sequences of letters between spaces and/or marks) increases with the length L of a text. For Grimm’s Tales we found the scaling law $N(L) = 22.8L^{0.46}$, *i.e.* reading the book we find permanently new words.

More empirical data on long texts and further studies of the statistical effects due to the finite length of the samples are needed in order to reach a more definite conclusion about the scaling properties.

* * *

The authors thank K. ALBRECHT, T. BOSENIUK, J. FREUND, H. HERZEL, G. NICOLIS and A. SCHMITT for fruitful discussions. Further we thank the *Project Gutenberg Etext* at Illinois Benedictine College for providing the ASCII files of the investigated literature.

REFERENCES

- [1] EBELING W. and NICOLIS G., *Europhys. Lett.*, **14** (1990) 191.
- [2] STANLEY H. E., *Introduction to Phase Transitions and Critical Phenomena* (Oxford University Press, New York, Oxford) 1971.
- [3] EBELING W. and NICOLIS G., *Chaos, Solitons & Fractals*, **2** (1992) 635.
- [4] PENG C. K., BULDYREV S. V., GOLDBERGER A. L., HAVLIN S., SCIORTINO F., SIMONS M. and STANLEY H. E., *Nature*, **356** (1992) 168.
- [5] SCHENKEL A., ZHANG J. and ZHANG Y., *Fractals*, **1** (1993) 47.
- [6] LI W. and KANEKO K., *Europhys. Lett.*, **17** (1992) 655.
- [7] SHANNON C. E., *Bell Syst. Tech. J.*, **30** (1951) 50.
- [8] EBELING W., PÖSCHEL T. and ALBRECHT K. F., submitted to *Int. J. Bifurcation & Chaos*.
- [9] HILBERG W., *Frequenz*, **44** (1990) 243.
- [10] HERZEL H., *Syst. Anal. Model. Simul.*, **5** (1988) 453.
- [11] LI W., *J. Stat. Phys.*, **60** (1990) 823.
- [12] NICOLIS J. S. and KATSIKAS A. A., in *Studies in Nonlinearity in Life Sciences*, edited by B. WEST (World Scientific, Singapore) 1992.
- [13] GRASSBERGER P., *Phys. Lett. A*, **128** (1989) 369; *IEEE Trans. Inf. Theory*, **35** (1989) 669.
- [14] HERZEL H., SCHMITT A. and EBELING W., to be published in *Chaos, Solitons & Fractals* (1993).
- [15] VOSS R. F., *Phys. Rev. Lett.*, **68** (1992) 3805.
- [16] SCHMITT A., HERZEL H. and EBELING W., *Europhys. Lett.*, **23** (1993) 303.
- [17] PENG C. K., BULDYREV S. V., GOLDBERGER A. L., HAVLIN S., SIMONS M. and STANLEY H. E., *Phys. Rev. E*, **47** (1993) 3730.