On the evolution of word usage of classical Chinese poetry

Liang Liu

Department of Statistics and Institute of Bioinformatics
University of Georgia

*Corresponding author:

Liang Liu (lliu@uga.edu)

Department of Statistics
Institute of Bioinformatics
University of Georgia
101 Cedar Street
Athens, GA 30602, USA
Tel: 706-542-3309
ABSTRACT

The hierarchy of classical Chinese poetry has been broadly acknowledged by a number of studies in Chinese literature. However, quantitative investigations about the evolution of classical Chinese poetry are limited. The primary goal of this study is to provide quantitative evidence of the evolutionary linkages, with emphasis on word usage, among different period genres for classical Chinese poetry. Specifically, various statistical analyses were performed to find and compare the patterns of word usage in the poems of nine period genres, including Shi jing, Chu ci, Han shi, Jin shi, Tang shi, Song shi, Yuan shi, Ming shi, and Qing shi. The result of analysis indicates that each of nine period genres has unique patterns of word usage, with some Chinese characters being preferably used by the poems of a particular period genre. The analysis on the general pattern of word preference implies a decreasing trend in the use of ancient Chinese characters along the timeline of dynastic types of classical Chinese poetry. The phylogenetic analysis based on the distance matrix suggests that the evolution of different types of classical Chinese poetry is congruent with their chronological order, suggesting that word frequencies contain useful phylogenetic information and thus can be used to infer evolutionary linkages among various types of classical Chinese poetry. The statistical analyses conducted in this study can be applied to the data sets of general Chinese literature. Such analyses can provide quantitative insights about the evolution of general Chinese literature.

SUBJECT KEYWORDS: evolutionary linguistics, classical Chinese poetry, phylogenetic tree

1. INTRODUCTION

Quantitative measurements have been commonly used in linguistic studies for understanding various language phenomena and language structures (Liu and Huang 2012). The analysis of affinity among Chinese dialects adopted correlation coefficients to quantify dialect similarity (Cheng 1991). Peng et al. (2008) utilized average distances, clustering coefficients, and the degree distribution to quantitatively compare the networks of Chinese syllables and characters. Moreover, it has become common practice in quantitative linguistics that language laws should be tested on empirical data using statistical methods. Increasing involvement of statistics and mathematics in linguistic studies has significantly changed the way of conducting scientific research for understanding important aspects of language. The origins and development of human language are perhaps the most fundamental questions in linguistics. A closely related but much more complicated question is how human language evolves over time. Various hypotheses regarding the origins and evolution of human language have been proposed and described in the context of probabilistic and biological models (Freedman and Wang 1996; Wang 2013). Since August Schleicher first introduced the representation of language families as an evolutionary tree, phylogenetic trees have been fundamental tools for understanding the history of language in the context of human evolution (Wang 1998a; Wang 1998b). Neighbor-joining (Saitou and Nei 1987) is one of the most commonly used methods for building phylogenetic trees in evolutionary linguistics. In addition, the efficiency of the algorithm searching for the best tree was greatly improved by matrix decomposition of phylogenetic trees (Qiao and Wang 1998). In this paper, phylogenetic trees are used as primary tools to investigate the evolution of poetic styles in word usage. Specifically, the similarity of word usage is measured by the average distance between the frequencies of Chinese characters in classical Chinese poems from nine period genres. A
neighbor-joining tree is then built from the matrix of similarity scores to illustrate the evolutionary linkage among nine period genres for classical Chinese poetry.

The earliest Chinese poem *tan ge* (谈歌) was first cited in an archer’s response to an inquiry of the king *Gou Jian* of Yue about the secret of accurate bow shooting. Although this short poem was recorded in a history book *wu yue chun qiu* (吴越春秋) dated back to 80 C.E., it has been speculated that the poem was actually passed down orally from Chinese primitive society (2600 B.C.E – 2100 B.C.E) and was documented in books much later by descendants. In general, Chinese poetry consists of two types of poetry, namely, classical Chinese poetry and modern Chinese poetry (Yip 1997). Classical Chinese poetry is characterized as written in traditional Chinese with certain traditional modes associated with particular historical periods (Hinton 2008). The tradition of classical Chinese poetry begins at least as early as the publication of *shi jing* (i.e., the Book of Songs), a collection of 305 poems from over two millennia ago (Watson 1984). Classical Chinese poetry continued to grow up to until May fourth 1919 movement (Grieder 1980), which is commonly considered as the stimulus of emergence of the modern Chinese poetry (Yeh 1991).

Over years, classical Chinese poetry has formed its unique style of rhythm and word usage (Zhong 2010). The earliest anthology of classical Chinese poetry is *shi jing*, in which poems or songs are predominantly composed of four-character lines developed during the period between Western Zhou Dynasty and the Spring and Autumn period (Dobson 1964). There are three chapters in *shi jing*, namely, folkway *feng*, elegance *ya*, and praise *song*. The chapter of folkway primarily features folk-songs, while the chapter of elegance includes songs from high-class officials and nobility. The songs in the chapter of praise were predominantly used to sing with dance in the monarch ritual of worship for ancestors. Another early anthology of Chinese poetry *chu ci*, known as the songs of the South, consists of poems particularly associated with the state of Chu in southern China. Most poems in *chu ci* are attributed to Qu Yuan and Song Yu. In contrast to *shi jing*, *chu ci* is typified by irregular line lengths (Hawkes 2011). The poems in *shi jing* and *chu ci* reflect the social and political status of pre-Qin period. A new form of classical Chinese poetry, known as *yue fu* style, was developed during the Han Dynasty (Birrell 1993). In contrast to *shi jing* and *chu ci*, *yue fu* poems are composed of five-character lines or seven-character lines (Liu 1966). The *yue fu* style poetry developed during the Han and the Jian'an period (the end of the Han Dynasty and the beginning of the Six Dynasties era) later became known as *gu ti shi* (i.e., ancient style poetry) (Watson 1971) in order to distinguish from the poetry developed during the Tang dynasty up to Qing dynasty, which are known as *jin ti shi* (i.e., new style poetry) (Yip 1976). Unlike its predecessor (i.e., ancient style poetry), new style poetry is being composed of five-character or seven-character lines with strict rules for the number of lines, rhyme, and a certain level of mandatory parallelism (Graham 1977). A study by Mair and Mei (1991) showed that tonal patterns of new style poetry rooted in Sanskrit prosody and poetics (Deo 2007).

Classical Chinese poetry continued to evolve in the context of social, political, and cultural changes over the past three thousand years (Olga 2013). The hierarchical structure of classical Chinese poetry has been broadly acknowledged by a number of studies in Chinese literature (Yip 1976). Two ancient anthologies, *shi jing* and *chu ci*, are on the top of the hierarchy, followed by *yue fu* style poetry (Han dynasty) and new style poetry (Tang dynasty - Qing dynasty), which represents the evolution of classical Chinese poetry along the lines of dynastic change. However, quantitative investigations about the evolution of classical Chinese poetry are exceedingly limited. The primary goal of this article is to provide quantitative evidence of the evolutionary
linkages, with emphasis on word usage, among different period genres (Table 2) of classical Chinese poetry.

2. Methods and Results

The data set of classical Chinese poems was collected from nine period genres in chronological order, including shi jing, chu ci, Han shi, Jin shi, Tang shi, Song ci, Yuan shi, Ming shi, and Qing shi (Table 2). Specifically, the data set consists of 305 poems from shi jing, 15 poems from chu ci, 675 Han shi selected from the Book of Han and the Records of the Grand Historian, 1821 Jin shi selected from the Book of Jin, 313 Tang shi from the Three Hundred Tang Poems (Sun 1763), 100 Song shi from Song Shi Xuan Zhu (Qian 1958), 651 Yuan shi from the book Yuan Shi Bie Cai Ji (Zhang 2012), 14602 Ming shi from the book Lie Chao Shi Ji (Qian 2007), and 27801 Qing shi selected from the book Qing shi Hui (Xu 1929). Three statistical analyses were performed: (1) the analysis of the distributions of word frequencies in the poems of nine period genres; (2) the significant analysis of word usage in the poems of nine period genres; (3) the similarity analysis on the evolutionary linkages among the poems of nine period genres.

In the analysis of the distribution of word frequencies, I calculated the frequency of each Chinese character occurred in the poems of nine period genres, respectively. The Chinese characters were ordered in accordance with their frequencies from highest to lowest. Moreover, twenty most frequently used Chinese characters (MFUCC) were selected for downstream analyses. The analysis consists of two types of comparison among the poems of nine period genres. The comparison among the distributions of word frequencies suggests that the word frequencies of shi jing and chu ci are significantly different from other period genres. The distributions of word frequencies in Han shi down to Qing shi are relatively flat, indicating that most words are used with similar frequencies in Han shi, Jin shi, Tang shi, Song shi, Yuan shi, Ming shi, and Qing shi (Figure 1). In contrast, some words in shi jing and chu ci are more frequently used than other words. This pattern is even more obvious in chu ci, where the word xi “兮” dominates the distribution of word frequencies (Figure 1). The high frequency of xi is one of the unique features of chu ci with regard to word usage. Many MFUCC of shi jing and chu ci are syncategorematic words, while most MFUCC of other period genres are notional words (i.e. noun). In addition, chu ci employs yu “余” to denote the first person, but other period genres use the word wo “我” as the first person. Finally, the word “兮” is included in the MFUCC of shijing, chu ci, Han shi, but not in Jin shi, Tang shi, Song shi, Yuan shi, Ming shi, and Qing shi (Figure 1). The most frequently used words (one, wind, cloud, sky) are shared by Jin shi, Tang shi, Song shi, Yuan shi, Ming shi, and Qing shi (Figure 1), which is an interesting pattern of post-Han shi in their preference towards particular words.

In the significant analysis of word usage among the poems of nine period genres, I calculated the frequency of each word in the poems of a particular period genre. I further calculated the frequency of the same word in the poems combined across all other period genres. The Binomial hypothesis test (Howell 2007) was adopted to evaluate if the word frequency in the poems of a particular period genre is significantly higher than other period genres. Since the test involves multiple comparisons, Bonferroni correction (Dunnett 1955) was carried out to adjust the p-value, leading to the final p-value P_{adj}. If P_{adj} ≤ 0.01, the frequency of the word in the poems of a particular period genre is significantly higher than other period genres, and thus this word is identified as a “characteristic word”. The result of the analysis indicates that Tang shi and Song shi have zero characteristic words. The number of characteristic words increases as it goes
towards the two ends (shi jing and Qing shi) of a spectrum of nine period genres on the x-axis of Figure 2. The poems of nine period genres can be divided into two groups in accordance with the number of characteristic words, where Tang shi and Song shi are the boundary points between two groups (pre-Tang group and post-Song group). Further analyses (see Figure 3) suggest that Tang shi and Song shi are more similar with the post-Song group (Yuan shi, Ming shi, and Qing shi). Thus, Tang shi and Song shi, together with Yuan shi, Ming shi, and Qing shi, form a poetry group, i.e., new style poetry. The first ten characteristic words ("榖", "覇", "駝", "儺", "殯", "讬", "讬", "譎", "騠") of shi jing are ancient Chinese characters with particular meanings that rarely occur in the poems of post-Qin period. Thus, it suggests that shi jing is inclined to use ancient Chinese characters. The first ten characteristic words ("後", "懐", "兮", "瞰", "些", "尪", "魂", "儶", "恆", "硃") of chu ci includes the word xi "兮", which is consistent with the conclusion of the previous analysis that frequent usage of word xi is an unique feature of chu ci. Another characteristic word of chu ci is qi “懐” (miserable), which reflects the miserable mood throughout the poems of chu ci. The characteristic words of Han shi, Jin shi, and Yuan shi include ("僶", "尷", "呪", "騠", " mlx", "騠", "騠", "騠", "騠", "騠", "騠", "騠", "騠", "騠", "騠", "騠", "騠"). ("騠", "騠", "騠", "騠", "騠", "騠", "騠", "騠", "騠", "騠", "騠", "騠", "騠", "騠", "騠", "騠", "騠", "騠", "騠", "騠", "騠", "騠", "騠", "騠"). respectively, which suggests that the number of ancient Chinese characters used in Han shi, Jin shi, and Yuan shi are less than that in shi jing. In contrast, no ancient Chinese characters are found in the first ten characteristic words of Ming shi ("沉", "燈", “著”,”初”,”初”,”初”,”初”,”初”,”初”,”初”,”初”,”初”,”初”,”初”,”初”,”初”,”初”,”初”,”初”,”初”,”初”,”初”,”初”,”初”,”初”,”初”,”初”,”初”,”初”). This result shows the tendency of using less ancient Chinese characters along the lines of dynastic change.

Finally, in the similarity analysis, I calculated the pairwise similarity score among the poems of nine period genres. The similarity score is measured by a distance function defined as the summation of differences between the frequencies of the words used in the poems of two period genres. Let \( x_{1k} \) and \( x_{2k} \) be the frequency of the \( k \)th word used in the poems of period 1 and period 2, respectively. The distance \( D \) between period 1 and period 2 is equal to \( D = \sum |x_{1k} - x_{2k}| \).

Since there are nine periods, the pairwise distances form a nine by nine distance matrix. The phylogenetic tree (Figure 3) was constructed from the distance matrix using the Neighbor-joining method (Saitou and Nei 1987). The branch lengths of the phylogenetic tree represent the distances among nine period genres. If two period genres are located nearby in the tree (for example, Ming shi and Qing shi in Figure 3), it indicates high similarity between the two period genres. Thus, the phylogenetic tree describes the evolutionary relationships among nine period genres for classical Chinese poetry. In figure 3, the tree built from the distance matrix is consistent with the dynastic order of nine period genres, except that the position of Song shi (Figure 3), which is located between Jin shi and Tang shi, conflicts with the chronological order of Jin dynasty, Tang dynasty, and Song dynasty. In general, the result of analysis indicates that the pattern of word usage of a particular type of poetry can be utilized to estimate the timing of a particular type of poetry, i.e., the period from which the poetry was developed. There are two major groups in the phylogenetic tree estimated from the distance matrix (Figure 3). The two groups formed in the tree are consistent with the classification of classical Chinese poetry (i.e., ancient style poetry and new style poetry) proposed by previous studies (Olga 2013).

3. Discussions
In this study, an extensive investigation was conducted on word usage of classical Chinese poetry. Various statistical analyses were performed to find and compare the patterns of word usage in the poems of nine period genres. The result indicates that each of nine period genres has unique patterns of word usage, with some Chinese characters being preferably used by the poems of a particular period genre. The analysis on the general pattern of word preference implies a decreasing trend in the use of ancient Chinese characters along the timeline of dynastic types of classical Chinese poetry. Word frequencies were used to measure the distance between different types of classical Chinese poetry. The phylogenetic analysis based on the distance measure suggests that the evolution of different types of classical Chinese poetry is congruent with their chronological order. This result indicates that word frequencies contain useful phylogenetic information and thus can be used to infer the evolutionary linkages among various types of classical Chinese poetry.

The present study focuses on the evolution of classical Chinese poetry. However, the statistical analyses conducted in this study can be applied to the data sets of general Chinese literature. Such analyses will provide insights about the evolution of general Chinese literatures. This study is an example of using statistical analysis to understand the evolutionary linkages among Chinese literatures.

Table 1: Classification and characteristics of classical Chinese poetry

| Period                   | Style     | Characteristics                              |
|--------------------------|-----------|---------------------------------------------|
| Western Zhou - the Spring and Autumn period | shi jing   | Four-character lines, from northern China   |
| Zhan guo                 | chu ci    | Irregular line lengths, from southern China |
| Han, Wei, Jin            | Yue fu (ancient style poetry) | Five-character or seven-character lines, unregulated |
| Tang - Qing              | new style poetry | Regulated poetry                           |

Table 2: The number of poems sampled from each period genre for classical Chinese poetry.

| Period                   | Source                             | number of sampled poems |
|--------------------------|------------------------------------|-------------------------|
| Western Zhou - Spring and Autumn period | shi jing                           | 305                     |
| Zhan guo                 | chu ci                             | 15                      |
| Han dynasty              | Book of Han and Records of the grand historian | 675                     |
| Jin dynasty              | Book of Jin                        | 1821                    |
| Tang dynasty             | Three hundred Tang Poems           | 313                     |
| Song dynasty             | Song Shi Xuan Ji                   | 100                     |
| Yuan dynasty             | Yuan Shi Bie Yang Ji               | 651                     |
| Ming dynasty             | Lie Chao Shi Ji                    | 14602                   |
| Qing dynasty             | Qing Shi Hui                       | 27801                   |
Figure 1: The distribution of word frequencies for each of nine period genres, including *shi jing*, *chu ci*, *Han shi*, *Jin shi*, *Yuan shi*, *Ming shi*, and *Qing shi*. The heights of bars represent relative frequencies (i.e., probabilities) of Chinese characters used in the poems of each period. In addition, the most frequently used words are listed in the box. Those words are ordered in accordance with their frequencies from maximum to minimum.
Figure 2: The number of characteristic words in the poems of nine period genres. The height of a bar represents the number of characteristic words.
Figure 3: The phylogenetic tree of nine period genres. The branch lengths represent the distances between the poems of nine period genres. The topology of the phylogenetic tree represents the evolutionary linkages among nine period genres. The tree includes two major groups: ancient style poetry and new style poetry.

REFERENCES

BIRRELL, A. (1993). Popular songs and ballads of Han China. Honolulu, University of Hawaii Press.

CHENG, C. C. (1991). "Quantifying affinity among Chinese dialects." Journal of Chinese Linguistics Monograph Series 3: 78-112.

DEO, A. S. (2007). "The metrical organization of Classical Sanskrit verse." Journal of Linguistics 43(1): 63-104.

DOBSON, W. A. C. H. (1964). "Linguistic Evidence and the Dating of the Book of Songs." Toung Pao (通报) 51(4): 322-334.

DUNNETT, C. W. (1955). "A multiple comparisons procedure for comparing several treatments with a control." Journal of the American Statistical Association 50(272): 1096–1121.

FREEDMAN, D. A. and W. S.-Y. Wang (1996). "Language polygenesis: a probabilistic model." Anthropological science 104(2): 131-138.
浅析中国古代诗歌文字使用的历史变化

中国古代不同类型诗歌之间存在着明显的演变关系。这一点已经被很多关于中国古代文学的研究所共识。然而对中国古代诗歌演变的量化研究却十分有限。这篇文章的主要目的是通过数据分析来量化中国古代诗歌在用字上的演变。具体地讲，运用不同的统计分析手段来比较和发现九种不同时期的古代诗歌（诗经，楚辞，汉诗，晋诗，唐诗，宋诗，元诗，明诗，清诗）在用字上的规律。研究结果表明，不同时期的古代诗歌都有各自独特的用字特点。特定时期的诗歌偏好使用特定的字。不同朝代的诗，按照从诗经到清诗的顺序，逐渐减少了对古汉字的使用。另外，研究结果表明不同时期诗歌的演化关系和它们的时间顺序是一致的。本文使用的统计分析可以应用到其他中国古代文学作品，从而对其他中国文学作品的演变提供量化的证据。

关键词：演化语言学，中国古代诗歌，进化树