Hindawi Publishing Corporation
Journal of Applied Mathematics
Volume 2014, Article ID 123983, 9 pages
http://dx.doi.org/10.1155/2014/123983

Research Article

Reconstruction of Uncertain Historical Evolution of the Polysyllablization of Chinese Lexis

Bing Qiu[1] and Jie Li[2]

[1] *College of Humanities and Social Sciences, Beijing Language and Culture University, Beijing 100083, China*
[2] *College of Software, Henan University, Kaifeng, Henan 475001, China*

Correspondence should be addressed to Bing Qiu; bingqiu@gmail.com

Received 9 February 2014; Revised 17 June 2014; Accepted 25 June 2014; Published 13 July 2014

Academic Editor: Guiming Luo

Polysyllablization, closely related to phonetics, semantics, and syntactics, is one of the fundamental trends in the development of Chinese lexis. However, with lots of uncertainties in the historical evolution of Chinese language, the quantitative modeling and reconstruction of polysyllablization remain open questions. Based on the *Comprehensive Dictionary of Chinese Words*, a mapping from the words to their time of occurrence is built. With the inverse mapping on random samples, the newly produced words with different numbers of syllables in different time periods are obtained. Finally the total quadratic variation minimization model is adopted to estimate the trend of polysyllablization. As a novel exploration in the computational linguistics, the results agree with the stage division of historical Chinese and answer some difficult questions related to polysyllablization in a quantitative manner.

1. Introduction

Language is always changing [1]. The change of language is variation over time in a language's lexical, phonetic, syntactic, and other features. Old English (from the mid-5th century to the mid-11th century), for instance, would be greatly different from Modern English (from the late 15th century to the present) [2]. Language changes in many and varied ways. It is interesting to note that lexical change is one of the most obvious and important types of language change. One may observe that new words are formed and old words are tagged as "obsolete." Slang terms, in particular, come and go every few years. Consequently, as far as a long historical period is concerned, the variation of lexis is a complicated issue, which is not only fundamental to historical lexicology but also related to many research fields ranging from historical linguistics [3], onomasiology, etymology, to sociolinguistics [4].

Polysyllablization is one of the fundamental trends in the development of Chinese lexis. It is a prevailing view nowadays in the academic circle that there are three stages in the evolution of Chinese [5–8], that is, Old Chinese, Middle Chinese, and Modern Chinese. Old Chinese was the commonly used language during the early and middle Zhou Dynasty [9] (around 1046–256 BC). Middle Chinese was the historical Chinese dialect which was phonologically recorded in *Qieyun* [10, 11], a rime dictionary first published in 601. Modern Chinese is mainly referred to as the form of Chinese varieties to the present. However, the boundaries of these stages are ambiguous. The periods between them are transitional phases. The vocabulary system of Old Chinese mainly consists of monosyllabic (i.e., single syllable) words, whereas that of Modern Chinese is mainly made up of polysyllabic (i.e., two or more syllables) words. The aforementioned evolution of Chinese lexis is called polysyllablization. Since each Chinese character represents a monosyllabic Chinese word or a morpheme, polysyllablization in the historical evolution of Chinese not only introduces the form expansion of words but also represents the combination of morphemes and meanings and depicts the coordination of grammatical relations, which reflect the inherent laws of Chinese phonetics, semantics, and syntactics. In addition, the importance of syllables in Chinese lies in their link to word recognition [12, 13]. Therefore, polysyllablization is one of the most principal research topics of Chinese language, which has attracted lots of attention from the academia.

Most of the traditional studies on the polysyllablization of Chinese lexis are qualitative, of which the conclusions are mainly based on personal experiences. In fact, the difference in the number of syllables between Old Chinese words and Modern Chinese words is obvious and easy to be noticed. However, the following questions are not easy to answer based on the qualitative conclusions.

(i) What was the quantitative degree of polysyllablization during a certain period, for example, Qin or Han Dynasty? When did the polysyllabic words begin to occupy a major position in the vocabulary system?

(ii) What was the speed of polysyllablization at a certain time point? When was the lexis polysyllablized most rapidly?

(iii) How were the evolving trends for different kinds of polysyllabic words (i.e., disyllabic words, three-syllable words and so forth)?

As a result, the related studies in recent years have put a particular emphasis on the adoption of quantitative methods. They usually focus on a specific range of language materials, for example, a specific chapter, a specific book, or all the books of a specific author. For these limited language materials, the numbers of monosyllabic words, polysyllabic words, and total words have been counted. The ratio of the number of polysyllabic words to that of total words, namely, polysyllabic word ratio, is then adopted to evaluate the degree of polysyllablization. In a sense, polysyllablization is studied in a quantitative manner. Although the statistics on the polysyllabic words in a specific range of language materials are closely related to polysyllablization, it is not the polysyllablization itself. The polysyllabic word ratio of a specific range of language materials seriously depends on the selected topic, the author's habit, and the length of the corresponding text materials, so it is only a local weight of polysyllabic words and cannot represent the global degrees of polysyllablization. However, it is indeed very difficult to count all words of various numbers of syllables through different time periods for the purpose of depicting the trend of polysyllablization.

The lexis, that is, the vocabulary system, of a language is related to many complex and dynamic features with lots of uncertainties. It is often very difficult to construct a mathematical model. To the best of our knowledge, the quantitative evaluation and modeling of the polysyllablization of Chinese are still open questions.

Hanyu Da Cidian [14] (literally *Comprehensive Dictionary of Chinese Words*, abbreviated as *CDCW* in the following text) is the most inclusive Chinese dictionary available. It has a diachronic coverage of the Chinese language, tracing its usage over three thousand years from Chinese classic texts to modern slangs. Researchers use *CDCW* as a reference book to search and examine specific words and develop related lexicological studies. At the macrolevel, however, the enormous amounts of information in *CDCW* have not yet been effectively mined. To draw an analogy, if we compare *CDCW* to a building, the existing studies are focusing on the bricks; however, the structure of the building as a whole has been neglected. The data mining of *CDCW* provides us with a novel approach to tackle the polysyllablization of Chinese language.

This paper aims to answer the aforementioned open questions related to polysyllablization. In particular, our contribution lies mainly in the following aspects.

(i) We propose a novel approach based on *CDCW* to study polysyllablization of the Chinese lexis in a quantitative and impersonal manner. The words in *CDCW*, as a whole, are regarded as the maximum vocabulary so far. We can obtain the time of occurrence for each word in *CDCW*. Once a mapping from the words to their time of occurrence is built, the inverse mapping leads us to the newly produced words with different numbers of syllables in different time periods.

(ii) We introduce several techniques in the practical workflow to handle various uncertainties, including the repetitions and exceptions in the entries of *CDCW* and inaccuracy of mapping from words to their time of occurrence. The statistical sampling method is adopted to avoid the huge workload to process hundreds of thousands of words. The total quadratic variation minimization model is adopted to estimate the trend of polysyllablization.

(iii) We obtain the statistical data and finally present the trend of polysyllablization of the Chinese lexis. We also discuss the results through comparison with the existing conclusions. As a novel exploration in the computational linguistics, especially for the Chinese language with such a long history and such a complex evolution, our approach is valuable to similar issues.

The rest of this paper is organized as follows: Section 2 briefly reviews related works. Section 3 introduces the formulation and the mathematical model. Section 4 presents the results and discusses the trends of polysyllablization. Finally, Section 5 concludes this paper.

2. Related Works

Since polysyllablization is among the most important rules in the development of Chinese vocabulary, in recent years, studies on Chinese polysyllabic words and polysyllablization have achieved rich results. For example, the syllabic structure of Mandarin Chinese is discussed in [15] based on the X-bar approach. From the viewpoint of the Buddhist and Taoist scriptures of Eastern Han Dynasty, the polysyllablization of Chinese vocabulary based on the new lexical items is discussed in [16]. Since most of polysyllabic words are disyllabic words, the development of disyllabic words in Chinese is discussed in [17]. Also, many quantitative surveys and analyses have been conducted on polysyllabic words in different monographs across different time periods. According to the statistics in [18], published papers on the vocabulary in Middle Chinese and Early Modern Chinese monographs have amounted to over 6,000. Generally, without uniform standards, scholars differ from each other on their subjective definitions of polysyllabic words; as a result, they have

obtained different statistical results about the number of the polysyllabic words in the same monograph. Take *Shi Shuo Xin Yu* (literally *A New Account of the Tales of the World*) for example, the number of polysyllabic words in it was said to be around 1,500 to 2,100 in different conclusions. Besides, due to the limitations of the subject, content, length of the book, and the authors mastery over the language, the vocabulary in one monograph alone cannot fully represent the complete picture of the its contemporary vocabulary system; as a matter of fact, sometimes there could be a great deviation. In other words, monographs are theme-based and scope-limited, so they can only reflect the vocabulary system from a restricted view rather than represent the whole. So far, many questions still remain unanswered, such as how polysyllablization evolved in the history of the Chinese vocabulary system.

CDCW is among the most important dictionaries in Chinese lexicological studies. Since the publication of its first volume in 1986, it has been put into wide academic application and achieved plentiful research results. *CDCW* has 50 million characters, containing 22.7 thousand Chinese characters and 375 thousand polysyllabic words. Wenkan Xu, one of the editorial board members of *CDCW*, said "Before the publishing of *CDCW*, there was no such a dictionary which contains both modern and ancient words, as well as words that were developed intermediately, and proves itself to be a confluence of the entire Chinese vocabulary, for the purpose of search and reference" (translated from Chinese according to [19]). Thus it is feasible for researchers either to observe the change of a single word from a diachronic perspective or to observe the semantic structure of a group of words from a synchronic perspective. Additionally, such a large-scale historical Chinese dictionary, which collects, organizes, and explains hundreds of thousands of ancient and modern words, is a very valuable corpus itself. Corpus linguistics is one of the most important fields nowadays [20]. *CDCW* is undoubtedly a huge corpus, and the reasonable utilization of *CDCW* will probably lead us to a novel approach, excluding the subjective prejudice of the researchers, to tackle the polysyllablization of Chinese.

3. Formulation and Mathematical Model

The vocabulary system has two characteristics, namely, integrity and dynamics. Integrity means that the vocabulary system is a macrolevel concept and refers to the sum of all the words in the language. Dynamics means that the vocabulary changes over time. Equivalently, new words appear and old words fade. Thus, in order to illustrate the polysyllablization of Chinese lexis, two key issues must be solved in accordance with these two characteristics of the vocabulary system. The first issue is to obtain the complete set of the words from Old Chinese to Modern Chinese. The second one is to determine the first-appearing and last-appearing time points of each word.

Unlike the English text in which sentences are sequences of words delimited by spaces, the word boundary in Chinese is fuzzy. A Chinese word may consist of one, two, or more Chinese characters (also referred to as *Hanzi*). There is no immediate way of deciding which characters in the text should be grouped into words. The studies on word segmentation have attracted lots of attention from the academia [21–23]. Once the words are obtained, the numbers of syllables are easy to be counted.

To assert the first-appearing and last-appearing time points of each word is complicated. A group of Chinese characters is possibly treated as a word at some time, but not considered as a word at another time. There are also various criteria to decide whether a group of characters is a word, most of which are personally prejudiced with lots of uncertainties.

We introduce a novel method to solve the aforementioned issues and analyze the polysyllablization of Chinese lexis in an impersonal and quantitative manner based on the data mining of *CDCW*. Briefly, we build a mapping from the words to their time of occurrence. The inverse mapping is used to count the newly formed words over different time periods. However, there are still lots of uncertainties to deal with. Thus we introduce a total quadratic variation minimization model and some related techniques to estimate the polysyllablization of Chinese lexis. The aforementioned issues finally lead to a constrained quadratic programming problem. The details are as follows.

3.1. Basic Idea. *CDCW* is the archive of Chinese vocabulary over time, containing 22.7 thousand Chinese characters (most of them are monosyllabic words) and 375 thousand polysyllabic words. Based on the entries of *CDCW*, the whole set of Chinese words over time can be obtained.

The structure of hundreds of thousands of entries in *CDCW* contains rich information and reflects many lexicological rules when examined at the macrolevel. Each entry has multiple properties, such as its pronunciation and explanation, which are directly listed in *CDCW*. In fact, more properties can be obtained indirectly with additional procedures. Among the indirect properties, the time of occurrence, which indicates when the word emerged, is of great significance to the studies of historical lexicology.

In *CDCW*, the definition of an entry usually contains several terms of meanings, most of which can be traced back to their earliest reference through the documentary evidences provided by *CDCW*. As long as the approximate time when the documentaries were published or the years when their authors lived are possible to be acquired, we will be able to date all the documentaries available for each term of meanings for the word entry, the earliest one of them reveals the time of occurrence of the word, that is, the time when the word emerged. Note that *CDCW* is a result of collective wisdom of more than 1000 scholars. Thus it offers us an authoritative and impersonal manner to deduce the time of occurrence for each word.

Here is an example to illustrate the procedure from the entry to its time of occurrence. For the word entry *Senlin* (Chinese word, literally, forest; here Chinese characters *Sen* and *Lin* both mean trees or woods), there is only one term of meaning. The earliest documentary evidence is a poem in Tang Dynasty. Thus, it is deduced that the word *Senlin* was composited during Tang Dynasty. We also know the author,

Xiji Cai, served as a junior officer in Luoyang City in 748. However, we know neither the exact year when he wrote the poem nor the birth year of him.

The inaccuracy in defining the publishing time of the ancient documentaries is not occasional. In fact, limited by the records of ancient history literatures, the exact years of the publication of ancient documentaries or even the exact living years of the related authors are difficult to resolve. For most of ancient documentaries, we only know the dynasties of their publication.

Figure 1 is a diagram showing how to investigate the time of occurrence entry by entry in *CDCW*. Suppose that the earliest documentary evidence for Entry 1 is found in Wei Dynasty, then its time of occurrence can be determined to be Wei Dynasty and it will be put under the corresponding chronological category. The same procedure applies to the rest of the entries. In theory, following this method, the time of occurrence for all entries can be determined exhaustively.

The above process is to determine the time of occurrence per word entry by exploring its earliest documentary evidence. However, in order to observe the dynamic evolution of the vocabulary system at the macrolevel, it is necessary to conduct another survey from the opposite direction, that is, to investigate words which occurred in the same time period. By gathering all words which share the same time span of occurrence in *CDCW*, we can obtain all the new words that emerged in the corresponding time period.

The evolution of the vocabulary system is a metabolic process. On one hand, new words have been emerging constantly and entering the vocabulary system; on the other hand, some old words have been withering away and falling out of the vocabulary system. The emergence of new words usually reflects the reform of the society, the occurrence of new things, the transformation of concepts, the evolution of the language itself, and so forth. The emergence of new words is an active and positive factor in the development of the vocabulary system and constitutes the subject of most lexicological studies. In fact, even when we talk about the fading of the old words, they do not completely exit but still exist as a historical preservation in the vocabulary system. Considering that the emergence of new words takes a much more important position than the extinction of old words, we can obtain the general trend of the diachronic evolution of Chinese vocabulary system at the macrolevel by investigating the new words categorized in different time periods according to their time of occurrence in *CDCW*.

Admittedly, omissions of entries and presence of errors can be found in *CDCW*, and it is also true that *CDCW* cannot include all the new words in a given time period. However, the definitions and explanations of words in *CDCW* are the concerted efforts of a great number of researchers and scholars, and the defects and mistakes in it only account for a trivial proportion in the total number of its entries. As the "archives of ancient and modern Chinese vocabulary," *CDCW* remains one of the most authoritative dictionaries. Therefore, it is not only feasible, but also conducive to excluding the subjective prejudice of the researchers and to use the time of occurrence per entry in *CDCW* for reference when studying the diachronic evolution of the Chinese vocabulary system.

Theoretically speaking, it is possible to exhaustively obtain the time of occurrence of all word entries in *CDCW*. However, considering the fact that *CDCW* contains as many as over 300 thousand entries, it would be too much work to analyze all the entries. Therefore, in the present study, under the guidance of statistical sampling theory, we only draw a moderate number of sample entries and then deduce the overall quantitative characteristics of the population.

3.2. Formulation under Ideal Condition. Let W denote the word set for the Chinese lexis, from Old Chinese to Modern Chinese. An element $w \in W$ is a Chinese word. The subsets of monosyllabic words, disyllabic words, and three-or-more-syllable (i.e., ≥3 syllables) words are written as $W^{(1)}$, $W^{(2)}$, and $W^{(3+)}$, respectively. The subset of polysyllabic words is written as $W^{(2+)}$ and satisfies $W^{(2+)} = W^{(2)} \cup W^{(3+)}$. Notice that disyllabic words are processed independently because disyllabic words are extremely important in Chinese. Let the cardinality of a set be denoted as $|\cdot|$. Thus the total number of the words is represented as $N = |W|$. Likewise, the number of words in $W^{(1)}$ is denoted as $N^{(1)} = |W^{(1)}|$ and so forth.

Let T represent the time (using year as the unit of time) for the evolution of Chinese language, which is in fact a range of integer number from t_b to t_e. The time length in years of T is given by $\|T\| = t_e - t_b + 1$.

As illustrated in Figure 1, there exists a function f which relates each word $w \in W$ to its time of occurrence $t = f(w) \in T$. With the denotation, $f^{-1}(t)$ represents all the words which first occurred in the year t. The symbol $n_t = |f^{-1}(t)|$ represents the total number of the words which first occurred in the year t. Considering that n_t is the number of the newly formed words in the year t, we define it as the speed of new word production.

The sequence $\{n_t\}$ constitutes a time series. Its partial sum sequence $\{c_t\}$ is the cumulative number of all the words which first occurred in or before the year t, satisfying

$$c_t = c_{t_{b^-}} + \sum_{\tau=t_b}^{t} n_\tau. \tag{1}$$

Here, $c_{t_{b^-}}$ means the initial cumulative number of all the words before time t_b.

In addition, let $f|_U : U \to T$ denote the restriction of f to the subset U of its domain W, which is defined by the same rule as f but with a smaller domain set U. Thus, $n_t^{(1)} = |f|_{W^{(1)}}^{-1}(t)|$ represents the total number of the monosyllabic words which first occurred in the year t. Likewise, we have $n_t^{(2)}$, $n_t^{(3+)}$ and also their partial sum $c_t^{(2)}$, $c_t^{(3+)}$, and so forth.

We introduce two indexes to evaluate the trend of polysyllablization. The first is the speed of new word production for polysyllabic words, which is, namely, $n_t^{(2+)}$. It is in fact the number of the newly formed polysyllabic words per year and indicates how fast the new polysyllabic words are generated into the vocabulary system. The second index is named the polysyllablization degree index (PDI). It is the ratio of the cumulative number of the polysyllabic words to that of all the

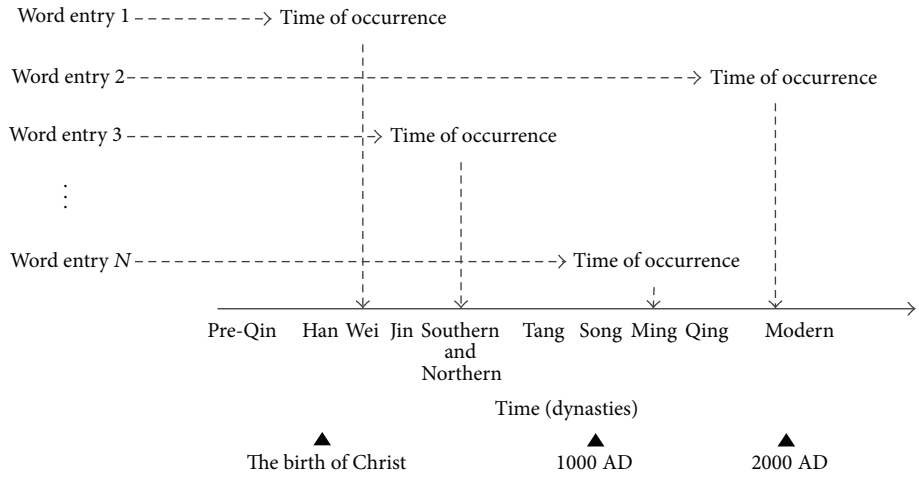

Figure 1: Determining the time of occurrence per word entry by the earliest textual evidence with *CDCW*.

words at the time t, and indicates the weight of the polysyllabic words in the then vocabulary system, which is defined as

$$\mathrm{PDI}(t) \triangleq \frac{c_t^{(2+)}}{c_t}. \tag{2}$$

3.3. Modeling with Uncertainties. However, there are still lots of uncertainties in the aforementioned formulation.

Firstly, there are a small amount of repetitions and exceptions in the entries in *CDCW*. The repetitions mean that two entries have different forms but the same meanings, which is similar to the case "color" and "colour" in English. Basically, they are just different forms of one word and should be counted as one word only. The exceptions mean that some entries are not valid Chinese words, such as the Japanese characters included in *CDCW*. Thus, the set of the whole entries of *CDCW*, E, is a superset of W; that is, $W \subset E$. The total number of entries in E is greater than that in W. The invalid entries, including repetitions and exceptions, should be preprocessed.

Secondly, there are above 300 thousand of entries in *CDCW*. They are too many to be handled one by one. The feasible way is to handle part of them with the statistical sampling method, which subsequently introduces probabilistic uncertainty. Assuming that we take a certain number of samples which are uniformly randomly chosen from the entries of *CDCW*, the set of the sampled entries is denoted as E_s, in which the set of valid entries (i.e., words) is denoted as W_s.

Based on the samples, the total number of word in *CDCW* can be estimated by

$$\widehat{N} = \widehat{|W|} = \frac{|W_s|}{|E_s|} \cdot |E| = \frac{|E|}{|E_s|} \cdot |W_s| = K \cdot |W_s|. \tag{3}$$

Here K is a constant factor and equals $|E|/|E_s|$.

Furthermore, the sequence $\{n_t\}$ can be estimated by the restriction of f to the subset W_s, which is similarly given by

$$\widehat{n}_t = K \cdot \left| f|_{W_s}^{-1}(t) \right|. \tag{4}$$

The cases to estimate $\widehat{n}_t^{(1)}$, $\widehat{c}_t$, and so on are similar and not listed here.

Thirdly, the precise value of the function f is usually difficult to decide. For most of the words, we can only know the dynasty of its occurrence rather than the time (accurate to year) of its occurrence. Thus the inaccuracy of time of occurrence will introduce lots of uncertainties.

Now consider that the set T can be divided into several segments, mathematically speaking, T can be expressed as the union of a number of disjoint sets D_i $(1 \le i \le n_d)$ as follows:

$$T = \bigcup_{1 \le i \le n_d} D_i, \tag{5}$$

where D_i is a set of consecutive integers beginning with $t_b^{(i)}$ and ending with $t_e^{(i)}$, that is, an integer interval with notation $[t_b^{(i)} \cdots t_e^{(i)}]$. The symbol n_d represents the number of the segments. Let D denote the set $\{D_1, D_2, \ldots, D_{n_d}\}$.

In fact, D_i represents a dynasty or a similar time period in Chinese history. Such partitions are caused by the uncertainty of the mapping from the words to their times of occurrence. The exact time of occurrence of many words is difficult or even impossible to resolve due to the obscureness in the ancient documentaries. Since dates are usually written in dynasties or reign titles in the ancient Chinese literatures, the exact dynasties (or time segments) can be determined for the first occurrence of many words. Thus the dynasties are the ordinary and natural implementation of the time segments.

Without loss of generality, we assume that the segments D_i are sorted by the lower endpoint, which satisfies

$$\begin{aligned} t_b^1 &= t_b, \\ t_e^i + 1 &= t_b^{i+1}, \quad \text{where } 1 \le i \le n_d - 1, \\ t_e^{n_d} &= t_e \end{aligned} \tag{6}$$

.

The rough version of the function f, which only maps from the word set W to the set of time segments (usually, dynasties) D, is written by

$$g : W \longrightarrow D \tag{7}$$

and satisfies

$$g(w) = d \Longleftrightarrow f(w) \in d. \tag{8}$$

Since the precise mapping from the words to their time of occurrence is usually unknown in the actual operation, we must estimate the parameters related to polysyllablization with the rough function g. It can be noted that $g^{-1}(D_i)$ represents all the words which firstly occurred in the time segment D_i. In fact, there exits the following relation:

$$g^{-1}(D_i) = \bigcup_{\tau \in D_i} f^{-1}(\tau), \quad \text{where } 1 \le i \le n_d. \tag{9}$$

Now that $f^{-1}(\tau)$ and $f^{-1}(\nu)$ are disjoint sets if $\tau \neq \nu$, we have

$$d_i = \left|g^{-1}(D_i)\right| = \sum_{\tau \in D_i} \left|f^{-1}(\tau)\right| = \sum_{\tau \in D_i} n_\tau, \quad \text{where } 1 \le i \le n_d. \tag{10}$$

Here, d_i represents the number of all the words which first occurred in the time segment D_i.

Based on the statistical sampling, the sequence $\{d_i\}$ can be estimated by the restriction of g to the subset W_s, which is similarly given by

$$\widehat{d}_i = K \cdot \left|g|_{W_s}^{-1}(t)\right|. \tag{11}$$

Let $d_i^{(1)}$ be the number of all the monosyllabic words which first occurred in the time segment D_i and so on. The cases to estimate $d_i^{(1)}$, $d_t^{(2)}$, and so forth are similar and not listed here.

Language changes over time. As a complex social phenomenon, language may change smoothly or unstably. We know little about the actual trends of the evolution of language. Without any prior knowledge or assumption, we will build a total variation minimization model [24, 25] for the evolution of the language. Other methods, such as maximum entropy [26, 27] method for the time series, should also be feasible and constructive. However, only the total variation minimization model is discussed in this paper, since it is simply based on the assumption that the language of today is not far from that of yesterday. Thus it is probably one of the safest estimation methods to start our exploration in such an open research field.

The total quadratic variation for a sequence $x = \{x_i\}$ $(1 \le i \le n)$ is defined as

$$V(x) \triangleq \sum_{i=1}^{n-1} (x_{i+1} - x_i)^2. \tag{12}$$

Our technique to estimate the sequences $n_t^{(1)}$, $n_t^{(2)}$, and $n_t^{(3+)}$ are similar. Generally, the case for the sequence $n_t^{(*)}$ is illustrated as follows, in which the symbol "$*$" can be replaced by any one of the symbols including "1," "2," or "3+." Basically, the issue to estimate the sequence $n_t^{(*)}$ is equivalent to a constrained optimization problem as follows:

$$\begin{aligned} &\text{minimize} \quad V\left(n^{(*)}\right) = \sum_{\tau=t_b}^{t_e} \left(n_{i+1}^{(*)} - n_i^{(*)}\right)^2 \\ &\text{subject to} \quad n_t^{(*)} \ge 0, \quad \text{where } t_b \le t \le t_e, \\ &\qquad \sum_{\tau=t_b^{(i)}}^{t_e^{(i)}} n_\tau^{(*)} = d_i^*, \quad \text{where } 1 \le i \le n_d. \end{aligned} \tag{13}$$

Here, the latter constraints can be similarly deduced as (10). Such a constrained optimization problem is a quadratic programming problem, which belongs to a relatively mature area and there are kinds of feasible techniques [28–30] to handle it. The problem can be solved in polynomial time with the ellipsoid method [31].

3.4. Work Flow. There are a series of steps to evaluate polysyllablization of the Chinese lexis based on *CDCW* in practice.

(1) Decide the endpoints of the time T and split the whole time into segments D_i.

(2) Obtain the set E of the whole entries of *CDCW* and classify the set E into $E^{(1)}$, $E^{(2)}$, and $E^{(3+)}$ by the number of syllables of each entry.

(3) Draw appropriate samples of entries, that is, $E_s^{(1)}$, $E_s^{(2)}$, and $E_s^{(3+)}$ from the sets $E^{(1)}$, $E^{(2)}$, and $E^{(3+)}$, respectively.

(4) Check each element e in the sets $E_w^{(1)}$, $E_s^{(2)}$, and $E_s^{(3+)}$ to obtain the sets of word samples, $W_s^{(1)}$, $W_s^{(2)}$, and $W_s^{(3+)}$.

(5) Estimate the total number of monosyllabic words $\widehat{N}^{(1)}$. Estimate $\widehat{N}^{(2)}$, $\widehat{N}^{(3+)}$ in the same way. Calculate the total number of words $\widehat{N} = \widehat{N}^{(1)} + \widehat{N}^{(2)} + \widehat{N}^{(3+)}$.

(6) Map each word in $W_s^{(1)}$, $W_s^{(2)}$, and $W_s^{(3+)}$ to its time of occurrence.

(7) Count the number of words in the same time segment and estimate $\widehat{d}_i^{(1)}$, $\widehat{d}_i^{(2)}$, and $\widehat{d}_i^{(3+)}$ according to (11).

(8) Solve the quadratic programming problem to obtain the estimation $\widehat{n}_t^{(1)}$, $\widehat{n}_t^{(2)}$, and $\widehat{n}_t^{(3+)}$.

(9) Calculate $\widehat{n}_t^{(2+)} = \widehat{n}_t^{(2)} + \widehat{n}_t^{(3+)}$ and $\widehat{n}_t = \widehat{n}_t^{(1)} + \widehat{n}_t^{(2+)}$. Also, calculate their partial sum sequences.

(10) Calculate the indexes to evaluate the trend of polysyllablization.

4. Analysis of the Polysyllablization

We use the CD-ROM version of *CDCW* published in 1998. This edition contains 27,989 character entries, 279,720 disyllabic word entries, and 63,587 three-or-more-syllable word

Table 1: Statistical data of words of various numbers of syllables.

i	Dynasties	Endpoints	Single syllable		Double syllables		Three or more syllables	
	d_i	$t_b^{(i)}$–$t_e^{(i)}$	$\widehat{g}_i^{(1)}$	$\widehat{d}_i^{(1)}$	$\widehat{g}_i^{(2)}$	$\widehat{d}_i^{(2)}$	$\widehat{g}_i^{(3+)}$	$\widehat{d}_i^{(3+)}$
1	pre-Qin	1045 BC–222 BC	274	7669	246	34406	55	3497
2	Qin, Western Han	221 BC–24 AD	44	1231	135	18881	33	2098
3	Eastern Han	25–219	56	1567	128	17902	22	1399
4	Three kingdoms, Jin	220–419	12	336	115	16084	23	1462
5	Southern and Northern	420–580	44	1231	238	33287	48	3052
6	Sui, Tang	581–907	23	644	275	38461	97	6168
7	5 dynasties, Song	908–1279	138	3862	243	33986	110	6995
8	Yuan, Ming, Qing	1280–1949	37	1036	391	54685	306	19458
		Total		17576		247692		44129

entries, from which we draw 1000 character entries, 2,000 disyllabic word entries, and 1,000 three-or-more-syllable word entries. Here we perform uniformly random sampling on all the entries without replacement. We use python, a computer script language, to perform the data sampling and processing.

The number of effective samples of monosyllabic words is 669 (66.9% of the drawn sample entries). According to the statistical sampling theory, the total number of effective monosyllabic words in *CDCW* is estimated to be about $27,898 \cdot 66.9\% \simeq 18,724$. The number of effective samples of the disyllabic words is 1,904 (95.2% of the drawn sample entries) and that of the words with three or more syllables is 876 (87.6% of the drawn sample entries). Likewise, the total number of disyllabic words in *CDCW* is estimated to be about $279,720 \cdot 95.2\% \simeq 266,293$ and that of three-or-more-syllable words is about $63,587 \cdot 87.6\% \simeq 55,702$. The total number of polysyllabic words in the Chinese lexis is about $266,293 + 55,702 = 321,995$. Thus, the number of polysyllabic words is greatly more than that of monosyllabic words up to now. Disyllabic words constitute the major part of the whole vocabulary.

We set the endpoints of the time axis to 1045 BC and 1949 AD. The former is approximately the beginning of Zhou Dynasty. The latter is the founding of the People's Republic of China. The time axis is split into 8 segments. We map all these effective word samples to their time of occurrence and count the number of samples belonging to each time segment.

The statistical data are listed in Table 1. The words which were newly formed after 1949 are ignored to avoid boundary problems. Therefore the sum in the table is a little less than the total number of words.

The data in Table 1 show the total number of the newly emergent words in each time period. For example, there were about 33,287 disyllabic words produced in Southern and Northern Dynasties and 38,461 in Tang Dynasty, the latter exceeding the former. However, the span of the former time segment was only 160 years and that of the latter was as long as 327 years. If we distribute the newly emergent words evenly among the years, the annual average for the Southern and Northern Dynasties is $33287/161 \simeq 206.8$ (word per year) and that for Tang Dynasty is $38461/327 = 117.6$ (word per year), only about half of the number in Southern and Northern Dynasties.

Notice that there were about 7,669 monosyllabic words produced in the pre-Qin period (1045 BC–222 BC). Part of them are generated during or before Shang Dynasty and known as Oracle Bone Script [32]. Several thousand bones and plastrons have been reconstructed and many thousands of texts have been studied. The texts contain over 30,000 distinct characters, which are thought to be variant forms of around 4,000 individual characters. Thus, the number of characters before Zhou Dynasty should be regarded as the initial cumulative number of the monosyllabic words. Now $c_{t_b}^{(1)}$ is set to be 4000, that is, the number of monosyllabic words at the beginning of pre-Qin period. We also subtract $c_{t_b}^{(1)}$ from the estimated number of the monosyllabic words in the pre-Qin period thereafter so as to make a revised estimation. The estimation on polysyllabic word is not revised because it is still argumentative whether there existed polysyllabic words in the Oracle Bone Script.

Figure 2 shows the final estimation on the speed of new word production. It illustrates that the speed is changeable. The speed of the new polysyllabic word production $n_t^{(2+)}$ is above that of monosyllabic words $n_t^{(1)}$ from the beginning of the time axis. It can be seen that the speed of new polysyllabic word production increased and reached two peaks at about 75 AD and 550 AD, that is, the Eastern Han Dynasty and Southern and Northern Dynasties. In the following years, new polysyllabic words still emerged, yet at a slower pace through Song, Yuan, Ming and Qing Dynasties. In fact, Southern and Northern Dynasties are treated as the center of Middle Chinese. Figure 2 shows that the Middle Ages are the critical transition period of polysyllablization. During Eastern Han and the following dynasties, the cultures from the western neighbors of China had a tremendous influence upon not only the Chinese traditional culture but also the Chinese language [33]. Linguistic evidences are available to reveal the early cultural exchange between China and India [34]. For example, the language contact [35] caused by the Buddhist texts translated from foreign languages (mainly Sanskrit) into Chinese was an important exterior influential factor to the evolution of the Chinese language. From this

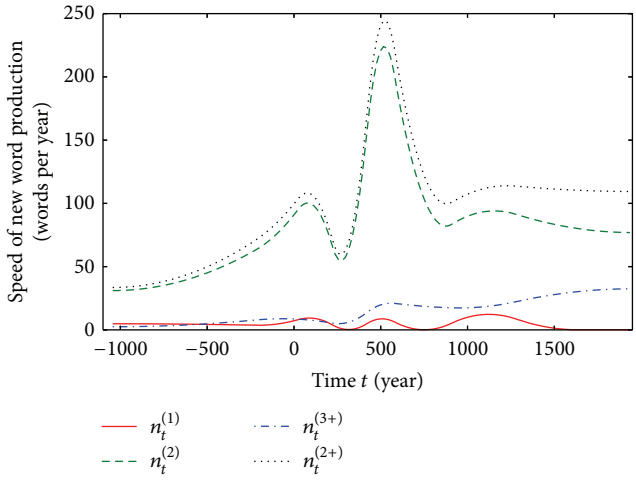

FIGURE 2: The speed of the word production.

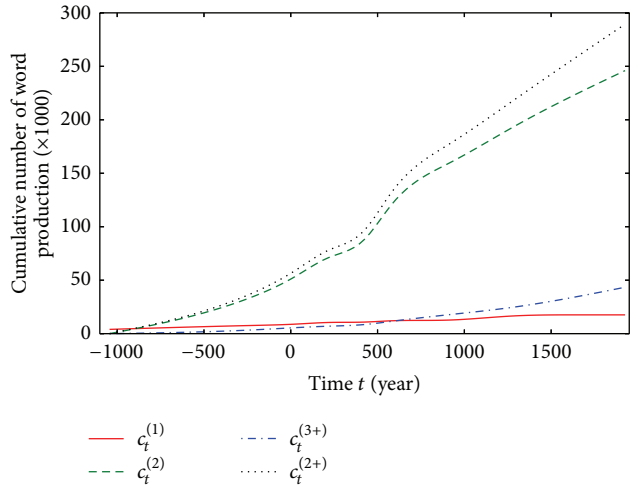

FIGURE 3: The cumulative number of the word production.

perspective, it can be concluded that the rising and falling in the above curves do reflect the evolution inside the Chinese language, agree with its historical stage division, and support the findings in other studies on the historical linguistics.

Figure 3 shows the final estimation on the cumulative number of new word production. It illustrates the increasing trend of the volume of Chinese vocabulary system. It is interesting to note that the amount of polysyllabic words $c_t^{(2+)}$ exceeds that of monosyllabic words $c_t^{(1)}$ before the ending of the first time segment. Thus polysyllabic words already were the major component of the lexis even in Old Chinese. However, back then, their frequency of use was very low, which has misled researchers to reach wrong judgments about the actual degree of the polysyllablization.

The trend that polysyllabic words became the major component of the Chinese lexis is also shown in Figure 4. It indicates that the weight of polysyllabic words in the Chinese lexis increased all along. At the ending of the time axis (1949 AD), about 94% of the Chinese words were polysyllabic.

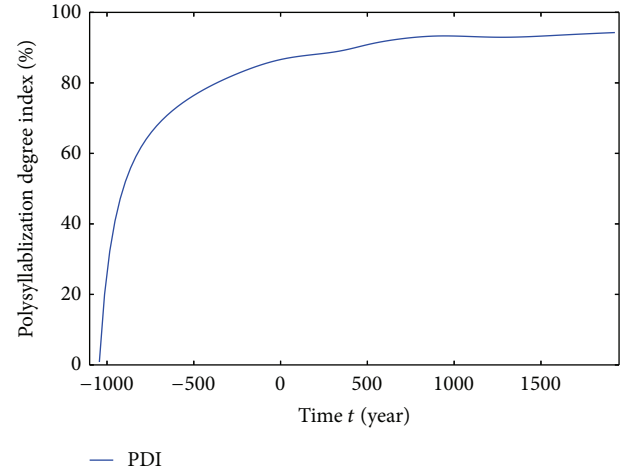

FIGURE 4: The quantitative trend of the polysyllablization degree index (PDI).

5. Conclusion

Polysyllablization is among the most important rules in the development of the Chinese lexis. Furthermore, due to the peculiarity of Chinese language, polysyllablization is not only a key issue in lexis, but also closely related to phonetics, semantics, and syntactics. However, the existing studies on polysyllablization are either in a qualitative manner or limited to a quantitative survey of specific language materials. The quantitative trend of polysyllablization is yet to be explored.

The lexis itself is an integral, dynamic, and complex system. With lots of uncertainties, its historical evolution is difficult to trace. As a novel exploration in the computational linguistics, we try to reconstruct the quantitative trend of polysyllablization of the Chinese lexis. A mapping from the words to their time of occurrence is built based on *Comprehensive Dictionary of Chinese Words*, a large-scale historical Chinese dictionary, which collects, organizes, and explains hundreds of thousands of ancient and modern words. Related formulation, mathematical model, and corresponding algorithms are introduced. We finally deduce the reconstruction to a constrained optimization problem based on the statistical sampling data. Such solution is really a data-mining procedure on the dictionary. It is not only feasible, but also conducive to excluding the subjective prejudice of the researchers. The results agree with the stage division of historical Chinese and answer some difficult questions related to polysyllablization in a quantitative manner.

Our research is only an initiatory exploration in this open research field. Our future research will focus on the revised mathematical models and algorithms, including statistical inference, maximum entropy estimation, and fuzzy mathematical theory based on more samples of word entries.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper. The authors also would like to thank Ms. Lu Yang and Ms. Di Lu for their assistance in the preparation of this paper. This work was supported by Beijing Higher Education Young Elite Teacher Project (Grant no. YETP0869) and Beijing Fund for the Humanities and Social Sciences (Grant no. 11WYC026).

References

[1] R. Lass, *Historical Linguistics and Language Change*, vol. 81, Cambridge University Press, 1997.

[2] R. M. Hogg and D. Denison, *A History of the English Language*, Cambridge University Press, 2006.

[3] B. D. Joseph and R. D. Janda, *The Handbook of Historical Linguistics*, Wiley Online Library, John Wiley & Sons, 2003.

[4] J. K. Chambers, *Sociolinguistic Theory: Linguistic Variation and Its Social Significance*, Blackwell, Cambridge, Mass, USA, 1995.

[5] E. G. Pulleyblank, "Lexicon of reconstructed pronunciation," in *Early Middle Chinese, Late Middle Chinese, and Early Mandarin*, UBC press, 1991.

[6] W. H. Baxter, *A Handbook of Old Chinese Phonology*, vol. 64, Walter de Gruyter, 1992.

[7] L. Sagart, *The Roots of Old Chinese*, vol. 184 of *Current Issues in Linguistic Theory*, John Benjamins, 1999.

[8] G. Edwin Pulleyblank, *Middle Chinese: A Study in Historical Phonology*, UBC Press, 2011.

[9] D. C. Twitchett, J. K. Fairbank, A. Feuerwerker, W. J. Peterson, K.-C. Liuv, and R. MacFarquhar, *The Cambridge History of China, Volume 1991*, Cambridge University Press, 1978.

[10] E. G. Pulleyblank, "Qieyun and yunjing: the essential foundation for chinese historical linguistics," *Journal of the American Oriental Society*, vol. 118, no. 2, pp. 200–216, 1998.

[11] L. Sagart, "The origin of Chinese tones," in *Proceedings of the Symposium/Cross-Linguistic Studies of Tonal Phenomena/Tonogenesis, Typology and Related Topics*, pp. 91–104, 1999.

[12] X. Zhou, W. Marslen-Wilson, M. Taft, and H. Shu, "Morphology, orthography, and phonology in reading Chinese compound words," *Language and Cognitive Processes*, vol. 14, no. 5-6, pp. 525–565, 1999.

[13] C. McBride-Chang, X. Tong, H. Shu, A. M.-Y. Wong, K. Leung, and T. Tardif, "Syllable, phoneme, and tone: psycholinguistic units in early Chinese and english word recognition," *Scientific Studies of Reading*, vol. 12, no. 2, pp. 171–194, 2008.

[14] Z. Luo, Ed., *Hanyu Da Cidian (Chinese Dictionary, Literally, Comprehensive Dictionary of Chinese Words)*, Hanyu Da Cidian Press, Shanghai, China, 1988, (Chinese).

[15] J. van de Weijer and J. Zhang, "An X-bar approach to the syllable structure of Mandarin," *Lingua*, vol. 118, no. 9, pp. 1416–1428, 2008.

[16] L. Yu and M. Gu, "Polysyllablization of chinese vocabulary based on the new lexical items in the buddhist and taoist scriptures of eastern han," *Journal of Sino-Western Communications*, vol. 5, no. 1, p. 225, 2013.

[17] S. Duanmu, "Stress and the development of disyllabic words in Chinese," *Diachronica*, vol. 16, no. 1, pp. 1–35, 1999.

[18] Z. Guo and Q. Yan, "A survey of vocabulary studies of special books in middle and modern Chinese," *Chinese Journal, Literally, Forward Position*, vol. 4, pp. 145–150, 2011 (Chinese).

[19] W. Xu, "Brief comments on the special features and worth of hanyu da cidian," *Cishu Yanjiu(Chinese Journal, literally, Studies on Dictionaries)*, vol. 1994, no. 3, pp. 36–45, 1994 (Chinese).

[20] D. Biber, S. Conrad, and R. Reppen, *Corpus Linguistics: Investigating Language Structure and Use*, Cambridge University Press, 1998.

[21] H. Li and B. Yuan, "Chinese word segmentation," in *Proceedings of the 12th Pacific Asia Conference on Language, Information and Computation*, pp. 212–217, Singapore, February 1998.

[22] N. Xue, "Chinese word segmentation as character tagging," *Computational Linguistics and Chinese Language Processing*, vol. 8, no. 1, pp. 29–48, 2003.

[23] J. K. Low, H. T. Ng, and W. Guo, "A maximum entropy approach to chinese word segmentation," in *Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing*, vol. volume, pp. 161–164, Jeju Island, Republic of Korea, 2005.

[24] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1–4, pp. 259–268, 1992.

[25] A. Chambolle, "An algorithm for total variation minimization and applications," *Journal of Mathematical Imaging and Vision*, vol. 20, no. 1-2, pp. 89–97, 2004.

[26] M. Palus, "Kolmogorov entropy from time series using information-theoretic functionals," *Neural Network World*, vol. 7, no. 3, pp. 269–292, 1997.

[27] C. Bandt and B. Pompe, "Permutation entropy: a natural complexity measure for time series," *Physical Review Letters*, vol. 88, no. 17, Article ID 174102, 2002.

[28] M. Frank and P. Wolfe, "An algorithm for quadratic programming," *Naval Research Logistics Quarterly*, vol. 3, pp. 95–110, 1956.

[29] P. Wolfe, "The simplex method for quadratic programming," *Econometrica*, vol. 27, pp. 382–398, 1959.

[30] P. T. Boggs and J. W. Tolle, "Sequential quadratic programming," *Acta Numerica*, vol. 4, no. 1, pp. 1–51, 1995.

[31] M. K. Kozlov, S. P. Tarasov, and L. G. Khachiyan, "The polynomial solvability of convex quadratic programming," *USSR Computational Mathematics and Mathematical Physics*, vol. 20, no. 5, pp. 223–228, 1980.

[32] D. N. Keightley, *Sources of Shang History: The Oracle Bone Inscriptions of Bronze Age China*, University of California Press, 1978.

[33] E. Zürcher, *The Buddhist Conquest of China: The Spread and Adaptation of Buddhism in Early Medieval China*, vol. 1, Brill Archive, 1959.

[34] Q. Zhu, *Some Linguistic Evidence for Early Cultural Exchange between China and India*, vol. 66 of *Sino-Platonic Papers*, Department of Asian and Middle Eastern Studies, University of Pennsylvania, 1995.

[35] S. G. Thomason and T. Kaufman, *Language Contact*, Edinburgh University Press, Edinburgh, UK, 2001.