International Journal of Education and Humanities
ISSN: 2770-6702 | Vol. 25, No. 1, 2026

A Corpus based Comparison of Human and Machine News Translation Styles

-- A case study of translations of Ministry of Foreign Affairs news

Yidong Ju

Shanghai International Studies University, Shanghai 201620, China

Abstract: This study compares the styles of news translations produced by human translators, neural machine translation system, and large language model. Using a corpus-based approach, it examines lexical, syntactic, and discourse features across the three sets of translations. Type–token ratios, mean word length, mean sentence length, and readability scores reveal no clear stylistic differences. The findings suggest that advances in machine translation have made human and machine translations difficult to distinguish on stylistic grounds in this news corpus.

Keywords: Large Language Models, Corpus, Translator Style.

1. Introduction

Large language models (LLMs) can understand and generate natural language and have attracted considerable interest for their performance in machine translation. Their implications for translation practice and research are widely recognized. Alongside their impact on the translation market, they have opened new avenues of academic researches and prompted researchers to revisit established questions. Hu and Li [1], for example, argue that LLMs have the potential to match human translation capabilities. A growing body of work has evaluated their translations from a range of perspectives, with literary translation receiving particular attention{citation}. Although these studies report strong performance, nonliterary texts remain less well studied. Translation style has also received relatively little attention compared with translation quality.

Research on translator style has likewise concentrated on literary translation, examining individual translators' characteristic choices at the lexical, syntactic, and discourse levels. Machine translation style, by contrast, has yet to receive comparable systematic attention. Lee [2] found that the literary translations produced by the most advanced neural machine translation systems available at the time differed stylistically from human translations. This raises two related questions: whether LLMs exhibit distinctive translation styles, and whether stylistic differences between human and machine translations also occur in nonliterary texts.

Although academics have not reached a consensus on how to define style, Hu and Xie [3] locate translator style in the approaches and linguistic patterns that recur across a translator's work. Their definition allows style to be investigated through measurable features of translated texts. Drawing on this approach, the present study compares the linguistic features of news translations produced by human translators, a neural machine translation system, and an LLM, in an effort to advance the understanding about machine translation style.

2. The characteristics of traditional embroidery art

2.1. Large Language Models in Translation

ChatGPT's translation performance has been a focus of academic research since its release. In an early study combining human assessment with automatic evaluation, Zhang and Zhao [4] found that ChatGPT outperformed neural machine translation in some domains. Although the relatively small size of corpus limited the study's representativeness, its evaluation design provided a useful point of reference for subsequent work on LLM translation.

Studies of literary translation have produced promising but varied results. Using Life, a literary work from Shaanxi, as a case study, Pang and Wang [5] compared an LLM with mainstream neural machine translation systems using BLEU and TER scores. ChatGPT performed better overall than DeepL Translator and Youdao Translate. Zhao et al. [6] drew on CSSCI-indexed journals in language and literature, books, and professional examination questions, and used BLEU and CHRF scores to compare several LLMs. ChatGPT outperformed Baichuan, Atom, and Aquila. Yet its advantage was not consistent across studies. In tests of Chinese-to-English translation using excerpts from Blossoms, several Chinese-developed LLMs surpassed ChatGPT on different measures [7]. Gao et al. [8] found that ChatGPT outperform neural machine translations in terms of poetry translation.

Taken together, these studies show that translation quality dominates research on LLMs, leaving other aspects of their output less fully explored.

2.2. Corpus-Based Studies of Translator Style

Mona Baker [9] established a corpus-based methodology for studying translator style through quantitative analysis of translated texts. By treating recurrent linguistic features as evidence of individual stylistic preferences, she provided a framework that continues to inform corpus-based research. Studies following this approach typically examine lexical, syntactic, and discourse patterns to identify translators' characteristic language choices. Much of these projects focus on individual translators.

One established approach is to compare translations of the same source text. Liu et al. [10], for example, identified distinct styles in four English translations of Hongloumeng. Joly favored long sentences, while Father Bonsall's translation was relatively easy to read. Yang Xianyi and Gladys Yang stayed closest to the source text, whereas David Hawkes and John Minford adhered most closely to target-language usage. In a comparison of Ku Hung-ming's and Arthur Waley's translations of the Analects, Lü and Chen [11] found that both conveyed the source information and its implicit semantic relationships effectively, but differed in lexical richness, sentence structure, and their treatment of culture-specific terms. Zhang and Zhu [12] examined two Chinese translations of The Kite Runner using lexical richness, part-of-speech frequency distributions, and high-frequency words. Li Jingyi's translation tended toward foreignization, while Li Jihong's favored domestication.

For translators who are also writers themselves, comparing translated and original works offers another way to investigate style. Zhu and Li [13] found that Ezra Pound's translations used a varied vocabulary, avoided excessively long words and sentences, and kept sentence structure from becoming overly complex. These features made the translations accessible and suited to the reading habits of native English speakers. Chen and Li [14] also found differences between Eileen Chang's style when translating her own works and when translating those of other writers.

Translator style can also be studied at the group level. Li and Li [15] treated China's official translation team as a collective translator and compared its translations of political texts with those produced by overseas translators. The official team's translations were found to align more closely with native English speakers' patterns of thinking. Liu [16] similarly identified stylistic differences between native and nonnative translators.

Despite this range of approaches, the literature remains centered on human translators, particularly those working with literary texts. Translator style in other domains, and the styles of machine translators, have received much less attention.

2.3. Research Gap and Questions

Existing research has expanded our understanding of both LLM translation capabilities and translator style, but has largely focused on literary translation rather than other practical text types. Studies of machine translation style remain scarce, with LLMs receiving particularly little attention. The present study addresses this gap by comparing news translations produced by an LLM and a neural machine translation system with human translations of the same texts. It answers questions as follows:

(1) Do LLM-generated news translations differ stylistically from human translations?

(2) Do neural machine translations of news differ stylistically from human translations?

(3) Do LLM-generated and neural machine translations of news differ stylistically?

3. Research Design

The study uses four purpose-built corpora: Chinese news texts published by the Ministry of Foreign Affairs (MFA), their official human translations, and translations of the same source texts produced by a neural machine translation system and an LLM. The corpora were compiled in the procedures as follows.

Text selection. News is often translated through a process that combines translation with editing, making it difficult to identify directly comparable source and target texts. Major online outlets such as People's Daily Online and Xinhua have official English websites, but these differ substantially from their Chinese counterparts in both structure and content. Their Chinese and English news items therefore could not be reliably matched for the present analysis. Some Ministry of Foreign Affairs news pages, however, link directly to corresponding English pages, with reciprocal links to the Chinese versions. Following Zhu and Hu [17], publication times and content correspondence were examined to establish the direction of translation. The English items consistently appeared slightly later than the Chinese versions and closely matched their content. These observations, together with the Ministry of Foreign Affairs's dedicated translation department, supported the use of the English texts as professional human translations of the Chinese news items. MFA news was therefore selected for the study.

Text collection. On July 26, 2025, a Python script downloaded all items available in the news section of the Ministry of Foreign Affairs's English website. Each page was checked for a link to a corresponding Chinese item, which was downloaded where available. The Chinese source texts and English translations were then aligned at the document level and saved as CSV files for further processing.

Text processing. The downloaded news texts contained no extraneous material requiring removal, so no additional cleaning was needed.

Text translation. Google Translate was selected to represent neural machine translation (NMT), and DeepSeek to represent LLMs. Google's adoption of NMT across its translation system in 2016 made Google Translate one of the earliest commercial products to deploy the technology at scale. Its architecture and translation performance make it an established benchmark for NMT. DeepSeek was chosen for its technical capabilities, prominence in Chinese-speaking contexts, and strong performance in processing Chinese and English. Access to the two systems was also a consideration.

Both systems provide web interfaces and APIs. Their APIs were used to translate the collected Chinese news items into English. DeepSeek declined to translate some items, possibly because it identified parts of their content as politically sensitive. Successfully translated items were retained, aligned at the document level, and saved as CSV files.

The resulting dataset consists of four corpora, each containing the same 131 news items. The source corpus holds the original Chinese texts. The human translation corpus contains the official English versions published on the Ministry of Foreign Affairs website, totaling 67,897 tokens. The Google Translate corpus contains 65,147 tokens, and the DeepSeek corpus contains 65,218. Because the three translation corpora share the same source texts and have similar token counts, they provide a comparable basis for stylistic analysis.

Python scripts were used to calculate lexical, syntactic, and discourse measures for each translation corpus, in an effort to examine style.

4. Results and Discussion

4.1. Lexical Features

Lexical measures are commonly used to identify

characteristic patterns in a translator's word choices. The analysis here focuses on type–token ratios and mean word length.

The type–token ratio (TTR), calculated as the number of distinct word types divided by the total number of tokens, measures lexical variation. A higher ratio generally indicates a more varied vocabulary. TTR has been widely used in translator style research since Baker introduced it to the field. Its usefulness declines, however, as corpus size increases [18]. The standardized type–token ratio (STTR) addresses this problem by dividing a corpus into fixed-length segments, usually of 1,000 words, calculating the TTR for each segment, and averaging the results. This provides a corpus-level measure that is less affected by text length.

Each corpus contains 131 news items aligned with their corresponding source texts. Python calculations showed that the average item contained approximately 500 tokens in all three translation corpora, with very few exceeding 1,000 tokens. Given these relatively short texts, the study introduces the average article type–token ratio (ATTR) as a supplementary measure of lexical richness. Following the averaging principle used in STTR, ATTR is calculated by averaging the individual TTRs of all 131 news items.

Mean word length is measured as the average number of letters per word. It serves as an indicator of lexical complexity: longer words tend to be associated with more complex vocabulary and greater reading difficulty.

Custom Python scripts calculated these measures for the human, NMT, and LLM translation corpora. Table 1 reports the results.

Table 1. Lexical Measures across the Three Translation Corpora

Parameter	MFA news Human translation	MFA news Neural machine translation	MFA news LLM translation
Token count	67897	65147	65218
Type–token ratio (TTR)	0.4824	0.4806	0.4845
Standardized type–token ratio (STTR)	0.4497	0.4433	0.4474
Average article type–token ratio (ATTR)	0.5370	0.5418	0.5450
Mean word length	5.80	5.93	5.78

Type–Token Ratios

Table 1 shows closely comparable values across the three corpora. The human translations have a TTR of 0.4824, an STTR of 0.4497, and an ATTR of 0.5370. The corresponding values are 0.4806, 0.4433, and 0.5418 for NMT, and 0.4845, 0.4474, and 0.5450 for the LLM.

None of the type–token measure reveals a marked difference between the corpora.

Mean Word Length

Mean word length is 5.80 in the human translations, 5.93 in the NMT outputs, and 5.78 in the LLM outputs (Table 1).

The LLM uses slightly shorter words than the human translators on average, while NMT uses slightly longer ones. Even the difference between the highest and lowest means is only 0.15 letters, indicating little variation in word length across the three corpora.

The lexical results suggest that the human and machine translations of Ministry of Foreign Affairs news show similar patterns of word choice. The nature of the material may help explain this convergence. Ministry of Foreign Affairs news is politically sensitive and allows little room for error, which may encourage conservative translation choices and constrain the scope for human translators to use a distinctive vocabulary.

4.2. Syntactic Features

Translation requires adjustments to sentence structure as meaning is transferred between languages. To examine whether such adjustments differ across the three sets of translations, the syntactic analysis focuses on mean sentence length.

Mean sentence length is the average number of words per sentence and provides an indication of syntactic complexity, with longer sentences generally associated with more complex structures. In this study, texts were first segmented into sentences with spaCy, after which the total number of words was divided by the number of sentences in each corpus. Table 2 presents the results.

Table 2. Mean Sentence Length across the Three Translation Corpora

Parameter	MFA news Human translation	MFA news Neural machine translation	MFA news LLM translation
Mean sentence length	28.29	27.13	26.94

Mean sentence length is 28.29 words in the human translations, 27.13 in the NMT outputs, and 26.94 in the LLM outputs (Table 2).

Human translations contain the longest sentences on average, followed by NMT and then the LLM. The two machine systems differ by just 0.19 words per sentence. Earlier research has distinguished short, medium, and long sentences, classifying those with more than 25 words as long. By this criterion, all three corpora favor long sentences. The 1.16-word difference between human translation and NMT is also modest given that both averages exceed 25 words.

On this measure, the three sets of translations show little

evidence of different syntactic styles.

4.3. Discourse Features

Readability matters in news translation because it affects how effectively information reaches readers. This study measure readability across three corpus using Flesch Reading Ease, Flesch–Kincaid Grade Level, and the Gunning Fog Index.

Flesch Reading Ease assesses readability using sentence length and syllable frequency, derived from sentence, syllable, and word counts. Scores are conventionally interpreted on a scale from 0 for extremely difficult texts to 100 for extremely easy texts.

Flesch–Kincaid Grade Level uses the same underlying sentence and word characteristics to estimate the educational level needed to understand a text. Higher values indicate that readers require more years of education.

The Gunning Fog Index estimates the educational level needed to read a text from its mean sentence length and the proportion of words with more than three syllables. Higher scores indicate greater reading difficulty.

All three readability measures were calculated with the Python textstat package. The results appear in Table 3.

Table 3. Readability Measures across the Three Translation Corpora

Parameter	MFA news Human translation	MFA news Neural machine translation	MFA news LLM translation
Flesch Reading Ease	18.78	15.24	18.51
Flesch–Kincaid Grade Level	17.36	17.87	17.37
Gunning Fog Index	20.26	0.81	20.25

Table 3 reports the following values. The reported Flesch Reading Ease scores are 18.78 for the human translations, 15.24 for NMT, and 18.51 for the LLM. Flesch–Kincaid Grade Levels are reported as 20.26, 17.87, and 17.37 for the human, NMT, and LLM corpora, respectively. The Gunning Fog Index is reported as 20.26 for the NMT corpus, 20.81 for the NMT corpus, and 20.25 for the LLM corpus.

The readability measures do not clearly distinguish the three sets of translations. Although the Flesch Reading Ease score for NMT differs from those of the other two corpora by more than three points, all three fall within the same readability category.

These results reveal no marked stylistic differences at the discourse level as measured by readability.

5. Conclusion

This study compared human, NMT, and LLM translations of Ministry of Foreign Affairs news using lexical, syntactic, and discourse measures. Following Hu and Xie, it examined recurring features across multiple texts produced by the same translator and extended this approach to news translation, in an effort to examine translators' style. The measures examined did not reveal marked stylistic differences. The translations were similar in lexical choice and sentence structure, and all were relatively difficult to read.

Lee reported that neural machine translation differed stylistically from human translation but was moving closer to it over time. The present findings are consistent with that trend. Both NMT and LLM produced news translations whose measured stylistic features closely resembled those of human translations. This convergence suggests progress in machine translation, while also indicating that the stylistic measures used here offer limited scope for distinguishing human from machine output. Other approaches to identifying machine translations therefore warrant investigation.

The findings should be considered in light of several limitations. All texts came from a single source, Ministry of Foreign Affairs news, whose political sensitivity may constrain stylistic choices and limit the generalizability of the results. The study examined only Chinese-to-English translation, leaving the reverse direction untested. Its analysis was also limited to common lexical, syntactic, and discourse measures.

Further research could test these findings using news from a wider range of sources and extend the comparison to English-to-Chinese translation and other language pairs. Exploring less commonly used corpus measures may also reveal stylistic differences that the measures examined here do not capture.

References

[1] Hu, K. B., & Li, X. Q. (2023). The development of translation studies in the context of large language models: Issues and prospects. Chinese Translators Journal, 44(6), 64–73, 192. (in Chinese)

[2] Lee, C. (2022). How do machine translators measure up to human literary translators in stylometric tests? Digital Scholarship in the Humanities, 37(3), 813–829. https://doi.org/10.1093/llc/fqac015

[3] Hu, K. B., & Xie, L. X. (2017). Corpus based research on translator style: Scope and approaches. Chinese Translators Journal, 38(2), 12–18, 128. (in Chinese)

[4] Zhang, W. Y., & Zhao, B. (2024). Has generative artificial intelligence opened a new era of machine translation? A comparative study of quality and reflections on translation education. Journal of Beijing International Studies University, 46(1), 83–98. (in Chinese)

[5] Pang, Y. X., & Wang, X. Y. (2023). A study of ChatGPT translation quality in the context of large language models: A case study of excerpts from Life, a literary work from Shaanxi. Modern English, (22), 67–70. (in Chinese)

[6] Zhao, X., Zhao, Z. X., & Sun, F. L. (2023). An evaluation of large language model performance in language and literature. Technology Enhanced Foreign Language Education, 57–65, 114. (in Chinese)

[7] Zhao, Y., Zhang, H., & Yang, Y. C. (2024). A comparative study of large language model quality in text translation: A case study of translating Blossoms. Technology Enhanced Foreign Language Education, 60–66, 109. (in Chinese)

[8] Gao, R., Lin, Y., & Zhao, N. (2024). Machine translation of Chinese classical poetry: A comparison among ChatGPT, Google Translate, and DeepL Translator. Humanities and Social Sciences Communications, 11(1), 1–10. https://doi.org/10.1057/s41599-024-02622-9

[9] Baker, M. (2000). Towards a methodology for investigating the style of a literary translator. Target. International Journal of Translation Studies, 12(2), 241–266. https://doi.org/10.1075/target.12.2.04bak

[10] Liu, Z. Q., Liu, C. P., & Zhu, H. (2011). An exploratory study of translator style in four English translations of A Dream of Red Mansions: Corpus based statistics and analysis. Chinese Translators Journal, 32(1), 60–64. (in Chinese)

[11] Lü, P. F., & Chen, D. S. (2021). A corpus based comparative study of translation styles in English versions of the Analects: A case study of the translations by Ku Hung-ming and Arthur Waley. Shanghai Journal of Translators, (3), 61–65. (in Chinese)

[12] Zhang, J. D., & Zhu, Y. F. (2020). A corpus based comparison of the styles of two translations of The Kite Runner. Technology Enhanced Foreign Language Education, (5), 50–57, 8. (in Chinese)

[13] Zhu, C. W., & Li, R. F. (2023). A corpus based investigation of Ezra Pound's translator style in English translations of Chinese classics. Foreign Language Education, 44(4), 75–82. (in Chinese)

[14] Chen, F., & Li, D. (2023). Patronage and ideology: A corpus-assisted investigation of Eileen Chang's style of translating herself and the other. Digital Scholarship in the Humanities, 38(1), 34–49. https://doi.org/10.1093/llc/fqac032

[15] Li, X., & Li, T. (2020). A corpus based comparative study of styles in the translation of political documents: A case study of English translations of the documents of the Third Plenary Session of the 18th CPC Central Committee. Foreign Language Learning Theory and Practice, (3), 74–84. (in Chinese)

[16] Liu, A. J. (2020). A corpus based comparative study of translator style: A case study of English translations of Zhu Ziqing's essays. Technology Enhanced Foreign Language Education, (4), 101–105, 16. (in Chinese)

[17] Zhu, Y. F., & Hu, K. B. (2014). Semantic preference and semantic prosody of bei constructions: A comparative study of translated and original news corpora. Journal of Foreign Languages, 37(1), 53–64. (in Chinese)

[18] Liang, M. C., Li, W. Z., & Xu, J. J. (2017). A course in corpus applications. Foreign Language Teaching and Research Press. (in Chinese)