A Study on the Diachronic Evolution of Ancient Chinese Vocabulary Based on a Large-Scale Rough Annotated Corpus

Yiguo Yuan
(Nanjing Normal University)
Bin Li
(Nanjing Normal University)

This paper makes a quantitative analysis of the diachronic evolution of ancient Chinese vocabulary by constructing and counting a large-scale rough annotated corpus. The texts from Si Ku Quan Shu (a collection of Chinese ancient books) are automatically segmented to obtain ancient Chinese vocabulary with time information, which is used to the statistics on word frequency, standardized type/token ratio and proportion of monosyllabic words and dissyllabic words. Through data analysis, this study has the following four findings. Firstly, the high-frequency words in ancient Chinese are stable to a certain extent. Secondly, there is no obvious dissyllabic trend in ancient Chinese vocabulary. Moreover, the Northern and Southern Dynasties (420-589 AD) and Yuan Dynasty (1271-1368 AD) are probably the two periods with the most abundant vocabulary in ancient Chinese. Finally, the unique words with high frequency in each dynasty are mainly official titles with real power. These findings break away from qualitative methods used in traditional researches on Chinese language history and instead uses quantitative methods to draw macroscopic conclusions from large-scale corpus.

Keywords: Ancient Chinese, Lexical Evolution, Quantitative Study, Corpus-based Analysis, Computational Linguistics

1. Introduction

The diachronic evolution of Chinese language is the object and task of the study of Chinese language history, which is of great significance to the standardization of modern Chinese and language education (Wang, 1980). However, the traditional study of Chinese language history only understands a vague outline on some issues rather than draws definite conclusions based on the in-depth study of a large number of language materials (Jiang, 1989). Quantitative research can often verify and supplement qualitative research and improve the accuracy of qualitative research. However, due to the lack of large-scale and deep-processing corpus of ancient Chinese, there is still a lack of research using quantitative analysis in the study of Chinese language history. So this paper tried to draw some conclusions about ancient Chinese history by using quantitative analysis methods.

As the building material of language and the collection of the smallest units that can be used independently in language, vocabulary is a dynamic collection, the evolution of which can reflect changes in social life. Meanwhile, it's easy to observe vocabulary from corpus for quantitative research. Therefore, the diachronic evolution of ancient Chinese vocabulary is regarded as the research object in this study. A large-scale rough annotated corpus about ancient Chinese language was built, with which this paper made statistics and data analysis in order to verify previous studies and analyze the macroscopic features of the diachronic evolution of ancient Chinese vocabulary preliminarily.

2. Background

The study on Chinese language history emphasizes sufficient materials. However, the nature of materials used in previous studies are various, which can be roughly summarized as example sentences, special books and corpus.

It's a common method to study Chinese language history with the use of example sentences, which often come from several common ancient books. These ancient books are studied thoroughly and simple retrieval can meet the needs of researches, so the processing of corpus they used is rarely carried out. For instance, Dong (2002) used plenty of example sentences when exploring the phenomenon of lexicalization. For example, she used sentences from *the Book of Rites* (《礼记》), *Mister Lv's Spring and Autumn Annals* (《吕氏春秋》), *Records of the Historian* (《史记》), *Lun Heng* (《论衡》) and other ancient books to illustrate the development of the nominalization of '作者' (author). Jin & Dong (2020) used a number of sentences from *the Doctrine of the Mean* (《中庸》), *the history of the later Han dynasty* (《后汉书》), *Records of the States in the Eastern Zhou Dynasty* (《东周列国志》) and other books to show the lexicalization process of '固执' (means 'stubborn'). It's obvious that these ancient books mentioned above are common, and example sentences from them could not show large-scale linguistic facts in ancient Chinese.

The characteristic of using a certain ancient book to study Chinese language history is that researchers studied special books more meticulously. But the processing of most of ancient books can only meet the needs of their own studies, and few complete processing of word segmentation and part-of-speech tagging were made. These studies are usually limited to a certain historical stage and have incomplete processing of corpus.

The characteristic of using corpus to study Chinese language history is that it can draw diachronic conclusions with a large span. However, according to different processing methods, the processing degree of a corpus is also different. One is by manual processing, the other is by machine. Of the two, the former can only meet the needs of researches. Li (2007) counted the dissyllabic words in the middle ancient Chinese corpus to study the characteristics of word formation. He just tagged the dissyllabic words in *Prescriptions for Fifty-two Diseases* (《五十二病方》), *Stories about Famous Women* (《列女传》), *Lun Heng* (《论衡》) and other 8 ancient books. Guo & Yang (2015) used an online ancient Chinese corpus which can only provide character-level retrieval to study the lexicalization of combined idioms. The latter can complete the processing of word segmentation and part-of-speech tagging, which is more helpful to quantitative analysis but rarely seen in the previous research on Chinese language history.

To sum up, few studies on Chinese language history use large-scale processed corpus. Even if some corpora were built, few were processed of word segmentation and part-of-speech tagging, but only tagged of the studied objects. Therefore, the former study of Chinese language history rarely drew macroscopic conclusions with quantitative analysis methods. In order to make up for this deficiency, this paper made a quantitative analysis of the evolution of ancient Chinese vocabulary based on a corpus of *Si Ku Quan Shu* (is abbreviated as *SIKU*) and a lexical analysis tool.

3. Methodology

It is necessary to construct a diachronic corpus for quantitative study on the diachronic evolution of vocabulary. First of all, the larger the scale of the corpus is, the closer the research object is to the real lexical appearance. As a large series compiled during the reign of Emperor Qianlong in the Qing Dynasty, *SIKU* covers four parts, including *Jing* (Confucian classics), *Shi* (history), *Zi* (philosophy) and *Ji* (literature), with rich categories. This paper selected a 731,852,425-character part of *SIKU* as the source of the corpus this study built. The distribution of the corpus in that four parts is shown in Table

1. It can be seen that the distribution of each part is relatively balanced, which is helpful for us to get objective results as far as possible.

Table 1. The Distribution of the Corpus This Paper Built

Parts of *SIKU*	Number of Characters	Proportion
Jing (Confucian Classics)	117,381,391	16.04%
Shi (History)	219,395,212	29.98%
Zi (Philosophy)	192,545,805	26.31%
Ji (Literature)	202,529,951	27.67%
Total	731,852,359	100.00%

Secondly, in order to construct a diachronic corpus, it is necessary to classify each text according to its time information. This paper classifies the texts according to 12 historical stages in ancient China: Pre-Qin Period, Han Dynasty, Three Kingdoms Period, Jin Dynasty, Northern and Southern Dynasties, Sui Dynasty, Tang Dynasty, Five Dynasties and Ten Kingdoms Period, Liao-Song-Jin Period, Yuan Dynasty, Ming Dynasty and Qing Dynasty.

Table 2. The Scale and Proportion of Corpus in Each Historical Stage

Serial Number	Historical Stage	Time Range	Number of Characters	Proportion
I	Pre-Qin Period	-221 BC	959,192	0.19%
II	Han Dynasty	202 BC-220 AD	4,979,274	1.00%
III	Three Kingdoms Period	220-280 AD	426,624	0.09%
IV	Jin Dynasty	265-420 AD	2,146,869	0.43%
V	Northern and Southern Dynasties	420-589 AD	6,407,417	1.29%
VI	Sui Dynasty	581-618 AD	425,494	0.09%
VII	Tang Dynasty	618-907 AD	23,467,927	4.72%
VIII	Five Dynasties and Ten Kingdoms Period	902-979 AD	2,218,116	0.45%
IX	Liao-Song-Jin Period	960-1079 AD	169,215,404	34.04%
X	Yuan Dynasty	1271-1368 AD	37,648,837	7.57%
XI	Ming Dynasty	1368-1644 AD	116,323,873	23.40%
XII	Qing Dynasty	1636-1912 AD	132,845,302	26.73%

In the process of classification, there are three special cases as follow: (1) If a text has no clear time information, it won't be put into this diachronic corpus. (2) The texts whose date of publication are inconsistent with their content are mostly found in the corpus of *Ji* (literature). This study determines their classification according to their content. The texts whose content spans more than two historical stages were not included in the diachronic corpus. For example, many poems written in the Tang Dynasty were compiled into books in the Song Dynasty, this research classified these poems as corpus in the Tang Dynasty. (3) There are many texts with annotation added by later generations. In order to make full use of the existing resources, this paper used regular expressions to match the content of annotation, and classified them into the original book and annotation. For instance, Guo Xiang annotated *Chuang Tzu* (a book written in the Pre-Qin Period) in the Jin Dynasty. Through our processing, the main body of this book was classified as the corpus in the Pre-Qin Period and the annotation was classified as the corpus in the Jin Dynasty. The scale and proportion of corpus in each historical stage are shown in Table 2 above. The total number of characters in this corpus is 497,064,329.

Finally, because the object of our research is ancient Chinese vocabulary, processing of word

segmentation is necessary for our corpus. Due to the large scale of the corpus, this paper uses the software developed by Cheng et. al (2020) for word segmentation and part-of-speech tagging of ancient Chinese language, whose F_1 value of word segmentation on test set can be up to 85.71%.

So far, the construction of the large-scale ancient Chinese diachronic annotated corpus needed by this study has been shown. The following will rely on this corpus to carry out the statistical work.

4. Results and Discussion

4.1. Word Frequency

Through word frequency statistics, this study sorted out the situation of high-frequency words in each historical stage. Figure 1 shows the diachronic change of the frequency rate of content words and function words with the highest frequency in each historical period. It's worth mentioning that the use of frequency rate instead of frequency is due to the inconsistency in the scale of the corpus in different historical stages. The frequency rate is calculated by formula (1).

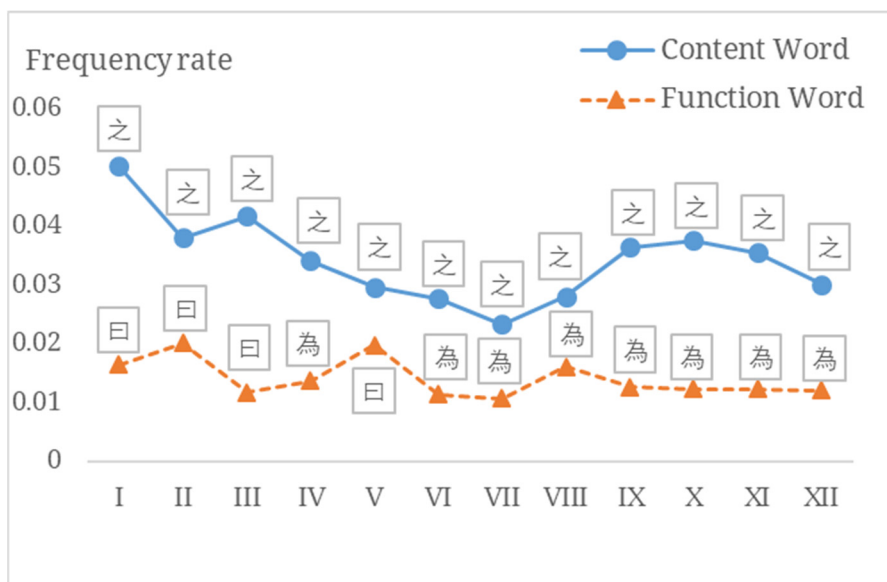$$f_i = \frac{n_i}{N} = \frac{n_i}{\sum_j n_j} \tag{1}$$

Figure 1. Diachronic Change of the Content Words and the Function Words with the Highest Frequency

From Figure 1, it can be seen that '之' (means 'of' in most cases) was always the most frequent function word in ancient Chinese. In modern Chinese, '的' (means 'of' in the vast majority of cases), which replaces '之', is still the most frequent function word. The frequency rate of '的' calculated by using the corpus of People's Daily in January 1998 is about 0.024, while the frequency rate of '之' in the corpus this study constructed is about 0.034. It can be seen that '之', as a function word, is stable in the diachronic development of ancient Chinese vocabulary.

From the view of content words, the content words with the highest frequency in ancient Chinese experienced the change from '曰' (means 'say' mainly) to '為' (means 'do' basically). Since the Sui Dynasty (VI), '為' had always been the most frequent content word, which is attributed to the universality of the meaning of '為'.

The content words and function words with the highest frequency in the above statistics are monosyllabic words. In order to present the results of word frequency statistics more comprehensively,

the frequency information of dissyllabic words with the highest frequency in each historical period is sorted out in this study, as shown in Figure 2. It can be seen that dissyllabic words with high frequency in ancient Chinese are also stable relatively.
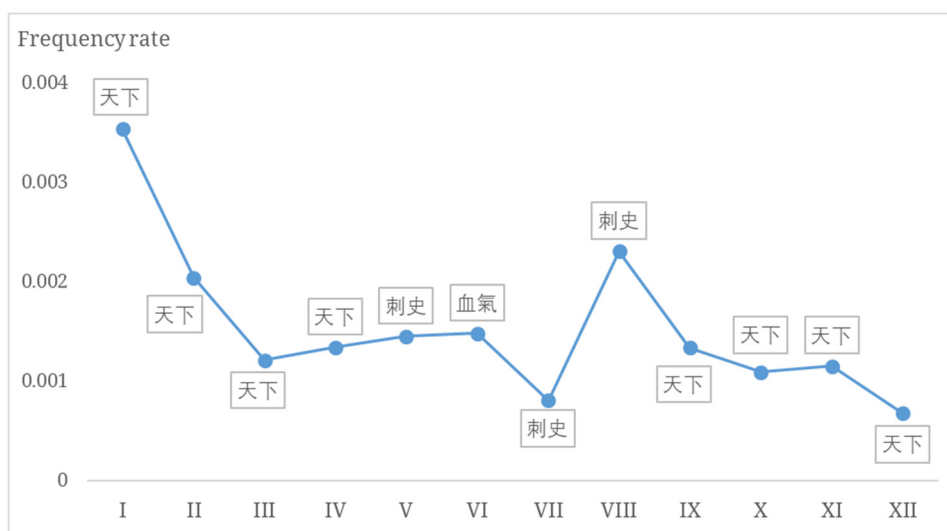

Figure 2. Diachronic Change of the Dissyllabic Words with the Highest Frequency

Among them, the most frequent dissyllabic word in the Sui Dynasty (VI) is '血氣' (a term of traditional Chinese Medicine which can be understood as 'blood and breath'). This phenomenon is due to the short existence of the Sui Dynasty and the lack of corpus in this stage. During this period, the word '血氣' mainly appeared in *Treatise on the Causes and Manifestations of Diseases* written by Chao Yuanfang. According to the data of other historical stages, '天下' (can be understood as the whole China or world) is the most frequent dissyllabic word in ancient Chinese, and '刺史' (feudal provincial or prefectural governor) also appears frequently in some dynasties.

'刺史' was established in the Han Dynasty (II) but not a fixed official position at the beginning, and its power was very small, so its occurrence number isn't very high at first. In the Northern and Southern Dynasties (V), the power of this position increased and began to have military power. It may be precisely because of its increased political influence that its frequency rate has increased. In the Han Dynasty, its frequency rate is 0.00010 which in the Northern and Southern Dynasties is 0.00146 in our corpus. '刺史' became a nominal position in the Song Dynasty (IX) and has been abolished since Yuan Dynasty (X). So we can see that it was no longer a very high frequency word in Figure 2 after the Liao-Song-Jin Period (IX). Its frequency rate is 0.00038 in the Liao-Song-Jin Period and 0.00011 in the Yuan Dynasty. '刺史' is a striking example of how the frequency of named entities, especially official position nouns, varies with their historical influence.

In order to investigate the stability of vocabulary in ancient Chinese, this study extracted the top 100 word types and 1000 word types in word frequency of the Pre-Qin corpus, counted the sum of their word frequency in each historical stage corpus, and obtained their proportion in word tokens. As can be seen from Figure 3, although the proportion of the most commonly used 100 and 1000 word types in the Pre-Qin Period declined in each historical stage, they were still relatively common. The proportion of the top 100 words remains above 30%, the proportion of top 1000 words stays above 60%. Their proportion does not fluctuate greatly and has diachronic stability.
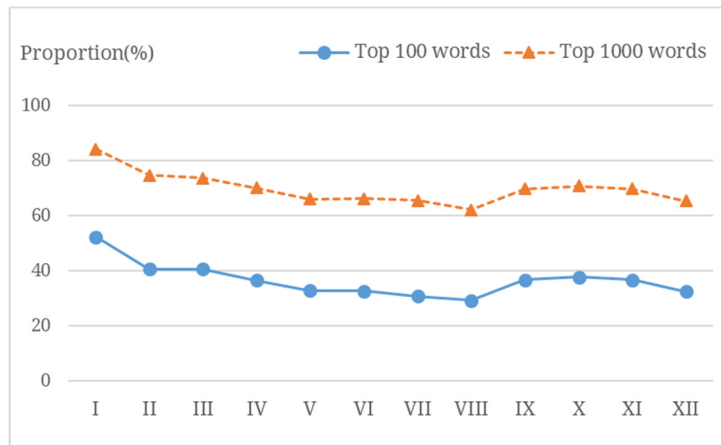
Figure 3. The Diachronic Change of the Proportion of the Top 100 and 1000 Word Types with the Highest Frequency in the Pre-Qin Period

4.2. Standardized Type/Token Ratio

Tokens are all the words in a given corpus while types are de-duplicated tokens. Baker (2000) thought that Type/Token Ratio (TTR) is a measure of the range and diversity of vocabulary used by a writer, or in a given corpus.

$$\text{TTR} = \frac{N_{type}}{N_{token}} \times 100\% \tag{2}$$

Formula (2) is the calculation process of TTR. Generally speaking, if the number of tokens in a corpus increases, the number of new types counted will decrease. The larger scale of the corpus, the lower TTR of the corpus. Therefore, TTRs of several corpora with difference scale are not comparable. In the same way, TTRs of different historical stages can't be compared due to the large differences in the scale of corpus in different stages. For example, the TTR of the least scale corpus is about 11.09%, while that of the most scale corpus is about 2.73%. However, the standardized type/token ratio (STTR) can solve this problem well. By calculating the TTR of every 1000 tokens, then calculating the mean value of these TTRs, this paper got the STTR data of each stage, which are shown in the figure below.

It can be seen that the three peaks of this line chart appear in the Northern and Southern Dynasties (V), Five Dynasties and Ten Kingdoms Period (VII) and Yuan Dynasty (X). Considering the scale of the corpus, the Five Dynasties and Ten Kingdoms Period corpus is too small to consider the peak value of this stage here. It probably means that the Northern and Southern Dynasties and Yuan Dynasty are the two periods with the most abundant vocabulary in ancient China, which may be related to the trend of ethnic integration. Minorities in the Northern and Southern Dynasties and Mongolians in the Yuan Dynasty brought a large number of new words and enriched the vocabulary of Chinese, which may be the reason for high STTRs in these stages.
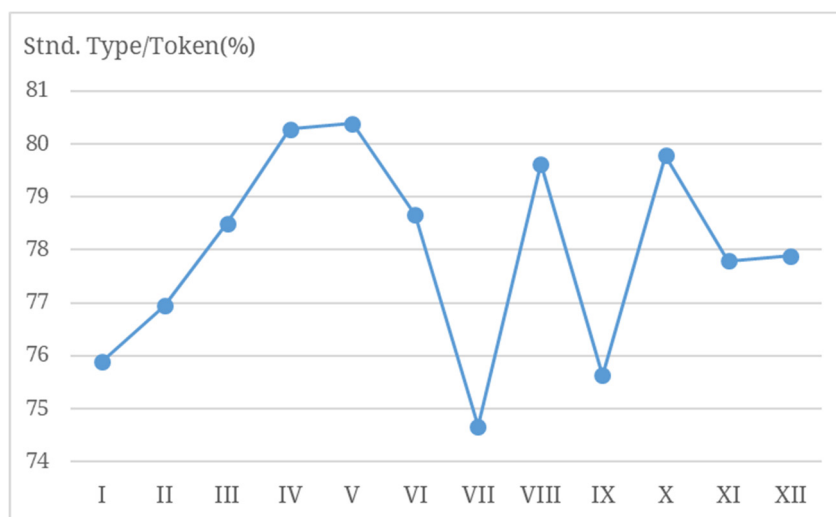
Figure 4. Diachronic Variation of the STTR

4.3. The Proportion of Monosyllabic Words and Dissyllabic Words

A prominent feature of ancient Chinese vocabulary is the predominance of monosyllabic words, while in modern Chinese, dissyllabic words account for the largest proportion. In order to discuss the diachronic change in the number of syllables of ancient Chinese vocabulary, this study makes a statistical analysis of the proportion of monosyllabic words (tokens) and dissyllabic words (tokens) in each stage corpus.

Table 3. The Proportion of Monosyllabic Words and Dissyllabic Words in the Corpus of Each Dynasty

Historical Stage	Proportion of Monosyllabic Words (%)	Proportion of Dissyllabic Words (%)
Pre-Qin Period	87.60	10.93
Han Dynasty	82.02	15.16
Three Kingdoms Period	82.85	14.96
Jin Dynasty	80.22	16.56
Northern and Southern Dynasties	76.13	19.10
Sui Dynasty	80.08	16.91
Tang Dynasty	78.24	17.45
Five Dynasties and Ten Kingdoms Period	71.98	21.60
Liao-Song-Jin Period	79.98	16.36
Yuan Dynasty	80.49	16.00
Ming Dynasty	81.14	15.58
Qing Dynasty	76.10	18.41
Average	79.74	16.58

In order to show the change of the proportion of monosyllabic words and dissyllabic words in each stage more intuitively, the data in Table 3 is represented by a line chart below. It can be seen that before modern times, the proportion of monosyllabic words and dissyllabic words in Chinese was relatively stable, and there was no obvious disyllabic trend.
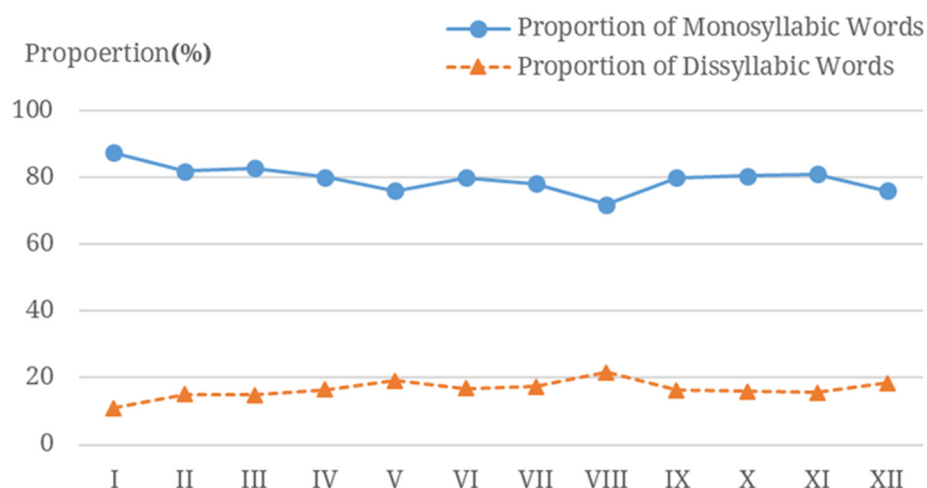
Figure 5. Diachronic Change of the Proportion of Monosyllabic Words and Dissyllabic Words

4.4 The Polysyllabic Words with High Frequency

Disyllabic words account for the majority of polysyllabic words in ancient Chinese. Although the frequency of polysyllabic words in ancient Chinese is relatively low, the polysyllabic words, especially the high-frequency polysyllabic words, contain abundant information because they were produced to express more complex concepts. Therefore, this study collected the top 20 polysyllabic words with the highest frequency in each period, and sorted out the unique high-frequency polysyllabic words in each period. This means that any one word in the following table only appears in the top 20 polysyllabic words with the highest frequency in only one period.

Table 4. The Unique High-frequency Polysyllabic Words in Each Period

Historical Stage	Word	English Interpretation
Pre-Qin Period	晏子	Yanzi who was a statesman and diplomat or the abbreviation of *Yanzi's Spring and Autumn Annals*
	孟子	Mencius who was a Confucian philosopher or *Mencius* which is a Confucian classics
	仁義	Benevolence and righteousness
	先王	Ancient saint king or dead emperor
Han Dynasty	匈奴	A nomadic people in the north of ancient China
	丞相	The highest minister who assisted monarchs
Three Kingdoms Period	足下	You (used to a superior or a peer)
	先帝	The dead father of the reigning emperor
Jin Dynasty	先主	The founding monarch
Northern and Southern Dynasties	徐廣	Xu Guang who was an official and scholar
	世祖	A posthumous title used for the founding monarch
	侍中	One of the highest ministers in this period
Tang Dynasty	主人	Host (a person who receives guests)
	左右	Attendant waiter
Five Dynasties and Ten Kingdoms Period	節度使	Governor of one or more provinces during Tang Dynasty
	侍郎	Assistant minister
	檢校	An official position having a formal name but no fixed position
Liao-Song-Jin Period	皇帝	Emperor
Yuan Dynasty	左氏	People surnamed Zuo or the author of *Zuo Zhuan*
	夫人	Wife or a women who was given a title or rank by the emperor
Qing Dynasty	知縣	County magistrate
	進士	A successful candidate in the highest imperial examinations

For the sake of table space, some words are omitted in Table 4. Unique words in the Sui Dynasty are all traditional Chinese medicine terms for the reason mentioned in 4.1. Therefore, they are not listed in Table 4. Similarly, the unique words of the Five Dynasties and Ten Kingdoms Period are not just those listed in Table 4. However, unique high-frequency polysyllabic words of the Ming Dynasty aren't found according to the word frequency statistics of the corpus used in this study.

It can be seen that most of these words are the names of official positions that are unique to some periods and have real power. For example, '節度使' was established in the late Tang Dynasty, and it had been a governor with local military and political power throughout the Five Dynasties and Ten Kingdoms Period. After that, this official position had no real power until it was abolished in the Yuan Dynasty. Except for official positions, other words also show the characteristics of each period. The unique high-frequency polysyllabic words in the Pre-Qin Period indicate that this is an era of ideological contention. As an unique word of the Han Dynasty, '匈奴' shows the theme of fighting against nomadic people invasion in this period. '世祖' reflects the frequent wars and regime changes in the Five Dynasties and Ten Kingdoms Period clearly. The Qing Dynasty was the heyday of the imperial examination system in the feudal era of China so that '進士' appears in the list of unique high-frequency polysyllabic words.

5. Conclusions

The main work of this study is to construct a large-scale and word-level-processing ancient Chinese corpus with the help of lexical analysis tools oriented to ancient Chinese, which is used to conduct a macroscopic quantitative study on the historical evolution of ancient Chinese vocabulary.

The conclusions of this study are as follows: (1) From the results of word frequency statistics, the high-frequency words of ancient Chinese are stable to a certain extent, and there is little difference between the high-frequency words used in corpus of different historical stages. In term of the grammatical function and meaning of words, function words tend to appear more often than content words even though they are few in number. In term of the number of syllables, the dominance of monosyllabic words had always been a feature of classical Chinese. To some extent, classical Chinese is a 'solidified' language, so we can see that the basic vocabulary used in the Pre-Qin Period (I) was still often used in the Qing Dynasty (XII). (2) According to the calculation results of the standardized type/token ratio, the Northern and Southern Dynasties and Yuan Dynasties are probably the two periods with the most abundant vocabulary in ancient Chinese, which may be related to the increase of new words brought by ethnic integration. (3) From the perspective of the number of syllables, the proportion of monosyllabic words and disyllabic words in Chinese had been relatively stable before modern times, and there was no obvious disyllabic trend.

Due to quantitative analysis, this study can draw conclusions that are difficult to be obtained by traditional research and these conclusions can be described with accurate data. For example, traditional researches using qualitative analysis methods may know that '之' is the most frequently used ancient Chinese function word, but it is a difficult problem to determine which word is the content word with the highest frequency in ancient Chinese. Because it's a very heavy work to count the whole vocabulary of ancient Chinese with frequency information only manually. Besides, word segmentation of ancient Chinese is an extremely time-consuming work, not to mention a large number of ancient Chinese books. Moreover, the application of statistical methods also provides a new perspective for the study of Chinese language history. For example, in order to study the lexical diversity of the corpus in each period, we counted STTR while traditional study tends to enumerate.

Because building an ancient Chinese corpus is a thick target of this study, the quality of the corpus needs to be further improved in the future work. In order to get more accurate data to describe the features of diachronic evolution of ancient Chinese vocabulary, the disequilibrium of the numbers of

these historical stages corpus is an objective difficulty to overcome for quantitative research. In addition, how to use other effective quantitative features to investigate the diachronic evolution of ancient Chinese vocabulary is also a direction in which this study needs to make efforts.

Acknowledgments

References

Baker, M. (2000). Towards a methodology for investigating the style of a literary translator. *International Journal of Translation Studies*, *12*(2), 241-266.

Cheng, N., Li, B., Ge, S., Hao, X. & Feng, M. (2020). A joint model of automatic sentence segmentation and lexical analysis for ancient Chinese based on BiLSTM-CRF model (in Chinese). *Journal of Chinese Information Processing*, *34*(4), 1-9.

Dong, X. (2002). Research of lexicalization of syntactic structure (in Chinese). *Studies in Language and Linguistics*, *3*, 56-65.

Guo, J., & Yang, E. (2015). A study on the lexicalization of combined idioms in mencius. In Lu, Q., & Gao, H. (Eds.), *Workshop on Chinese Lexical Semantics,* (pp. 307-319). Cham: Springer.

Jiang, S. (1989). Review and prospect of the study of Chinese language history (in Chinese). *Language Teaching and Linguistic Studies*, *2*, 124-129.

Jin, H., & Dong, Y. (2019). Investigation on the lexicalization process and causes of "Guzhi". In Hong, J. F., Zhang, Y., & Liu, P. (Eds.), *Workshop on Chinese Lexical Semantics*, (pp. 275-283). Cham: Springer.

Li, S. (2007) The development of mid-ancient Chinese word formation from dissyllabic word data (in Chinese). *Journal of Ningxia University (Social Science Edition)*, *3*, 1-7.

Wang, L. (1980). *The Manuscript of Chinese History* (in Chinese). Beijing: Zhonghua Book Company.

THE AUTHORS

Yiguo Yuan is a MA student at School of Chinese Language and Literature at Nanjing Normal University, China with emphasis in computational linguistics.

Bin Li is an Associate Professor at School of Chinese Language and Literature at Nanjing Normal University, China with emphasis in computational linguistics.

THE AUTHORS' ADDRESSES

First and Corresponding Author
Yiguo Yuan
MA student
School of Chinese Language and Literature
Nanjing Normal University
122 Ninghai Rd, Gulou District, Nanjing, CHINA
E-mail: lexcliff1023@gmail.com

Co-author
Bin Li
Associate Professor
School of Chinese Language and Literature
Nanjing Normal University
122 Ninghai Rd, Gulou District, Nanjing, CHINA
E-mail: libin.njnu@gmail.com