Hindawi
Complexity
Volume 2018, Article ID 9362468, 12 pages
https://doi.org/10.1155/2018/9362468

WILEY | Hindawi

Research Article

Quantifying Evolution of Short and Long-Range Correlations in Chinese Narrative Texts across 2000 Years

Heng Chen[1] **and Haitao Liu**[1,2,3]

[1]*Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies, Guangzhou 510420, China*
[2]*Department of Linguistics, Zhejiang University, Hangzhou 310058, China*
[3]*Ningbo Institute of Technology, Zhejiang University, Ningbo 315100, China*

Correspondence should be addressed to Haitao Liu; lhtzju@yeah.net

Received 15 August 2017; Accepted 20 December 2017; Published 8 February 2018

Academic Editor: Hiroki Sayama

We investigate how short and long-range word length correlations evolve in Chinese narrative texts. The results show that, for short-range word length correlations, no significant linear evolutionary trend was found. But for long-range correlations, there are two opposite tendencies for two different regimes: the Hurst exponent of small-scale (box size n ranges from 10 to 100) word length correlations decreases over time, and the exponent of large-scale (box size n ranges from 101 to 1000) shows an increasing tendency. The increase of word length is corroborated as an essential regularity of word evolution in written Chinese. Further analyses show that a significant correlation coefficient is obtained between Hurst exponents from the small-scale correlations and mean word length across time. These indicate that word length correlation evolution possesses different self-adaptive mechanisms in terms of different scales of distances between words. We speculate that the increase of word length and sentence length in written Chinese may account for this phenomenon, in terms of both the social-cultural aspects and the self-adapting properties of language structures.

1. Introduction

As a result of human evolution, [1, 2] language is closely related to the evolution of human physical being and the increasing need for effective communication [3]. Many studies show that language can be usefully described as a complex system [4–7], with hierarchical structure in terms of syntactic organization [8, 9], from morphemes to words, phrases, and sentences. At each level, language is constantly evolving to adapt to human needs and constraints.

Word is the fundamental unit of language, which is arranged and structured, according to syntactic principles, to form phrases, sentence, and texts [8]. Dependency grammar even holds words as the only key units of sentences, linked together to form syntactic structures of different sizes [10]. As a result, lexical features may not only reflect linguistic properties at the level of words, but also throw light on syntactic patterns at the level of phrases and sentences. One lexical feature that has drawn much attention is word length, which has been extensively studied, especially in quantitative linguistics [11–13]. Piantadosi et al. (2011) reported that word length is intimately related to the information that words transmit; Garcia et al. [14] found that longer words are more likely to express abstract things. Köhler [15] points out that word length may reflect the properties of its basic language units-words. Many recent researches have been devoted to word length distributions in texts [16, 17], such as the famous Zipf's law of word length distribution. However most of them are concerned with words per se. Only very recently has the syntagmatic dimension been paid attention. That is, word length has been investigated in terms of word sequences, not individual words [15, 18]. There are several methods to investigate into word length in sequences, including word length entropies [19, 20] (Papadimitriou, 2010; Grotjahn, 1979), word length correlations [21, 22], word length repetitions [23], and the latest word length motifs [15, 18, 24].

In a recent study, Chen and Liang [24] explored the evolution of word length motifs in written Chinese, and the

results show that there is a tendency that the distributions are more concentrated on some certain motif patterns. Here in this study, we intend to diachronically explore the patterns of word length correlations in written Chinese narrative texts, which may shed much light on the evolution of not only words, but also the syntagmatic organizations of words, that is, the syntactic patterns. Chinese is suitable for such a diachronic exploration [25, 26], because it is one of the most archaic living languages with many continuous written records [27].

Then, what is word length correlation? If word sequences in a text are mapped onto word length (measured in syllables) sequences, patterns may be found in the sequences of word lengths, which is reflected by word length correlation [28]. Furthermore, the patterns involve self-similarity among word length sequences, which can be estimated through word length correlations and explained by fractal and cascade effects in narrative texts [21].

In a written text, there are temporal correlations among segments between short and long distances [29, 30]. These correlations can be found in word length series, word frequency series, unicode (of word/character) series, and so on. Using the natural visibility graph method, Guzmán-Vargas et al. [29] studied the correlation of word lengths in large texts from 30 ebooks in the English language from the Gutenberg Project. The existence of long-range correlations has also been explored by Montemurro and Pury [31] and Bhan et al. [32] in English and Korean texts using Hurst's exponents. Accordingly, in this study, we define short-range correlations as distances between words that are no longer than 10 and long-range as distances that are larger than 10, which are displayed in Section 2 in the following.

Word length correlations may reflect some fundamental and universal features of human communications, closely related to the structural patterns of communication [31]. However, the word length patterns of local word length correlations [19] do not exhibit the universality of ordering as proposed by Montemurro and Zanette [33]. Entropy provides a rigorous measure of the degree of order (especially in a short range) in symbolic sequences (Lesne, Blanc and Pezard, 2009); [19]. For example, Grotjahn (1979) explored word length entropies in Goethe's "Erlkönig." The problem of assigning a value to the entropy of language has inspired research since the seminal work by Shannon [34]. As for the long-range word length correlations, the detrended fluctuation analysis (DFA for short) is frequently used [35].

However, diachronic changes in word length correlations are so far unexplored, especially for Chinese, which is what this paper is committed to. Specifically, we will explore the following questions in this study.

Question 1. How do the word length distributions and the Zipfian n-gram word length block distributions evolve over time in Chinese written narrative texts?

Question 2. How do short-range word length correlations (calculated through Relative Entropy D in this study) and long-range word length correlations (calculated through detrended fluctuation analyses) evolve over time?

Question 3. What are the implications of the evolution of word length correlations, and if it is related to the structural patterns of word length distributions? To test these, the Pearson Analyses will be used to detect if there are significant correlations between evolution of mean word length and parameter values of word length correlations.

The rest of this paper is organized as follows. Section 2 describes the materials and methods used in this study; Section 3 gives the results and discussions of the diachronic investigations. Section 4 has conclusions. To anticipate, this study may give us a much more in-depth understanding of written Chinese word length ordering patterns.

2. Materials and Methods

2.1. Materials. Our diachronic study includes six historical periods, which are the pre-Qin period, the Northern and Southern Dynasties, Song dynasty, Ming dynasty, Qing dynasty, and the contemporary period. There are 10,000 Chinese characters in each time period (average number of words: 7568), and the details of the corpus are in Table 1.

Since there is no reliable word segmentation tool for ancient Chinese, we have to manually segment the ancient Chinese texts. For the contemporary texts we use ICTCLAS 2008 to segment words. The critical problem of manual segmentation is the tokenizing standards, especially the standard to distinguish between words and phrases. In order to keep consistence, we used the modern tokenization standards when we encounter these kinds of problems in manual segmentation. For example, in order to handle the ancient Chinese texts precisely, we referred to a lot of historical linguistic works concerning ancient Chinese lexicon, such as *Hanyu Da Cidian (the big dictionary of Chinese)*. After word segmentation, we programmed a java software to extract statistics concerning frequency distribution of word lengths. We use MATLAB to test if the relation between word length distribution and mean words frequency obeys power laws.

2.2. Word Length Distribution and the Zipfian N-Gram Word Length Distribution. We construct the time sequences of word length by mapping each text to a sequence of numbers w_i, $i = 1, 2, 3, \ldots, N$, where every number represents the length of the respective word. The resulting sequences consist of integers, with the minimum being 1 and the maximum being the length of the longest word in the specific language corpus. An example of this sequence is shown in Figure 1 (the words are segmented with blank spaces).

As can be seen, the mapping of texts onto time series is achieved by replacing every word with its length. The word length sequences are then studied to derive frequency distributions, the n-gram entropies as defined below, and the detrended fluctuation analyses, which will be introduced in the following section.

2.3. N-Gram Word Length Entropies and Short-Range Word Length Correlations. Word length entropy is an indicator of the uniformity of the data [36]. It is generally considered that the larger the entropy is, the more uniform the data are. For

Table 1: Diachronic corpus details[1].

Time period	1 Work	2 Work	3 Work	4 Work	5 Work	6 Work
Texts	*MèngZǐ* (Mencius) *Lǚshìchūnqīu* (Mister Lv's Spring and Autumn Annals)	*Shìshuōxīnyǔ* (A New Account of the Tales of the World) *Yánshì Jiāxùn Shū* (Mister Yan's Family Motto)	*Niǎn Yùguānyīn* (Grinding Jade Goddess of Mercy) *Cuòzhǎncuīníng* (Wrongfully Accused of Ying Ning) *Jiǎntiēhéshang* (A letter from a monk)	*Shíèrlóu* (Twelve Floors) *Wúshēngxì* (A Silence Play)	*Nàhǎn* (Yelling) *Pánghuáng* (Hesitating)	*Xīndàofózhī* (The Buddha Knows Your Mind) *Huíménlǐ* (A Wedding present)
Scale (characters)	141,864	94,729	11,220	233,430	91,705	12,980
Selected text scale (characters)	10,000	10,000	10,000	10,000	10,000	10,000
Time span	B.C. 3th-B.C. 2th	A.D. 4th-A.D. 5th	A.D. 12th-A.D. 13th	A.D. 16th-A.D. 17th	Pre-A.D. 20th	A.D. 21th

[1]The narrative texts from Time Period 1 to 4 are from http://www.gushiwen.org/; the narrative texts of Time Period 5 are from http://yuedu.163.com/; the X*inwen Lianbo* texts in Time Period 6 are from http://tv.cctv.com/lm/xwlb/.

Chinese: "老廖 一 整夜 都 睁 着 眼睛 坐 在 值班室 的 电话机 旁。"
Word length sequence (Chinese): 2 1 2 1 1 1 2 1 1 3 1 3 1.
Pinyin:
(lǎo liào) (yī) (zhěng yè) (dōu) (zhēng) (zhe) (yǎn jīng)
(zuò) (zài) (zhí bān shì) (de) (diàn huà jī) (páng)
English (correspondence to words):
"(Laoliao) (one) (night) (all) (open) (auxiliary "*zhe*") (eyes) (sit) (on)
(duty office) (particle "*de*") (telephone) (side)."
English:
Laoliao sat beside the telephone with his eyes open all night long in the duty office.

n-grams:

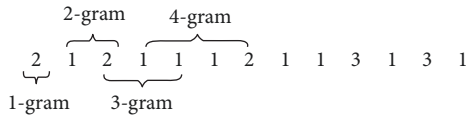

Figure 1: The word length sequence derived from a sentence.

example, if all four word length classes (e.g., word length class 1 means all the words that have word length 1, and so forth) in written Chinese (usually there are 4 word length classes in Chinese) have the same number of words (in terms of types or tokens), then the word length entropy has the maximum of 2. Mathematically, word length entropy cannot exceed 2.

The Shannon like word length entropy (WLE) can be calculated with the following formula:

$$H(i) = -\sum_{i \in X} p(i) \ln\left(p(i)\right), \tag{1}$$

where i refers to word length classes and $p(i)$ refers to the words probability of word length class x. The *n-gram* word length entropy defined in this paper is block entropy, which is extended in this definition [19]. The exact definition of word length entropy is as follows.

For the word length sequence w_i, $i = 1, 2, 3, \ldots, N$, we set $K = N - n + 1$ and define the *n-grams as* $n_j = (w_j, \ldots, w_{(j+n-1)})$ where $j = 1, 2, \ldots, K$. The formula for calculating the Shannon like entropy is

$$H_n = -\sum_{(s_1,\ldots,s_n)} p^{(n)}(s_1, \ldots, s_n) \ln\left(p^{(n)}(s_1, \ldots, s_n)\right), \tag{2}$$

where $p^{(n)}(s_1, \ldots, s_n)$ denotes the probability of each *n-gram* $(s_1, \ldots, s_n)$, calculated as

$$\frac{\text{No. of } n\text{-grams } (s_1, \ldots, s_n) \text{ encountered when gliding}}{\text{Total No. of } n\text{-grams}}. \tag{3}$$

(Refer to [19])

Here we take the sentence in Figure 1 as an example. There are 12 2-gram sequences in the word length sequence "2 1 2 1 1 1 2 1 1 3 1 3 1." Their frequencies and probability distributions are shown in Table 2. And the Shannon like 2-gram entropy H_n can be calculated as

$$\begin{aligned} H_n = &-(0.25 * \ln 0.25 + 0.25 * \ln 0.25 + 0.1667 \\ &* \ln 0.1667 + 0.1667 * \ln 0.1667 + 0.1667 \\ &* \ln 0.1667) = 1.589. \end{aligned} \tag{4}$$

Table 2: The frequencies and probability distribution of 2-gram word length sequences of the sentence in Figure 1.

2-gram		Frequency	Probability
1	1	3	0.2500
2	1	3	0.2500
1	2	2	0.1667
3	1	2	0.1667
1	3	2	0.1667
2	3	0	0.0000
3	2	0	0.0000
2	2	0	0.0000
3	3	0	0.0000

On the basis of n-gram sequences or blocks, we can estimate short-range word length correlations. Since the n-gram ($n = 2, 3, 4$) word length entropies contain information of word length ordering within sentences in texts, we shuffled the words in texts and obtained the shuffled word length entropy H_s (using the above method). By definition, the shuffled word length sequence should have more uniform n-gram distributions and larger entropy. Therefore, the short-range correlations of word length sequences can be defined as $D = H_s - H$, where the quantity D is the decrease of entropy due to the ordering of words (Refer to [19]).

2.4. Detrended Fluctuation Analysis and a Generalization of the Hurst Exponent. Detrended Fluctuation Analysis was proposed by Peng et al. [37] to analyze the statistical self-correlation in a time sequence which may span a long memory. The obtained power law exponent is similar to the Hurst exponent, except that DFA may also be applied to signals whose underlying statistics (such as mean and variance) or dynamics are nonstationary. The calculation process is as follows.

Given a word length time sequence $x(t)$, whose length is N, first, the cumulative sum or profile $y(t)$ is calculated:

$$y(t) = \sum_{i=1}^{t} [x(i) - \overline{x}] \quad (i = 1, 2, 3, \ldots, N), \tag{5}$$

where $x(i)$ indicate the word length of the i-th word in time series $x(t)$ and $\overline{x}$ is the mean value of all the word lengths in time series $x(t)$, calculated with the following formula:

$$\overline{x} = \frac{1}{N}\sum_{t=1}^{N} x(t). \quad (6)$$

Then, the word length time sequence $x(t)$ is divided into m nonoverlapping blocks by length n, where $m = [N/n]$ (rounded). Then a local least squares straight line fit is calculated by minimizing the least squared errors within each interval. Let $y_n(t)$ be the series of straight line fits.

Next, the root-mean-square deviation from the trend, the fluctuation, is calculated:

$$F(n) = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left[y(t) - y_n(t)\right]^2}. \quad (7)$$

Take the sentence in Figure 1 for an example, the word length sequence is "2 1 2 1 1 1 2 1 1 3 1 3 1":

$$\overline{x} = 20/13 \approx 1.538.$$

We set the block size $n = 4$; then $m = [N/n]$ (rounded) = 3; the three blocks are (2 1 2 1) (1 1 2 1) (1 3 1 3).

In the first block, that is, (2 1 2 1),

$$y(t) = (2-1.538)+(1-1.538)+(2-1.538)+(1-1.538) = -0.152.$$

Then a local least squares straight line fit is calculated by minimizing the least squared errors within each block.

$$y_1(t) = 1.8,\ \ y_2(t) = 1.6,\ \ y_3(t) = 1.4,\ \ y_4(t) = 1.2.$$

In the second block,

$$y(t) = (1-1.538)+(1-1.538)+(2-1.538)+(1-1.538) = -1.152,$$ and then

$$y_5(t) = 1.1,\ y_6(t) = 1.2,\ y_7(t) = 1.3,\ y_8(t) = 1.4.$$

And in the third block,

$$y(t) = (1-1.538)+(3-1.538)+(1-1.538)+(3-1.538) = 1.848,$$

$$y_9(t) = 1.4,\ y_{10}(t) = 1.8,\ y_{11}(t) = 2.2,\ y_{12}(t) = 2.6.$$

Finally,

$$F(n) = F(4) = (((-0.152-1.8)\text{^}2+(-0.152-1.6)\text{^}2+(-0.152-1.4)\text{^}2+(-0.152-1.2)\text{^}2+(-1.152-1.1)\text{^}2+(-1.152-1.2)\text{^}2+(-1.152-1.3)\text{^}2+(-1.152-1.4)\text{^}2+(1.848-1.4)\text{^}2+(1.848-1.8)\text{^}2+(1.848-2.2)\text{^}2+(1.848-2.6)\text{^}2)/12)\text{^}1/2 = 1.464.$$

The above calculations are repeated over all time scales (i.e., block sizes n); thus the relationship between $F(n)$, the average fluctuation, as a function of box size n, can be obtained. Then a straight line on this *log-log* graph indicates statistical self-correlation expressed as $F(n) \sim n^{\alpha}$. The scaling exponent α is calculated as the slope of a straight line fit to the *log-log* graph of n against $F(n)$ using least squares method. The exponent α is a generalization of the Hurst exponent. When the exponent is between 0 and 1, the result is Fractional Brownian motion, with the precise value giving information about the series self-correlations: $\alpha < 1/2$ means anticorrelated, $\alpha \approx 1/2$ means uncorrelated, and $\alpha > 1/2$ means correlated; moreover, the closer the value toward 1 the greater the correlation.

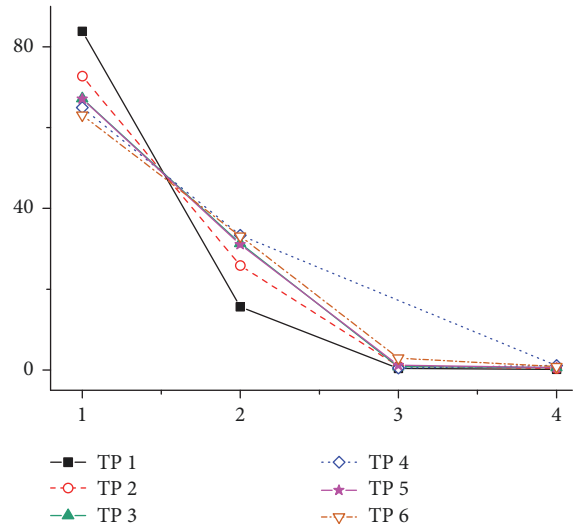

Figure 2: Word length distributions across six time periods (The abscissa represents word length and the ordinate represents percentage of word tokens).

3. Results and Discussion

The results of short/long-range word length correlations are given in this section. The evolution in word length correlation is closely related to that of word length. As a result, this section begins with the evolution of word length distributions and Zipfian *n-gram* word length distributions.

3.1. Evolution of Word Length Distributions and Metrics. The evolution of word length distributions (based on word tokens) is displayed in Figure 2. It should be noted that since words longer than 4 (in Time Periods 5 and 6) have an extremely little share (smaller than 1/‰), we pooled the data for better comparison between different time periods.

As can be seen from Figure 3, the frequency of words whose length is 1 has steadily decreased, and the frequency of words whose lengths are longer than 1 has increased. Table 3 shows the static mean word lengths (SMWL, which is calculated on the basis of the number of word types) and dynamic mean word lengths (DMWL, which is calculated on the basis of the number of word tokens) in different periods.

Figure 3 shows four measures of the distributions shown in Figure 2: (a) entropy, (b) standard deviation, (c) skewness, and (d) kurtosis.

The first measure is entropy. As can be seen in Figure 3(a), the diachronic changes in entropy point to increasing uniformity of word length frequency distribution.

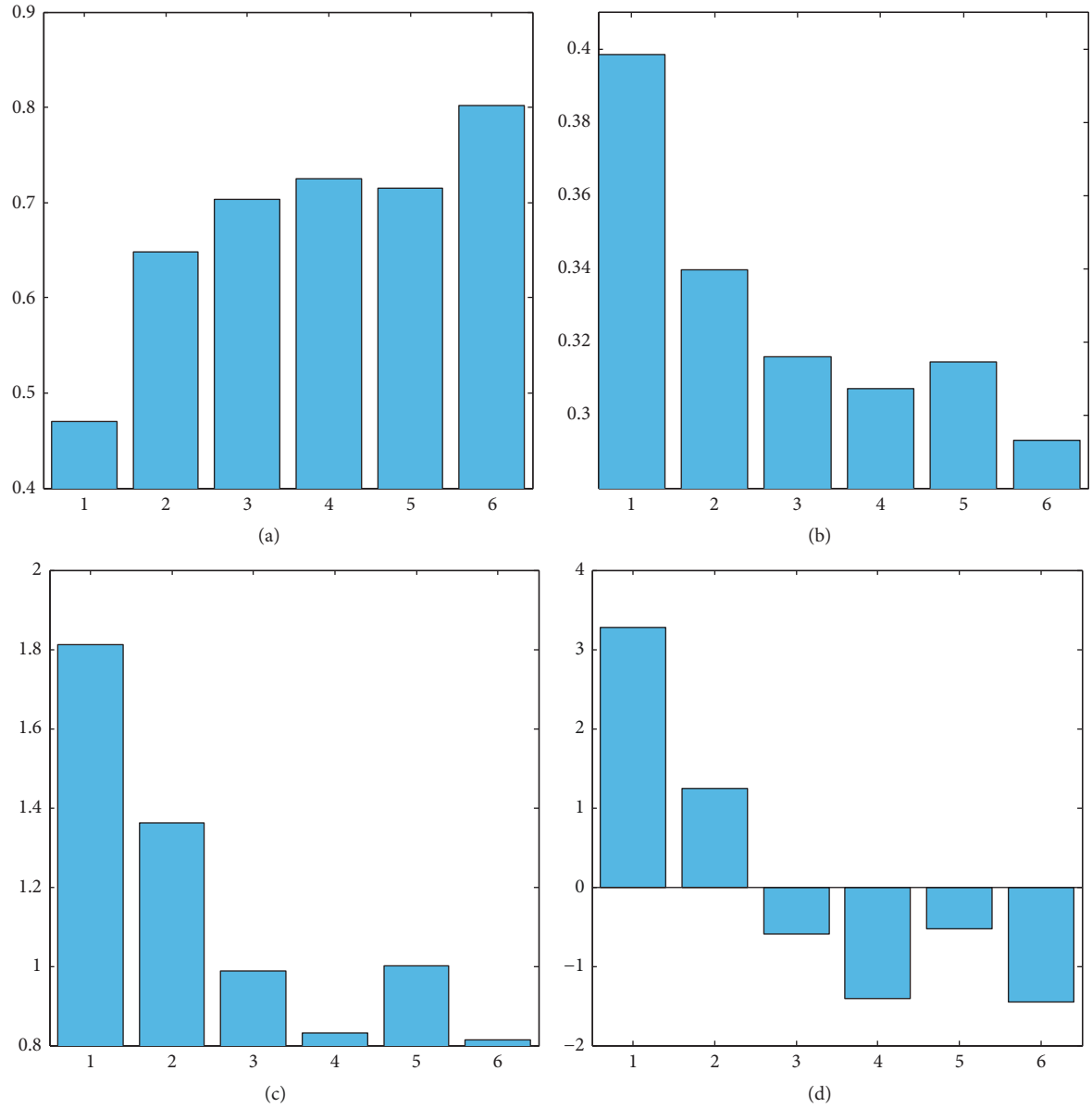

Figure 3: Four indexes changes with age: (a) entropy, (b) standard deviation, (c) skewness, and (d) kurtosis.

Table 3: Evolution of mean word length in written Chinese narrative texts across 2000 years (adapted from [27]).

Type	Time periods					
	1	2	3	4	5	6
SMWL	1.4754	1.6289	1.6808	1.7186	1.7188	1.8139
DMWL	1.1688	1.2909	1.3604	1.3784	1.3531	1.4163

The second measure is standard deviation, which quantifies the variation or dispersion of data. It can be seen from Figure 3(b) that the standard deviation of dynamic word length distributions has diachronically decreased, which means the increase of uniformity of the distributions.

The third measure is skewness, which characterizes the asymmetry of the distribution. If the value of skewness is zero, the distributions are perfectly symmetric. Otherwise, large positive values of skewness indicate the distributions with long tails at the right side of the mean value, and large negative values of skewness indicates the distributions with long tails at the left side of the mean value. It can be seen from Figure 3(c) that the values of skewness have diachronically decreased, which means that the distributions have become more symmetric over time or, rather, that there have been more and more long words.

Lastly, kurtosis is used to measure the degree of data aggregation in the center, which is related to the peakedness and tailedness of the distribution. Large values (kurtosis > 3) suggest distributions with high, steep peaks, and long, thick tails [19]. For normal distribution, the kurtosis coefficient is 0; the positive kurtosis coefficient indicates that the observed

data is more centralized and has a longer tail than the normal distribution; the negative kurtosis coefficient indicates that the data is less centralized and have a shorter tail than the normal distribution. As can be seen in Figure 3(d), there is a high and steep peak in the distribution of Period 1. In the distributions of other periods, it can be seen that the kurtosis value has decreased, which indicates increasingly even distributions.

We can conclude from Figure 3 that, diachronically, there is an increasing use of multicharacter words in Chinese, and the distribution of word with different lengths has become more uniform over time.

Generally speaking, we can see a diachronic trend toward more complexity at the lexical level. At the beginning, words were dominantly monosyllabic. But this dominance has declined unrelentingly throughout the 6 periods, with bisyllabic words increasing rapidly. At the same time, we can see that frequency of words with 3 or more syllables has also slightly increased. These changes lead to evener distributions of word length. Languages usually evolve into more complexity to match human needs. At the lexical level, it is apparently reflected by the increasing word length. The number of monosyllable words was sufficient at a very early stage. However, with the evolution of civilization, there has been constant demand for new words. Monosyllable words cannot be coined at will, because, on one hand, Chinese are ideographic, and on the other hand, the biphoneme combinations are limited. So the easiest way out is to combine two monosyllable words into a single bisyllable word, which have rapidly increased to become the most frequent in the second period. Theoretically, the frequently used monosyllable words in Chinese may provide more than forty million different bisyllable words, which probably suffice the needs for complex communication. However, according to our statistics of Lancaster Corpus of Mandarin Chinese (http://ota.oucs.ox.ac.uk/scripts/download.php?otaid=2474) [38], only about 0.2294% of potential bisyllable words are valid, that is, actually used. Of course, language users have diverse needs and constraints, not simply the task of information encoding, and as a result, longer words have also slightly increased [39]. However, it is clear that the increase of words with 3 or more syllables is much slower than that of bisyllable words, which probably have matched the increasing complexity of communication. The decrease in the frequency of monosyllable words and the increase in that of longer words, especially of the bisyllable words, naturally lead to evener distributions of word length.

Our results indicate that the increase of word length is an essential regularity of word evolution in written Chinese. Bochkarev et al. [40] claim that sociocultural factors may have influence on the changes of mean word length in a relatively short time period. Our results suggest that the increase of mean word length can also be a primary result of word evolution in the long run. The disyllabic trend in Chinese word evolution is highlighted in previous studies, e.g., Packard [39]. However, the modern Chinese Vernacular Movement in the early 20th century has changed this language to some extent, resulting in simplifications in grammar and Europeanized sentences [27, 41]. The simplification and the Europeanization in Chinese grammar may account for, to a considerable degree, why mean word length in TP 5 slightly decreases. As suggested by Bochkarev et al. [40], sociocultural factors may have influence on the changes of mean word length in a relatively short time period; therefore, this phenomenon may be interpreted as a result of the modern Chinese Vernacular Movement.

Although Chinese word evolution can be influenced, to some extent, by social-cultural aspects, there seems to be a persistent trend for word length to increase in evolution. As suggested by many studies, the primary reason may be that, with social developments, the increase of word length is inevitable because it is more efficient to express new meanings through combining existent words than to coin new ones. This result corroborates the "error limit" theory put forward by Nowak et al. [42], which claims that the increase of new concepts results in longer words.

On the whole, the increase of word length adds great redundancy to the lexical system: only about 0.2294% potential bisyllable words are valid. Therefore, in the lexical system, along with the increase of word length, the tones and the characters all become simplified for balance [41]. The simplification of Chinese characters is continuous and systematic despite the fact that some characters may become complex, such as "上, 王," which can be distinguished from "二, 玉" especially when they are handwritten now. These changes in Chinese lexical system indicate that Zipf's "principle of least effort" and the "error limit" theory play critical roles in human communications.

3.2. Evolution of n-Gram Word Length Entropies and Short-Range Word Length Correlations. In above section, we found that word length distributions of the six periods largely differ in the tail, that is, the distribution of long words. In this section, we focus on not only the distributions of lengths of individual words, but also the distributions of n-gram word length sequences in sentences.

Figure 4 displays the Zipfian distributions of n-gram word length sequences in six periods.

It can be seen from Figure 4 that the differences of n-gram word lengths of different time periods also lie in high ranks, and they evolve gradually with time. However, as for low ranks, what they have in common is that the most frequent n-grams in all six time periods are "11," "111," and "1111."

The n-gram word length entropies of six periods as well as their differences to the shuffled ones can be seen in Table 4. It should be noticed that all the values are normalized to [0-1].

From Table 4, we can see that the 2-gram, 3-gram, and 4-gram word length entropies of written Chinese has increased over time, which means that the distributions of n-gram word length sequences tend to be uniform. This tendency indicates the diversification of word length collocations in sequences, which may bear on the diachronic increase of word length in Chinese [27].

The changes of Relative Entropy D across 6 periods are presented in Figure 5.

We can see from Figure 5 that there is no evident linear correlation between D and different periods. Furthermore, Pearson Correlation analysis is employed to examine if there

Table 4: Evolution of 2-, 3- and 4-gram word length entropies compared with the entropies of the shuffled texts (with normalized values).

Time period	2-gram			3-gram			4-gram		
	$H1$	$Hs1$	$D1$	$H2$	$Hs2$	$D2$	$H3$	$Hs3$	$D3$
1	**0.00**	**0.00**	0.12	**0.00**	**0.00**	0.04	**0.00**	**0.00**	0.10
2	0.52	0.53	**1.00**	0.52	0.53	0.85	0.51	0.53	0.88
3	0.68	0.69	0.97	0.68	0.69	0.75	0.68	0.70	0.68
4	0.75	0.76	0.90	0.74	0.76	**1.00**	0.74	0.76	**1.00**
5	0.73	0.73	0.01	0.73	0.73	**0.00**	0.73	0.73	**0.00**
6	**1.00**	**1.00**	**0.00**	**1.00**	**1.00**	0.08	**1.00**	**1.00**	0.07

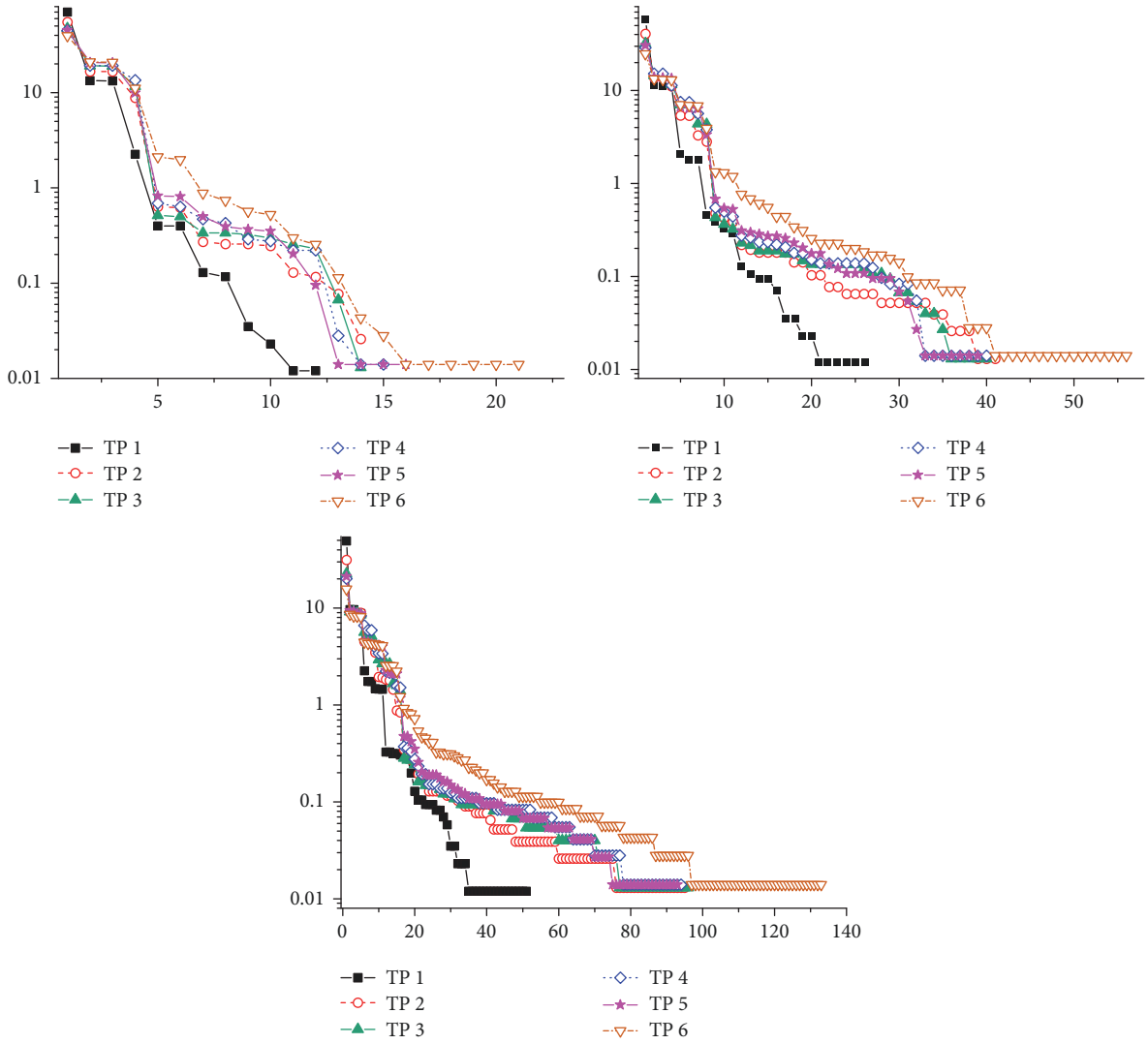

Figure 4: Zipfian distribution of 2-, 3-, and 4-gram word length block frequencies (the abscissa represents rank and the ordinate represents log frequency).

is a relationship between D and SMWL but results in no significant correlation.

3.3. Evolution of Hurst Exponents and Long-Range Word Length Correlations. In stochastic processes, chaos theory and time sequence analysis, detrended fluctuation analysis (DFA) is a method to analyze the long-range correlation in time series. One advantage of the DFA is that it can effectively filter out the trend in the sequence and is suitable for long-range power law correlation analysis of nonstationary time series. The exponents of six periods for blocks whose sizes n range from 10 to 1000 are, respectively, 0.5950, 0.5837, 0.5646, 0.5449, 0.5662, and 0.5828, and no obvious linear trend can be detected.

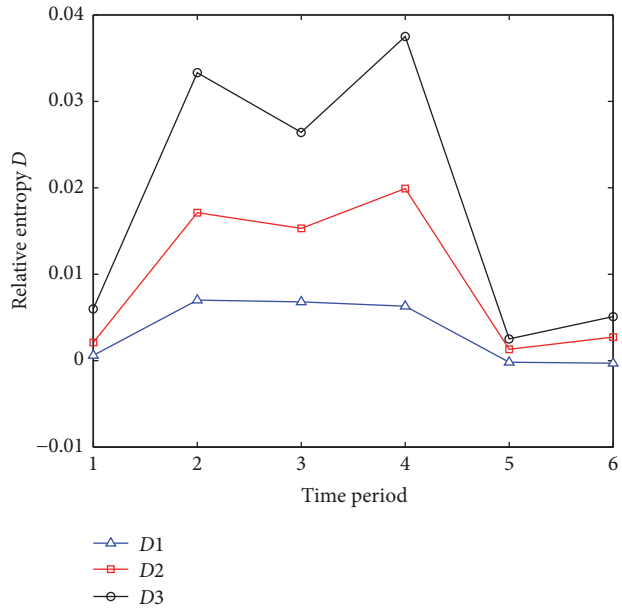

Figure 5: Relative Entropy D of different periods ($D1$ refers to the Relative Entropy of 2-gram word length block, and so forth).

The power law distribution $F(n)\sim n^{\alpha}$, of long-range word length correlations, when n ranges from 10 to 100, is somewhat different from the distribution when n ranges from 101 to 1000. The distributions can be seen in Figure 6. The small-scale long-range (with n ranging from 10 to 100) word length sequences correlations roughly reflect correlations at sentence or paragraph level; the large-scale long-range (with n ranging from 101 to 1000) word length sequences correlations roughly correspond to correlations at level of a paragraph or a section.

Figure 6 indicates power law distributions in all six periods, with quite high R^2. For all 6 periods, the Hurst exponents of two regimes of are between 0.5~0.7, indicating, despite different periods, weak but stable long-range word length correlations. We fit linear functions to the exponents of the two regimes, which is presented in Figure 7.

As can be seen in Figure 7, in terms of small-scale, the long-range correlation seems to have decreased over time, while in terms of large-scale, it seems to have increased. This may have to do with the tendency toward more complexity in the evolution of languages. Sentence length of Chinese has increased diachronically, usually exceeding 20 or 30 today. In other words, small-scale blocks sometimes may not include an entire sentence, which may diachronically become more frequent with increasing complexity of sentences. However, large-scale blocks normal can include complete sentences. Word length has much with syntactic properties of words, which are used in regular syntactic patterns. Such patterns can be captured at the level of sentence. So the repeated patterns will be more likely to appear in blocks containing complete sentences. This might be one reason for the two trends observed in Figure 7.

Similarly, we test if the Hurst exponents (two regimes, i.e., Slope 1 and Slope 2) in long-range correlations in Figure 7 correlate with SMWL. Pearson Correlation analyses are again employed, and a significant correlation coefficient is obtained at the 0.01 level (−0.957, 2-tailed) for Slope 1 and SMWL, but no significant one between Slope 2 and SMWL. This indicates that word length evolution may have more influence on the small regime scale first.

4. Conclusions

In this paper, we investigate the diachronic changes of word length correlations in Chinese narrative texts. The diachronic changes in entropy, standard variance, skewness, and kurtosis indicate that the Chinese word length has been steadily growing and multisyllable (character) words have been increasing. As suggested by many studies, the primary reason may be that as society develops, the increase of word length is inevitable because it is more efficient to express new meanings through combining signs rather than creating new signs. The increase of new concepts results in longer words, which also corroborates the "error limit" theory.

In addition, increase of n-gram word length entropies means evener distributions of different n-gram word length sequences, that is, the more diversification of word length collocation patterns, which may bear on the diachronic increase of word length in Chinese. Moreover, we observe that as n (in n-gram) becomes larger, the values of D also increase, which means that word length correlations increase with the distance between words.

However, in terms of N-gram word length sequences, the exponent does not present consistent increasing or decreasing short-range word length correlations. In contrast, two opposite tendencies have been found in long-range word length correlations: the exponent of small-scale (block size n ranges from 10 to 100) word length correlations seems to have decreased over time, while the exponent of large-scale (block size n ranges from 101 to 1000) seems to have increased over time. This phenomenon may have much to do with the evolutionary tendency toward long and complex sentence. Nevertheless, the exponent from large-scale is greater than its counterparts except in Time Period 1. Chen et al. [27] tests show that there are two jumps in the evolutionary process of word length: one is from TP 1 to TP 2, and the other is from TP 2 to TP 6. We speculate that the shortest word length and sentence length in TP 1 may account for the exception. In TP 1, basically, one character/syllable corresponds to one words, and the sentences are extremely short [27].

Lastly, for the relationship between mean word length and short-range word length correlations, the Pearson Correlation analysis shows that there is no significant correlation between them. However, for the two regimes in long-range correlations, two opposite cases are found: a significant correlation coefficient is obtained at the 0.01 level for Slope 1 (of small-scale) and SMWL, but no significant one between Slope 2 (of large-scale) and SMWL. Thus, we speculate that word length evolution may have influence on word length correlations in small regime scale first. In a forthcoming paper [43], we find that the lexical word cooccurrence (i.e., 2-gram words) network in written Chinese evolves to be greater in global level. And the connections of words in microlevel

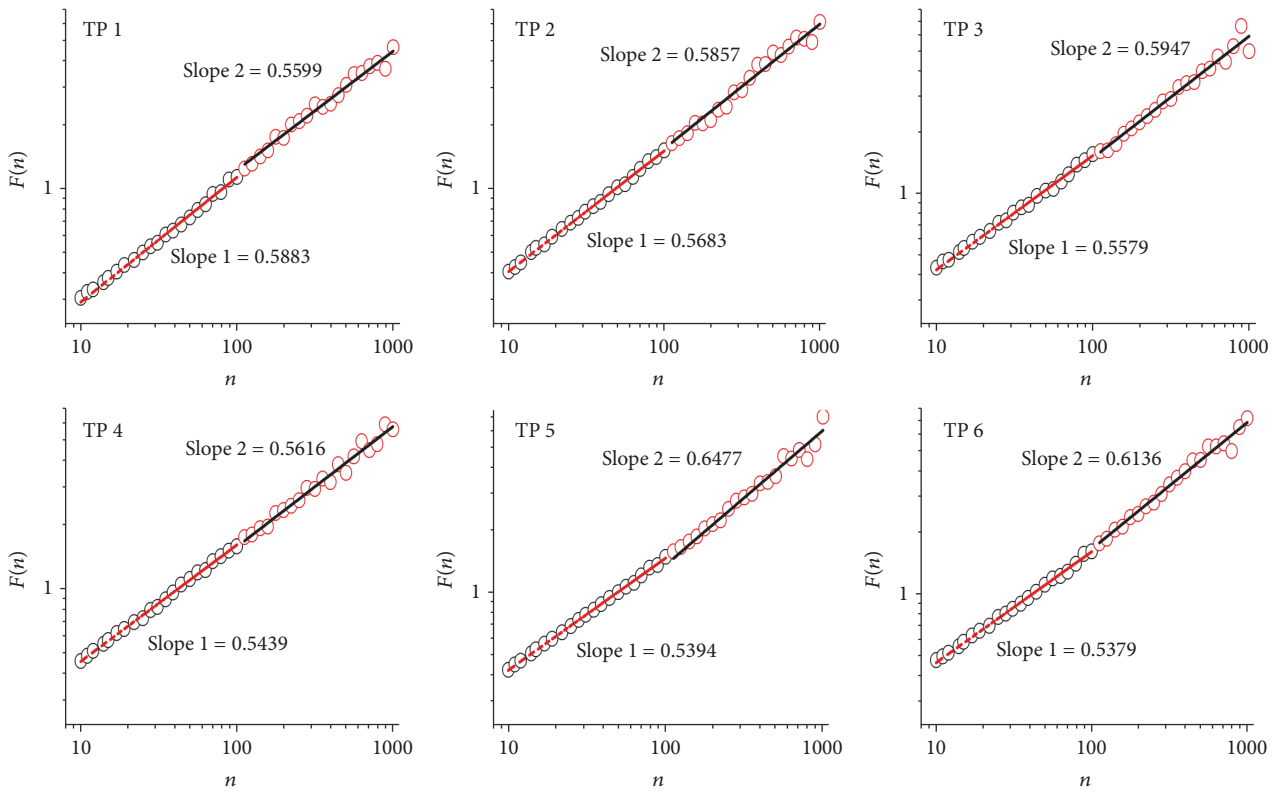

Figure 6: DFA analyses of Chinese word length sequence across six time periods.

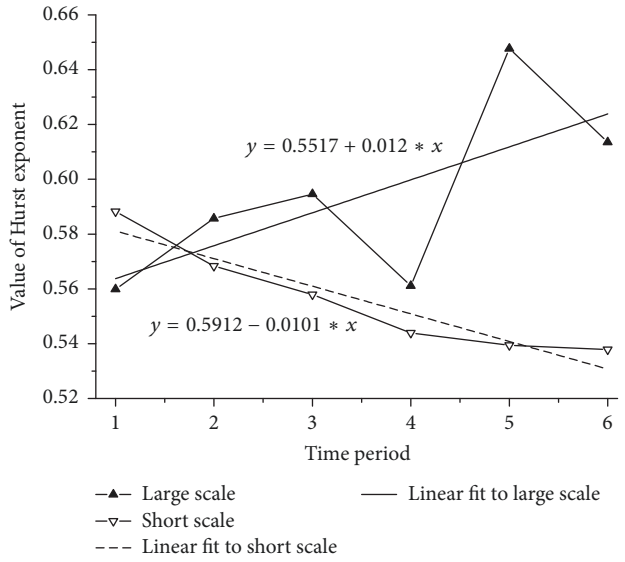

Figure 7: Evolution of two regimes of long-range word length correlations in written Chinese.

are continually weakening; the number of words in mesolevel communities increased significantly. This means that more and more words tend to be connected to the medium-central words and form different communities. These indicate that the connections of words among sentences or larger levels are expanding. Nevertheless, the deep interrelationships between them still need investigating.

This study also has room for future improvement. In this study, different methods are used to measure short and long-range word length correlations, which makes it hard to compare the values between them directly. Moreover, further researches are needed in the future to clarify if the findings in this study could generalize to other languages with different lexical systems, and if the results could be influenced by boundary conditions such as genres, registers, and text scales.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors thank Chunshan Xu for polishing the article. This study was partly supported by the National Social Science Foundation of China (Grant no. 17AYY021 and Grant no. 12&ZD224), the MOE Project of the Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies, and the Fundamental Research Funds for the Central Universities (Program of Big Data PLUS Language Universals and Cognition, Zhejiang University).

References

[1] M. A. Nowak, J. B. Plotkin, and V. A. Jansen, "The evolution of syntactic communication," *Nature*, vol. 404, no. 6777, pp. 495–498, 2000.

[2] R. V. Solé, B. Corominas-Murtra, S. Valverde, and L. Steels, "Language networks: Their structure, function, and evolution," *Complexity*, vol. 15, no. 6, pp. 20–26, 2010.

[3] S. Pinker, *The Language Instinct*, Harper Collins, New York, NY, USA, 2000.

[4] W. A. Kretzschmar, *The Linguistics of Speech*, Cambridge University Press, 2009.

[5] D. L. Freeman and L. Cameron, "Research methodology on language development from a complex systems perspective," *Modern Language Journal*, vol. 92, no. 2, pp. 200–213, 2008.

[6] À. Massip-Bonet and A. Bastardas-Boada, Eds., *Complexity Perspectives on Language, Communication and Society*, Springer, 2012.

[7] Q. Lu, C. Xu, and H. Liu, "Can chunking reduce syntactic complexity of natural languages?" *Complexity*, vol. 21, pp. 33–41, 2016.

[8] N. Chomsky, *Syntactic Structures*, The Hague, Paris, France, 1957.

[9] S. Miyagawa, R. Berwick, and K. Okanoya, "The emergence of hierarchical structure in human language," *Frontiers in Psychology*, vol. 4, p. 71, 2013.

[10] H. Liu, C. Xu, and J. Liang, "Dependency distance: A new perspective on syntactic patterns in natural languages," *Physics of Life Reviews*, vol. 21, pp. 171–193, 2017.

[11] R. Köhler, "Synergetic linguistics," in *Quantitative linguistics. An international handbook*, R. Köhler, G. Altmann, and R. G. Piotrowski, Eds., pp. 760–774, de Gruyter, Berlin, Germany, 2005.

[12] K.-H. Best, *The Distribution of Word and Sentence Length*, Glottometrika 16, Wissenschaftlicher Verlag Trier, Trier, Germany, 1997.

[13] G. Wimmer, R. Köhler, R. Grotjahn, and G. Altmann, "Towards a Theory of Word Length Distribution," *Journal of Quantitative Linguistics*, vol. 1, no. 1, pp. 98–106, 1994.

[14] D. Garcia, A. Garas, and F. Schweitzer, "Positive words carry less information than negative words," *EPJ Data Science*, vol. 1, no. 1, pp. 1–12, 2012.

[15] R. Köhler, "The frequency distribution of the lengths of length sequences," in *Favete linguis. Studies in honour of Viktor Krupa*, J. Genzor and M. Bucková, Eds., pp. 145–152, Slovak Academic Press, Bratislava, Slovakia, 2006.

[16] P. Grzybek, "History and methodology of word length studies," in *Contributions to the science of text and language: Word length studies and related issues*, vol. 31, pp. 15–90, Springer, Dordrecht, The Netherlands, 2006.

[17] H. Chen and H. Liu, "How to Measure Word Length in Spoken and Written Chinese," *Journal of Quantitative Linguistics*, vol. 23, no. 1, pp. 5–29, 2016.

[18] R. Köhler, "Word length in text. A study in the syntagmatic dimension," in *Jazyk a jazykoveda v pohybe*, S. Mislovicová, Ed., pp. 416–421, Veda, Bratislava, Slovakia, 2008.

[19] M. Kalimeri, V. Constantoudis, C. Papadimitriou, K. Karamanos, F. K. Diakonos, and H. Papageorgiou, "Word-length entropies and correlations of natural language written texts," *Journal of Quantitative Linguistics*, vol. 22, no. 2, pp. 101–118, 2015.

[20] W. Ebeling and T. Pöschel, "Entropy and long-range correlations in literary English," *EPL (Europhysics Letters)*, vol. 26, no. 4, pp. 241–246, 1994.

[21] S. Drożdż, P. Oświęcimka, A. Kulig et al., "Quantifying origin and character of long-range correlations in narrative texts," *Information Sciences*, vol. 331, pp. 32–44, 2016.

[22] M. Kalimeri, V. Constantoudis, C. Papadimitriou, K. Karamanos, F. K. Diakonos, and H. Papageorgiou, "Entropy analysis of word-length series of natural language texts: Effects of text language and genre," *International Journal of Bifurcation and Chaos*, vol. 22, no. 9, Article ID 1250223, 2012.

[23] E. G. Altmann, J. B. Pierrehumbert, and A. E. Motter, "Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words," *PLoS ONE*, vol. 4, no. 11, Article ID e7678, 2009.

[24] H. Chen and J. Liang, "Chinese word length motif and its evolution," in *Motifs in Language and Text*, H. Liu and J. Liang, Eds., De Gruyter, Berlin, Germany, 2017.

[25] W. G. Boltz, *The Origin and Early Development of the Chinese Writing System*, vol. 78, American Oriental Society, New Haven, Conn, USA, 1994.

[26] V. H. Mair, Ed., *The Columbia History of Chinese Literature*, Columbia University Press, New York, NY, USA, 2001, The Columbia History of Chinese Literature (Columbia University Press.

[27] H. Chen, J. Liang, and H. Liu, "How does word length evolve in written Chinese?" *PLoS One*, vol. 10, no. 9, Article ID E0138567, 2015.

[28] W. Ebeling and A. Neiman, "Long-range correlations between letters and sentences in texts," *Physica A: Statistical Mechanics and its Applications*, vol. 215, no. 3, pp. 233–241, 1995.

[29] L. Guzmán-Vargas, B. Obregón-Quintana, D. Aguilar-Velázquez, R. Hernández-Pérez, and L. S. Liebovitch, "Word-length correlations and memory in large texts: A visibility network analysis," *Entropy*, vol. 17, no. 11, pp. 7798–7810, 2015.

[30] T. Yang, C. Gu, and H. Yang, "Long-range correlations in sentence series from a story of the stone," *Plos One*, vol. 11, no. 9, Article ID e0162423, 2016.

[31] M. A. Montemurro and P. A. Pury, "Long-range fractal correlations in literary corpora," *Fractals*, vol. 10, no. 4, pp. 451–461, 2002.

[32] J. Bhan, S. Kim, J. Kim, Y. Kwon, S.-I. Yang, and K. Lee, "Long-range correlations in Korean literary corpora," *Chaos, Solitons & Fractals*, vol. 29, no. 1, pp. 69–81, 2006.

[33] M. A. Montemurro and D. H. Zanette, "Universal entropy of word ordering across linguistic families," *PLoS ONE*, vol. 6, no. 5, Article ID e19875, 2011.

[34] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. 27, no. 4, pp. 623–656, 1948.

[35] G. Şahin, M. Erentürk, and A. Hacinliyan, "Detrended fluctuation analysis in natural languages using non-corpus parametrization," *Chaos, Solitons & Fractals*, vol. 41, no. 1, pp. 198–205, 2009.

[36] I.-I. Popescu, G. Altmann, J. Grzybek et al., *Word Frequency Studies*, Mouton de Gruyter, Berlin, Germany, 2009.

[37] C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger, "Mosaic organization of DNA nucleotides," *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, vol. 49, no. 2, pp. 1685–1689, 1994.

[38] H. Liu, *An Introduction to Quantitative Linguistics*, The Commercial Press, Beijing, China, 2017.

[39] J. L. Packard, Ed., *New Approaches to Chinese Word Formation: Morphology, phonology and the lexicon in modern and ancient Chinese*, vol. 105, Walter de Gruyter, Berlin, Germany, 1998.

[40] V. Bochkarev, A. Shevlyakova, and V. Solovyev, "Average word length dynamics as indicator of cultural changes in society," https://arxiv.org/abs/1208.6109.

[41] L. Wang, *A Manuscript of Chinese Language History*, The Chinese Publishing House, Beijing, China, 2004.

[42] M. A. Nowak, D. C. Krakauer, and A. Dress, "An error limit for the evolution of language," *Proceedings of the Royal Society B Biological Science*, vol. 266, no. 1433, pp. 2131–2136, 1999.

[43] H. Chen, X. Chen, and H. Liu, "How does language change as a lexical network? An investigation based on written Chinese word co-occurrence networks," *PloS One*, 2018, In Press.