Toward a Quantitative Approach to the Specialness of Poetry: Taking Chang Hen Ge and Chang Hen Ge Zhuan as a Case Study

Cong Zhang¹ and Rui Yuan²

Abstract
The differences of linguistic features between Chang Hen Ge (Ge) and Chang Hen Ge Zhuan (Zhuan) have rarely been mentioned in the relevant fields. Nevertheless, these differences can best highlight the specialness of poetry, for the two works were written contemporaneously by two friends on the same subject, in distinct styles. This article employs quantitative methods and indicators to provide empirical evidence for the specialness of Ge through comparisons between the two. The results show that, on the premise of expressing the same subject in different styles, Ge does have certain linguistic characteristics compared with Zhuan. Its particularity is reflected not only in fewer repeat characters and words but also in their richness, as well as in the use of more content words and fewer function words. Moreover, all of these characteristics have had a great influence on Ge’s artistic level and dissemination. Through this study, we hope that our methods provide a new perspective and shed some light on this area.

Keywords
poetry, Chang Hen Ge, Chang Hen Ge Zhuan, quantitative methods, character and word frequency

Introduction
In the field of literary research, there have been numerous attempts to obtain new understanding by means of corpus and quantitative methods (e.g., Algee-Hewitt et al., 2015; Burrows, 1987; Chesnokova et al., 2017; Cox & Brandwood, 1959; Elliott, 2016; Fabb & Halle, 1908; Fleay, 1874; Hoover, 2007; Matthews & Merriam, 1993; Mendenhall, 1887; Miall, 1988; Van Peer et al., 2012; Yule, 1939). Unfortunately, those methods are less frequently used in studies related to ancient Chinese poetry. The present study adopts quantitative methods to two Chinese texts with the same subject yet different style, from Tang Dynasty. By investigating the quantitative difference between these texts, we try to provide some insight into the specialness of poetry.

In 806 AD, Bai Juyi 白居易 (772–846), a great poet from China’s Tang Dynasty, had a discussion with his friend Chen Hong 陳鴻 (?–?), who is a novelist, about the love story of Tang Xuanzong 唐玄宗 (685–762) and Yang Guifei 杨贵妃 (719–756). To record that story and their feelings toward it during the discussion, Bai Juyi wrote a long narrative poem named Chang Hen Ge 長恨歌 (The Song of Everlasting Regret), and Chen Hong wrote a prose narrative entitled Chang Hen Ge Zhuan 長恨歌傳 (The Biography of the Song of Everlasting Regret) in parallel. Chang Hen Ge (Ge, for short) was widely disseminated as soon as it was published; however, Chang Hen Ge Zhuan (Zhuan), for short) is not only rather unknown to the public but also considered far less significant than the former by the scholars from the field of ancient Chinese literary history.

Prior studies of these two classics involve literature reviews, textual distinguished, text analysis, translation studies, artistic appreciation, and so on. For instance, in Britain and the United States, more than a dozen versions of Ge have emerged during the past 100 years. Early English sinologist Waley (1949), American scholars Levy (1957) and Kroll (1990), and others also discussed the theme of this poem. Nonetheless, the discussion in this area has no further progress at the moment, which is incommensurable with the enthusiasm of Chinese and Japanese scholars on the themes, styles, and techniques of the works. In Japan, scholars represented by Matsuura (1981), Monoda (2011), and Shimosada (2009) insisted on the idea that Ge is meant to celebrate love...
and Zhuan concentrated on Allegorical critique. In China, the interpretation of the theme of Ge has changed with the development of Chinese society. Zhou (2003) summarized the research on the subject of the works since the 20th century into five categories, analyzed the relationship between Ge and Zhuan, and further explained the artistic characteristics of poetry. Zhou’s research is highly representative of this field, but he did not investigate Ge’s and Zhuan’s unique language characteristics. Deng and Gao (2005) compared the difference in the narrative methods between the two, but it still settled on the conclusion that the poetry and the prose narrative borrowed from each other and blended, blurring the stylistic boundaries of the two. Many of the follow-up researchers (e.g., Xu, 2020) also followed this theory and took two as examples to present the phenomenon of “Narrativized Poetry” and “Poetic narratives” in the literary world of the middle Tang Dynasty. All these viewpoints have taken note of the polystylism of the Tang Dynasty, but have neglected the basic differences in language characteristics between poetry and prose narratives. As a result, the prevalence of Ge and the low-keyed mystery of Zhuan also became difficult to clarify along with the abovementioned “fuzzy theory.”

The comparison of the linguistic features of the two can not only provide new perspectives and evidence for previous studies but also reveal the connection between the language, artistic level, and dissemination of poetry and prose narrative to some extent. To compare Ge and Zhuan, researchers in this field first need to answer two questions: whether Ge and Zhuan are two parts of one work, or two separate works, and which version of Zhuan is the authoritative one. Regarding the discussion on their relationship, the scholars hold two views: One is put forward by Chen (1947), who believes that Ge and Zhuan are two pieces of a unified literary composition. And the other is proposed by Xia and Chen (1949), who deems that the two are independent of each other. Many contemporary scholars (Hu & Yanrong, 2008; Shimosada, 2009) are in favor of Xia’s viewpoint. Regarding the second question, Hu and Yanrong (2008) examined that a version called Wen Yuan Ying Hua 文苑英華 (Finest Blossoms in the Garden of Literature) was the closest to the original version of Chen Hong’s Zhuan. In this article, we use the above version of Zhuan as data material to show the specialness of Ge through quantitative comparison.

Despite all the studies we have mentioned earlier, unfortunately, they did not answer the following questions:

On the premise of expressing the same subject in different styles within the same period of time,

a. What are the main differences in linguistic features between Ge and Zhuan?

b. And how have these differences influenced their artistic level and dissemination?

These questions are of great significance to interpreting of Ge and Zhuan, but they have not been investigated in the field of ancient Chinese literature yet. To fill that gap, this article adopts the methods and indicators from corpus linguistics and quantitative linguistics to examine the relationship between Ge and Zhuan. Lexical statistics and frequency analysis are demonstrated to show their differences, through which we explore the advantages of Ge and summarize the unique charm and special characteristics of poetry.

**Methods and Statistical Indicators**

Frequency is one of the most important issues and the basis for quantitative research. The frequency of linguistic units in texts and their relationship to the frequency-related indicators are all somewhat different. By investigating the frequency distribution of characters and words in Ge and Zhuan, we can quantitatively compare the richness of the use of characters and words between them and provide some empirical evidence for the specialness of poetry. In this section, we briefly introduce the survey procedures, methods, indicators, and terminology used in this article.

Before conducting the investigation, we manually typed both texts into the computer and carefully checked. As we all know, there is no interspace between characters as well as between words in Chinese texts, which poses difficulties for quantitative statistics. To solve this problem, we use R (version 3.5.0) and an R package named “JiebaR” to segment the characters and words in these two texts automatically. Although the accuracy of automatic word segmentation is very high now, it still is not 100% accurate. Therefore, we also manually proofread the results of automatic segmentation. Through the steps described above, we have ensured the accuracy of the data of the texts. Then, we conduct the following investigations:

First, we investigate the frequency of the Chinese characters in Ge and Zhuan. Chinese characters are one of the basic units of the Chinese language. To some extent, the frequency of Chinese characters can reflect the uniqueness and characteristics of distinct styles. By investigating the frequency distribution and richness of characters in Ge and Zhuan, we can compare poetry with prose narratives under the same subject and reveal the specialness of poetry from a quantitative perspective.

Next, we focus on the words in Ge and Zhuan. Compared with Chinese characters, the quantitative features of words can tell us more about the features of the texts. They can tell us the subject of the text, the usage percentage of the content words and the function words, etc. The rank-frequency (taking the words of the text into a sequence from the highest to the lowest frequency) of words can tell us the distribution characteristics of the vocabulary, help us introduce diversified quantitative indicators, and distinguish between linguistic characteristics of poetry and prose narratives.

Last but not least, we describe the quantitative indicators and relative terms used in this article:
Type refers to the number of unique characters or words in the text.

Token refers to the sum of the frequency of occurrence of all unique characters or words in the text.

Type/Token Ratio (TTR) is a commonly used frequency measurement indicator in corpus linguistics (Xiao & Hu, 2015) for measuring the richness (density) of characters and words in a text. The formula used to calculate TTR is as follows:

\[
TTR = \left( \frac{\text{Type}}{\text{Token}} \right) \times 100\%.
\]

With Formula 1, we are able to compare the characters richness and vocabulary richness in Ge and Zhan.

The h-point is the fuzzy boundary of the content words and function words (Popescu, Mačutek, & Altmann, 2009). Every text has the h-point (or h-points). In each text, the words before its h-point are usually function words, and the words after the h-point are mainly content words. The h-point has been validated in 20 languages. We can calculate it by the following formula:

\[
h = \begin{cases} 
   r, & \text{if there is an } r = f(r) \\
   \frac{f(a)r_b - f(b)r_a}{r_b - r_a} = f(a) - f(b), & \text{if there is no } r = f(r) 
\end{cases}
\]

where \( r_a \) and \( r_b \) refer to any rank of two neighboring words in a rank-frequency distribution, and \( f(a) \) and \( f(b) \) refer to the corresponding word frequencies of \( r_a \) and \( r_b \) respectively.

The investigation of the h-points in Ge and Zhan allows us to have a good understanding of the use of both content words and function words. Moreover, high-frequency words before h-point are also worth of our investigation.

The a-index is based on the h-point and can be used to distinguish different styles. Usually, the texts we analyze are of varying lengths. To quantitatively compare these texts, it is important to move away from the influence of text length on indicators. Popescu, Altmann, et al. (2009) presented the a-index as a solution. The equation is as follows:

\[
a = \frac{N}{h^2},
\]

where \( h \) refers to the h-point in the text and \( N \) refers to the length of the text (measured by the number of words).

This function has also been validated in 20 languages. Huang and Liu (2017) and Zhang and Liu (2019) have verified the ability of this indicator to distinguish between different styles.

**Results**

**The Characters Frequency in Ge and Zhan**

Different literary styles have diverse distribution features in terms of character frequency, even if the subject is the same. In Ge, there are 467 unique Chinese characters, which means typeGe = 467. These characters altogether occur 846 times, which means tokenGe = 846. Zhan has 626 unique Chinese characters that appear 1,388 times, consequently, typeZhan = 626, tokenZhan = 1,388.

As shown in Figure 1, the higher the frequency, the lower the number of characters. There are 289 characters in Ge and 367 characters in Zhan that only occur once. Ge’s character frequency mainly distributes within 10 times, and the only character that exceeds 14 times is Bu 不 (no). Considering that most ancient Chinese poems are shorter than Ge, we argue that the character frequency of most ancient Chinese poems is less than 10. In contrast to Ge, Zhan’s character frequency is mainly within 17, except for Shang 上 (up), which occurs 21 times, Bu 不 (no) with 24 times, and Zhi 之 (of, this, etc.) with 26 times.

We employ log-likelihood test to examine the number of characters that only occur once in the two texts. The result shows that the difference between them is statistically significant (\( p < .001 \)). Accordingly, on the premise of expressing the same subject, characters in Ge occurred significantly fewer than Zhan. Nevertheless, the same test method shows that the most used character of the two texts has no statistical significance (\( p = .859 \)). This is possibly due to the fact that Ge is a long narrative poem.

Even with a long narrative poem like Ge, the length of its text is still shorter than that of Zhan, and they told the same story. Here, we use the total frequency of characters to measure the text length of the two, that is, Ge’s text length equals tokenGe’s value, which is 846, and Zhan’s text length is 1,388. The text length of the two differs by 524 characters. In this case, the poet needs richer and more unique characters to stay consistent with what the prose narrative is telling. As a consequence, Ge’s character richness should be higher than Zhan’s character richness.

We use TTR to measure the character richness of the two texts. The larger the result of TTR, the higher the character richness of the text is, and vice versa. According to the foregoing formula, the results of Ge’s and Zhan’s TTR are as follows:

\[
TTR_{Ge} = 55.2\%.
\]

\[
TTR_{Zhan} = 45.1\%.
\]

In the light of the results, the character richness of Ge is higher than that of Zhan by over 10%. This explains why poetry is shorter but has the same narrative power as a prose narrative in telling a story with the same content.

**The Word Length and Word Frequency in Ge and Zhan**

As a fundamental unit in the human language system, word is the smallest unit that can be freely used in written discourse and is more important than characters in Chinese...
literary works such as poetry, prose narrative, and fiction. The arrangement and combination of words help the text form various unique styles. Quantitatively, high-frequency words demonstrate the subject of the text, and the number of hapax legomena reflects the richness of the vocabulary in the text. Unlike characters, words come in different lengths, and the ratio between monosyllabic words and disyllabic words has always been a research subject of great concern in Chinese linguistics (Zhang & Liu, 2019). Therefore, in this section, we first examine the word lengths of Ge and Zhuan.

We use syllables as a measure of word length. Chinese characters and syllables are inherently one and the same, for one Chinese character generally represents only one syllable. Thus, the word length of a monosyllabic word is 1, and the word length of a disyllabic word is 2. For example, the word length of 博不 (no) is 1, and the word length of 君王 (emperor) is 2. After investigating Ge and Zhuan, we obtained the following results: There are 186 monosyllabic words in Ge with 275 occurrences in total and 218 disyllabic words with 233 occurrences in total. There are 297 monosyllabic words in Zhuan with 562 occurrences and 292 disyllabic words totaling 333 occurrences. To compare these results directly, we divide them by the total number respectively and convert the numbers into decimal forms.

As shown in Figure 2, in Ge, the monosyllabic words accounted for 0.43 and the disyllabic words accounted for 0.5; the occurrences of monosyllabic words accounted for 0.51 and the occurrences of disyllabic words accounted for 0.43. These results mean that in Ge, the number of monosyllabic words is fewer than that of disyllabic words, but the monosyllabic words are used more frequently than the disyllabic words. In Zhuan, the monosyllabic words accounted for 0.47, the disyllabic words accounted for 0.46, the occurrences of monosyllabic words accounted for 0.6, and the occurrences of disyllabic words accounted for 0.35. The results show that in Zhuan, both the occurrences and number of monosyllabic words are more than disyllabic words. Overall, the occurrences of monosyllabic words are dominant in both Ge and Zhuan, which is consistent with most linguists’ view (e.g., Wang, 2004; Zhang & Liu, 2019) of ancient Chinese in the Tang Dynasty.

Figure 1. The frequency distribution of the characters in Ge and Zhuan.
The comparison between them shows that Ge has fewer monosyllabic words and more disyllabic words than Zhuan. And in the respect of frequency, Ge has fewer occurrences of monosyllabic words and more occurrences of disyllabic words than Zhuan. In the Tang Dynasty, the function words were mostly monosyllabic, and the disyllabic words were mostly content words. Hence, Zhuan makes more use of function words than Ge.

The word frequency distribution of Ge and Zhuan is the next feature to investigate. In Ge, there are 434 unique words (type$_{Ge}$ = 434), and these words altogether appear 538 times (token$_{Ge}$ = 538). In Zhuan, type$_{Zhuan}$ = 630, and token$_{Zhuan}$ = 942.

As shown in Figure 3, the frequency distribution of words is similar to that of characters. The higher the frequency, the lower the number of words, and most of the words in both texts are hapax legomena. Of all the words in Ge, 371 only appear once. And they account for 85.5% of the total number of words; the word with the most occurrences in Ge is Junwang 君王 (emperor), and it appears 6 times. In Zhuan, there are 506 hapax legomena, which account for 80.3% of the total number of words; Zhi 之 (of, this, etc.) is the word with the most occurrences and appears 24 times in Zhuan.

Log-likelihood test shows that the difference between the hapax legomena of Ge and Zhuan is statistically significant ($p < .001$), and the most frequently occurring words is marginally significant ($p = .051$). We can conclude that the word frequencies of Ge and Zhuan are significantly different.

The proportion of hapax legomena in two texts has all risen to over 80%. This indicates that both texts have high vocabulary richness. As the gap between the two proportions is relatively small, we use TTR again to examine the vocabulary richness Ge and Zhuan. Based on the abovementioned formula, we obtained the following results:

$$\text{TTR}_{Ge} = 80.7\%.$$ \hspace{1cm} (6)

$$\text{TTR}_{Zhuan} = 66.9\%.$$ \hspace{1cm} (7)

The character richness of Ge is higher than that of Zhuan by over 13.8%, which means the vocabulary richness of poetry is higher than prose narratives when telling the same story. And this gap is even wider than the difference in character richness between the two. Again, the results explain why poetry, with the same content, is shorter but has the same narrative power as a prose narrative in telling a story.

The h-Point and a-Index of Ge and Zhuan

There are several quantitative indicators based on rank-frequency distribution of words. And h-point and a-index are two of them. We first visualize Ge’s and Zhuan’s rank-frequency distribution with the help of R.

As can be seen in Figure 4, the rank-frequency distribution of Ge differs from that of Zhuan. Although Ge and Zhuan overlapped around rank 60 and 120, the frequency of words in Zhuan is significantly higher than that of Ge, in general.

In the previous section, we discussed the ratio of the content words and function words in Ge and Zhuan from the perspective of word length. In this section, we first continue to investigate the distribution of content words and function words through h-point. As mentioned before, h-point is the fuzzy boundary between content words and function words. The words before h-point are usually function words in the text, and the words after h-point are mainly content words in the text. According to the formula, we find,

$$h_{Ge} = 5.$$ \hspace{1cm} (8)

$$h_{Zhuan} = 8.5.$$ \hspace{1cm} (9)

These results mean that the function words used in Zhuan are more than that of Ge. Next let us examine high-frequency words before the h-point in Ge and Zhuan.

In Table 1, we listed all the high-frequency words before the h-point of the two texts. Both the number and occurrences of function words in Ge are fewer than that of Zhuan. Overall, although Ge and Zhuan express the same subject, there is a considerable difference in the ratio of content words to function words between the two.

While many scholars (e.g., Hu & Yanrong, 2008; Shimosada, 2009; Xia & Chen, 1949) agree that Ge and Zhuan belong to different styles and are not two parts of the same work, the debate still exists. Thus, it is necessary to investigate the stylistic differences between the two and to provide some empirical evidence for this question. Based on h-point, we introduce indicator a-index as a stylistic differentiation. We have observed the sequence of the rank-frequency distributions of Ge and Zhuan. It can be seen that the rank-frequency distribution of the two is different. But because of the diversity of their text length, we cannot directly compare h-points from each text. The a-index overcomes that limitation and is therefore an ideal indicator for distinguishing the literary styles. Zhang and Liu (2019) compared the a-index of several modern Chinese literary styles and found that the difference between the average a-index of modern Chinese prose
narratives (11.04) and Chinese prose (8.86) style is 2.18, and they are statistically significant. As for Ge and Zhuan, we obtained the results according to the formula:

\[ a_{Ge} = 21.52. \]  \( (10) \)
\[ a_{Zhuan} = 13.04. \]  \( (11) \)

The difference between Ge and Zhuan is 8.48. Unfortunately, due to insufficient sample sizes, we could not statistically test the \( a \)-index of the two. But compared with the previous result of Zhang and Liu (2019), we argue that Ge and Zhuan have a big difference in terms of literary style and they are not two segments of a unified literary composition.

**Discussion and Conclusion**

From the results obtained by the quantitative analysis, we can see the differences in character frequency, word frequency, rank-frequency distribution, \( h \)-point, and \( a \)-index in Ge and Zhuan. The premise of these differences is that the two works are created in the same historical background, with the same subject and even similar titles. All these results also manifest their differences in style, artistic level, and popularity.
According to the meticulous classification of ancient Chinese style, Ge belongs to the Gexing Ti 歌行體 (Style of Gexing), which is the combination of the style of Ge 歌 and the style of Xing行. The style of Ge歌 is just like modern prose essays, which express the feelings of the poet freely, mingle with various mixed statements, and have no fixed direction or subject. On the contrary, the style of Xing行 lays emphasis on keeping a record of events and narrating the story according to the sequence of events. Compared with metrical poetry, Ge is freer of forms and syllables, retains the narrative characteristics of Yuefu 楯府 (Songs of the Music Bureau), and has similarities with other styles of narrative literature. Despite this, the poetic attributes of Ge歌 still determine that it surpasses the meticulousness and exquisiteness of Zhuan在 in terms of wording, rhythm and lyrics, and lyricism.

Ge is much more limited in text length, which is better for the diffusion. But it must compensate for this disadvantage in some way, to express the same scale of information as Zhuan. To start with, Ge reconciles the limited text length and information by expressing richer meanings through the words with “four strong”: strong exaggeration, strong emotion, strong symbolism, and strong impact.

First, using imaginary figures rather than actual figures is a principal counting method of Ge. As summed up in the size of the harem, Ge chooses the imaginary writing of “Jia Li Sanqian Ren” 佳麗三千人 (three thousand harem) like “Bai Fa San Qian Zhang” 白髮三千丈 (White hair trails three thousand Zhang) in the fifteenth “Qiupu Ge” 秋浦歌 (Song of Qiupu) by Li Bai, “Fei Liu Zhi Xia San Qian Chi” 飛流直下三千尺 (Its torrent dashes down three thousand feet from high) in Li Bai’s “Wang Lushan Pubu” 望廬山瀑布 (Watching the Lu Mountain Falls), and other similar statements. The exaggerated techniques they used are exactly the same; on the other hand, Zhuan always writes the numbers accurately, like “Three Furen 三夫人, Nine Pin 九姊妹, Twenty-seven Shifu 二十七世婦, Eighty-one Yuqi 八十一御妻,” etc. This feature of Zhuan embodies the characteristics of legendary prose narratives in the Tang Dynasty.

Second, there are many words with tragic meanings in the vocabulary of Ge, whereas Zhuan uses neutral words rather than tragic words. For example, “Gu Deng” 孤燈 (lonely lamp-wick), “Ye Yu” 夜雨 (night rain), “Wu Tong” 梧桐 (Chinese parasol), and other words in Ge render the feeling of cold and sadness of Tang Xuanzong’s return to the palace all alone. However, in Zhuan, no such words appear. These emotive words satisfy the special needs of poetry for artistic conception and make the work more appealing.

Third, compared with prose narratives, poetry has a stronger sense of symbolism and ambiguity. Therefore, the actual things in Zhuan are often replaced by Referential words in Ge; in addition, anaphora, cataphora, and symbols are just common techniques for making poetry. For example, “Ning Zhi” 凝脂 (creamy skin) refers to the skin. “Furong Zhang” 芙蓉帳 (lotus-flower curtain) refers to sexual intercourse. “Yuyang Pigu” 渔陽鼙鼓 (war drums) refers to the rebel force. “Huaidean Wei Di” 花細委地 (hair-pin fallen to the ground) refers to the death. Even with the reference of historical figures and historical events, Ge also uses referential words, for example, “Han Huang” 漢皇 refers to Tang Xuanzong and “Lie Tu” 列土 refers to be knighted. These subtle expressions have the potential meaning of avoiding mentioning the respectable and can also support the popular topic of “celebrate love” in the academic field. On the contrary, in Zhuan, the royal family members, such as Tang Xuanzong and Wu Shu Fei 武淑妃 (699–737), were directly named according to historical archetypes. They also directly expressed about Yang Guifei’s marriage to Shou Wang 森王 (720–775). And all these match the historical style of legendary prose narratives in the Tang Dynasty. Besides, image words with a strong symbolic meaning are further summarized in Ge, which is conducive to the creation of a unique artistic conception. For example, “Qingyun” 青雲 (blue cloud), “Xianle” 仙樂 (heavenly music), and “Jiuchong” 九重 (nine-fold), which refer to high places and heaven, appear repeatedly, creating a fairy world and inducing readers’ illusions, implying that Yang Guifei becomes an immortal. This unique language technique is not incarnate in Zhuan.

Fourth, the strong contrast of words with strong impact refers to the emphases of the author’s emotions and the shock experience brought by readers’ reading. First, Ge often uses synonyms or antonyms in the antithesis of sentences to achieve emphasis, such as “Bi” 碧 (blue) and “Qing” 青 (green) in “Shujiang Shui Bi Shushan Qing” 蜀江水碧蜀山青 (The rivers were deep blue, and the Sichuan

### Table 1. High-Frequency Words Before the h-Point in Ge and Zhuan.

| High-frequency words of Ge | Frequency | High-frequency words of Zhuan | Frequency |
|----------------------------|-----------|-------------------------------|-----------|
| 君王 (emperor)             | 6         | 之 (of, this, etc.)           | 24        |
| 在 (at, in etc.)           | 6         | 於 (at, in etc.)              | 16        |
| 下 (down)                  | 5         | 為 (for)                      | 13        |
| 無 (without)               | 5         | 其 (such)                     | 13        |
| 生 (grow, alive)           | 5         | 有 (have)                     | 12        |
|                           |           | 上 (up)                       | 11        |
|                           |           | 以 (with, by etc.)            | 10        |

|                |           | Frequency |
|----------------|-----------|-----------|
| 風 (wind)      |           | 50        |
| 雲 (cloud)     |           | 45        |
| 雨 (rain)      |           | 40        |
| 雪 (snow)      |           | 35        |
| 霞 (cloud)     |           | 30        |
mountains were green), “Baifa” 白髮 (gray-haired) and “Qing’e” 青娥 (green-clad), “Xin” 新 (new) and “Lao” 老 (old) in “Liyuan Dizi Baifa Xin, Jiaofang A’jian Qing’e Lao” 裂禦子弟白髮新，椒房阿監青娥老 (Actors, although still young, began to have hair gray. Eunuchs and waiting maids looked old in palace deep), etc. Only poetry can express this kind of emphasizing and contrasting words smoothly and naturally, and the “four strong” has greatly improved the artistic level of Ge. Second, Ge uses more content words and few function words than Zhuan, and tries to avoid repetition and maximizes vocabulary richness. The avoidance of repetition of words in metrical poetry is the basic requirement for composition. Although the lyrics of Ge are relatively loose in this aspect, the flexibility and richness of the language of poetry still encourage the author to use synonym conversion and other techniques to avoid repetition of words actively. This creates a sense of freshness, reduces reading fatigue, and is also better for dissemination. Finally, Ge uses more poly-syllabic words, alliterations, and assonance than Zhuan, which not only satisfies the rhythm of poetry but also improves the artistic level as well as the spread of Ge.

When an author chooses a style to convey one subject, he or she is also choosing a particular linguistic style. By comparing Ge and Zhuan, which expressed the same subject in poetry and prose narrative respectively, we found the two works have the following differences with respect to the linguistic features:

a. Characters and words in Ge repeat significantly fewer than in Zhuan.
b. The character and vocabulary richness of Ge are higher than that of Zhuan.
c. Ge tends to use more content words and fewer function words than Zhuan.

With the α-index, we also provide empirical evidence for the viewpoint that Ge and Zhuan are different in terms of literary style and they are not two segments of a unified literary composition.

These differences are one of the driving forces behind Ge’s higher artistic level and wider dissemination than Zhuan. Ge needs to give the same amount of information as Zhuan, but its text length is limited and it has a more fixed format than Zhuan. The author of Ge must put more efforts into the use of words and expressions. Bai Juyi has done this by expressing richer meanings through the words with “four strong,” reducing repetition, increasing vocabulary richness, and using more polysyllabic words, alliterations, and assonance.

Although the investigation of the differences between Ge and Zhuan is representative, it is not representative of all works that share the same subject but differ in style. We will continue to collect similar works and further the investigation.

Declaration of Conflicting Interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work is supported by the MOE’s (Ministry of Education) Humanities and Social Sciences Foundation of China (19YJC740115).

ORCID iD
Cong Zhang (https://orcid.org/0000-0001-5616-0626)

References
Algee-Hewitt, M., Heuser, R., & Moretti, F. (2015). On paragraphs. Scale, themes, and narrative form. Universitätsbibliothek Johann Christian Senckenberg.
Burrows, J. F. (1987). Word-patterns and story-shapes: The statistical analysis of narrative style. Literary & Linguistic Computing, 2(2), 61–70.
Chen, Y. (1947). Chang Hen Ge Jianzheng 《長恨歌》箋證 (元白詩箋證稿之一) [Annotation of The Song of Everlasting Regret]. Qinghua Xuebao 清華學報/Tsing Hua Journal, 14(1), 1–34.
Chesnokova, A., Zvyagin, S., Viana, V., Jandre, J., Rumbeahe, A., & Ribeiro, F. (2017). Cross-cultural reader response to original and translated poetry: An empirical study in four languages. Comparative Literature Studies, 54(4), 824–849.
Cox, D. R., & Brandwood, L. (1959). On a discriminatory problem connected with the works of Plato. Journal of the Royal Statistical Society. Series B (Methodological), 21, 195–200.
Deng, Q., & Gao, C. (2005). Chang Hen Ge Yu Chang Hen Ge Zhuan 《長恨歌》與《長恨歌傳》 [The Song of Everlasting Regret and the biography of the Song of Everlasting Regret]. Xibei Shida Xuebao (Shehui Kexue Ban) 西北師大學報 (社會科學版)/Journal of Northwest Normal University “Social Science Edition”, 3, 37–41.
Elliott, J. (2016). Vocabulary decay in category romance. Digital Scholarship in the Humanities, 31(2), 321–332.
Fabb, N., & Halle, M. (2008). Meter in poetry: A new theory. Cambridge University Press.
Fleay, F. G. (1874). On metrical tests as applied to dramatic poetry. Transactions of the New Shakespeare Society, 1–39.
Hoover, D. L. (2007). Corpus stylistics, stylistometry, and the styles of Henry James. Style, 41(2), 174–203.
Hu, K., & Yannong, W. (2008). Chang Hen Ge De Xu Yu Zhuang 《長恨歌》序與傳 [The preface and biography of The Song of Everlasting Regret]. Shehui Kexue Zhuanxian 社會科學戰線/Social Science Front, 5, 89–95.
Huang, W., & Liu, H. (2017). Cipin Fenbu Canshu Keyi Xifen Han yi Xinwen Yuti Ma 詞類分佈參數可以細分漢語新聞語體嗎? [Can parameters in word frequency distribution distinguish Chinese news genres?]. Yuanxiao Xueyu Yu Yanjiu 語言教育與研究/Language Teaching and Linguistic Studies, 4, 79–86.
Kroll, P. W. (1990). Po Chü-i’s “song of lasting regret” a new translation. Tang Studies, 1990, 97–105.
Levy, H. S. (1957). The career of Yang Kuei-Fei. T’oung Pao, 45(1), 451–489.
Zhang, C., & Liu, H. (2018). Cipin Jiliang Zhibiao Yu Hanyu Yuti Yanhua [Word frequency indicators and Genre Variation in Chinese]. Waiyu Jiaoxue/Foreign Language Education, 2, 9–20.

Zhang, C., & Liu, H. (2019). Chinese evolution in recent 150 years: A diachronic study of word frequency in the Gospel of Mark. Journal of Chinese Linguistics, 47(2), 497–530.

Zhou, X. (2003). Chang Hen Ge Yanjiu《長恨歌》研究 [The Study of The Song of Everlasting Regret]. Bashu Press.