ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202503_8(3).0008

A Corpus-Based Study on the Characteristics of Machine Translation in Historical Texts: Taking An Excerpt from the First Biography in Volume 101 of *the Book of Jin*

Zihan Cui[1], Jie Xu[2,*]

[1]Hong Kong Baptist University, Hong Kong, HKG, China

[2]Xuzhou Vocational College of Industrial Technology, Xuzhou Jiangsu 221000, China

[*]Corresponding Author: 715274556@qq.com

Abstract

With the widespread application of corpus technology, machine translation quality has significantly improved. While machine translation offers advantages such as speed and convenience, it also has limitations. This paper focuses on selected excerpts from the "Book of Jin," comparing human and machine translations. By establishing a corpus, this study examines the differences between human and machine translations in literary texts. The findings indicate that although the quality of machine translation has improved, it still struggles to fully grasp many cultural nuances, often providing insufficient contextual information and relying too heavily on literal translations. To enhance translation efficiency, the paper recommends integrating machine translation with human translation while continuously improving machine translation quality.

Keywords

Corpus, Machine Translation, Book of Jin.

1. Introduction

The English translation of ancient Chinese texts is crucial for the preservation of traditional Chinese culture and for enhancing the global understanding of Chinese history. The "Book of Jin," one of the classic works in the "Twenty-Four Histories," is known for its detailed records and rich documentary resources, serving as an essential reference for scholars engaged in historical and literary studies, both domestically and internationally. With the advancement of corpus technology, the quality of machine translation has been continuously improving, gradually demonstrating its significant role in translation practice. However, despite the speed and relatively low difficulty of machine translation, it does have its limitations.

This study is based on the first section of the "Book of Jin," combining both human and machine translations to establish a corpus for comparative analysis. It explores the evolution of machine translation within Chinese literature. The results indicate that although the quality of machine translation has improved, many cultural nuances remain difficult to accurately comprehend. Furthermore, there are often deficiencies in interpreting cultural backgrounds, and there is an excessive reliance on literal translations. To address these issues, this paper proposes a new approach: the organic integration of machine translation and human translation to enhance translation efficiency.

2. The Current State of Machine Translation Research

Machine translation is a key technology for modern cross-linguistic communication. Hutchins (1986) defined it as "the technology that uses computers to translate text from one natural language to another." [5] Since the 20th century, machine translation technology has gradually

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202503_8(3).0008

evolved from an initial phase of slow exploration and development into a stage of mature application. In recent years, the rise of neural machine translation technologies, represented by the Transformer model, has significantly improved translation quality, enabling translations that exhibit capabilities comparable to those of human translators. Additionally, Hassan et al. (2018) noted that this method integrates various algorithms and incorporates human evaluation, resulting in translation quality in the news domain that approaches human levels.

Over the past 11 years, the number of publications in the field of machine translation in China has generally shown a year-on-year increase, reflecting the growing importance and maturity of academic research in this area.

When translating classical historical works, machine translation faces unique challenges for two main reasons: First, due to the scattered nature of translations of Chinese historical texts, scholars find it difficult to construct a comprehensive corpus of Classical Chinese. Second, works like the "Book of Jin," which possess rich cultural characteristics, require a deep understanding of their cultural connotations and contextual nuances for accurate English translation—an area where current machine translation technologies still fall short.

To address these challenges, this project will focus on enhancing the quality and diversity of text corpora, improving the understanding of cultural contexts, and developing personalized translation strategies for classical texts. This exploration aims to establish a new human-machine integration model, providing fresh insights to tackle the aforementioned issues.

3. Corpora and Research Methods

The "Twenty-Four Histories" has been translated into multiple languages and has had a significant impact on the world. However, there has been relatively little translation of the "Book of Jin," which makes it particularly challenging to construct a relevant corpus. This study employs corpus analysis tools such as AntConc and Wordsmith to conduct a quantitative analysis from the perspectives of vocabulary, syntax, and text. Basic information is presented in Table 1:

Table 1. Basic information of the corpus

	Manual translation	Baidu translation	Deepl translation	Google translation
type	279	265	254	250
token	868	805	839	801
TTR	32.14%	32.92%	30.27%	31.21%
STTR	0.820	0.746	0.776	0.768

4. Research Process and Results

4.1. Comparison of Lexical Density

The Type/Token Ratio (TTR) is commonly used to measure the lexical density of a text. When dealing with large volumes of text, the Standardized Type/Token Ratio (STTR) is considered a more accurate measure of lexical density[8]. As shown in Table 1, the TTR and STTR values of the human-translated text are higher than those of other machine-translation versions. This indicates that human translation employs a more diverse vocabulary. In stark contrast, the lexical density and richness of vocabulary in machine translations are noticeably lacking. This may be attributed to the fact that machines still struggle to process literary materials in a perfect and diverse manner, resulting in translations that exhibit relatively mechanical characteristics. For example, in the translation of "炎光再阐," the differences between the human translation and Baidu's translation are illustrated in Table 2:

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202503_8(3).0008

Table 2. Mechanical characteristics of machine translation

Original text	Manual translation	Baidu translation
炎光再阐	successively rekindled the glory of our flame.	Yan Guang elaborated again.

Through comparison, it can be observed that in the translation of "炎光," the human translator rendered it as "flame." Although this does not explicitly indicate its implication for "汉德"(the merits of the Han dynasty), it is sufficient to convey a sense of brilliance and radiance to English-speaking readers. On the other hand, the machine translation only identifies "炎光" as the subject of "再阐" (Zai Chan, "rekindled") without explaining its deeper meaning. Additionally, the machine translation awkwardly translates "阐" as "elaborated," failing to accurately convey the original intent.

4.2. Comparison of Average Word Length

Average word length refers to the average number of letters contained in the words of a corpus. This is an important metric for assessing the complexity and readability of text. Using WordSmith 9.0, we conducted a statistical analysis of words containing 1 to 15 letters in human translations and various machine translation texts. The results are shown in Figure 1.

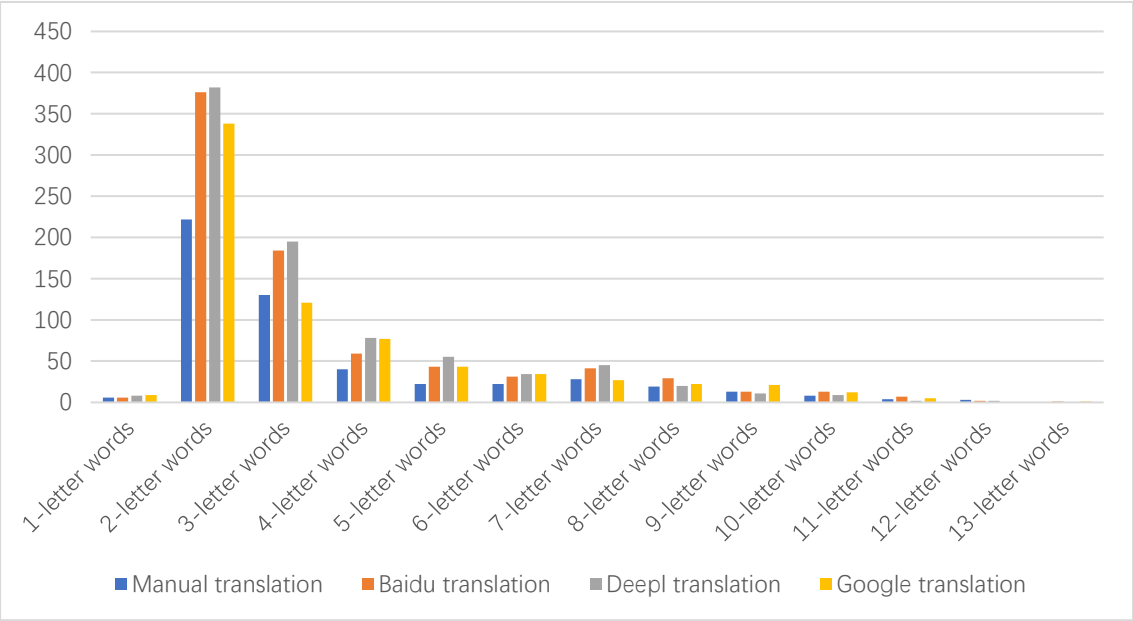

Figure 1. Comparison of Average Word Length

From the chart, it can be observed that for high-frequency word lengths, 2-letter and 3-letter words are the most common across all translation methods. In contrast, the frequency of low-frequency word lengths, such as 1-letter words and very long words (e.g., words with more than 10 letters), is quite low, with almost all translation methods following this trend. When comparing different translation methods, the advantages of human translation become more apparent as the number of letters in the words increases. This indicates that human translators tend to use more complex vocabulary when handling classical Chinese texts. Although machine translation is relatively close to human translation in the frequency of short words, it still exhibits significant shortcomings when considering the cultural connotations of the translated text.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202503_8(3).0008

4.3. Keyword comparison

By analyzing keywords, the lexical differences in style and content between the two corpora can be revealed.[1] Using AntConc software, a comparative analysis of the keywords from human translations, DeepL translations, and Google translations was conducted. The results are presented in Table 3.

Table 3. Comparison of keywords in different translations

	Manual translation	Baidu translation	Deepl translation	Google translation
1	emperor	emperor	emperor	emperor
2	great	great	world	great
3	three	dynasty	dynasty	three
4	Han	contry	gods	Han
5	dynasty	three	Han	dynasty
6	cao	divination	three	Cao
7	emperors	CaoCao	Tsung	old
8	old	Shu	repentance	divination
9	world	state	yuan	divine
10	divination	spread	cheng	people

The keywords across the four translations exhibit similar results. For instance, keywords such as "emperor," "Cao," "great," "three," and "Han" are all names and titles found in the original text. The frequent occurrence of proper nouns in the keywords highlights the rich cultural background and complex references within the article, a common feature in Chinese narratives [9]. In both human and machine translations, the most frequently occurring keyword is "emperor." In human translation, the seventh keyword is "emperors," reflecting the original text's tendency to omit the posthumous titles of two emperors and combine them for the sake of rhyme. However, the machine translations fail to capture this subtlety, sometimes resulting in erroneous translations, as detailed in Table 4:

Table 4. Comparison of machine translation error characteristics

	Manual translation	Baidu translation	Deepl translation	Google translation
元成多僻, 哀平短祚	under [Emperors] Yuan and Cheng, deviations were many; [Emperors] Ai and Ping saw brief reigns,	Yuan Chengduo was reclusive, Ai Ping was short-lived,	Yuan Cheng was more remote, mourning the short reign of Ping	Yuan Cheng was eccentric, Ai Ping was short-lived,

This situation partly arises from the lack of punctuation in ancient Chinese, such as commas, periods, and semicolons. Modern edited versions, while adhering to the original text's rhyme scheme, place greater emphasis on preserving cultural connotations rather than simplifying understanding. Similarly, this paper adopts a similar approach when discussing machine translation. However, in the actual translation process, translators should strive to avoid such issues.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202503_8(3).0008

4.4. Comparison of average sentence length

Average sentence length refers to the mean length of sentences in a text. Along with the standard deviation of sentence length, it serves as an important metric for evaluating a translator's style. [1] The standard deviation of sentence length reflects the variability in sentence lengths: a high standard deviation indicates significant variation in sentence lengths, while a low standard deviation suggests that the sentence lengths are relatively consistent[6]. Translators primarily employ two translation methods: one is literal translation, which maintains the original form and often results in shorter sentences; the other is free translation, which focuses on clearly conveying the text's meaning, leading to longer sentences [4].

By comparing the data in Table 5, it can be observed that human translations tend to have longer sentences with more variability. However, the average sentence length is lower than that of the English corpus (20.09). This indicates that the translator, while accurately conveying the original information, provides more explanations to better assist readers in understanding the source text. Both Baidu and Google translations tend to have shorter sentence lengths and adopt a literal translation approach, which can easily lead to misunderstandings among readers, such as excessive literalism in the translations. Deepl's translations show greater variability in sentence length. Since Chinese is a topic-prominent language and English is a subject-prominent language, the implicit semantic relationships in Chinese sentences need to be explicitly expressed when translated into English. However, the chosen machine translation methods have all proven inadequate, highlighting the grammatical and semantic shortcomings of machine translation. Notably, the average sentence length of the Deepl translations is closer to that of the human translations.

Table 5. Comparison of average sentence length and sentence standard deviation

	Manual translation	Baidu translation	Deepl translation	Google translation
average sentence length	54	41.88	47.44	42.29
sentence standard deviation	18.79	13.46	20.43	11.71

Sample1:

Original text:

太宗孝文皇帝重以明德，升平汉道。世宗孝武皇帝拓土攘夷，地过唐日。中宗孝宣皇帝搜扬俊乂，多士盈朝。

Manual translation:

Emperor Xiaowen, our illustrious forebear, furthered this legacy with his luminous virtue, elevating and pacifying the Han path.

Emperor Xiaowu, our sovereign ancestor, expanded our territories and repelled the barbarians, surpassing the realms in Tang’s days.

Emperor Xiaoxuan, our central progenitor, sought out and elevated the talented, filling the court with virtuous scholars.

Baidu translation:

Emperor Shizong Xiaowu expanded the land and spread the barbarians, and the land exceeded the Tang Dynasty.

Emperor Xiaoxuan, the middle emperor, searched for and raised handsome mowers, and many scholars filled the court.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202503_8(3).0008

In manual translation, titles such as "Taizong" and "Shizong" were appropriately omitted to maintain consistency with the original text, with the translator opting to directly translate the posthumous titles. However, Baidu Translate failed to distinguish between temple names and posthumous titles, translating them directly. Nevertheless, Baidu Translate did recognize the rhyming structure of the original text, which undoubtedly marks a significant advancement in the development of machine translation.

Sample2:

Original text:

是我祖宗道迈三王，功高五帝，故卜年倍于夏、商，卜世过于姬氏。

Manual translation:

Thus, our ancestral path has outstripped that of the Three Sovereigns, and our achievements have towered above the Five Emperors, hence our reigns have outlasted those of Xia and Shang, and our lineage has surpassed that of the Ji family.

Baidu translation:

It is my clan Dao Mai three kings, and high merit five emperors, so the divination years than Xia, Shang, divination too Ji's.

This translation is not only clear in meaning but also structurally sound. The human translation effectively conveyed the semantic relationship between the phrases "Three Sovereigns" and "Five Emperors" in the original text, indicating that the term "Three Sovereigns" broadly encompasses the "Five Emperors." This approach is clear and logically structured. In contrast, Baidu Translate rendered these terms as "kings" and "emperors," which neglects the cultural context and introduces common historical inaccuracies. This method of translation reduces readability and does not align with the reading and writing conventions of the target language audience.

5. Research Results and Analysis

With the advancement of technology, the quality of machine translation has been steadily improving. However, its performance in translating literary works remains unsatisfactory. Research indicates that machine translations tend to have lower lexical density and richness, and a smaller vocabulary, reflecting a noticeable mechanical nature. In analyzing literary works, machine translations often exhibit overly simplified language structures and excessive use of literal translation, which do not align with the reading habits of the target audience.

Regarding key terms, both machine and human translations include similar key terms, many of which are proper nouns. These terms typically carry diverse cultural connotations and complex references, fitting the typical narrative style of Chinese stories. However, due to the high reliance on literal translation in machine translation, there is an overuse of verbs such as "emperor," resulting in translations that appear overly verbose.

In terms of sentence length, human translations generally feature longer average sentence lengths, which accurately convey the original information while providing explanations and clarifications, making it easier for the target language readers to understand. On the other hand, machine translations lack the explanation of cultural and contextual information, although they have made progress in understanding rhetorical devices such as rhyme.

6. Conclusion

From the above analysis, it is evident that although the current level of machine translation has significantly improved, it has not been well applied in the field of literature. Therefore, in future work, it is necessary to continuously refine the relevant literary corpora to improve translation

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202503_8(3).0008

accuracy. Additionally, for handling large volumes of literature, it is recommended to adopt a combination of human and machine translation to enhance translation efficiency.

Acknowledgments

[Funding Project] Horizontal Project "Design and Service of ERP Systems" (Project Number: XGY2025B003)

References

[1] Baker M. Towards a methodology for investigating the style of a literary translator[J]. Target. International Journal of Translation Studies, 2000, 12(2): 241-266.

[2] Chen Zhibo. Research on Machine Translation of Traditional Chinese Medicine Classics Based on Instance Corpus: A Case Study of Four-Character Terms in "Huangdi Neijing" [J]. English Square, 2024(18): 15-18.DOI:10.16723/j.cnki.yygc.2024.18.001.

[3] Hassan H, Aue A, Chen C, et al. Achieving human parity on automatic chinese to English news translation [J]. arXiv preprint arXiv:1803.05567, 2018.

[4] Huo Yuehong. A Comparative Study of Translator's Style Based on Corpus [J]. Journal of Dalian University of Technology (Social Sciences), 2010, 31(02): 111-115.DOI:10.19525/j.issn1008-407x.2010.02.024.

[5] Hutchins W J. Machine translation: past, present, future[M]. Chichester: Ellis Horwood, 1986.

[6] Jiang Qiuyong. A Study on the Translator's Style in the English Version of "Decoded" Based on Corpus [J]. Overseas English,2019(04):27-28.

[7] Li Yao. A Study on the Quality of Machine Translation of Literary Works Based on Corpus: A Case Study of "Chronicle of a Blood Merchant" [J]. Overseas English,2021(18):39-40+42.

[8] Liang Maosheng, Li Wenzhong, Xu Jiajin. Corpus Application Tutorial [M]. Beijing: Foreign Language Teaching and Research Press.,2010:9-10

[9] Wang Yuan. A Study on the Translator's Style of "Fortress Besieged" Based on Corpus [J]. Journal of Jiamusi Vocational Institute.,2019(03):190-191.