Statistical Physics and Mathematics for Complex Systems

December 2011 Vol.56 No.34: 3717–3722
doi: 10.1007/s11434-011-4752-0

English and Chinese language frequency time series analysis

DENG WeiBing[1,2,3*], WANG DuJuan[2], LI Wei[2,4] & WANG Qiuping Alexandre[1,3]

[1] *LUNAM Université, ISMANS, Laboratoire de Physique Statistique et Systèmes Complexes, 44, Ave. Bartholdi, 72000 Le Mans, France;*
[2] *Complexity Science Center, Institute of Particle Physics, Central China Normal University, Wuhan 430079, China;*
[3] *LUNAM Université, Universite du Maine, LPEC, UMR CNRS 6087, 72085 Le Mans, France;*
[4] *Max-Planck-Institute for Mathematics in the Sciences, Leipzig 04103, Germany*

Received April 9, 2011; accepted May 20, 2011

English and Chinese language frequency time series (LFTS) were constructed based on an English and two Chinese novels. Methods of statistical hypothesis testing were adopted to test the nonlinear properties of the LFTS. Results suggest the series exhibited non-normal, auto-correlative, and stationary characteristics. Moreover, we found that LFTS follow the power law distributions, and thereby we investigated the fractal structure, long range correlation, and intermittency, which indicated the self-similarity features of LFTS, and also provided hints that human societies are likely to share some universal properties.

language time series, self-similarity, Hurst exponent, long range correlation, intermittency

Citation: Deng W B, Wang D J, Li W, et al. English and Chinese language frequency time series analysis. Chinese Sci Bull, 2011, 56: 3717–3722, doi: 10.1007/s11434-011-4752-0

Composed of a number of different words, written human language texts, such as novels, poems, and essays, are simply normal examples of complex systems in nature [1–3]. In recent years, along with the development of complex network theory, physicists have shown great interest in analyzing the characteristics of written human language texts from the complex network perspective.

For example, Masucci and Rodgers [3] found the existence of different functional classes of vertices, and noted the significance of second order vertex correlations in English written human language networks. Li and Zhou [4], on the other hand, analyzed the Chinese character system, supposing that radicals comprised nodes and that two nodes were linked if they could form a character or part of one. Their results revealed that character networks displayed small-word properties and showed non-Poisson degree distributions. Liu [5] built a Chinese semantic network based on a treebank with semantic role annotation and then investigated its global statistical properties. Liu and Li [6] also explored 15 linguistic complex networks based on the dependency of the syntactic treebanks of 15 languages. Yu et al. [7] described a series of identification experiments and rating experiments on the influences of the distance, spectral shape, and relative amplitude of the first two formants of the phonetic quality of /γ/.

In addition to the network point of view, however, time series analysis is also an important method for extracting information from signals related to real world complex systems. By analyzing such signals, we can better understand the underlying properties of complex systems.

Thus, time series analysis methods have also been used to investigate written human language texts [8–11]. Currently, there are two ways to map a text into a time series. One counts the number of letters of each word, namely word length l, while time t refers to the position of the word in the document, i.e. the first word is considered to appear at time t=1, the second at time t=2, etc. By mapping word length to time in this way, length time series $l(t)$ are constructed. The second way calculates the probability of appearances of each word in a text, namely the frequency f, while time t refers, again, to the position of the word; thus frequency time series $f(t)$ are constructed.

*Corresponding author (email: wdeng@ismans.fr)

In the current study, we attempted the second way of text mapping in an experiment to map the English and Chinese language text into frequency time series (www.cuiweiju.com, www.marxists.org). Two Chinese novels, *A Q Zheng Zhuan* (AQC) and *Kun Lun Shang* (KLS), and an English translated version of *A Q Zheng Zhuan* (AQE) were selected. The frequency of each word in the three novels was calculated, and each word was replaced by its corresponding frequency throughout the novels; the language frequency time series $f(t)$ was thus constructed.

As is well known, English and Chinese are two of the commonly used languages of the world, and both play important roles in international communication. In this paper, through analyzing the self-similarity of language frequency time series (LFTS), we attempt to find similar characteristics between the English and Chinese written human language texts, while also making comparisons between the two types of text. Furthermore, we relate the findings to language study and human thinking styles.

By employing statistical hypothesis testing, we first put our emphasis on investigating several nonlinear properties of the LFTS, such as the JB test, the autocorrelation test, and the unit root test. These tests showed the foremost properties of the LFTS, on the basis of which we discussed the scaling properties of the LFTS, such as frequency distribution, fractal behavior, long range correlations, and intermittency.

1 Statistical hypothesis testing of the language frequency time series

In the literature, hypothesis testing is generally called confirmatory data analysis, the results of which are deemed to have statistical significance if they are unlikely to have occurred by chance. When such tests are available, we may discover whether a second sample is significantly different from the first. Such decisions are normally made using null-hypothesis tests, which, assuming that the null hypothesis is true, determine the probability of observing a value for the test statistic that is at least as extreme as the value actually observed.

In the following section, we considered the *Jarque-Bera* test, the autocorrelation test, and the unit root test of the LFTS, respectively.

1.1 *JB* test

It is apparent that the mean value, standard deviation, skewness, and kurtosis are simply normal parameters which describe the characteristics of time series. Skewness and kurtosis reflect the asymmetry degree and the convergence degree of the return series, respectively; for a standard normal distribution, the skewness is 0 and the kurtosis is 3.

Here, we introduce the *Jarque-Bera* test [12], along with the null hypothesis that the data are drawn from a normal distribution, that is, the skewness and kurtosis values are 0 and 3, respectively. The statistical quantity of the *JB* test is

$$S = \frac{\frac{1}{n}\sum_{i=1}^{n}(f_i - \overline{f})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(f_i - \overline{f})^2\right)^{3/2}}, \tag{1}$$

$$K = \frac{\frac{1}{n}\sum_{i=1}^{n}(f_i - \overline{f})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(f_i - \overline{f})^2\right)^{2}}, \tag{2}$$

$$JB = \frac{n}{6}\left(S^2 + \frac{(K-3)^2}{4}\right), \tag{3}$$

where S is the skewness, K is the kurtosis, n is the statistics of the sample, and JB is the quantity for the JB test. The results of the JB test are shown in Table 1.

The results show that the skewness values of the language frequency time series are all non-zero, and that the kurtosis values are much larger than 3, indicating that they exhibit the leptokurtic. At the 5% significance level, the p values all equal 0, and thus we can reject the null hypothesis following the normal distribution, which implies that the language frequency time series are not random. In other words, when we express our ideas with text, some words may be used more often than others, such as prepositions and conjunctions, etc.

1.2 Autocorrelation test

Autocorrelation is the cross-correlation of a signal with itself. As a test, it is a mathematical tool to find repeating patterns, such as the presence of a periodic signal that has been buried under noise, or the identification of a missing fundamental frequency in a signal implied by its harmonic

Table 1 *JB* test results of the LFTS

Sample	Statistics	Mean	SD	Skewness	Kurtosis	*JB*	*p* value*
AQC	21118	0.007	0.010	2.145	7.381	28441	0
AQE	17204	0.009	0.013	1.981	6.523	19975	0
KLS	23270	0.005	0.009	2.699	9.848	64069	0

* Statistical significance at 5% level.

frequencies [13].

The autocorrelation coefficient denotes the correlation degree of a language frequency time series $f(t)$, $f(t-1)$, ..., $f(t-k)$ in different periods, and is defined as

$$AC_k = \frac{\sum_{t=k+1}^{T} [f(t) - \overline{f}][f(t-k) - \overline{f}]}{\sum_{t=1}^{T} [f(t) - \overline{f}]}. \tag{4}$$

The values of AC_k range from −1 to 1, and the larger the absolute value of AC_k, the stronger the correlation in the language frequency time series.

The *Ljung-Box* test [14], on the other hand, is a type of statistical test that tests whether any of a group of autocorrelations of a time series is different from zero. Instead of testing the randomness at each distinct lag, it tests the "overall" randomness based on a number of lags. The null hypothesis of the *Ljung-Box* test is that the language frequency time series is random, and the test statistic is the Q value, which is defined as

$$Q_{LB} = T(t+2)\sum_{j=1}^{k} \frac{AC_j^2}{T-j}, \tag{5}$$

where t is the observation of the sample, and AC_j is the autocorrelation coefficient. As in Table 2, one can observe that the Q values are large and the p values are all 0, and thus we can conclude that there exist strong autocorrelations in the language frequency time series, which indicates that the later time series strongly correlate with the former ones.

1.3 Unit root test

In statistics, the unit root test is used to test whether time series variables are non-stationary using an autoregressive model [15]. Generally, the stationary time series should satisfy the condition that the mean value be almost constant at any time point in the time series, and that the autocorrelation function at two different time points is only relative to the time interval of the two time points Δt.

A well-known test that has been validated in large samples is the Augmented Dickey Fuller (ADF) test, which uses the existence of a unit root as the null hypothesis. The more negative the ADF statistic is, the stronger the rejection of the hypothesis that there is a unit root at some level of confidence.

Table 2 Autocorrelation test of the LFTS

Sample	AC*	*Q* value*	*p* value*
AQC	0.004	167.7	0
AQE	0.002	416.38	0
KLS	0.011	229.2	0

* At 11th order of lag.

As seen in Table 3, one can find that the values of the ADF are much smaller than the critical values of the three different significance levels 1%, 5%, and 10%, and thus we can conclude that the language frequency time series are all stationary time series. Furthermore, it might be convenient for us to analyze the other properties of the language frequency time series.

The results of the autocorrelative and stationary properties tests suggest that the structures of human language texts might be similar across sentences, with only some insignificant fluctuations.

2 Self-similarity of the language frequency time series

The above statistical hypothesis tests demonstrate that language frequency time series are non-normal and stationary, and that later time series correlate with former ones; these results suggest that language frequency time series might display self-similarity characteristics.

Therefore, in this section we discuss in detail the distributions of the language frequency time series, and also employ detrended fluctuation analysis (DFA) to investigate the self-similarity of language frequency time series characteristics, such as fractal structure, long range correlation, and intermittency, etc.

Moreover, through such investigations we try to identify commonly shared characteristics of written human language, while also comparing the different properties of the English and Chinese languages.

2.1 Distribution of the language frequency time series

As seen in Figure 1, the distributions of the language frequency time series all follow the shifted power law (SPL) [16] function:

$$P(>f) \propto (f + f_0)^{-\gamma}. \tag{6}$$

Resorting to the original data, we discover that words with a large frequency tend to form binary structures, and are more likely to constitute phrases or short sentences. Also, one can observe that the two Chinese language frequency time series almost overlap with each other, while they are different from the English frequency time series. This difference may be partially attributed to the different grammar rules of English and Chinese.

Table 3 Unit root test of the language frequency time series

Sample	ADF	1%	5%	10%	*p* value
AQC	−105.4	−3.43054	−2.86151	−2.566795	0.0001
AQE	−105.907	−3.43056	−2.86152	−2.5668	0.0001
KLS	−91.1179	−3.4305	−2.86149	−2.56679	0.0001

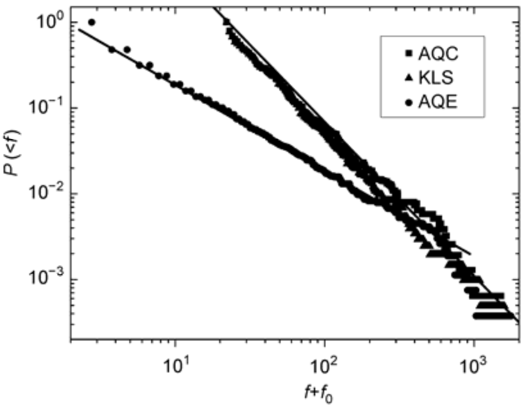

Figure 1 Distribution of the language frequency time series.

2.2 Fractal structure of the language frequency time series

The Hurst exponent [17,18] is generally used to measure the long term memory of time series, such as autocorrelations, etc. The value 0<*H*<0.5 indicates a time series with a negative autocorrelation, which means that a word with a small frequency value will probably be followed by a word with a large frequency value. The value 0.5<*H*<1, on the other hand, indicates a series with a positive autocorrelation, such as a word with a large frequency value followed by another word with a large frequency value. Finally, a value of *H*=0.5 indicates a true random walk of the time series which has no memory of previous values.

Here, taking the standard deviation of the language frequency time series as a new time series, we employed the detrended fluctuation analysis (DFA) method to calculate the Hurst exponent, that is, to calculate the standard deviation of the new time series:

$$std(t) \sim t^H. \tag{7}$$

Std(*t*) is the standard deviation of the standard deviation time series, *t* is the time scale, and *H* is the Hurst exponent. The calculation steps are as follows.

(1) The language frequency time series sample *N* is divided into *n* bins, with the length of every bin being *T*=*N*/*n*, and then the standard deviation *s*(*j*) is calculated in all non-overlapping bins of length *T*, which denotes the fluctuations in every bin.

$$s(j) = \sqrt{\frac{1}{T-1}\sum_{i=1}^{T}(f_i - \overline{f})^2}, j = 1, 2, \cdots n, \tag{8}$$

$$\overline{f} = \frac{1}{T}\sum_{i=1}^{T} f_i. \tag{9}$$

(2) *s*(*j*) forms a new time series. The sum of the fluctuations trace in step *t* is

$$x(t) = \sum_{j=1}^{t} s(j), t = 1, 2, \cdots n. \tag{10}$$

The increment is $\Delta x(t_0)=x(t+t_0)-x(t)$, *t* is an original value, and t_0 is the increment.

(3) The standard deviation of the increment $\Delta x(t_0)$ is calculated as

$$std(t_0) = \sqrt{\overline{\Delta x^2(t_0)} - \overline{\Delta x(t_0)}^2}. \tag{11}$$

(4) The least squares method has been applied to calculating the Hurst index *H*, $Std(t_0)$~t_0^H.

The results of the detrended fluctuation analysis of the standard deviation is shown in Figure 2, and the Hurst exponent *H* is calculated using least-squares regression, with $Std(t_0)$~t_0^H. As seen in Figure 2, the values of the Hurst exponent *H* are 0.53, 0.57, and 0.61 for AQC, AQE, and KLS, respectively, which are all between 0.5 and 1; thus one can observe the persistent behavior that exists in the language frequency time series. The larger the Hurst exponent, the stronger the persistent behavior, and thus we can conclude that the persistent behavior for the English language is stronger than that of the Chinese language due to the flexible mechanisms of English.

While the fractal dimensions *α* are all less than 2, which suggests the language frequency time series possess the characteristics of the fractal structure and the long term memory, structures of written human language appear to have similar properties in different sentences. The fractal time series is self-similar in essence, that is, the series may have some similar statistical characteristics in the different time scales, and the probability distributions of the series still retains the same profile even if the time scale changes.

2.3 Long range correlation of language frequency time series

The detrended fluctuation analysis (DFA) [19,20] is a scaling analysis method that can be also applied to quantifying the long range power law correlations in a time series analysis; the scaling exponent is used to clarify the time series that appear to be long memory processes or 1/*f* noise.

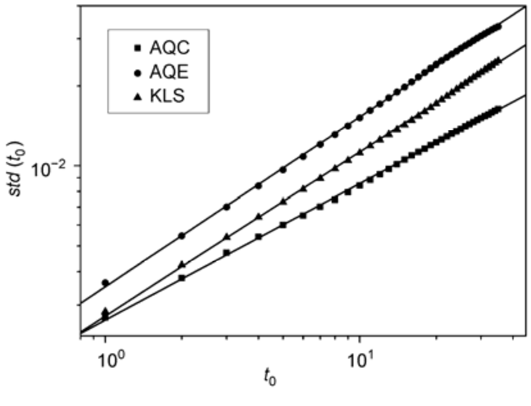

Figure 2 Fractal structure of language frequency time series, by investigating the Hurst exponent.

Supposing f_i (i=1, 2, $\cdots$ N) is a time series of length N, the calculation approach of the DFA follows five steps.

(1) The time series to be analyzed (with N samples) are integrated.

$$y(j)=\sum_{i=1}^{j}[f_i-\langle f\rangle], \qquad (12)$$

$$\langle f\rangle=\frac{1}{N}\sum_{i=1}^{N}f_i. \qquad (13)$$

(2) The integrated time series are divided into boxes of equal length S, and a least squares line is fit to the data in each box of length S, which represents the trend in that box; then the y coordinate of the straight line segments is denoted by $y_s(k)$.

(3) The integrated time series $y(k)$ is detrended by subtracting the local trend $y_s(k)$ in each box; then the root mean square fluctuation of this integrated and detrended time series is calculated by

$$F(S)=\sqrt{\frac{1}{N}\sum_{k=1}^{N}[y(k)-y_s(k)]^2}. \qquad (14)$$

(4) Repeat the calculation steps (1)–(3) for different box sizes S to characterize the relationship between $F(S)$ and the box size S.

(5) The scaling exponent β is calculated as the slope of a straight line fit to the log-log graph of S against $F(S)$.

$$F(S)\propto S^{\beta}. \qquad (15)$$

The scaling exponent β is similar to the Hurst exponent, according to which we can divide the time series into three categories. (1) If β=0.5, there is no correlation at all, and the time series follows random walking. (2) If $\beta>0.5$, a persistent long-range power-law correlation exists in the time series. (3) If β<0.5, power-law anti-correlation is presented in the time series.

The results of the detrended fluctuation analysis of the language frequency time series are presented in Figure 3, with $F(S)\propto S^{\beta}$, slope β equal to 0.61 for AQC and KLS, while the result of AQE was 0.63. Thus we can observe that the persistent long range power law correlation exists in the time series, and that the persistent trend is stronger for the English language, which is consistent with the results of the research on the fractal structure. The persistent long range power law correlations may also imply that self-similarity definitely presents in the time series.

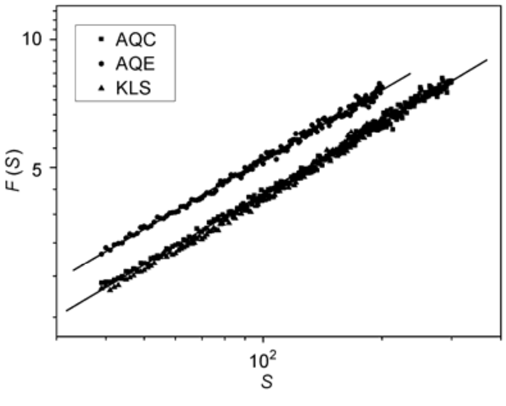

Figure 3 Long range correlation analysis of the language frequency time series.

2.4 Intermittency in the language frequency time series

Intermittency has been observed in many time series, which indicates that observations can differ dramatically depending on the timing. For example, fast increases in heart rate, which results from physiological activity, might exhibit intermittency. This is an essential property of the system that has been broadly used to characterize databases. The concept of intermittency we consider here also has connections with the concept of multi-scaling, or multi-fractal, in the stochastic processes.

We first divide the value range of the frequency Δ into M intervals δ, with M = Δ/δ. n is the number of frequencies that fall in δ of one event. Here, "one event" refers to the ensemble of frequencies in a single paragraph. Then, the q-order scaled factorial moment is defined as

$$F_q=\frac{\langle n(n-1)...(n-q-1)\rangle}{\langle n\rangle^q}, \qquad (16)$$

where the brackets indicate an average over all paragraphs.

If power-law scaling follows

$$F_q\sim\delta^{-\varphi_q},\ \varphi_q>0, \qquad (17)$$

then we can conclude that intermittency behavior has been observed.

Calculating the second-order scaled factorial moment F_2, we show the relationship between F_2 and the intervals δ in Figure 4; the power-law scaling is also shown in double-logarithm scale, with

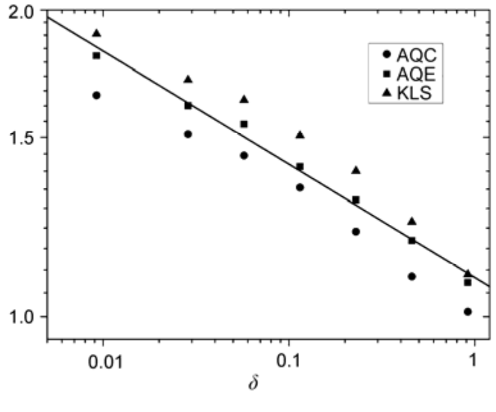

Figure 4 Intermittency behavior of the frequency time series using the scaled factorial moment method.

$$F_2 \sim \delta^{-\nu}, \ \nu > 0. \quad (18)$$

Thus, we can draw the conclusion that intermittence phenomena exist in the frequency time series, which might further show evidence that the frequency time series possesses self-similarity characteristics.

3 Discussion and conclusion

We have executed both English and Chinese language frequency time series analyses, which demonstrate that language frequency time series show non-normal and autocorrelative properties, and that the distributions of the language frequency time series exhibit the shifted power law format. Based on the above analyses, further investigation will be needed on the fractal behavior, long range correlations, and intermittency of the language frequency time series. In this study, we found that the fractal structure, persistent long range correlation, and intermittency exist in the language frequency time series, which suggests that human societies are apt to possess some commonly shared characteristics, such as self-similarity. However, whether these properties are tenable for language length time series is unclear. Future work should pay attention to such issues.

This work was supported by the National Natural Science Foundation of China (10647125, 10635020, 10975057 and 10975062), the Program of Introducing Talents of Discipline to Universities (B08033) and the PHC CAI YUAN PEI Program (LIU JIN OU [2010] No. 6050) (2010008104).

1 Zipf G K. Human Behavior and the Principle of Least Effort. Reading, MA: Addision-Wesley, 1949

2 Calderia S M G, Lobao T C P, Andrade R F S, et al. The network of concepts in written texts. Eur Phys J B, 2006, 49: 523–529

3 Masucci A P, Rodgers G J. Network properties of written human language. Phys Rev E, 2006, 76: 026102

4 Li J Y, Zhou J. Chinese character structure analysis based on complex networks. Physica A, 2007, 380: 629–638

5 Liu H T. Statistical properties of Chinese semantic networks. Chinese Sci Bull, 2009, 54: 2781–2785

6 Liu H T, Li W W. Language clusters based on linguistic complex networks. Chinese Sci Bull, 2010, 55: 3458–3465

7 Yu S Y, Chen Y D, Wu J R. Spectral integration and perception of Chinese back vowel /γ/. Sci China Inf Sci, 2010, 53: 2300–2309

8 Schenkel A, Zhang J, Zhang Y C. Long range correlation in human writtings. Fractals, 1993, 1: 47–57

9 Amit M, Shemerler Y, Eisenberg E, et al. Language and codification dependence of long range correlation in texts. Fractals, 1994, 2: 7–13

10 Ausloos M. Equilibrium and dynamic methods when comparing an English text and its Esperanto translation. Physica A, 2008, 387: 6411–6420

11 Kosmidis K, Kalampokis A, Argyrakis P. Language time series analysis. Physica A, 2006, 370: 808–816

12 Carlos M J, Anil K B. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. Econom Lett, 1980, 6: 255–259

13 Baum C F. An Introduction to Modern Econometrics Using Stata. College Station, Texas: Stata Press, 2006

14 Ljung G M, Box G E P. On a measure of a lack of fit in time series models. Biometrika, 1978, 65: 297–303

15 Mills T C. Time Series Techniques for Economists. New York: Cambridge University Press, 1990

16 Chen Y Z, He D R. A study on some urban bus transport networks. Physica A, 2007, 376: 747–754

17 Martinis M, Knezevic A, Krstacic G, et al. Changes in the Hurst exponent of heartbeat intervals during physical activity. Phys Rev E, 2004, 70: 012903

18 Carbonea A, Castelli G, Stanley H E. Time-dependent Hurst exponent in financial time series. Physica A, 2004, 344: 267–271

19 Peng C K, Buldyrev S V, Havlin S, et al. Mosaic organization of DNA nucleotides. Phys Rev E, 1994, 49: 1685–1689

20 Ivanova K, Ausloos M. Application of the detrended fluctuation analysis (DFA) method for describing cloud breaking. Physica A, 1999, 274: 349–354