A Statistical Study of lyrics of "Huar" from Northwest China

Yan Xu[1], Qianqian Tang[2*] and Ying Yuan[3]

[123]School of information science, Beijing Language and Culture University, Beijing, Beijing, 100083, China

Abstract. As an intangible cultural heritage, "Huar" from Northwest China is a folk song created and shared by many ethnic groups in Gansu, Qinghai and Ningxia provinces. It is a precious card of Chinese national culture. However, current research on "Huar" is mainly based on qualitative methods. This paper uses statistical methods to study the lyrics of "Huar". First, the word frequency of the lyrics of "Huar" is analysed statistically. Then, the lyrics of Hezhou Huar and Taomin Huar are compared and analysed from the perspective of quantitative linguistics ("Huar" mainly includes Hezhou Huar and Taomin Huar). By comparing three quantitative indicators, it is concluded that the lexical richness of Taomin Huar is higher than that of Hezhou Huar. Based on the frequency distribution of parts of speech, the similarities and differences of the use of parts of speech between Hezhou Huar and Taomin Huar are found. This paper uses statistical methods to analyse "Huar", which has certain research value and social value.

1 Introduction

"Huar" is a regional folk song widely popular in Gansu, Qinghai, Ningxia, Xinjiang and other provinces and regions in China. It is a regional folk song performed in Chinese by nine echic groups, including Han, Hui, Tibetan, Tu, Salar, Dongxiang, Bao'an, Yugur and Mongolia. It is a landmark oral art in folk songs of western China and even Chinese folk songs [1]. "Huar" has a history of hundreds of years. In 2009, it was rated as one of the world intangible cultural heritage projects by UNESCO. Originated from ancient river state and spread in the Ming Dynasty, it was widely spread through the silk road. "Huar" is the best way for people to express joys and sorrows. It comes from life; Its melody is high and generous, melodious and graceful; Its lyrics are improvised with emotion. It is a wonderful flower in Chinese and world folk songs.

But as a world-class cultural card in China, "Huar", a folk culture brand with local color and flavor, has faced an awkward situation in recent years. In the 1980s, a large number of folk singers once emerged in the northwest "Huar", which had a glorious history and a broad mass base. However, since the 1990s, the number of "Huar" singers in Northwest China has gradually decreased, and there are many problems in the inheritance and development of "Huar". There are few high-level "Huar" singing talents. At present, China has evaluated 11 national representatives and 50 provincial representatives of "Huar". Because most of them are dead and old, in view of the current grim situation, the digitization and networking of "Huar" performance art is particularly important. As a school with rich cultural heritage, Beijing Language and Culture University shoulders the important mission of spreading and vigorously carrying forward the excellent traditional culture and promoting the development of Chinese excellent traditional culture.

The study of "Huar" has experienced the period of germination, formation, development and prosperity, and its research is becoming more and more comprehensive. But it is mainly qualitative. At present, there is little research on the text features of "Huar" from the perspective of statistics and metrology. This paper studies the lyrics of "Huar" by statistical methods. Firstly, the frequency of the lyrics of "Huar" is analyzed statistically. Then, the lyrics of Hezhou and Taomin Huar are compared and analyzed from the perspective of quantitative linguistics ("Huar" mainly includes Hezhou Huar and Taomin Huar). Through the comparison of three measurement indexes, it is concluded that the lexical richness of Taomin Huar is higher than that of Hezhou Huar; Through the statistics of the usage frequency of parts of speech, we find the similarities and differences in the use of parts of speech between Hezhou Huar and Taomin Huar, and further verify that there are obvious differences in the use of auxiliary words, location words, numerals and adverbs between the two schools through clustering experiments. It is of great significance to study “Huar” in Northwest China by statistical methods.

2 Related work

From the beginning of the 21st century to now, the research on "Huar" has entered a prosperous period. Some folklore scholars, anthropologists, language scholars and social scholars have joined the research team of "Huar", and the scope of the research has been expanded accordingly. Some high-level research has also been set up. For example, Caoqiang was established in 2009 by the

*201921198622@stu.blcu.edu.cn

National Social Science Fund Project "Huar" language folk custom research. Scholars began to study the flowers from many angles. Based on the previous studies on "Huar", the author studies it from the perspectives of anthropology, musicology, folklore, sociology and communication. The representative monographs and papers are: the general theory of Chinese "Huar" by wuyulin [2], Taomin Huar research by qixiaoping: the "Huar" survey of Tiancun in the perspective of living space, and "a brief analysis of the embodiment of Hehuang agricultural folk custom in flowers" by wangyoufu. Zhuxiaofeng has "Taomin 'Huar' research in the sociological perspective" and Cao Qiang 'the word use of the words of the Huar' lyrics from the perspective of communication science.

In the research on "Huar", school research is a very important topic. In his essay on the characteristics, schools and rules of "Huar", Mr. Liu Kai said: "the 'Huar' are generally recognized as two schools: Qinghai, Linxia (formerly known as Hezhou) and Ningxia "Huar". This group of "Huar" is commonly known as "Hehuang Huar" because it is popular in the Yellow River areas of Gansu, Ningxia and Qinghai and the coastal areas of Huangshui River in Qinghai. "Huar" in Taomin area of Gansu Province is another school. After many studies, Professor wuyulin agreed with the division, and called these two schools "Hezhou type Huar" and "Taomin type Huar".

Zhang Xiaojin studies "Huar" from the perspective of Quantitative Linguistics [3]. The study uses the method of quantitative linguistics to investigate the word frequency distribution and the word character of the rhyme, thus verifying the internal law of the "Huar" text; And the author analyzes "Huar" and the May 4th New Poetry Movement, and finds that there is a close relationship between them. Therefore, the measurement method can not only provide objective verification for the traditional qualitative method, but also find some text features and rules that are difficult to find by traditional methods. "Huar", as a special northwest folk song, is not only rich in content but also in various forms. Its characteristics and internal rules are worth studying and exploring. The method of quantitative linguistics will open the door to the new world for the study of flowers. But the research on "Huar" is mainly qualitative. At present, the research on the text features of "Huar" from the perspective of measurement is still rare.

3 Research based on statistics

Quantitative linguistics is a branch of linguistics that studies the structure and development of language through quantitative methods based on real corpus. With the help of the method of quantitative linguistics, we can extract the data of various language units from the text, and make a quantitative analysis of their characteristics, so as to get an objective and empirical conclusion.

3.1 Corpus

3.1.1. Data Acquisition. Due to the scarcity of "Huar" text resources, data acquisition is very difficult. First of all, we found five professional books of "Huar" from the Superstar Digital Library, and then generated Huar corpus from the books. The process includes three steps: OCR recognition of the text in the book, python extraction of "Huar" text, and manual proofreading.
3.1.2. Data cleaning. Remove unnecessary markers in the text, such as "male:" and "female:" etc., and de-duplicate the texts.
3.1.3. Word segmentation and part of speech tagging. Use a professional tool Jieba to carry out word segmentation and part of speech tagging on the texts. At the same time, we extract 3324 "Huar" vocabulary and from books and papers, and constructed a dictionary of "Huar", to improve the accuracy of word segmentation and part of speech tagging.

Based on the above steps, we got a small Huar corpus of the following size below:

Table 1. The size of Huar corpus.

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

3.2 word frequency analysis

According to the statistics of the word frequency of the lyrics of "Huar", we arrange them according to the frequency from large to small(See Figure 1):

Figure 1. The frequency of words in Huar's Lyrics.

According to the statistics of usage frequency, nouns account for the highest proportion of "Huar", especially those with gender characteristics, such as "a ge", "ga mei", "mei mei", "daughter-in-law", etc.

3.3 lexical richness

Vocabulary richness is an index to measure the vocabulary size of a text. In this section, three parameters are selected to measure the vocabulary richness: the type token ratio, the proportion of notional words, and the relative repetition rate. The type token ratio is the ratio of type to example in text; The proportion of notional words is based on the H-point estimation of the ratio of notional words in the text; The repetition rate is the estimation of text repetition based on word frequency. The relative repetition rate is the standardized form of repetition rate, which makes the calculation results of different texts comparable; The three parameters are proportional to lexical richness. The measurement results of the two types of texts are as follows:

Table 2. Measurement of vocabulary richness of Hezhou Huar and Taomin Huar.

school	Type case ratio	Proportion of notional words	Relative repetition rate
Hezhou Huar	0.1534	0.5966	0.9074
Taomin Huar	0.1897	0.6687	0.9368

From this table, the three indexes show that the vocabulary richness of Taomin Huar is slightly higher than that of Hezhou Huar.

3.4 usage of word

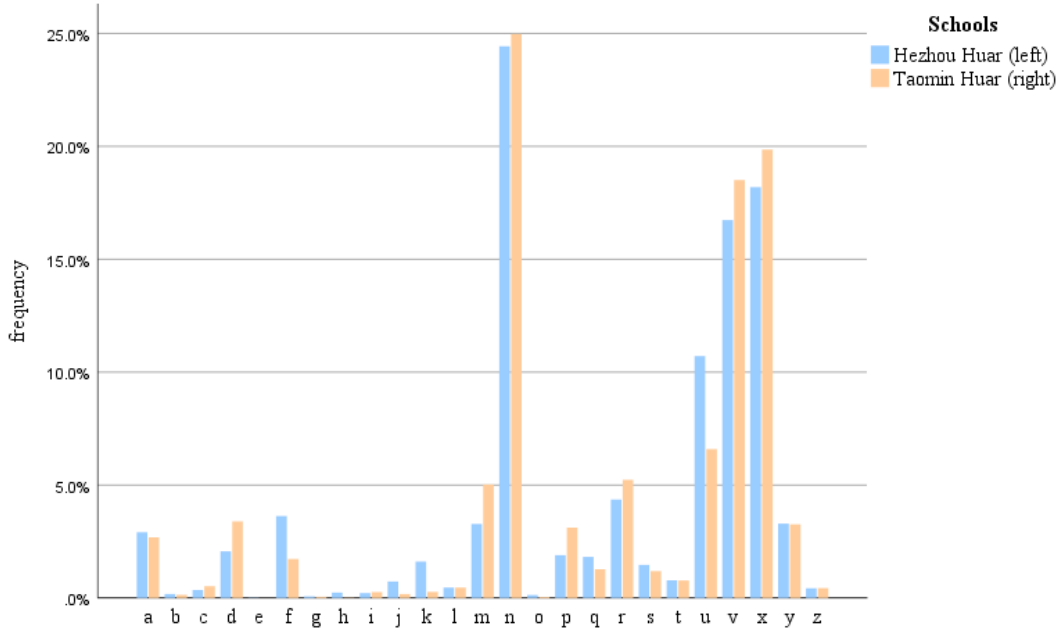

Figure 2. Frequency distribution of parts of speech.

Research shows that the frequency of specific words in text can distinguish different text styles. For example, Liu Ying and Xiao Tianjiu [4] show that the frequency of using nouns, verbs, pronouns, adjectives, adverbs, prepositions, auxiliary words, numerals, directional words and conjunctions can effectively distinguish the novels texts of Jin Yong and Gu Long. Is there any difference in the usage frequency of specific parts of speech between Hezhou and Taomin Huar? Can these differences distinguish the text styles of the two schools? In order to answer these questions, this section first investigates the distribution of parts of speech of Hezhou and Taomin Huar, calculates the usage frequency of each part of speech, and draws the frequency distribution of parts of speech, as follows:

From this figure, the usage of the words of Hezhou and Taomin Huar is both the same and different. The similarities are as follows:

- Firstly, nouns appear most frequently. This is because the rhetoric of "Fu Bi Xing" is widely used in "Huar", which makes people use a lot of "images" when they create "Huar" Image is mainly expressed as noun in language, so there are abundant nouns in "Huar".
- Secondly, there are many verbs and few adjectives in Huar texts. By calculating the busman coefficient of two kinds of text, we find that both texts show strong initiative.
- Thirdly, there exist a large number of functional words.

The difference is that there are obvious differences in the usage frequency of the seven parts of speech, namely, auxiliary words, verbs, orientation words, numeral words, adverbs, prepositions and pronouns between Hezhou Huar and Taomin Huar.

4 Conclusion

The research on "Huar" in northwest China generally adopts qualitative method, and seldom uses statistical or quantitative linguistic methods. This paper uses statistical methods to study the lyrics of "Huar". Firstly, the frequency of the lyrics of "Huar" is statistically analyzed. It is found that nouns account for the highest proportion of "Huar", especially those with gender characteristics, such as "a ge", "ga mei", "mei mei", "daughter-in-law", etc; Then from the perspective of quantitative linguistics, this paper makes a comparative analysis of the lyrics of Hezhou Huar and Taomin Huar; Based on the statistics of the usage frequency of parts of speech, the similarities and differences of the use of parts of speech between Hezhou Huar and Taomin Huar are found. This paper uses statistical methods to analyze "Huar", which has research value and social value.

Acknowledgments

This research 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. Zhao Z.F.(1989), General theory of Huar. Qinghai People's publishing house, Xining.
2. Wu Y.l.(2008), General theory of Chinese Huar. Ningxia people's publishing house, Yinchuan.
3. Zhang X.J., Liu H.T. (2017) Quantitative characteristics of Chinese folk song "Huar". Journal of Ningxia University (HUMANITIES AND SOCIAL SCIENCES),39 (05): 76-80
4. Liu Y., Xiao T.J.(2014). Research on Metrological style of dream of Red Mansions. Journal of dream of Red Mansions, 2014 (04): 260-281