# A Corpus-based Comparative Study between Machine Translation and Manual Translation—Taking the translation of *Farewell to Cambridge* as an example

**Yueqin Wu[1] , Hongyu Hu[2, *]**

School of Foreign Languages, Wuhan University of Technology, Wuhan, China

*Corresponding author: 1013318864@qq.com

**Keywords:** Corpus; Comparative study; Translation of *Farewell to Cambridge*; ChatGPT

**Abstract:** The paper aims to compare and analyze ChatGPT translations and manual translations of Xu Zhimo's poem *Farewell to Cambridge* through the self-built corpus from the dimension of vocabulary, sentence and discourse respectively. As to the dimension of vocabulary, this paper will deepen the illustrations and corpus from lexical diversity, lexical density, lexical frequency and word length. As to the dimension of sentence, figures and illustrations will be explored from mean sentence length. As to the discourse, readability and understandability will be developed in details. According to the collected data, this paper attempts to put forward the limitations of machine translation in translating literary works and the irreplaceability of human translation so as to help Chinese literature "go out" better.

## 1. Introduction

With the development of science and technology, machine translation software like ChatGPT has occupied an important place in translating modern Chinese literary works. This study will build a self-built corpus based on three manual translations and three ChatGPT translations of Xu's *Farewell to Cambridge*. It will explore the differences between the manual translation and the ChatGPT translation from three major dimensions: vocabulary, sentence and discourse so as to study the linguistic features and overall style of manual translator and machine translator. Moreover, the shortcomings of the machine in translating literary works will be listed with some linguistic suggestions for the improvement of translation software in the future.

## 2. Corpus-based Study

This study will build its own monolingual translation corpus of the poem *Farewell to Cambridge* to avoid the subjectivity. The corpus collected includes three manual translations and three machine translations. The translators and the titles of the six translations are shown in Table 1. After collecting the corpus, lots of software are used for a series of operations. These six translations will be studied and analyzed from the three aspects of vocabulary, sentence and discourse.

Table 1. English and Chinese comparison of the text titles of the corpus used

| Translation | Translator | Title |
|---|---|---|
| ChatGPT1 | ChatGPT | *Leaving Cambridge Again* |
| ChatGPT2 | ChatGPT | *Leaving Again for Cambridge* |
| ChatGPT3 | ChatGPT | *Leaving Cambridge Again* |
| M1 | Xu Jingcheng | *Farewell to Cambridge Again* |
| M2 | Edward Connynham | *On Leaving Cambridge Again* |
| M3 | Xu Yuanchong | *Adieu, Cambridge!* |

### 2.1 The Dimension of Vocabulary

#### 2.1.1 Lexical diversity

Token refers to the total number of words in the text, and type refers to the number of tokens that are not counted twice. "Type/token ratio is a measure of the range and diversity of vocabulary used by a writer, or in a given corpus. A high type-token ratio means that the writer uses a wider range of vocabulary” [1]. According to the data in Table 2, the average number of tokens among the three machine translations is 173, while the average number of tokens in the three manual translations is 164. In terms of the number of types, the average number of machine translations is 106, and the average number of manual translations is 103. In terms of TTR, the average TTR of the three machine translations is 61.39%, and the average TTR of the three manual translations is 62.43%. In comparison with the lexical diversity of the machine translation, the lexical diversity of the manual translation is higher, indicating that the manual translation is richer in vocabulary.

Table 2. Statistics of tokens and types of the selected corpus

| Translation | Token | Type | TTR(%) |
|---|---|---|---|
| ChatGPT1 | 171 | 106 | 61.99% |
| ChatGPT2 | 171 | 109 | 63.74% |
| ChatGPT3 | 178 | 104 | 48.43% |
| M1 | 148 | 100 | 63.29% |
| M2 | 162 | 106 | 64.43% |
| M3 | 174 | 103 | 48.86% |

### 2.1.2 Lexical density

“Lexical density is the percentage of lexical as opposed to grammatical items in a given text or corpus of texts” [2]. According to Table 3, the lexical density of the manual translation is greater, indicating that the machine translation reduces the difficulty of reading to some extent. According to Liu, the lexical density of original English poems is 48.3% [3]. The average lexical density of the manual translation is 55.81% which is higher than that of the reference corpus which means that the manual translation achieves the level of authenticity.

Table 3. Lexical density statistics

| Translation | ChatGPT1(%) | ChatGPT2(%) | ChatGPT3(%) | M1(%) | M2(%) | M3(%) |
|---|---|---|---|---|---|---|
| Noun | 28.64% | 31.21% | 29.21% | 32.48% | 32.43% | 24.99% |
| Adj. | 9.94% | 7.41% | 8.43% | 9.44% | 10.84% | 6.21% |
| Adv. | 7.02% | 8.09% | 7.87% | 10.19% | 7.23% | 9.60% |
| Full verb | 9.36% | 8.67% | 6.74% | 8.92% | 11.44% | 12.43% |
| Lexical words | 44.97% | 44.49% | 42.24% | 61.14% | 62.04% | 44.24% |

### 2.1.3 Lexical frequency

This study will analyze from the part of speech and high-frequency words. According to Fig. 1, the words with the highest frequency in the six translations except M1 are DT, NN, and IN. Specifically, as shown in Fig. 2, the four words with the highest frequency in the three machine translations are the, I, a, and in, while the high-frequency words are different in three manual translations. According to Olohan and Zhu & Li, the five most frequently used words in the TEC and BNC were “the, of, and, to, and a” [4, 5]. As shown in Table 4 below, compared with the high-frequency words in the three machine translation corpora, the high-frequency words in the three manual translation corpora are closer to these two authoritative corpora.

| | ChatGPT1 | | | ChatGPT2 | | | ChatGPT3 | | | M1 | | | M2 | | | M3 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | POS | FREQ | Prop | POS | FREQ | Prop | POS | FREQ | Prop | POS | FREQ | Prop | POS | FREQ | Prop | POS | FREQ | Prop |
| 1 | DT | 28 | 16.47% | NN | 29 | 16.96% | NN | 33 | 18.86% | NN | 33 | 21.02% | NN | 31 | 19.14% | NN | 34 | 19.43% |
| 2 | NN | 28 | 16.47% | DT | 28 | 16.37% | DT | 29 | 16.57% | PP | 16 | 10.19% | DT | 19 | 11.73% | DT | 22 | 12.57% |
| 3 | IN | 23 | 13.53% | IN | 25 | 14.62% | IN | 25 | 14.29% | RB | 16 | 10.19% | IN | 18 | 11.11% | IN | 20 | 11.43% |
| 4 | PP | 17 | 10.00% | PP | 16 | 9.36% | JJ | 14 | 8.00% | DT | 15 | 9.55% | JJ | 17 | 10.49% | RB | 16 | 9.14% |
| 5 | JJ | 16 | 9.41% | NNS | 14 | 8.19% | PP | 14 | 8.00% | IN | 14 | 8.92% | NNS | 13 | 8.02% | PP | 15 | 8.57% |
| 6 | NNS | 12 | 7.06% | RB | 13 | 7.60% | RB | 13 | 7.43% | JJ | 14 | 8.92% | RB | 12 | 7.41% | VV | 12 | 6.86% |
| 7 | RB | 11 | 6.47% | JJ | 12 | 7.02% | NNS | 11 | 6.29% | NNS | 11 | 7.01% | PP | 11 | 6.79% | JJ | 11 | 6.29% |
| 8 | NP | 9 | 5.29% | NP | 11 | 6.43% | NP | 8 | 4.57% | NP | 7 | 4.46% | NP | 10 | 6.17% | MD | 8 | 4.57% |
| 9 | VVP | 7 | 4.12% | VVP | 7 | 4.09% | VBP | 5 | 2.86% | TO | 6 | 3.82% | VVG | 9 | 5.56% | NNS | 6 | 3.43% |
| 10 | VVG | 4 | 2.35% | VVZ | 3 | 1.75% | VVG | 5 | 2.86% | VV | 5 | 3.18% | CC | 3 | 1.85% | NP | 6 | 3.43% |
| 11 | VBZ | 3 | 1.76% | VBZ | 2 | 1.17% | TO | 3 | 1.71% | VVG | 4 | 2.55% | VV | 3 | 1.85% | TO | 5 | 2.86% |
| 12 | VVD | 2 | 1.18% | VVD | 2 | 1.17% | VBZ | 3 | 1.71% | CC | 3 | 1.91% | VVN | 3 | 1.85% | CC | 3 | 1.71% |
| 13 | TO | 2 | 1.18% | VVG | 2 | 1.17% | CC | 2 | 1.14% | VVP | 3 | 1.91% | VBP | 2 | 1.23% | VBZ | 3 | 1.71% |
| 14 | VVZ | 2 | 1.18% | CC | 1 | 0.58% | VV | 2 | 1.14% | MD | 2 | 1.27% | VBZ | 2 | 1.23% | VVP | 3 | 1.71% |
| 15 | CC | 1 | 0.59% | JJR | 1 | 0.58% | VVD | 2 | 1.14% | VBZ | 2 | 1.27% | VVD | 2 | 1.23% | VVD | 2 | 1.14% |
| 16 | JJR | 1 | 0.59% | MD | 1 | 0.58% | VVP | 2 | 1.14% | VVD | 2 | 1.27% | VVP | 2 | 1.23% | VVG | 2 | 1.14% |
| 17 | MD | 1 | 0.59% | TO | 1 | 0.58% | JJR | 1 | 0.57% | JJR | 1 | 0.64% | JJR | 1 | 0.62% | VVZ | 2 | 1.14% |
| 18 | VBP | 1 | 0.59% | VBP | 1 | 0.58% | MD | 1 | 0.57% | VB | 1 | 0.64% | MD | 1 | 0.62% | WDT | 2 | 1.14% |
| 19 | VV | 1 | 0.59% | VV | 1 | 0.58% | VVN | 1 | 0.57% | VBP | 1 | 0.64% | RP | 1 | 0.62% | VB | 1 | 0.57% |
| 20 | WRB | 1 | 0.59% | WRB | 1 | 0.58% | WRB | 1 | 0.57% | VVZ | 1 | 0.64% | TO | 1 | 0.62% | VVN | 1 | 0.57% |
| 21 | | | | | | | | | | | | | VBD | 1 | 0.62% | WRB | 1 | 0.57% |

Fig. 1 Lexical frequency ranking of the POS used in corpus text

| | ChatGPT1 | | | ChatGPT2 | | | ChatGPT3 | | | M1 | | | M2 | | | M3 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | HF | FREQ | Prop | HF | FREQ | Prop | HF | FREQ | Prop | HF | FREQ | Prop | HF | FREQ | Prop | HF | FREQ | Prop |
| 1 | the | 18 | 10.59% | the | 18 | 10.53% | the | 22 | 12.36% | the | 10 | 6.33% | a | 9 | 5.56% | the | 15 | 8.57% |
| 2 | i | 10 | 5.88% | i | 10 | 5.58% | i | 8 | 4.49% | i | 8 | 5.06% | the | 8 | 4.94% | i | 8 | 4.57% |
| 3 | a | 10 | 5.88% | a | 8 | 4.68% | a | 7 | 3.93% | my | 6 | 3.80% | i | 7 | 4.32% | a | 6 | 3.43% |
| 4 | in | 6 | 3.53% | in | 6 | 3.51% | in | 6 | 3.37% | to | 6 | 3.80% | in | 5 | 3.09% | as | 6 | 3.43% |
| 5 | of | 3 | 1.76% | my | 3 | 1.75% | of | 5 | 2.81% | in | 5 | 3.16% | queitly | 5 | 3.09% | my | 5 | 2.86% |
| 6 | with | 3 | 1.76% | with | 3 | 1.75% | am | 3 | 1.69% | quietly | 4 | 2.53% | bright | 3 | 1.85% | to | 5 | 2.86% |
| 7 | as | 2 | 1.18% | arrived | 2 | 1.17% | gently | 3 | 1.69% | a | 3 | 1.90% | green | 3 | 1.85% | of | 4 | 2.29% |
| 8 | cambridge | 2 | 1.18% | as | 2 | 1.17% | is | 3 | 1.69% | as | 3 | 1.90% | river | 3 | 1.85% | will | 4 | 2.29% |
| 9 | came | 2 | 1.18% | cambridge | 2 | 1.17% | leaving | 3 | 1.69% | farewell | 3 | 1.90% | waving | 3 | 1.85% | green | 3 | 1.71% |
| 10 | dreams | 2 | 1.18% | depart | 2 | 1.17% | my | 3 | 1.69% | leave | 3 | 1.90% | are | 2 | 1.85% | in | 3 | 1.71% |

Fig. 2 Ranking of high frequency words in the used corpus text

Table 4. The frequency of high-frequency words compared with TEC corpus

| HF | ChatGPT1 | ChatGPT2 | ChatGPT3 | M1 | M2 | M3 |
|---|---|---|---|---|---|---|
| the | 10.43 | 10.43% | 12.36% | 6.33% | 4.94% | 8.47% |
| and | 0 | 0 | 0 | 0.63% | 0.62% | 0 |
| to | 1.17% | 0.48% | 1.69% | 3.80% | 0.62% | 2.86% |
| of | 1.74% | 1.17% | 2.81% | 0.63% | 1.23% | 2.29% |
| a | 4.84% | 4.68% | 3.93% | 1.90% | 4.46% | 3.43% |

### 2.1.4 Word length

"Longer words encode more linguistic units and thus take more effort to process" [6]. Mean word length reflects the complexity of the words in the text. Word length standard deviation reflects the difference between the length of each word in the text and the average word length of the text. In addition, the larger the standard deviation, the greater the difference between the length of each word in the text. From the data in Table 5, it can be seen that the average word length of the three ChatGPT translations is 4.37, while the average word length of the three manual translations is 4.28, indicating that the complexity of the words used in the manual translations is lower. The standard deviation of word length in the machine translations is higher than that in the manual translations, indicating a relatively large difference between the word lengths of each word in the texts which may make the translation lack the strong rhythm of the original.

Table 5. The mean word length and its standard deviation of the selected corpus

| Translation | Mean word length | Word length std.dev |
|---|---|---|
| ChatGPT1 | 4.36 | 2.38 |
| ChatGPT2 | 4.40 | 2.36 |
| ChatGPT3 | 4.24 | 2.21 |
| M1 | 4.30 | 2.19 |
| M2 | 4.42 | 2.21 |
| M3 | 4.04 | 2.08 |

## 2.2 The Dimension of Sentence

Olohan pointed out that "mean sentence length could be also a measure for comparison" [4]. The longer the sentence, the more difficult the text is and the lower the readability is. The total number of sentences in the original text is 8, and the machine translation is closer to it in terms of the total number of sentences. According to the data in Table 6, the mean sentence length of the machine translation is higher than that of the manual translation, indicating that the readability of the machine translation is lower than that of the manual translation.

Table 6. The mean sentence length and the standard deviation of sentence length

| Translation | Number | MLS | Sentence length std.dev |
|---|---|---|---|
| ChatGPT1 | 7 | 6.86 | 2.91 |
| ChatGPT2 | 8 | 6.88 | 2.17 |
| ChatGPT3 | 8 | 7.74 | 2.38 |
| M1 | 10 | 6.80 | 2.30 |
| M2 | 13 | 6.23 | 2.49 |
| M3 | 14 | 6.71 | 2.23 |

### 2.3 The Dimension of Discourse

#### 2.3.1 Understandability

The understandability of discourse is analyzed based on three lists. Baseword 1 and 2 are derived from GSL, each containing 1000-word families, covering 87% of the words in the English text. Baseword 3 is derived from the AWL, containing 470-word families. According to Fig. 3, the average proportion of the machine translation in Baseword 1 is 64.04% and that of the manual translation is 64.02%. In Baseword 2, the average proportion of the machine translation is 13.27%, and that of the manual translation is 13.69%. In Baseword 3, the average proportion is 9.36% for the machine translation and 6.32% for the manual translation. It indicates that the machine translation is not as understandable as the manual translation in the discourse.

| Word list | ChatGPT1 | | | ChatGPT2 | | | ChatGPT3 | | | M1 | | | M2 | | | M3 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Token/% | Type/% | Family | Token/% | Type/% | Family | Token/% | Type/% | Family | Token/% | Type/% | Family | Token/% | Type/% | Family | Token/% | Type/% | Family |
| 1 | 113/66.08 | 54/50.94 | 49 | 106/61.99 | 52/47.71 | 47 | 114/64.04 | 50/48.08 | 44 | 94/59.87 | 47/47.00 | 42 | 107/66.05 | 62/58.49 | 54 | 121/69.14 | 55/53.40 | 53 |
| 2 | 21/12.28 | 20/18.87 | 17 | 24/14.04 | 21/19.27 | 18 | 24/13.48 | 22/21.15 | 19 | 23/14.65 | 19/19.00 | 17 | 28/17.28 | 20/18.87 | 15 | 16/9.14 | 14/13.59 | 11 |
| 3 | 15/8.77 | 13/12.26 | 10 | 14/8.19 | 13/11.93 | 10 | 18/10.11 | 14/13.46 | 11 | 12/7.64 | 9/9.00 | 7 | 10/6.17 | 10/9.43 | 9 | 9/5.14 | 9/8.74 | 9 |
| Not listed | 22/12.87 | 19/17.92 | | 27/15.79 | 23/21.10 | | 22/12.36 | 18/17.31 | | 28/17.83 | 25/25.00 | | 17/10.49 | 14/13.21 | | 29/16.57 | 25/24.27 | |
| Total | 171 | 106 | 76 | 171 | 109 | 75 | 178 | 104 | 74 | 157 | 100 | 66 | 162 | 106 | 78 | 175 | 103 | 73 |

Fig. 3 Range of statistical results of the selected corpus

#### 2.3.2 Readability

GFS can be used to analyze the reading difficulty of a discourse and its formula is as follows: 0.4[(words/sentences+100(complex words/words)]. "The higher the index score, the more difficult it is to represent the discourse" [7]. According to Table 7, the GFS of the machine translation is much higher, which shows that the machine translation is more difficult to read than the manual translation. FKRE index is also used to calculate the readability, and its formula is as follows: 206.835-1.015(total words/total sentences)-84.6(total syllables/total words). The difficulty of the discourse is inversely proportional to the FKRE score. According to Table 8, the average FKRE score of machine translation is 62.24, and the average FKRE score of manual translation is 67.18, which also provides evidence that manual translation is more readable.

Table 7. GFS of the selected corpus

| Translation | ChatGPT1 | ChatGPT2 | ChatGPT3 | M1 | M2 | M3 |
|---|---|---|---|---|---|---|
| GFS | 11.41 | 9.95 | 9.80 | 7.00 | 5.48 | 5.89 |
| Average GFS | 10.39 | | | 6.12 | | |

Table 8. FKRE of the selected corpus

| Translation | ChatGPT1 | ChatGPT2 | ChatGPT3 | M1 | M2 | M3 |
|---|---|---|---|---|---|---|
| FKRE | 53.87 | 55.02 | 77.82 | 75.23 | 73.55 | 52.77 |
| Average FKRE | 62.24 | | | 67.18 | | |

### 2.4 Results

Machine translator and human translator have their own advantages and disadvantages from the three dimensions of vocabulary, sentence and discourse. The machine translation tries to make the translation one-to-one correspondent which makes the readability and understandability low. Besides, its translation always maintains neutrality and realism which reflects the translator's style of "taking the original as the center". While human translators integrate their consciousness into the translation, implement their translation views in the translation process, and embrace a lyrical charming style of the original. Therefore, the translation of human reflects the "human-centered" translator's style, considering both the original writer and the target audience.

## 3. Conclusion

By means of the comparative translation study of this paper, it is undeniable that even if machine translation is on the rise, it still cannot completely replace manual translation in the translation of literary works. More technical suggestions for machine translation software are needed to be put forward to improve translation software. Since this study only focuses on the translation of literary works, scientific works of machine translation and manual translation need to be deepened to make the research holistic.

## References


[1] Baker, Mona. Towards a methodology for investigating the style of a literary translator. Target, 2000, (2): 241-266.

[2] Baker, Mona. Corpora in translation studies: an overview and some suggestions for future research. Target, 1995, (2): 223-243.

[3] Liu Xiaogang. Corpus-based research on translator's style of *Pi Pa Xing*. Hebei University, 2017.

[4] Olohan, Maeve. Introducing corpora in translation studies. New York: Routledge, 2004.

[5] Zhu Zhaowei, Li Runfeng. Corpus-based research on Pound's translator's style in translating Chinese classics into English. Foreign Language Education, 2023, 44(4): 75-82.

[6] Victor Kuperman, Sascha Schroeder, Daniil Gnetov. Word length and frequency effects on text reading are highly similar in 12 alphabetic languages. Journal of Memory and Language, 2024, (135):1-15.

[7] Li Minjie, Zhu Wei. A study on the style of English translation of *A Dream of Red Mansions* based on parallel corpora. Journal of SCUN (Humanities and Social Sciences Edition), 2012, 32(02): 177-180.