Possible Chaotic Structures in the Turkish Language with Time Series Analysis

Avadis Hacınlıyan[1], Murat Erentürk[1], and Gökhan Şahin[1]

[1]Department of Physics and Department of Information Technologies,
Yeditepe University, İstanbul, Turkey
{ahacinliyan, merenturk, sahin}@yeditepe.edu.tr

Abstract *— The possibility of chaotic structures in Turkish and English texts, as well as the possibility of using the pseudo-invariants in a reconstructed phase space as identifying characteristics for languages is investigated. Texts of length up to 83000 in both languages have been analysed. Two alternatives for the dependent variable in a time series analysis have been used. Word frequencies based on a corpus have been one alternative inspired by Zipf's law. The other alternative is based on assigning values to the letters in a word as inspired by a random walk. A positive maximal lyapunov exponent has been observed. Values of this exponent are different for the two languages. This and differing detrended fluctuation analysis results for the two languages for either parametrization imply that our analysis methods can point to differences in languages.*

1. INTRODUCTION

The structure of natural languages has recently become an important field of research following the observation that texts written in natural languages obey laws that approximately obey rules of fractal geometries[1,2]. This behavior is characteristic of many other systems in nature including many forms of music. On the other hand, time series analysis methods[3,4] have become an import tool in analysing fractal structures using a one dimensional signal in time. Unfortunately, there are several ways in which one can generate a one dimensional time signal based on a literary text.

A natural language is a hierarchy of structures that involve both a sound and a meaning. The simplest structures consist of the letters of the alphabet and the syllables; they will have a contribution to the sound but not to the meaning. Structures higher in the hierarchy such as words, sentences and paragraphs will contribute to the meaning. In order to be an effective communication medium, a language must be patterned. It is expected that the patterns will involve self similarity[2,5] and there should be a certain characteristic distance or distances between words that significantly contribute to the meaning[6]. It would be tempting to see time series analysis can give us an estimate on both the possibility of self similarity and the existence of such a window, since chaotic behaviour is related to the long time predictability of a system .

The first issue involved in an attempt to analyze a natural language as a time series is constructing a time series from a text. A meaningful dependent variable is needed for the time series analysis. Two different dependent variables have been used in this work: i) The frequencies derived from a corpus and ii) a variable derived from values assigned to the letters constituting a given word. The resultant time series are analised via nonlinear time series analysis and detrended fluctuation analysis (DFA) [6].

DFA has become an important tool in analysing long-term correaltions in nonstationary time series[6,7] such as organization of DNA nucleotides, heartbeat time series, long-range weather forecasting, economical time series and solid-state dynamics [8-19]. DFA is reported to have advantages over conventional methods (e.g.,spectral analysis and Hurst analysis). It permits the detection of intrinsic self-similarity embedded in a seemingly non stationary time series, and also avoids the spurious detection of apparent self-similarity, which may be an artifact of extrinsic trends [20-24].

2. TIME SERIES ANALYSIS

Time series analysis of a one dimensional signal in time requires a phase space reconstruction and analysis. Sequences of dependent values are investigated via nonlinear time series analysis as described in references 3 and 4 using the TISEAN software package[3].

Details of the phase space reconstruction from the scalar $s(k)$, where k means the k^{th} time step, follows the well known procedure. Details will be only given to the extent needed to set the notation. Time delay vectors $\vec{y}(k)$ given by

$$\vec{y}(k) = [s(k), s(k+\tau), \ldots, s(k+(d-1)\tau)] \quad \vec{y} \in R^d \tag{1}$$

where τ denotes the delay time and d denotes the embedding dimension are constructed. There are no clear cut rules for their determination since a limited range of data is available and noise is present. The time delay is found from [10] the first zero of the linear autocorrelation function given by

$$C_1(\tau) = \frac{\frac{1}{N}\sum_{m=1}^{N}[s(m+\tau)-\bar{s}][s(m)-\bar{s}]}{\frac{1}{N}\sum_{m=1}^{N}[s(m)-\bar{s}]^2} \tag{2}$$

where

$$\bar{s} = \frac{1}{N}\sum_{m=1}^{N} s(m) \tag{3}$$

Another method for the determination of the delay time is to find the first minimum of the average mutual information. This can be used as if it were a nonlinear correlation function given by [13],

$$I(\tau) = I_{AB} = \sum_{a_i,b_j} P(s(n+\tau), s(n)) \log_2 \left[\frac{P(s(n), s(n+\tau))}{P(s(n+\tau))P(s(n))} \right] \tag{4}$$

Here $P(s(n),s(n+\tau))$ is the joint probability that if at time $ns(n)$ is measured; then at time $n+\tau$, $s(n+\tau)$ is measured and $P(s(n))$ is the probability of measuring $s(n)$ [10 and 14].

Since we are interested only in detecting the presence of chaotic behavior determining the maximal Liapunov exponent is sufficient. This is calculated by the formula

$$S(\Delta n) = \frac{1}{N} \sum_{n_0=1}^{N} \ln \left(\frac{1}{|u_n(\vec{s}_{\,0})|} \sum_{\vec{s}_{\,n} \in u(\vec{s}_{\,n_0})} |\vec{s}_{\,n_0+\Delta n} - \vec{s}_{\,n+\Delta n}| \right) \tag{5}$$

Here $\vec{s}_{\,n_0}$ is the embedding vector, chosen as a reference point. We select all the neighbors with distance smaller than ε, (denoted by $u_n(\vec{s}_{\,0})$), and average over the distances of all neighbors to the reference point at time Δn. If $S(\Delta n)$ shows a linear robust increase for Δn then the slope is estimated as the maximal Liapunov exponent.

3. DETRENDED FLUCTUATION ANALYSIS

In order to compute the scaling exponent α from nonstationary time series denoted by $x(i)$ $[i = 1,....,N]$, the time series is integrated[6] first:

$$y(k) = \sum_{i=1}^{N} x(i) - \overline{x}$$

Here $\overline{x}$ is the the average value of the series x(i), and i ranges between 1 and N.

$y(k)$ is next divided into n boxes of equal length. A line ($y_n(k)$) is fitted to each of the boxes by a least squares fit. As the next step the time series is detrended by subtracting the local trend ($y_n(k)$) from the integrated time series. The root-mean square fluctuation of the detrended series, F(n) is computed as :

$$F(n) = \sqrt{\frac{1}{N} \sum_{k=1}^{N} [y(k) - y_n(k)]^2}$$

$F(n)$ is calculated for all n. The slope of the graph of $\log(F(n))$ versus $\log(n)$ is the scaling exponent α. This slope α is related to the 1/f spectral slope, m by the relation , m= $2\alpha - 1$.

4. ANALYSIS OF TEXTS

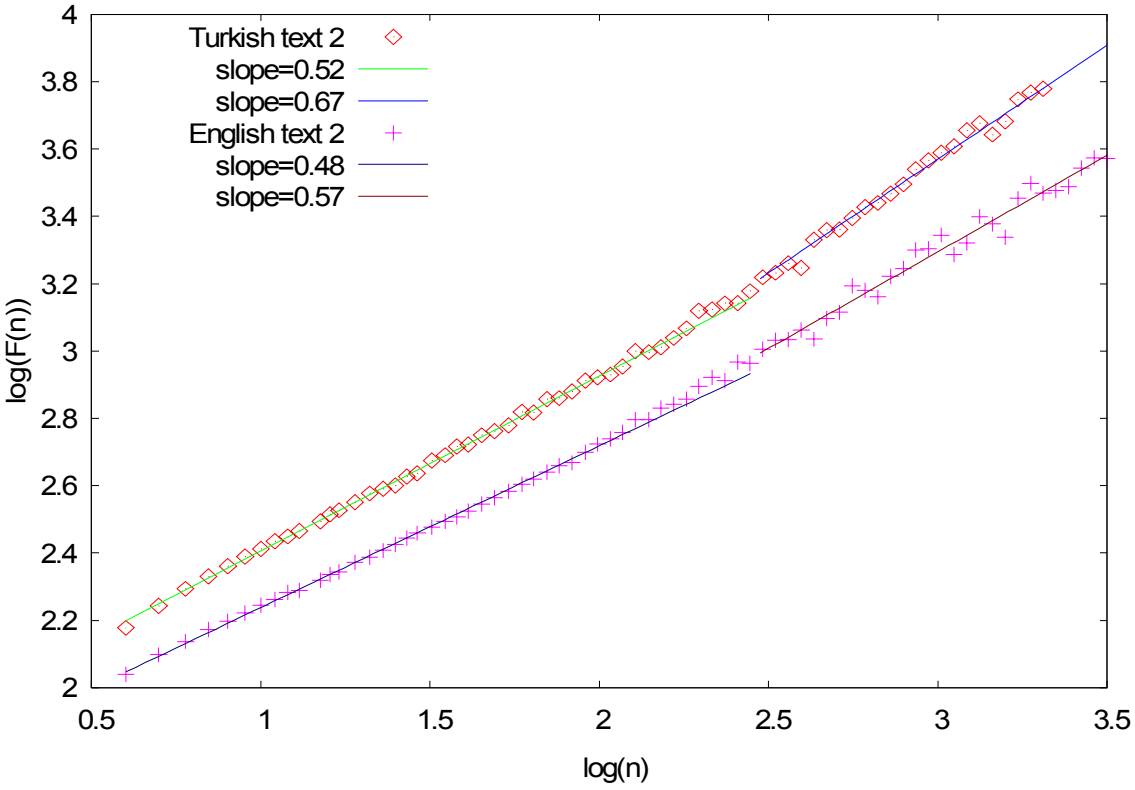
As an example of the proposed analysis method, two Turkish and two English texts (will be referred as Turkish text 1, Turkish text 2, English text 1, English text 2) were used. Turkish text 1 and English text 1 are independent of each other whereas Turkish text 2 and English text 2 are translations of each other. The time series is constructed from all the texts both using the frequencies from the corpuses and using derived variable from values assigned to the letters constituting a given word. In the latter case DNA Random Walks [25] served as a main source of inspiration. The methodology is as follows: Each word is accepted as a single step in a random walk. The length of each step (word) is determined from the letters as

$$s(n) = \sum_{i=1}^{N} y(i)$$

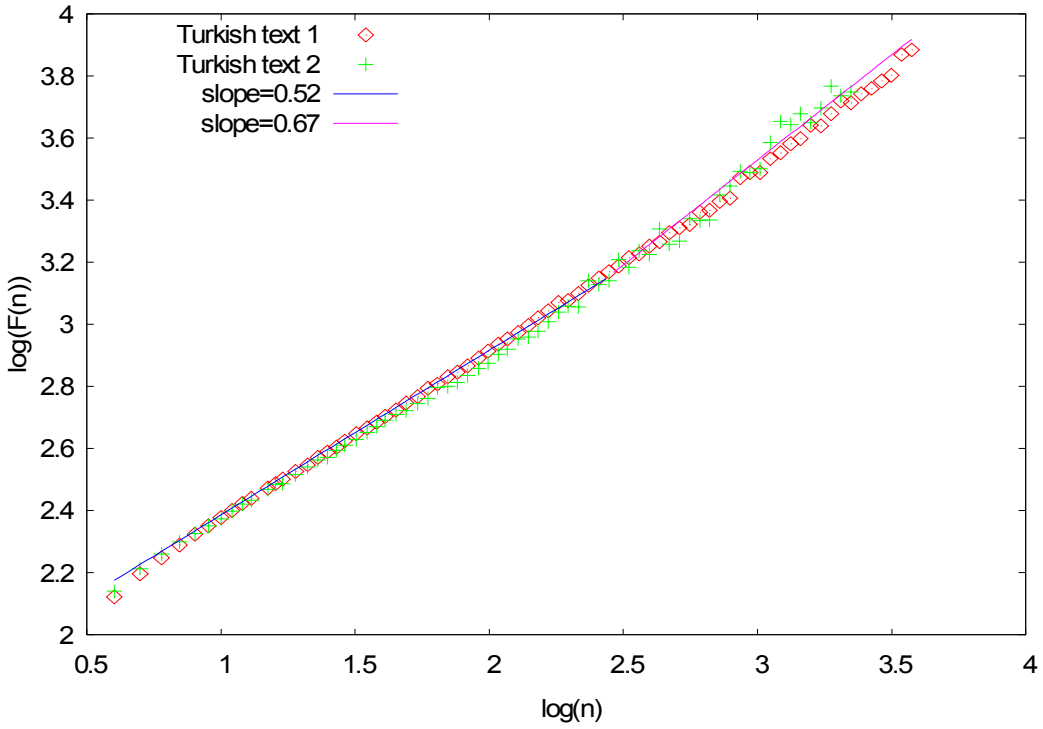
Here N is the length of the word and $y(i)$ is the ASCII value of the corresponding letter. After the time series is constructed, the scaling exponent and Lyapunov exponent are found.

4.1. Analysis using a variable derived from letters

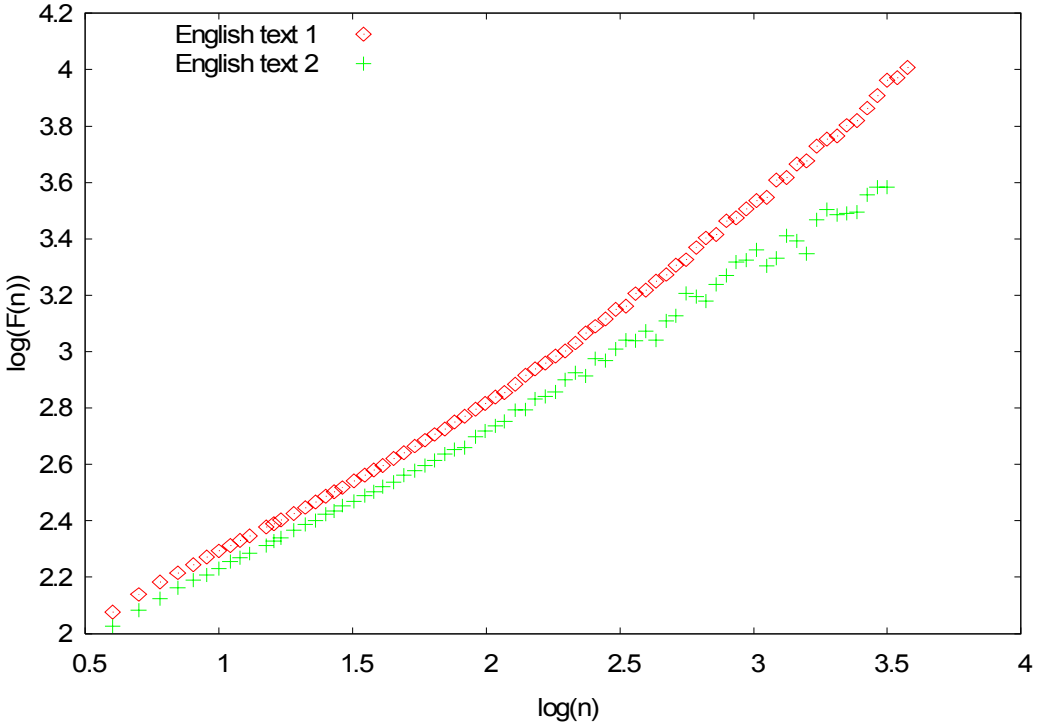
The detrended fluctuation analysis of English text 2 and Turkish text 1 (which are translations of each other) are presented below. The breakdown in the slopes show changes of the correlation properties. The slope of Turkish text is approxiamately 10% higher. The difference of correlation relations in the two regions for either analysis is clearly observed.



In order to further verify that the correlation properties are not of the texts but the languages the same analysis is applied to two turkish texts with different contexts. Below is the result of analysis where the correlation properties of the different turkish texts are quite near.



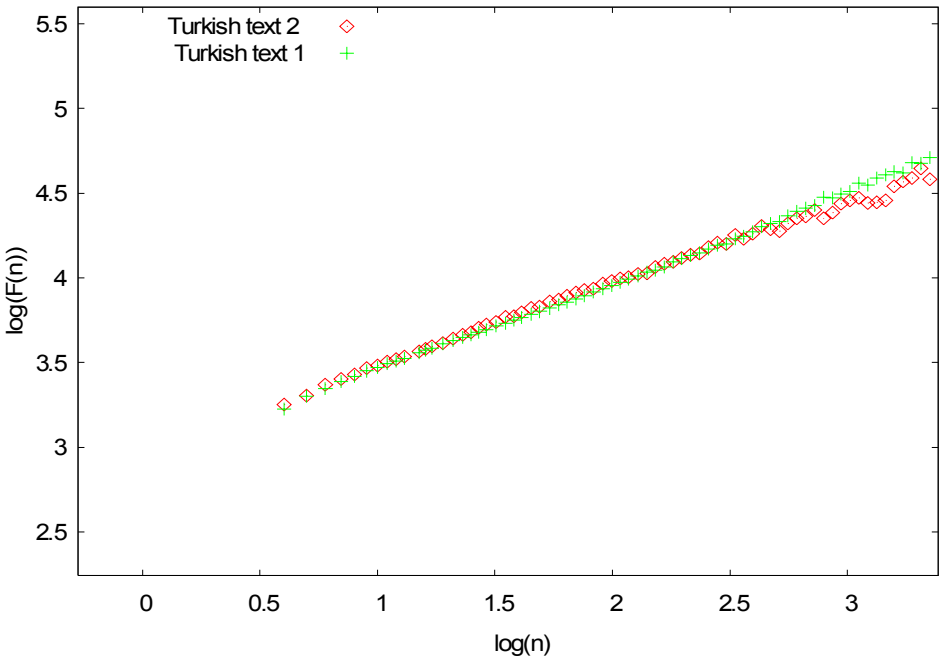
The same analysis applied to English texts is presented in the next figure. Again the correlation properties (the slopes) are similar.

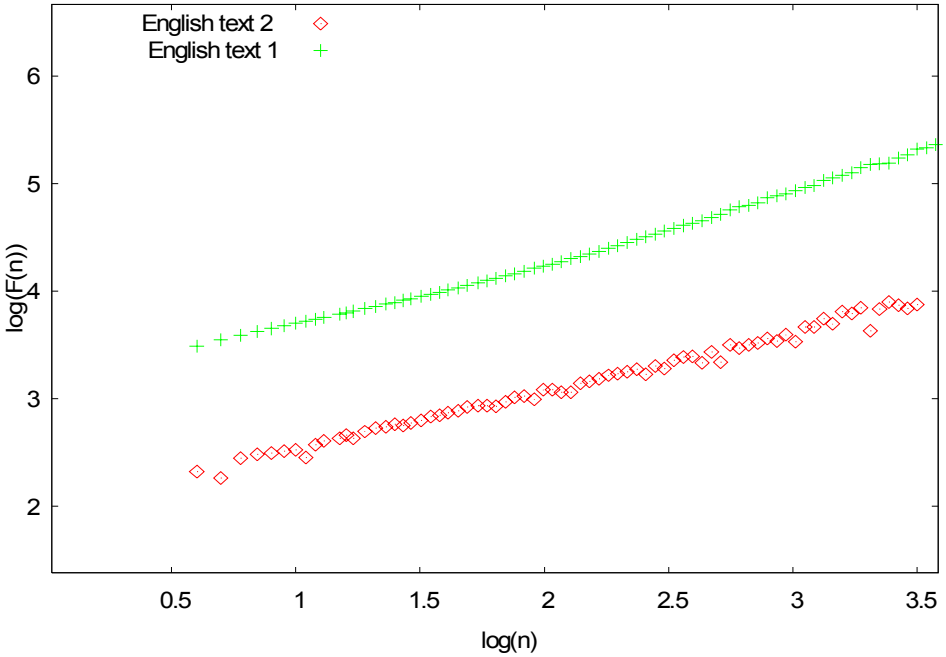


As a last check, the correlation properties of a random text using *Steganography* ciphering [26] (in English) was checked, and it is found that for the randomized text the correlation properties do not resemble the correlation properties of the English text.

4.2. Analysis Using Corpus

Detrended fluctuation analysis is now applied on the same texts, using frequencies derived from a corpus. The results are therefore independent of the previous parametrization. Below are the results for different Turkish texts, and different English texts, all results parelel to the previous section.





4.3. Lyapunov Exponents

The time series analysis is applied to the dependent variable derived by using the frequencies in corpus. The first results imply significantly different magnitudes for the maximal lyapunov exponents in Turkish (of order 0.20) and in English texts (of order 0.018). For a more reliable conclusion more sets of data and collection of frequencies from a very large set of data and using the collected frequencies for the segments of data, in order to eliminate the corpus dependence would be useful. Still, there is evidence both for a positive maximal lyapunov exponent and a possible difference between the two languages.

TEXT	Lyapunov Exponent
English Text 1	0.012
English Text 2	0.018
Turkish Text 1	0.2
Turkish Text 1	0.2

5. CONCLUSION

Zipf's laws relate the frequencies to the rank so that frequencies based on a corpus are natural candidates for a dependent variable; unfortunately this choice depends on the corpus. A random walk model is based on the progression from one word to the next, so that it is relatively more localized. Detrended fluctuation analysis results reveal differences between the two languages. This is independent of the two parametrizations used in this work.

Time series analysis using the corpus based parametrization yields a positive maximal liapunov exponent. The value of this exponent is different in the two languages. This indicates that word frequencies are an important tool for using a time series analysis on a natural language. Words constitute one of the smallest units in a natural language that carry meaning. Both the Turkish and English language shows evidence of chaotic behavior when word frequencies are used.

It is natural to expect a window of length up to eight words (the length of a sentence) where meanings of nearby words are expected to affect each other to the greatest extent. We think that this is why a time series analysis based on frequencies is meaningful but a time series analysis based on the random walk inspired model is relatively less feasible.

It is known that sounds in all natural languages exhibit fractal structure, but the evidence presented here along the lines that a time series can be derived from a written language would be of interest in many fields including cryptography.

There are indications that both word frequencies based on a corpus and a parametrization based on assigning values to letters can serve as blueprints for distinguishing languages. However we have used two languages that have significantly different grammatical structures. Different verb positions in the two languages or the presence of articles before nouns in English, much wider usage of suffixes in deriving Turkish words are examples of this difference. Finally, corpus quality is a factor that can also affect these results.

6. REFERENCES

[1] Hřebiček, Ludek, "Text as a Self-Similar Structure", *Text in Communication: Supra-Sentence Structures*, Universitätsverlag Dr. N. Brockmeyer, Bochum, 1992, pp. 91-96

[2] Hřebiček, Ludek, "Text Levels, Language Constructs, Constituents and the Menzerath-Altmann Law", Quantitative Linguistics Vol. 566, Wissentschaftlicher Verlag Trier (1995).

[3] Kantz, H.and T.Schreiber, *Nonlinear Time Series Anaylsis*, Cambridge University Press, Cambridge, 1997.

[4] Abarbanel, H.D.I., R.Brown, J. J. Sidorowich L.S. Tsimring, *The analysis of observed chaotic data*, Revs. Of Modern Phys., Vol. 65, No. 4, pp. 1331-1392, October 1993.

[5] Hřebiček, Ludek, "Text in Communication", Quantitative Linguistics Vol. 48, Universitätsverlag Dr. N. Brockmeyer, Bochum, 1992,

[6] C.-K. Peng, S.V. Buldyrev, S. Havlin, M. Simons, H.E. Stanley, A.L. Goldberger, Phys. Rev. E **49** 1685 (1994).

[7] J.W. Kantelhardt, E. Koscielny-Bunde, H.H.A. Rego, S. Havlin, and A. Bunde Physica A **295**, 441 (2001).

[8] S.V. Buldyrev, A.L. Goldberger, S. Havlin, R.N. Mantegna, M.E. Matsa, C.-K. Peng, M. Simons, H.E. Stanley, Phys. Rev. E **51** (1995) 5084.

[9] C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, R.N. Mantegna, M. Simons, H.E. Stanley, Physica A **221** (1995) 180.

[10] S.V. Buldyrev, N.V. Dokholyan, A.L. Goldberger, S. Havlin, C.-K. Peng, H.E. Stanley, G.M. Viswanathan, Physica A **249** (1998) 430.

[11] C.-K. Peng, J. Mietus, J.M. Hausdorff, S. Havlin, H.E. Stanley, A.L. Goldberger, Phys. Rev. Lett. **70** (1993) 1343.

[12] C.-K. Peng, S. Havlin, H.E. Stanley, A.L. Goldberger, Chaos **5** (1995) 82.

[13] K.K.L. Ho, G.B. Moody, C.-K. Peng, J.E. Mietus, M.G. Larson, D. Levy, A.L. Goldberger, Circulation **96** (1997) 842.

[14] G.M. Viswanatha, C.-K. Peng, H.E. Stanley, A.L. Goldberger, Phys. Rev. E **55** (1997) 845.

[15] C.K. Peng, J.M. Hausdorff, S. Havlin, J.E. Mietus, H.E. Stanley, A.L. Goldberger, Physica A **249** (1998) 491.

[16] Y.H. Liu, P. Cizeau, M. Meyer, C.-K. Peng, H.E. Stanley, Physica A **245** (1997) 437.

[17] P. Cizeau, Y.H. Liu, M. Meyer, C.-K. Peng, H.E. Stanley, Physica A **245** (1997) 441.

[18] M. Ausloos, N. Vandewalle, P. Boveroux, A. Minguet, K. Ivanova, Physica A **274** (1999) 229.

[19] M. Ausloos, K. Ivanova, Physica A **286** (2000) 353.

[20] Buldyrev, S. V., A. L. Goldberger , S. Havlin, C. K. Peng , H. E. Stanley and M. Simons, Biophys. J., Vol.65, pp.2673-2679, 1993.

[21] Ossadnik, S. M., S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, C. K. Peng, M .Simons, H.E Stanley, Biophys. J., Vol. 67, pp. 64-70, 1994.

[22] Hausdorff, J. M., C. K. Peng, Z. Ladin, J. Y. Wei, A. L.Goldberger, J. Appl. Physiol., Vol. 78, pp. 349-358, 1995.

[23] Hausdorff, J. M., P. Purdon, C. K. Peng, Z. Ladin, J. Y. Wei, A. L. Goldberger, J. Appl. Physiol., Vol. 80, pp. 1448-1457, 1996.

[24] Hu, K., P.Ch. Ivanov, Z. Chen, P. Carpena, H. E. Stanley, "Effect of nonstationarities on detrended fluctuation analysis", Phys. Rev. E., Vol. 64, pp. 011114-15, 2001.

[25] Alexandre Rosas, Edvaldo Nogueira, Jr., and Jose F. Fontanari, PHYSICAL REVIEW E **66**, 061906 (2002).

[26] http://www.fourmilab.ch/javascrypt/stego.html