A Comparative Genre Analysis of Chinese Folk Song Huar Based on Digital Humanities

Qianqian Tang¹, Yan Xu²* and Ying Yuan³

¹²³ School of information science, Beijing Language and Culture University, Beijing, Beijing, 100083, China
*201921198622@stu.blcu.edu.cn

Abstract. As an intangible cultural heritage, the folk song Huar is a precious name card of Chinese national culture. Its protection and inheritance are of great significance for promoting multi-ethnic cultural identity and enhancing cultural confidence. Genre research is an important topic in the study of Huar. However, current Huar genre researches are mainly based on qualitative methods. Digital Humanities is an emerging discipline that applies computing and digital technology to the study of the humanities. Comparing to traditional methods, Digital Humanities is a more accurate and efficient method of style analysis on literary texts. Through the methods of Digital Humanities, we can objectively show and verify the similarities and differences between Huar genres, and make up for the lack of objectivity in the traditional Huar genre research. Therefore, in this paper we combined qualitative and quantitative methods to make a comparative style analysis on Hehuang Huar and Taomin Huar from two aspects, namely rhythm and sentence pattern. Furthermore, we explored the reasons in depth behind these surface rules from the perspectives of music, literature, nationality and history. We simultaneously use hierarchical clustering to verify the discrimination of genres on specific style features. The result of this research shows that the research perspective and methods of digital humanities can provide objective conclusions for Huar researches, which can provide data support and scientific basis for researchers or enthusiasts to distinguish, appreciate and study Huar.

1. Introduction
Huar is a distinctive folk song, which is popular in northwest China, such as Gansu, Qinghai, Ningxia and Xinjiang. It is composed and performed in Chinese by numerous ethnic groups, including Han, Hui, Dongxiang, Salar, Tibetan, Mongolian and so on. The singing content of Huar is all-embracing, and the popularity of Huar has spanned hundreds of years, a wide range of regions and various ethnic groups. It possesses great cultural tolerance and artistic vitality. On May 20th, 2006, with the approval of the State Council of China, Huar was included in the first batch of China's national intangible cultural heritage list. As an intangible cultural heritage, Huar has great artistic and cultural value, and is a precious name card of Chinese national culture. The protection and inheritance of Huar is of great significance for promoting multi-ethnic cultural identity and enhancing cultural confidence. However, in recent years, the study of Huar has hit a bottleneck. Researchers do not have enough understanding of the necessity of interdisciplinary research, which leads the research on Huar into the dilemma of "single subject" isolation.

Digital humanities, which originated from humanistic computing in the late 20th century, is a new discipline that applies computing or digital technology to the study of humanities. Thanks to the rapid
development of computer technology and Internet technology, digital humanities has developed rapidly, from the initial field of linguistics, to literature, history, music, film and other fields, and its concepts and boundaries are still being continuously extended and developed.

A typical symbol of the application of digital humanistic in literature is that Bingzao Chen from the University of Wisconsin first used computer technology to make a statistical analysis based on the word frequency of the text of A Dream of Red Mansions, to discuss the author issue which has been debated for a long time in the field of Redology [1]. This research has brought new perspectives and methods for the style analysis of literary works, and show researchers the giant potential of digital humanities on literary research. Since then, more and more researchers have attempted to apply digital humanities to the analysis of literary style and literary genres. For example, Ying Liu and Tianju Xiao use quantitative methods to systematically compare the style of Jin Yong's and Gu Long's novels from multiple aspects, such as the average paragraph length, word length dispersion, notional words, function words, high-frequency words, part of speech and so on [2]. Chulin Fan and Ying Liu extracted 387 text style features from Lu Xun's letters, novels and essays, and made a comparative analysis of the text styles of the three literary styles through multi-dimensional analysis [3]. Professors Haitao Liu and Xiaxing Pan verified the difference and correlation between new poetry and modern poetry through clustering experiments based on the usage of parts of speech [4].

The study of Huar genres is an important topic in the study of Huar. Huar is generally divided into two genres, namely "Hehuang Huar" and "Taomin Huar". "Hehuang Huar" refers to Huar that is popular in the geographical region centred on the confluence of the Yellow River and Huang River, including Qinghai Huar, Gansu Linxia Huar and Ningxia Huar. Taomin Huar is another type of Huar that is popular in Taomin region of Gansu Province. Different geographical and cultural environment endows Hehuang Huar and Taomin Huar with their distinctive styles and features. Then how are the similarities and differences between these two genres reflected in the texts? So far, there are no researchers who have made a comparative study of the text styles of the two Huar genres from the perspective of digital humanities. Therefore, in this paper, we established a small Huar genre database. Based on this database, we will use both quantitative and qualitative research methods to measure and compare the styles of the two Huar genres from the perspectives of rhythm and sentence structure. We will simultaneously make hierarchical clustering experiments to verify the discrimination to Huar genres on specific style features.

2. Construction of Huar Text Database

2.1. Data Collection

Digital text resources of Huar are quite rare, therefore we struggled a lot to collect the digital text data of Huar. The collecting process consists of the following four steps.

- Step 1: Search resources in the Superstar Digital Library. We finally found five professional books which record a large amount of Huar lyrics [5-9].
- Step 2: Use Python to extract and initially clean the Huar text.
- Step 4: Proofread manually.
- Step 5: Data cleaning. Remove unnecessary markers in the text, such as "male:”, "female:”, etc., and de-duplicate the data.

2.2. Word segmentation and POS-tagging

In order to compare the word segmenting results of several commonly-used tools, first of all, we made a manual word segmentation on a random sample of 50 pieces of data. In addition, we manually sorted out all the vocabularies recorded in a professional Huar dictionary, which is called Huar Folk-custom dictionary. Furthermore, we extracted Huar vocabularies from notes of the five books mentioned above and Huar papers, and manually collated them. Finally, we constructed a Huar dictionary with 3324 entries, and used it to improve the accuracy of word segmentation and POS-tagging results.
Based on the segmented samples and the Huar dictionary, we compared the accuracy of the results of each tool. We finally used Jieba for the word segmentation and POS-tagging of the text. Through all the steps, we constructed a small Huar Database, which is of the following size below:

| Genres     | Texts | Words | Characters |
|------------|-------|-------|------------|
| Hehuang Huar | 1909  | 65385 | 109263     |
| Taomin Huar | 570   | 41387 | 73036      |
| Total      | 2479  | 106772| 182299     |

3. Prosodic Feature Analysis

Huar has poetic characteristics. Its lyrics are very particular about the beauty of rhyme. The characteristics of the last characters of each sentence can often reflect the rhyme characteristics of the work. Therefore, in this section we will analyse the distribution and the linear combination of the finals of the last character of each sentence, so as to explore the rhyming preference and rhyming schemes of Hehuang Huar and Taomin Huar.

3.1. Frequency distribution of finals of last character in each line

The distribution of finals of the last character in each line can reflect the rhyme preference of the work. As Figure 1 shows, the distribution of rhyme finals of Hehuang and Taomin Huar is roughly similar. Kaikou Hu and Qichi Hu finals are mainly used, followed by Hekou Hu finals, Cuokou Hu syllables are rarely used. Both Kaikou Hu and Qichi Hu belong to open-mouth rhyme. Open-mouth rhyme sounds loud, and is very suitable for the free and unbridled ballad Huar, which is sung in open fields and mountains. However, there are still differences between the two Huar genres on the rhyming preference. Hehuang Huar prefer Qichi Hu, especially finals "i" and "iao", whose frequency is obviously higher than that in Taomin Huar. By contrast, besides of finals "i" and "iao", Taomin Huar also uses finals beginning with "e" heavily, which is obviously different from Hehuang Huar in this point.

Figure 1. Frequency distribution of finals of last characters of each line.
3.2. High-frequency rhymes analysis
In order to figure out the reason for this difference in rhyming, we found out the high-frequency rhymes of the two types of Huar and draws bubble diagrams. The size of the bubbles represents the frequency of the rhymes. It can be seen from the figure below that a large number of function words are used as rhymes in Huar. In Hehuang Huar, the frequency of function words "liǎo" and "li" accounts for more than 30% of the rhymes, which is a typical rhyming feature of Hehuang Huar. This phenomenon shows up as massive use of Qichi Hu in Hehuang Huar at finals level, which is in consistent with the high-pitched and melodious singing characteristics of Hehuang Huar. The three most common function words used as rhymes in Taomin Huar are "liǎo", "ne", and "li" in order of frequency. The extensive use of "ne" is a clear distinction from Hehuang Huar, which can explain the frequent use of finals "e". "li" is a word commonly used in Song and Yuan dynasty, which is more classical and graceful, while "ne" is more vernacular. Using “ne” more frequently than “li” reflects the simple musical personality of Taomin Huar.

![Figure 2. High-frequency rhymes of Hehuang Huar.](image)

![Figure 3. High-frequency rhymes of Taomin Huar.](image)

3.3. Linear combinations of finals of last character in each line
The linear combination of finals of last character in each sentence can reflect the rhyming schemes of Huar. Finals consist of three parts: head vowel, main vowel and tail vowel. Generally speaking, it could be regarded as rhyme as long as the main vowels are the same or similar. In this section, we classified the finals groups in Hehuang Huar and Taomin Huar separately according to Chinese New Rhyme (14 Rhymes), and listed them in table 2 by the order of frequency.

| Genres        | Monorhyme | Alternating Rhyme | No Rhyme in First Line |
|---------------|-----------|-------------------|------------------------|
| Hehuang Huar  | 28.91%    | 2.73%             | 2.73%                  |
| Taomin Huar   | 24.99%    | 1.33%             | 1.33%                  |

It can be seen from the table above that the rhyme scheme of Hehuang Huar is more flexible by contrast. Besides monorhyme, alternating rhymes are widely used in Hehuang Huar, which is a typical feature of Hehuang Huar. This rhyming pattern can create a leaping beauty of rhythm. Monorhyme is the main rhyme scheme of Taomin Huar, among which there exist a phenomenal that first lines don't rhyme. This can be attributed to the existence of "three characters head" in Taomin Huar, which is a
stylized set of two lines that is located in the beginning of a piece of Huar and is meant to determine the rhyme of the following lyrics. Three characters head does not rhyme, which causes the phenomenal that some first lines don't rhyme.

4. Analysis of sentence structure features

According to Professor Yulin Wu, Hehuang Huar and Taomin Huar have their own characteristics on sentence structure. Therefore, in this section we will analyse the distribution of sentence length of Huar. Furthermore, we will extract multiple sentence structure features, and use the method of hierarchical clustering to explore on which aspects of sentence structure can be distinguished between the two Huar genres.

4.1. Frequency distribution of sentence length

From the figure, we discovered that eight-character sentences are the most frequently used sentence type in Hehuang Huar, accounting for 45.4%, and is followed by nine-character sentences and seven-character sentences. Seven-character sentences account for the vast majority of the sentences in Taomin Huar, with a proportion of 63.8%. It is worth noting that special three-character sentences account for a large proportion of Taomin Huar, second only to seven-character sentences, while three-character sentences are not very common in Hehuang Huar. These differences in sentence structures can be explained from the following two aspects:

A. Different nationalities of creators and singers

The main creators and singers of Taomin Huar are the Han nationality, thus its sentence structure has been influenced by traditional seven-character poems. However, the rhythm of Hehuang Huar, which has a minority style, is dominated by three syllables. Based on this feature, a large number of eight-character sentences are created, which is a distinct feature of Hehuang Huar in sentence structure.
B. Different singing styles and forms

Taomin Huar are generally sung in the form of duet or antiphon, which is a great test of people's improvisation and coping skills. Under the pressure of time, some formulas that do not have much practical meaning usually are used as the first sentence, to fix the rhyme of the subsequent lyrics quickly. "Three-character head" is a common type of formula, so there exist plenty of three-word sentences in Taomin Huar.

4.2. Hierarchical clustering based on sentence structure features

In order to find out the significant differences between different genres of Huar in sentence structure, we divided the text into eight sub-databases according to the source and the genre of the texts. Upon these sub-databases, we extracted five sentence features separately, including the average sentence length, the mode of sentence lengths, the average number of sentences, the mode of sentence number and sentence length dispersion, on which we carried on a series of clustering experiments. The specific results are as follows:

![Figure 5. Hierarchical clustering based on the average sentence length and the mode of sentence lengths.]

The letters along the vertical axis in the figure refers to eight sub-databases, among which the letters are the initial letters of each book, and "0" represents Hehuang Huar while "1" represents Taomin Huar. As can be seen from the figure, sub-databases with "0" and "1" are perfectly divided into different categories, which proves that there are significant differences in sentence length feature between Hehuang Huar and Taomin Huar.

5. Conclusion

From the perspective of digital humanities, this paper combines quantitative and qualitative research method to compare and analyze the text styles of Hehuang Huar and Taomin Huar, from rhythm and sentence structure two aspects.

(1) Through the analysis of the distribution and linear combination of the finals of last characters of sentences, we discovered the similarities and differences in rhyme preference and rhyme scheme between Hehuang Huar and Taomin Huar. The two genres of Huar both prefer open-mouth rhyme. Hehuang Huar is inclined to use Qichi Hu, and uses "liǎo", and "li" extensively as end rhyme, which
contributes to the high-pitched and melodious singing style of Hehuang Huar. Taomin Huar prefers to use Kaikou Hu, especially finals beginning with "e". This feature is mainly reflected in the massive use of the character "ne" at the end of lines, which reflects the simple musical personality of Taomin Huar.

In terms of rhyme scheme, the pattern of Hehuang Huar is relatively flexible. There are two typical rhyming patterns in Hehuang Huar, namely monorhyme and alternating rhyme. Alternating rhyme can bring a kind of leaping rhythmic beauty, which has become one of the typical features of Hehuang Huar. By contrast, monorhyme is the main rhyme scheme of Taomin Huar. There are also some cases in Taomin Huar when first line doesn’t rhyme, which is attributed to the massive existence of “three-character head”.

(2) By drawing the frequency distribution diagram of sentence length of two Huar genres, we discovered that eight-character sentences are the most frequently used sentence type in Hehuang Huar, while seven-character sentences account for the vast majority of the sentences in Taomin Huar, followed by the special three-character sentences. These differences can be attributed to different nationalities of singers and different singing forms of the two Huar genres.

(3) Through hierarchical clustering based on various sentence structure features, and it was found that there were significant differences in sentence length features between the two genres.

The result of this research shows that the research perspective and methods of digital humanities can provide objective conclusions for Huar researches, which provides data support and scientific basis for researchers or enthusiasts to distinguish, appreciate and study Huar.

Acknowledgments
This work is supported by the Fundamental Research Funds for the Central Universities, and the Research Funds of Beijing Language and Culture University (Approval number: 21YCX175), by "Construction of Ningxia Huar digital promotion platform" project of China National Arts Fund (No. 2018-a-03-(177)-0521).

References
[1] Chen, B.Z, Hu, Q. (2002) On the Later Forty Chapter of the Novel. A Dream Of Red Mansions, 03: 267–282.
[2] Liu, Y., Xiao, T.J. (2014) A Stylistic Comparison Between Jin Yong’s Novels and Gu Long’s Novels. Journal of Tsinghua University (Philosophy and Social Sciences), 05: 135–147.
[3] Fan, C., Liu, Y. (2020) A Comparative Study on Three Genres of Lu Xun Based on Multidimensional Analysis. Journal of Chinese Information Processing, 10: 94–104.
[4] Liu, H.T., Pan, X.X. (2015) Quantitative Properties of Chinese Contemporary Poetry. Journal of Shanxi University (Philosophy and Social Science Edition), 02: 40–47.
[5] Wu, Y.L. (2008) General Introduction to Chinese "Huar". Ningxia people's Publishing House, Yinchuan.
[6] Long, Y. (2005) Extraction of "Huar": the Hezhou Huar volume. Gansu Culture Publishing House, Lanzhou.
[7] Long, Y. (2005) Extraction of "Huar": the Lianhuashan Huar volume. Gansu Culture Publishing House, Lanzhou.
[8] Zhu, Z.L. (2002) "Huar" about love. Dunhuang Literature Publishing House, Lanzhou.
[9] Xue, L., Ke, Y.(2012) Selected songs of "Huar" in Northwest China. Qinghai people's Publishing House, Qinghai.