International Journal of Education and Humanities
ISSN: 2770-6702 | Vol. 16, No. 2, 2024

Comparative Study of Chinese Translations of "Gitanjali" based on Natural Language Processing

Taochen Huang*

School of Public Management and Law, Fujian Agriculture and Forestry University, Fuzhou, Fujian, 350002, China
* **Corresponding Author's Email:** 1912100533@mail.sit.edu.cn

Abstract: This paper employs the HanLP natural language processing model to conduct a quantitative text analysis of four Chinese translations of "Gitanjali". The analysis examines text features such as character and sentence count, lexical properties, lexical density, and high-frequency words. The results indicate that, except for Wu Yan's translation, the text features of the other translations are relatively similar in terms of data distribution. Wu Yan's translation has higher character and sentence count, and usage of lexical properties than the other translations, but it has the lowest lexical density, indicating a lower information density. The study also reveals that the choice of lexical properties and high-frequency words in each translation is influenced by the translator's personal language preferences. Although the types of high-frequency words are relatively fixed, their frequencies vary significantly. Based on these findings, the study further uncovers the intrinsic logic and structural relationships among different text features, offering significant insights for the study of specific literary and translation texts, as well as broader translation and linguistic research. This paper demonstrates the effectiveness of HanLP in literary and translation research, highlighting the potential of natural language processing models in large-scale text analysis.

Keywords: Translation Studies; Gitanjali; Chinese Translations; Comparative Analysis; Natural Language Processing; HanLP.

1. Introduction

"Gitanjali" is the masterpiece of Indian poet Rabindranath Tagore, representing the pinnacle of his poetic creation. "Gitanjali" consists of 103 prose poems, directly translated as "Song Offerings." It uses poetry as an offering to the god, with the poet himeself expressing, in the first-person narrative perspective, the communication and fusion with the ineffable "you," referring to the god, and the desire for spiritual fulfillment, along with various complex emotions. The poems are imbued with a strong sense of pantheism. Additionally, they contain profound reflections on reality, reflecting the poet's concerns and hopes for the future of his homeland and his nation. Originally written in Bengali, Tagore personally translated "Gitanjali" into English during his journey to Europe in 1912. After being brought to London and read by several literary and artistic friends (such as Sir William Rosenstein and W.B. Yeats, who spoke highly of this anthologies and pushed for its dissemination), Gitanjali was rapidly published and became a sensation throughout European literature, thas had an unprecedented impact worldwide, which has also left an indelible mark on the history of human literature[1]. The research on Chinese translations of "Gitanjali" is relatively scarce. Zeng Qiong pointed out in her 2012 paper "Tagore in World Literature: Introduction and Research on 'Gitanjali' " that there is a enormous lack of studies on the Chinese translation of "Gitanjali", with most existing research focusing on the comparison between Chinese translations and the original English version[2]. The earliest relevant study was conducted by Liang Fang in 2016, titled "Tracing the Beautiful Footsteps of Translators: A Microscopic Comparison of Chinese Translations of "Gitanjali" [3]. In this paper, the author utilized three Chinese translations of "Gitanjali" and selected the first poem in this anthologies for quantitative analysis, observing its features in terms of word count, sentence count, Lexical properties, and lexical density. While this study pioneered quantitative text analysis of "Gitanjali", it was limited by the lack of effective statistical tools at the time and only analyzed the data of one poem, thus lacking reflection and analysis on the overall text features of Chinese translations of "Gitanjali". Furthermore, in 2018, LIU Jia-chang conducted a study titled "Corpus-based Comparative Study on Chinese Versions of Gitanjali-A Case Study on Bingxin's and Bai Kaiyuan's Versions", using a corpus to analyze Chinese translations of "Gitanjali"[4]. This research method can be traced back to Hu Xianyao's paper "A Study on the Word Features of Chinese Translated Novels Based on Corpus"[5]. While LIU Jia-chang's analysis based on the ICTCLAS corpus was impressive and represented the first comprehensive comparative analysis of the Chinese versions, the study was limited in scope. The author only investigated aspects such as the token-type ratio, average sentence length, and idiom usage, lacking statistical analysis of other text elements.

Based on the aforementioned circumstances, this paper will focus on the horizontal comparison of Chinese translations of "Gitanjali". Four Chinese translations from different eras will be selected for analysis: the 1955 translation by Bing Xin, the 1986 translation by Wu Yan, the 2010 translation by Wang Li, and the 2019 translation by Xiao Xingzheng, the selection standard is based on the course of the eras[6-9]. HanLP, a natural language processing model, will be employed to conduct quantitative text analysis on these four translations, observing their data distribution features in terms of character and sentence count, lexical properties, lexical density, and high-frequency words. The aim of this study is to present as comprehensively as possible the differences in textual features among these translations, thereby perfecting the comparative research among Chinese translations of "Gitanjali" and filling the research gap in the field. Additionally, this paper hopes that the quantitative text analysis method based on HanLP used in this study can inspire more scholars, and the related experimental design

logic and methods can be more widely applied in literary research, opening a new door for literary studies.

2. Methods for Quantitative Text Analysis

2.1. Explanation of Translation Selection

The selection of Chinese translations of "Gitanjali" in this study is based on the initial publication date of these translations and whether they are representative. The initial publication date reflects the completion time of the translation, indicating the era in which the translator lived. Since the evolution of eras is the basic logic of this comparative analysis, it is taken as the first criterion for sample selection. On the other hand, due to limitations in time and resources, it is not possible for this paper to analyze all Chinese translations of "Gitanjali". To ensure that the results of quantitative text analysis reflect the general textual features of "Gitanjali" translations, the representativeness of the Chinese translations is another important reference criterion.

Based on the criterion of representativeness, this article comprehensively considers the influence, sales, reader reviews, and recognition of each translation. As mentioned earlier in the review of the history of Chinese translation of "Gitanjali", Bing Xin's translation is the first complete translation of "Gitanjali" in China, and it is also the most widely influential and highly recognized translation. Wu Yan's translation is the second complete translation of "Gitanjali" after Bing Xin's translation, and its recognition and influence is second only to Bing Xin's translation. For these reasons, Bing Xin's translation and Wu Yan's translation were the first translations selected for this study. Subsequently, Xiao Xingzheng's translation was chosen. This is the most recent Chinese translation of "Gitanjali" available on the market. Considering that some translations of "Gitanjali" published after 2010 are of mediocre quality, Xiao Xingzheng's translation stands out as a high-quality translation, receiving high praise from both readers and professionals. Therefore, this article considers Xiao Xingzheng's translation to be the most representative translation in recent years and should be included. The selection of Shen Huan and Wang Li's translation (hereinafter referred to as the Wang Li's translation) is controversial, as its influence and translation quality are not satisfactory. On the other hand, the translation by Bai Kaiyuan during the same period surpasses the Wang Li translation in all aspects. However, although the Bai Kaiyuan translation primarily references the English version during translation, it also adjusts the translation based on the Bengali version. This dual-language source in translation differs from other translations, considering the significant impact that changes in the source language of the same text can have on the translation (target language)[10]. To prevent errors in text analysis due to differences in source languages, the Wang Li translation was chosen as an alternative.

Furthermore, the initial publication dates of the four translations selected in this article are 1955, 1986, 2010, and 2019, respectively, with intervals ranging from 9 to 30 years, which basically conforms to the comparative analysis logic based on the evolution of eras in this paper. Although the publication date of the Wang Li translation in 2010 is relatively close to that of the Xiao Xingzheng translation in 2019, the Bai Kaiyuan translation from 2006 seems to better meet the selection criteria based on the publication date. However, for the reasons mentioned earlier, the Wang Li translation was chosen to prioritize ensuring the accuracy and reliability of the data obtained in this study.

2.2. Quantitative Text Analysis Tool

This study utilizes HanLP in the Python environment to perform statistical analysis on the textual features of the full text of the Chinese translations of "Gitanjali". Given the particularities of text statistics and the size of the full text, many text analysis objects often cannot directly yield the desired results using traditional statistical tools or corpora, or they require manual intervention and the assistance of other tools. This makes the experiment challenging and difficult to ensure the accuracy of statistical data. HanLP, leveraging the advantages of natural language processing, can quickly and directly generate data for more complex text analysis objects such as lexical density and high-frequency words. This not only ensures the accuracy of the data but also enhances the efficiency of the experiment. Using HanLP also allows for more complex and detailed linguistic analysis of the text, making it an extremely effective statistical tool for translation studies. However, the use of HanLP or other natural language processing models for translation and literary research is currently scarce in academia, which is also the innovative aspect of this study.

2.3. Experimental Design for Quantitative Text Analysis

This paper will conduct a quantitative text analysis of four Chinese translations of "Gitanjali" from four aspects: the number of characters and sentences, lexical property, lexical density, and high-frequency words. To present the results more clearly, this study designates the four Chinese translations of "Gitanjali" as T1-T4 in chronological order of their publication. The 1955 translation by Bing Xin is marked as T1, the 1986 translation by Wu Yan as T2, the 2010 translation by Wang Li as T3, and the 2019 translation by Xiao Xingzheng as T4. The study comprises the following steps:

Step 1: Sentence and Character Count. Calculate the number of sentences and characters in each of the four translations (T1-T4). This yields eight sets of data: Sentence Number T1 - Sentence Number T4 and Character Number T1 - Character Number T4.

Step 2: Lexical Property Calculation. Analyze the lexical properties of each translation, focusing on nouns, verbs, adjectives, adverbs, pronouns, and conjunctions. This produces 24 sets of data, for example, Noun T1 - Noun T4 and Verb T1 - Verb T4.

Step 3: Total Word Count. Determine the total number of words in each translation, resulting in four sets of data: Word Number T1 - Word Number T4.

Step 4: Content Words Count. Calculate the number of content words (including nouns, adjectives, verbs, numbers, pronouns, excluding adverbs) in each translation, resulting in four sets of data: Content Words T1 - Content Words T4.

Step 5: Lexical Density Calculation. Compute the lexical density of each translation using the formula:

$$\text{Lexical Density} = \frac{\text{Number of Content Words}}{\text{Total Number of Words}} \quad (1)$$

This yields four sets of data: Lexical Density T1 - Lexical Density T4.

Step 6: High-Frequency Words Analysis. Identify and count the top five high-frequency nouns in each translation.

3. Comparative Analysis of Textual Factors in Chinese Translations of "Gitanjali"

3.1. Sentences and Characters

Character count and sentence count are among the most fundamental textual factors when examining translated texts. The data characteristics implicit in these elements can significantly impact the overall comprehensive analysis of the translated texts. Therefore, this paper will first conduct a data analysis of the character count and sentence count. The statistics for the character count and sentence count of the four Chinese translations of "Gitanjali" (T1-T4) are shown in Table 1 below:

Table 1. Character Count and Sentence Count of the Four Chinese Translations of "Gitanjali"

	Number of sentences	Number of characters
T1	567	19073
T2	578	21529
T3	564	18962
T4	563	19011

Note. T1: Bing Xin's translation, T2: Wu Yan's translation, T3:Wang Li's translation, T4:Xiao Xingzheng's translation

From Table 1, it can be seen that both the number of characters and sentences in the translations by Bing Xin, Wang Li, and Xiao Xingzheng generally maintain the same standard level, with no significant differences in their data. The 1986 translation by Wu Yan, however, is an exception. It has a higher number of sentences and characters compared to the other three translations, with the difference in the number of characters being particularly notable. Wu Yan's translation has 2,456 more characters than Bing Xin's, which has the second-highest character count, while the differences in character counts among the other three translations are minor, ranging from 49 to 111 characters.

Due to the nature of translated texts, the amount of information contained in the translated text is fixed, meaning it is equivalent to the amount of information in the original text. Although this can vary to some extent depending on the translation strategies of different translators, on a macro level, the difference in information content between the translated text and the original text is minimal, as translators cannot deviate from the framework of the original text and create entirely new content. Based on the above, Wu Yan's translation has more characters and sentences than other translations while the amount of information in the text is fixed, it can be initially inferred that the information density of Wu Yan's translation text is relatively lower. This inference will be further confirmed in the subsequent analysis.

3.2. Lexical Properties

This chapter will discuss the preferences of the four translations in the use of different lexical properties. Lexical properties are a crucial factor in analyzing translated texts as they reflect the translators' choices and tendencies in different grammatical components. This study has counted the occurrences of all nouns, verbs, adjectives, adverbs, pronouns, and conjunctions in the four translations. The relevant statistics are shown in Table 2 below:

Table 2. Lexical Properties Statistics

	Noun	Verb	Adjective	Adverb	Pronoun	Conjunction
T1	1959	2768	593	1896	1660	260
T2	2194	2908	681	2036	1719	340
T3	2006	2703	523	1750	1637	308
T4	1930	2682	621	1624	1659	239

Note. T1: Bing Xin's translation, T2: Wu Yan's translation, T3:Wang Li's translation, T4:Xiao Xingzheng's translation

When analyzing lexical properties, this paper first observes the evolution trends of lexical properties selection in each translation over time. From a broad perspective, the choice of the five lexical properties across the four translations does not vary significantly, with the numbers fluctuating within a stable range. Wu Yan's translation has the highest usage of each lexical property, which is logical given that it contains the most characters among all translations. Focusing on eachlexical property, except for adjectives and pronouns, the other four lexical properties generally follow the same evolutionary pattern. There is a sharp increase from Bing Xin's 1955 translation to Wu Yan's 1986 translation, followed by a gradual decrease and stabilization in Wang Li's 2010 translation and Xiao Xingzheng's 2019 translation. Adjectives and pronouns, however, follow a different pattern: they increase from 1955 to 1986, decline sharply from the 1986 peak to 2010, and then rise again in 2019, returning to levels similar to 1955.

Regarding the preference for lexical property in each translation (based on comparisons between translations), Bing Xin's translation shows no distinct preference for any part of speech. Each part of speech in this translation has quantities similar to those in other translations. This could be because it is the earliest and generally recognized most authoritative Chinese translation of "Gitanjali", influencing subsequent translations to varying degrees. Xiao Xingzheng's translation is the closest to Bing Xin's in terms of lexical property selection, with differences in quantities within 86 for most lexical properties. This is likely because Xiao Xingzheng's first exposure to the Chinese translation of "Gitanjali" was Bing Xin's version, which he mainly referred to during his translation process, as mentioned in the preface of Xiao Xingzheng's translation[9]. However, Xiao Xingzheng's translation uses very few adverbs, 126 fewer than Wang Li's, which uses the second fewest, and nearly 400 fewer than Wu Yan's, which uses the most. This is the most significant difference in part-of-speech usage among all translations.

3.3. Lexical Density

Lexical density can be used to measure the amount of information contained in a text. Generally, the higher the lexical density, the more information the text contains, and the richer its vocabulary. Baker suggests that lexical density is the ratio of content words to the total number of words, expressed as a percentage. The higher the proportion of content words, the greater the information content of the translation[11]. In this paper, Baker's definition is used to calculate lexical density, as the formula described in Chapter 3.

There is ongoing debate in the academic community about which lexical properties should be classified as content words. Scholars like Biber argue that content words include nouns, verbs, adjectives, and adverbs[12]. On the other hand, Wang Li and Zhu Dexi suggest that content words include nouns,

verbs, and adjectives, but not adverbs[13]. Considering that "Gitanjali" is a poetic work in which adverbs are less likely to appear on their own as words with real meaning, this paper classifies adverbs as function words. Thus, only nouns, verbs, and adjectives are counted as content words, following the second definition mentioned above.

The number of content words and the lexical density for the four Chinese translations of "Gitanjali" are shown in Table 3 below.

Table 3. Number of Content Words and Lexical Density

	Number of content words	Lexical density
T1	7152	53.35%
T2	7712	50.13%
T3	7060	53.88%
T4	7097	52.88%

Note. T1: Bing Xin's translation, T2: Wu Yan's translation, T3:Wang Li's translation, T4:Xiao Xingzheng's translation

From Table 3, it can be seen that although Wu Yan's translation has the highest number of content words among the four translations, its lexical density is the lowest, averaging three percentage points lower than the other translations. This means that despite using more content words, the information content in Wu Yan's translation is relatively lower. This confirms the hypothesis made earlier in this paper that, due to the fixed information content of the original text, Wu Yan's translation, with more characters, contains less information. Based on this, it is further inferred that the more characters a translation has, the lower its lexical density and the less information it contains. The other three translations are very close in terms of number of content words and lexical density, with Wang Li's translation having the fewest content words and the highest lexical density.

3.4. High-Frequency Words

The analysis of high-frequency words helps to understand the translator's preferences for certain specific words during the translation process . By counting and analyzing the high-frequency words in the four Chinese translations of "Gitanjali", it can reveal the consistency and differences in the translators' word choices. This study will first count the top five nouns with the highest frequency in each translation and then compare and analyze the usage of these high-frequency words. The relevant data is shown in Table 4.

Table 4. High-Frequency Words Count

	HFW 1	HFW 2	HFW 3	HFW 4	HFW 4
T1	时候(52)	生命(42)	世界(33)	心 (32)	人 (32)
T2	生命(36)	时候(34)	心 (34)	人 (33)	世界(27)
T3	生命(49)	心 (25)	世界(17)	光明(17)	时间(16)
T4	生命(51)	心 (39)	人 (28)	世界(27)	光明(22)

Note. T1: Bing Xin's translation, T2: Wu Yan's translation, T3:Wang Li's translation, T4:Xiao Xingzheng's translation, HFW: high-frequency word.

The number in parentheses means the total number of occurrences of the word.

The top five high-frequency words that appear across all translations are as follows: 时候/时间 (day), 生命 (life), 心 (heart), 人 (human), 世界 (world), and 光明 (light). The terms 时候 and 时间 both correspond to the word "day" in the original text, and thus they are classified together. Among these, the words cn-life, cn-heart, and cn-world appear as high-frequency words in all translations. The word cn-life appears most frequently, being the highest frequency word in three of the translations except for Bing Xin's version. The word cn-day appears in all translations except Wang Li's translation, and the word cn-light only appears in Wang Li and Xiao Xingzheng's translations.

From a comparative perspective, the high-frequency words among all translations do not vary much in terms of types, but there are significant differences in their frequencies of use. Overall, the top five high-frequency words across the four translations are limited to six types, indicating a strong constraint on the variety of high-frequency words in translation texts. Bing Xin's and Wu Yan's translations share the same high-frequency words. In comparison, both Wang Li's and Xiao Xingzheng's translations include the word cn-light, which is absent in the other two, Wang Li's translation does not include the word cn-human, and Xiao Xingzheng's translation does not include the word cn-day. The number of occurrences of high-frequency words shows significant variation, such as the word cn-day appearing 52 times in Bing Xin's translation but only 16 times in Wang Li's translation, such differences are evident in Table 4.

4. Conclusion

Using HanLP to quantitatively analyze the textual features of four Chinese translations of "Gitanjali", this paper demonstrates the data distribution characteristics in terms of character and sentence counts, lexical properties, lexical density, and high-frequency words. It also explores the intrinsic logic and structural connections between these textual features. Overall, except for Wu Yan's translation, the data characteristics of the other translations are relatively similar, with no significant differences. Wu Yan's translation has higher counts in characters, sentences, and the use of lexical properties, making it the translation with the most content words, but it has the lowest lexical density. Based on the case of Wu Yan's translation, this paper makes an important inference: because the amount of information in the translated text is constrained by the original text, the more characters or words a translation contains, the lower its lexical density and the less information it conveys. Generally, the richness of vocabulary is also positively correlated with lexical density.

In terms of internal textual features, Bing Xin's translation is similar to Xiao Xingzheng's in terms of lexical property choices, but Xiao Xingzheng uses very few adverbs. Wang Li's translation uses fewer adjectives and pronouns, while Wu Yan prefers using verbs. To further investigate the translators' language preferences, this paper also examines the high-frequency words (nouns only) in each translation. The top five high-frequency words across the translations were limited to a small set, showing consistency in word choice due to the original text. This indicates that, although the overall distribution characteristics of textual features do not show significant differences, the internal specifics of each textual element are influenced by language preferences of the translators, giving each translation its unique characteristics.

Presenting textual features through data makes it easier to see their intrinsic logic and the structural connections between different textual features. This not only has significant value for the study of specific literary and translated texts but also has broad implications for more extensive translation and linguistic research. Additionally, the quantitative text analysis process in this paper proves the effectiveness of the HanLP

natural language processing model in literary and translation studies. Natural language processing models have the capability to handle large volumes of text, offering vast potential for broader and more extensive literary and translation research.

References

[1] Yang, J. (2017). The key promoter of Tagore's Nobel Prize in Literature. Literature and History World, (06), 74-79.

[2] Zeng, Q. (2012). Tagore in world literature: The translation and research of "Gitanjali". Foreign Language Teaching, (04), 82-85.

[3] Liang, F. (2016). Following the beautiful footsteps of translators: A comparative study of Chinese translations of "Gitanjali". Journal of Harbin University, (08), 81-86.

[4] Liu, J. C., & Ren, X. F. (2018). Corpus-based Comparative Study on Chinese Versions of Gitanjali-A Case Study on Bingxin's and Bai Kaiyuan's Versions. Journal of Literature and Art Studies, 8(1), 37-42.

[5] Hu, X. (2007). A corpus-based study on the lexical features of Chinese translated novels. Foreign Language Teaching and Research, (03), 214-220+241.

[6] Tagore, R. (2008). Gitanjali (Bing Xin, Trans.). Yilin Translation Publishing House. (Original work published 1912).

[7] Tagore, R. (1990). Gitanjali (Wu Yan, Trans.). Shanghai Translation Publishing House. (Original work published 1912).

[8] Tagore, R. (2010). Gitanjali (Shen Huan & Wang Li, Trans.). China Pictorial Publishing House. (Original work published 1912).

[9] Tagore, R. (2019). Gitanjali (Xiao Xingzheng, Trans.). Yunnan people's Publishing House. (Original work published 1912).

[10] Baruah, M. (2014). Translation Theories and Translating Assamese Texts. Translation Today, 5, 8(2), 5-30.

[11] Baker, M. (1995). Corpora in translation studies: An overview and some suggestions for future research. Target. International Journal of Translation Studies, 7(2), 223-243.

[12] Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (2000). Longman grammar of spoken and written English.

[13] Hu, X., & Zeng, J. (2009). A corpus-based study on the grammatical markers explicitation in translated novels. Foreign Language Research, (05), 72-79.