A Comparison of natural (english) and artificial (esperanto) languages. A Multifractal method based analysis

J. Gillet∗ and M. Ausloos†
GRAPES, B5a Sart-Tilman, B-4000 Liège, Belgium
(Dated: February 3, 2008)

We present a comparison of two english texts, written by Lewis Carroll, one (Alice in wonderland) and the other (Through a looking glass), the former translated into esperanto, in order to observe whether natural and artificial languages significantly differ from each other. We construct one dimensional time series like signals using either word lengths or word frequencies. We use the multifractal ideas for sorting out correlations in the writings. In order to check the robustness of the methods we also write (!) (consider?) the corresponding shuffled texts. We compare characteristic functions and e.g. observe marked differences in the (far from parabolic) \( f(\alpha) \) curves, differences which we attribute to Tsallis non extensive statistical features in the frequency time series and length time series. The esperanto text has more extreme values. A very rough approximation consists in modeling the texts as a random Cantor set if resulting from a binomial cascade of long and short words (or words and blanks). This leads to parameters characterizing the text style, and most likely in fine the author writings.

PACS numbers:

As soon as modern fractals appeared in order to describe physical objects, it was evident that some generalization was in order: multifractals spurred up, e.g., since obviously a fractal dimension \( D \) is not enough to describe an object [1] [2]. The more so in non equilibrium systems, characterized by some unusual dynamics. Through a generator and from an initiator one can produce a fractal object with a given dimension. How to produce realistic and meaningful multifractal models is a challenge. Do they really exist [3]? Do multifractal model exist nowadays [4]? These questions come in parallel with the measurement of the fractal dimension, ... and its distribution. One question of interest is whether the apparently multifractal nature of an object is due to its finite size or to a complex dynamical feature or something else! Some attempt in this direction results from observation of multifractal features in meteorology and climate studies [5] [6] [7], but also in many other fields [8], like mathematical finance [4] [9] [10] [11] [12] [13].

Let us recall that one has basically to obtain a \( D(q) \) function or the \( f(\alpha) \) spectrum, where \( q \) represents the degree of some moment distribution of some variable, and \( \alpha \) is some sort of critical exponent at phase transitions, also called the Holder exponent: \( f(\alpha) \) being its distribution.

There is a need for experimental work leading to reliable \( D(q) \) and \( f(\alpha) \) data, before modeling. Interesting pioneering data should be here recalled : see work on DLA [14] [15], DNA [16] [17], SOI [18], NAO [19], .... It appears that most of the time some “signal is either directly a time series or is transformed into a time series; more generally, the signal is called a text, because it can be decomposed through level thresholds which can be thought to be a set of characters taken from an alphabet. Here below we take real texts in fact as the source of experimental observations and follow the multifractal ideas to make an analysis of such texts. The main question concerns whether multifractals are indeed found in real texts; a question raised in [20]; another is whether the technique of analysis can give some insight on a logical construction [21], from which stems the possible connection of such ideas with coding, transcription, machine translation, social distances, network properties of languages, ... [21] [22] [23] and more classically in physics about identifying coherent structures in spatially extended systems [24]. We examine two classical english texts but also one translation, into an unusual language, and the corresponding shuffled texts. We focus on how local/structural properties develop into global ones.

Since Shannon [25], writings and codings are of interest in statistical physics. Writings are systems practically composed of a large number of internal components (the words, signs, and blanks in printed texts). In terms of complexity investigations, writings which are a form of recorded languages, like living systems, belong to the top level class involving highly optimized tolerance design, error thresholds in optimal coding, and financial markets [27]. Relevant questions pertain to the life time, concentration, distribution, .. complexity of these. One should distinguish two main frameworks. On one hand, language developments seem to be understandable through competitions, like in Ising models, and in self-organized systems. Their diffusion seems similar to percolation and nucleation-growth problems taking into account the existence of different time scales, for inter- and intra- effects; this is the realm of anthropology. The second frame originates from more classical linguistics studies; it pertains to the content and meanings of words and texts. Concerning the internal structure of a text, supposedly char-
acterized by the language in which it is written, it is well known that a text can be mapped into a signal, of course through the alphabet characters. However it can be also reduced to less abundant symbols through some threshold, like a time series, which can be a list of +1 and -1, or sometimes 0. In fact, laws of text content and structures have been searched for a long time ago by Zipf and others, see many refs. in [26]. Through a looking glass (TLG) [39]; - to our knowledge few comparisons have been presented about letter distribution can be found in [20].

Interestingly writings can be thought as social networks [23]. Social networks have fractal properties [28]: most usually they should be multifractals; one can thus imagine/consider that a text which is a form of partially self-organized social network (for words) due to grammatical and style constraints present multifractal features. The properties of such texts taken as signals have already been examined, e.g., a multifractal analysis of Moby Dick letter distribution can be found in [20].

Even though we recognize such a pioneering paper, we stress that sentences made of words, not letters, are translated. Thus we present below an original consideration in this respect, i.e. the analysis and results about a translation between one of the most commonly used language, i.e. english, and a relatively recent language, i.e. esperanto. Esperanto is an artificially constructed language [29], which was intended to be an easy-to-learn lingua franca. Statistical analyses seem to indicate that esperanto’s statistical proportions are similar to those of other languages [30]. It was found that esperanto’s statistical proportions resemble mostly those of German and Spanish, and somewhat surprisingly least those of French and Italian. English seems to be the intermediary case.

Comparison of different languages (writings) arising from apparently different origins or containing different signs, e.g. greek [31], turkish [32], chinese [33],... even somewhat artificial languages, like those used for simulation codes on computers [34] has also been made. To our knowledge few comparisons have been presented about written texts translated from one to another language [35] [40] and in particular from the point of view of LROC in words.

The text used here was chosen for its wide diffusion, freely available from the web [37] and as a representative one of a famous scientist, Lewis Caroll, i.e. Alice in wonderland (AWL) [38]. Moreover knowing the special (mathematical) quality of this author mind, and some, as we thought a priori, some possibly special way of writing, another text has been chosen for comparison, i.e. Through a looking glass (TLG) [39]; - to our knowledge only available in english (on the web). This will allow us to discuss whether the differences, if any, between esperanto and english, are apparently due to the translation or to the specificity of this author work. Previous work on the english AWL version (AWLeng) should be mentioned [40], but pertains to a mere Zipf analysis.

In Sect. [1] we present some elementary facts on these texts and briefly expose the methodology, i.e. we emphasize that we distinguish between “frequency time series (FTS) and “length time series (LTS). We recall the multifractal technique for this specific application. In order to check the robustness of the method we also invent (write !) (or consider) the corresponding shuffled texts, to which we apply the same technique of analysis. Therefore, in Sect. [1] we present the somewhat unusual results, and discuss them in Sect. [1] For the simplicity of the discussion we very roughly approximate the texts as resulting from a binomial cascade of short and long words. We obtain parameters in fine characterizing the text style and the author’s writings. We observe that such multifractals have a deep connection to Tsallis non extensive statistics as pointed out in ref. [41] in another framework.

I. DATA AND METHODOLOGY

For our empirical considerations we have selected the two texts here above mentioned downloading them from a freely available site [37], resulting obviously into three files, called. Next, we have removed the chapter heads. All our analyses are carried over this reorganized file for each text. Thereafter, we have shuffled these texts, in the files, without taking into account the punctuation [42].

There are two ways to construct a time series from such documents

1. Take a document of N words. Select all the different words. Count the frequency f of appearance of each word in the document. The time of “appearance is played by the rank position of the word in the file. We map the word frequency to a time series f(t). Such a time series is called the frequency time series (FTS).

2. Take a document of N words. Consider the length l (number of letters) of each word. Record where each word of length l is located in the text; the time is played by the position of the word in the document, i.e. the first word is considered to be emitted at time t= 1, the second at time t = 2, etc. A time series l(t) is so constructed. We refer to such a time series as the length time series (LTS).

Obviously there is a large number of ways to map a text to a time series, but in the present study we only consider the above two since some physical meaning can be thought to arise in the mapping. As indicated in [43] the length of the word is associated with speaker effort, meaning that the longer the word the higher the effort required to pronounce it. The frequency of the word is also associated with the hearer effort as frequently used words require less effort to be understood by the hearer.
Generalized fractal dimension $D(q)$

FIG. 1: $D(q)$ for (a) FTS (b) LTS of the original texts

Generalized critical exponent $f(\alpha)$

FIG. 3: $f(\alpha)$ for (a) FTS (b) LTS of the indicated texts

Generalized fractal dimension $D(q)$

FIG. 2: $D(q)$ for (a) FTS (b) LTS of the shuffled indicated texts

Generalized critical exponent $f(\alpha)$

FIG. 4: $f(\alpha)$ for (a) FTS (b) LTS of the shuffled indicated texts
II. RESULTS

The results of the FTS and LTS multifractal analysis for the three main texts and their shuffled corresponding ones are shown in Figs. 1-4 (a-b).

A mere perusing of the graphs indicate that the multifractal approach is in good order, e.g. since $D(q)$ is not a single point, and should allow one to observe interesting LRO correlations and local ones.

A. $D(q)$ plots: Figs. 1-2

In FTS, the generalized fractal dimension has a similar set of values for both english texts, decaying from ca. 1.2 to 1.0 for $q$ increasing but negative; $D(q)$ decays slowly for $q$ positive, barely reaching a value 0.95 for $q = 80$ (Fig. 1). The value of $D(q)$ is much greater along the negative $q$ axis, for AWLesp but is identical to the other two for $q > 0$. In LTS, even though the form of $D(q)$ is that to be expected, it has to be stressed that the AWLeng and AWL esp are very quantitatively similar, but markedly differ from TLGeng. This already indicates that one can observe the high creativity of the author through these two books. Moreover the translation effect on style is observe the high creativity of the author through these two books. Moreover the translation effect on style is much better seen on FTS than LTS.

The shuffled texts (Fig.2) remarkably have the same $D(q)$ values; their range and variations being similar to those of the real texts. Slight quantitative differences occur, more markedly for the shuffled AWL esp, but along a Baaysian reasoning these can be attributed to the finite size of the sample.

By the way,

$$C_1 = \left. \frac{d\tau(q)}{dq} \right|_{q=1}$$

is a measure of the intermittency lying in the signal $y(n)$; it can be numerically estimated by measuring $\tau_q$ around $q = 1$. In all cases the value of $C_1$ is close to unity. Some conjecture on the role/meaning of $C_1$ is found in Ref. [45]. From some financial and political data analysis it seems that $H_1$ is a measure of the information entropy of the system. The same can be thought of here.

B. $f(\alpha)$ plots: Figs. 3-4

The $f(\alpha)$ spectra are shown in Figs. 3-4. They are markedly non symmetric, as was found for DLA [13] [15], with very high positive skewness, i.e. for $q \leq 0$. Interestingly, the esperanto text curve behaves differently from the english texts, in FTS, though TLGeng is different in the LTS case; the shuffled texts $f(\alpha)$ spectra behaving in a very similar qualitative and quantitative way. However the shuffling does not fully symmetrize the spectra.

It is interesting to observe that the $f(\alpha)$ curve is very sharp: it originates from negative values for $\alpha$ less than...
1.0; reaches a maximum (=1.0) at 1.0, at the maximum so called box dimension, and decays rapidly for \( \alpha \) positive; \( f(\alpha) = 0 \) at \( \alpha = 1.2 \) and 1.3 respectively for AWLeng and TLGeng; the maximum is also reached at (1.0, 1.0) and the spectrum spans the (narrow) interval 0.90: 1.25 for AWLesp on the \( f(\alpha) =0 \) line. It is worth noticing that the values are reasonable in view of their correspondence to the fractal dimensions. On the other hand, the sharpness indicates a high lack of uniformity of the texts LROC.

### III. DISCUSSION WITH CONCLUSION

In summary, one can observe similarities between the original and shuffled texts and their translations; see Table I for summarizing the similarities seen through \( D(q) \) and \( f(\alpha) \). The English texts look more similar with each other than with respect to the esperanto translation. On the other hand, one physical conclusion arises from the above: the existence of a multifractal spectrum found for the examined texts indicates a multiplicative process in the usual statistical sense for the distribution of words length and frequency in the text considered as a time series. Thus linguistic signals may be considered indeed as the manifestation of a complex system of high dimensionality, different from random signals or systems of low dimensionality such as the financial and geophysical (climate) signals. Finally, the \( f(\alpha) \) curve represents the measurable aspects of the word networks, be they considered through LTS or FTS. Our work confirms that texts could be seen as networks indeed [23].

Before suggesting a physical model describing the construction of a writing, let us consider implications from the \( f(\alpha) \) spectrum in some detail. The not fully parabolic, to say the least, \( f(\alpha) \) curve indicates non uniformity and strong LROC between long words and small words, evidently arising from strong short range correlations between these. In some sense this is expected for classical writings. It is usually known that the left (right) hand side of the \( f(\alpha) \) curve corresponds to fluctuations of the \( q \geq 0 \) \( (q \leq 0) \)-correlation function. In other words, they correspond to fluctuations in small (large) word distributions. Therefore the distribution of small and long words should be examined in further work in order to observe these local correlations, e.g. through a detrended fluctuation analysis.

Moreover, in order to characterize the writings, texts and/or authors, we propose a very rough approximation/model, i.e. let us consider (assume!) that the writings are made of only two types of words: small and large [16, 17], appearing through some recursive process. In so doing one can consider the behavior of the atypical \( f(\alpha) \) curve as originating from a binomial cascade of short and long words, on a support [0,1], with an arbitrary contraction ratio \( r_i \) and a weight \( w_i \) for the word in each successive subinterval, as for a multifractal Cantor set construction [11]. For an arbitrary number \( n \) of subintervals the generalized fractal dimension (or rather \( \tau(q) \)) is obtained from

\[
\sum_{i=1}^{n} w_i^q r_i^{-\tau} = 1.
\] (9)

The formula is easily generalized for random contraction ratios and weights. However it simplifies for the case of a simple binomial cascade, i.e.

\[
w_i^q r_i^{-\tau} + w_i^q r_2^{-\tau} = 1.
\] (10)

Whence the extremal \( \alpha \) values read \( \alpha_- = \log(w_2)/\log(r_2) \) and \( \alpha_+ = \log(w_1)/\log(r_1) \), from which the weights and ratios can be estimated by inversion (table I), thereby suggesting the author’s somewhat systemic way used in his/her writings.

The physics connection can be obtained if one relates the \( f(\alpha) \) curve extremal points through their physical meanings [48], i.e.

\[
\frac{1}{1-Q} = \frac{1}{\alpha_+} - \frac{1}{\alpha_-}
\] (11)

where \( \alpha_- \) and \( \alpha_+ \) are the extremes of the range of support for the (positive) multifractal spectrum \( f(\alpha) \) and \( Q \).
(instead of the usual $q$) is used to represent the parameter arising in the non extensive description of statistical physics \cite{[11]}; see values in table \cite{[11]}. By extension it is a measure of the attractor dimension or the number of so called degrees of freedom. It is obvious from the table that $Q$ varies between 4 and 7, with interesting differences between the LTS and FTS cases, LTS’s $Q$ being systematically smaller, in the original or shuffled texts. Notice that the value of $Q$ is more extreme though with the same order of magnitude in the case of the esperanto text for both types of series.

Finally we re-emphasize the remarkable difference for the esperanto text (Fig. 3a) with the english texts in the FTS analysis. Linguistics input should be searched at this level and is left for further discussion. The origin of differences between TLG and AWL needs more work at the linguistic level. However we have indicated the interest of the multifractal scheme in providing a measure of these correlations, thus a new measure of an author's style. This suggests a (binomial, at first) cascade model containing parameters characterizing (or reflecting, at least) the text style, and most likely in fine the author writings. It remains to be seen whether the $f(\alpha)$ curve and the (to be generalized) binomial cascade model, with the weight and ratio parameters hold through in other cases, and can characterize authors and texts, - and in general time series. Moreover the multifractal method should additionally be able to distinguish a natural language signal from a computer code signal \cite{[11]} and help in improving translations by suggesting perfection criteria and indicators of text qualitative values.

Acknowledgements The authors would like to thank D. Stauffer for as usual fruitful discussions .... This work has been supported by European Commission Project CREEN FP6-2003-NEST-Path-012864.

\[1\] T.C. Halsey, M.H. Jensen, L.P.Kadanoff, I. Procaccia, and B.I. Shraiman, Fractal measures of strange attractors, Phys. Rev. A 33 (1986) 1141-1151.
\[2\] B.J. West and W. Deering, The lure of modern science : fractal thinking, (World Sci., River Edge, NJ, 1995).
\[3\] Th. Lux and M. Ausloos, Market Fluctuations I: Scaling, Multi-scaling and their Possible Origins, In The Science of Disaster : Scaling Laws Governing Weather, Body, Stock-Market Dynamics, A.Bunde, J. Kropp, and H.-J.Schellnhuber, Eds. (Springer Verlag, Berlin, 2002) pp. 377-413.
\[4\] Th. Lux , The Markov-Switching Multifractal Model of Asset Returns: Estimation via GMM Estimation and Linear Forecasting of Volatility, Journal of Business and Economic Statistics (in press).
\[5\] K. Ivanova, N. Gospodinova, H.N. Shirer, T.P. Ackerman, M.A. Mikhailov, M. Ausloos, Multifractality of Cloud Base Height Profiles, cond-mat/0108395.
\[6\] K. Ivanova, H.N. Shirer, E.E. Clothiaux, N. Kitova, M.A. Mikhailov, T.P. Ackerman, and M. Ausloos, A case study of stratus cloud base height multifractal fluctuations, Physica A 308 (2002) 518-532.
\[7\] R.G. Kavasseri and R. Nagarajan, A multifractal description of wind speed records, Chaos, Solitons and Fractals, 24 (2005) 165-173.
\[8\] J.W. Kantelhardt, S.A. Zschiegner, E. Koscielny-Bunde, A. Bunde, S. Havlin and H.E. Stanley, Multifractal detrended fluctuation analysis of nonstationary time series, Physica A 316, 87 - 114 (2002).
\[9\] M. Ausloos and K. Ivanova, Multifractal nature of stock exchange prices, Comp. Phys. Commun. 147 (2002) 582-585.
\[10\] J. Alvarez-Ramirez, M. Cisneros, C. Ibarra-Valdez and A. Soriano, Multifractal Hurst analysis of crude oil prices,Physica A 313 (2002) 651-670.
\[11\] P Oswiecimka, J Kwapien, S Drozdz, Wavelet versus detrended fluctuation analysis of multifractal structures, Phys. Rev. E 74, 016103 (2006) (17 pages).
\[12\] E. Canessa, Multifractality in time series, J. Phys. A: Math. Gen. 33 (2000) 3637-3651.
\[13\] K. Ivanova and M. Ausloos, Low q-moment multifractal analysis of Gold price, Dow Jones Industrial Average and BGL-USD exchange rate, Eur. Phys. J. B 8 (1999) 665-669; Err. 12, 613 (1999).
\[14\] S. Schwarzer, J. Lee, A. Bunde, S. Havlin, H. E. Roman and H. E. Stanley, Minimum growth probability of diffusion-limited aggregates , Phys. Rev. Lett. 65, 603-606 (1990).
\[15\] J. Lee, S. Havlin, H.E. Stanley and J. E. Kiefer, Hierarchical model for the multifractality of diffusion-limited aggregation , Phys. Rev. A 42 (1990) 4832-4837.
\[16\] C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, M. Simons and H. E. Stanley, Statistical properties of DNA sequences, Physica A 221 (1995) 180-192.
\[17\] A. Rosas, E. Nogueira, and J. F. Fontanari, Multifractal analysis of DNA walks and trails, Phys. Rev. E 66, 061906 (2002) [6 pages].
\[18\] M. Ausloos and F. Petroni, Tsallis non-extensive statistical mechanics of El Nino southern oscillation index, Physica A 373 (2007) 721 - 736.
\[19\] C. Collette and M. Ausloos, Scaling Analysis and Evolution Equation of the North Atlantic Oscillation Index Fluctuations, Int. J. Mod. Phys. C 15, 1353 - 1366 (2005).
\[20\] A.N. Pavlov, W. Ebeling, L. Molgedey, A. R. Ziganshin and V. S. Anishenko, Scaling features of texts, images and text series, Physica A 300 (2001) 310-324.
\[21\] David B. Saakian, Error threshold in optimal coding, numerical criteria, and classes of universalities for complexity, Phys. Rev. E 71, 016126 (2005).
\[22\] V Zlatic, M Bozicevic, H Stefancic, M Domazet, Wikipedias: Collaborative web-based encyclopedias as complex networks, Phys. Rev. E 74, 016115 (2006).
\[23\] A. P. Masucci and G. J. Rodgers, Network properties of written human language, Phys. Rev. E 74, 026102 (2006) (8 pages).
\[24\] Cosma Rohilla Shalizi, Robert Haslinger, Jean-Baptiste Rouquier, Kristina Lisa Klinkner, and Cristopher Moore, Automatic filters for the detection of coherent struc-
ture in spatiotemporal systems, Phys. Rev. E 73, 036104 (2006) (16 pages)

[25] C. Shannon, A mathematical theory of communications, Bell. Syst. Tech. J. 27, pp. 379-423, 623-656 (1948); see also Prediction and entropy of printed English, Bell Syst. Tech. J. 30 (1951) 50-64.

[26] J. Gillet and M. Ausloos, A Comparison of natural (english) and artificial (esperanto) languages. I. Zipf and Grasseberg Procaccia method based analysis, submitted

[27] M. Amit, Y. Shmerler, E. Eisenberg, M. Abraham and N. Shnerb, Language and Codification Dependence of Long-Range Correlations in Texts, Fractals 2 (1994) 7 - 13

[28] Marin Boguna, Romualdo Pastor-Satorras, Albert Daz-Guila, and Alex Arenas, Models of social networks based on social distance attachment, Phys. Rev. E 70, 056122 (2004)

[29] M. Boulton, Zamenhof, Creator of Esperanto. London: Routledge, Kegan & Paul. (1960).

[30] Bill Manaris, Luca Pellicoro, George Pothering, Harland Hodges, Investigating Esperanto’s statistical proportions relative to other languages using neural networks and Zipf’s law, Proceedings of the 24th IASTED (International Association Of Science And Technology For Development ) international conference on Artificial intelligence and applications, held in Innsbruck, Austria, pp. 102 – 108 (ACTA Press, Anaheim, CA, USA, 2006)

[31] N. Hatzigeorgiou, G. Mikros, G. Carayannis, Word Length, Word Frequencies and Zipf’s law in the greek language, J. Quant. Ling. 8 (2001) 175-185; George Mikros, Nick Hatzigeorgiou, George Carayannis, Basic Quantitative Characteristics of the Modern Greek Language Using the Hellenic National Corpus, Journal of Quantitative Linguistics, 12 (2005) 167 - 184

[32] G. Dalkilic and Y. Cebi, Zipf’s Law and Mandelbrot’s Constants for Turkish Language Using Turkish Corpus, Lect. Notes Comput. Sci. 3621 (2004) 273-282.

[33] R Rousseau, Qiaoqiao Zhang , Zipf’s data on the frequency of Chinese words revisited, Scientometrics, 24 (1992) 201-220.

[34] M Zemankova, CM Eastman -, Comparative lexical analysis of FORTRAN code, code comments and English text, ACM Southeast Regional Conference Proceedings of the 18th annual Southeast regional conference, Tallahassee, Florida, 193 - 197 (1980)

[35] When finalizing the writing of this paper we became aware of related work of interest [36].

[36] D.R. Amancio, L. Antiqueira. T.A.S. Pardo, L. da F. Costa, O.N. Oliveira Jr., M. G. V. Nunes, Complex networks analysis of manual and machine translations. Translations between English, Spanish and Portuguese are used to find structural differences between human and computerized translations. Int. J. Mod. Phys. C 19 (2008)

[37] Project Gutenberg (National Clearinghouse for Machine Readable Texts) (one mirror site is at UIUC ). Project Gutenberg (2005), accessed Sep. 25, 2005. gutenberg.org

[38] L.W. Carroll, Alice’s Adventures in Wonderland, Macmillan (1865); http://www.gutenberg.org/etext/11

[39] L.W. Carroll, Through the Looking Glass and What Alice Found There, Macmillan (1871) ; http://www.gutenberg.org/etext/12

[40] D.M.W. Powers, Applications and explanations of Zipf’s laws, in New Methods in language Processing and Computational natural language learning (ACL, XXX..., 1998)pp. 151-160.

[41] A. Erzan and J.P. Eckmann, q-analysis of Fractal Sets, Phys. Rev. Lett. 78 (1997) 3245-3248.

[42] The shuffle algorithm is found on wikipedia. The first “data point is exchanged with a following one, its location chosen from a generated random number. The second data point is exchanged with a following one, chosen from another random number, etc. The random number generator was check to lead to a rather uniform distribution, for a number between 0 and 1. The algorithm was applied ten times on the texts to get the final shuffled texts we used.

[43] K. Kosmidis, A. Kalampokis, P. Argyrakis, Language time series, Physica A 370 (2006) 808-816

[44] A. Chhabra, R.V. Jensen, Direct determination of the $f(\alpha)$ singularity spectrum, Phys Rev Lett 62 (1989) 1327 - 1330.

[45] N. Vandewalle and M. Ausloos, Coherent and random sequences in financial fluctuations, Physica A 246, 454-(1997)

[46] We apologize to all authors for this simplification of their work. Yet, in some sense this is equivalent to considering that the (+/-1) Ising model describes all magnetic features. Notice that this assumption on writings has been successfully made in the study of blogs, where words occurring with an equivalent frequency are considered to be identical for studying the blogs statistical properties.

[47] R. Lambiotte, M. Ausloos, M. Thelwall, Word statistics in Blogs and RSS feeds: Towards empirical evidence, J. Informetrics 1 (2007) 277 - 286.

[48] T. Arimitsu and N. Arimitsu, Analysis of turbulence by statistics based on generalized entropies, Physica A 295 (2001) 177.

[49] C. Tsallis, Nonextensive statistics: theoretical, experimental and computational evidences and connections, Braz. J. Phys. 29 (1999) 1.