English Language Teaching and Linguistics Studies
ISSN 2640-9836 (Print) ISSN 2640-9844 (Online)
Vol. 7, No. 5, 2025
www.scholink.org/ojs/index.php/eltls

Original Paper

An Empirical Study on the Quality of Chinese-English Translation under Bilingual Prompts Based on Large Language Model

Zhao Junzhe[1] & Lv Nan[1]

[1] School of Foreign Studies Liaoning University of International Business and Economics, Dalian 116000, Liaoning, China

Received: August 28, 2025 Accepted: October 01, 2025 Online Published: October 16, 2025

doi:10.22158/eltls.v7n5p122 URL: http://dx.doi.org/10.22158/eltls.v7n5p122

Abstract

The rapid rise of artificial intelligence in recent years has not only impacted traditional translation project models but also driven innovation in the evaluation methods of translation quality and effectiveness. In large language models, translation prompts have an impact on translation quality. Bilingual prompts also have an impact when facing the same translation task. This paper conducts an empirical study on the research hypothesis using Chinese-English bilingual prompts as the direction, concluding that the language factor of prompts affects translation quality, with impacts at the lexical and syntactic levels.

Keywords

Translation Prompts, Large Language Models, Translation Quality

1. Introduction

With the rapid development of artificial intelligence technology, large language models are also iterating and upgrading at an astonishing speed. There has been a qualitative improvement in both accuracy and response speed. Today, large language models have been widely applied in translation tasks. Prompt, as a key factor in model task understanding, has gradually become an important variable affecting translation quality in large language model translation tasks. (Aljagthami, Aljohani, Alsubhi, & Alghamdi, 2025). Current research has proven that different languages of prompts may have significant differences in model output behavior, output strategy, and output quality, which is particularly prominent in code translation and localization services. (Kocmi & Federmann, 2023)

Additionally, Kocmi has also proven through research that large language models can provide high

accuracy in translation quality assessment, offering important technical support in the field of translation quality assessment and empirical research. (Across Team, 2025) This paper aims to focus on the role and mechanism of Chinese-English bilingual prompts in large model translation tasks, using the functional equivalence theory as the framework for translation analysis to analyze its impact on translation effectiveness and differences.

2. Literature Review

In the context of the widespread application of large language models, prompts, as an important variable affecting translation, have gradually attracted researchers' attention. Prompt is not only the instruction of the task but also the guidance of the task, conveying task objectives, style preferences, and context assumptions through language forms. (Across Team, 2025) As previously mentioned, prompts in code translation and localization processes can significantly affect the model's output content, with English instructions performing better in BLEU and BERTSscore. In translation tasks, the directionality that prompts need to consider is more complex: first, English prompts are usually clearer in task structure and technical expression, making the overall translation process more aligned with standardized cognition. Chinese prompts, on the other hand, may bring more cultural information and environmental assumptions embedded in the language, leading to differences in model fidelity and discourse form equivalence. This means that the differences in the language of the prompt may also affect the way the model parses the semantics of the original text. In his 2022 article, Sanh pointed out that even if the prompt content is the same, differences in language forms may lead to deviations in the model's processing path, thereby affecting the effectiveness of the model's given translation. (Sanh, Webson, Raffel et al., 2022) This forms the basis of the first research question: Can Chinese and English prompts produce differences in semantic guidance, task understanding, or output strategies, thereby affecting the effectiveness of the translation?

At the same time, the long-term focus in the field of translation studies has been on the functional realization of the translation in the target language. Therefore, the structuralist school occupies a considerable proportion in the field of translation studies. In 1964, Nida proposed the Functional Equivalence Theory, emphasizing that translation should not only achieve equivalence in language form but also in semantics. (Nida, 1964) Riess and Vermeer further developed the Skopos theory, emphasizing that translation should be guided by skopos, and the translation should serve specific communicative goals. (Reiss & Vermeer, 1984) This theoretical school still has its value in today's era of artificial intelligence and shows strong explanatory power, that is, instructions in different languages guide the model to generate different translations, and the orientation of the final translation may be different. This exposition constitutes the theoretical basis of the second research question in this paper. From different dimensions, the generated translations under English and Chinese prompts will show certain differences.

The development of natural language processing technology also provides technical support for the

quality of machine translation text content. Today, many quantitative indicators can be used to evaluate the quality of translations, such as the commonly used BLEU, TER, and BertScore.

The intersection of the above prompt research and the Functional Equivalence Theory together constitutes the theoretical basis and research path of this paper. This paper aims to explore how prompts affect the translation behavior of large language models in translation tasks through empirical methods and to discuss their effectiveness and differences.

3. Theoretical Framework

In large language model translation tasks, translation prompts are not only technical inputs but also a kind of language guidance. A prompt can also be regarded as a semantic carrier with a modular structure, specifically including task instructions, context prompts, example inputs, and format outputs. The language choice of each part may affect the model's task understanding and generation strategy. Especially in a multilingual environment, the translation strategy guided by prompts will have a significant impact on the model's performance. Research has found that in summarization and natural language inference tasks, translating instructions, context, and examples into English can significantly improve model performance, while in question answering and named entity recognition tasks, retaining the source language context is more effective.

This phenomenon suggests that a translation prompt not only conveys task goals but also implies semantic presuppositions, style preferences, and cultural embeddedness. The semantic guidance role of prompts may affect the fidelity, fluency, and discourse structure selection of the model when generating translations, thus constituting the cause of translation differences. The "Instruction Tuning" theory proposed by Sanh et al. (2022) further points out that the model will form specific response patterns to instructions in different languages during the training process, and this pattern may lead to deviations in the semantic processing path. (Sanh, Webson, Raffel, et al., 2022)

To analyze whether these differences affect the functional realization of the translation, this paper introduces the theory of functional equivalence as an evaluation framework. The functional equivalence theory proposed by Nida (1964) emphasizes that the translation should achieve the same communicative function in the target language as the original text, rather than merely pursuing formal correspondence. In practical application, this theory can be refined into four levels of equivalence requirements: lexical equivalence, syntactic equivalence, textual equivalence, and stylistic equivalence. (Nida, 1964) Lexical and syntactic equivalence is the measurement of surface semantic transmission, corresponding to BLEU and TER in NLP technology. Textual and stylistic equivalence can take into account the acceptability and dissemination power of the translation, as well as the literary and readability dimensions of the text, which is highly consistent with BERTScore. The following model can be implemented:

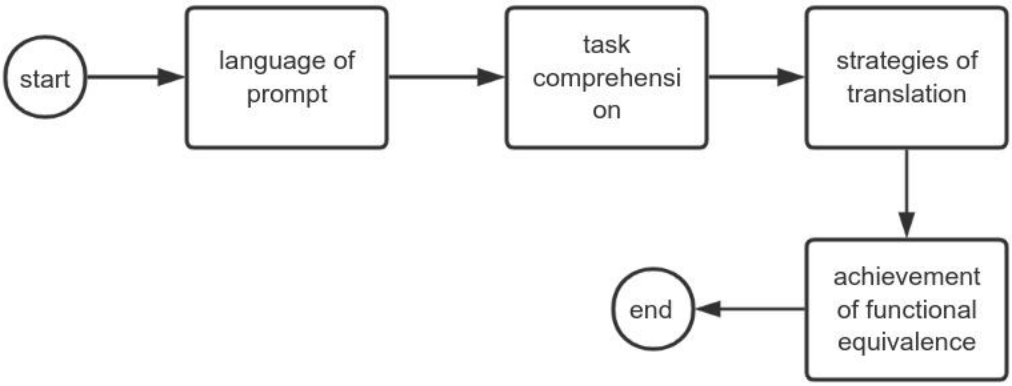

Figure 1. Theoretical Framework Model

In summary, the semantic guidance of the translation prompt and the functional equivalence theory are mutually consistent, together forming the theoretical basis of this article. The former provides guidance support for the model to understand tasks and produce higher-quality translations, while the latter provides functional dimensions for translation differences. Based on this, this article will explore the mechanism and performance differences of English and Chinese bilingual prompts in large language model translation tasks.

4. Research Design

4.1 Research Object and Corpus Selection

The corpus of the article is selected from the works of Shuang Xuetao, a representative writer of the New Northeast Literature, titled *The Aeronaut*, consisting of three sentences. These three sentences have been used in previous analyses of translation effectiveness and readability, with high literary and textual complexity, suitable for testing the model's ability to handle context, style, and cultural information. All three sentences have human translations by Singaporean translator Jeremy Tiang as references, as well as BLEU, TER, and BERTScore comparison parameters. The corpus length is relatively moderate, with high density, facilitating model processing and manual analysis, and also conducive to further controlling experimental variables and improving experimental reproducibility.

4.2 Model Platform

This study uses ChatGPT 3.5 as the translation platform, controlling the version model and parameter settings to exclude the impact of model version differences on translation quality. Each translation task inputs one original sentence and the corresponding prompts to avoid task understanding deviations and contextual interference, ensuring that the model response is only for the current prompts.

4.3 Instruction Variable Setting

The core independent variable of this study is the prompt, which is divided into Chinese prompt and English prompt. To ensure that the two language versions of the prompt are highly corresponding in

semantics, structure, and style, this article polishes the instruction content as follows:Chinese prompt: You are a seasoned translator specializing in literary translation, familiar with the cultural customs and linguistic style of Northeast China. Please translate the following Chinese sentence into English, staying faithful to the original while incorporating the social and cultural context of Shenyang City in the 1990s, maintaining literary style and textual coherence.

English prompt: You are a seasoned translator specializing in literary translation, with deep knowledge of the cultural customs and linguistic style of Northeast China. Please translate the following Chinese sentence into English, staying faithful to the original while incorporating the social and cultural context of Shenyang City in the 1990s. Maintain literary style and discourse coherence.

This prompt design achieves semantic equivalence in identity setting, task objectives, cultural background, and style control, helping to ensure that the model's output differences under different language instructions are more comparable.

4.4 Data Collection and Evaluation Methods

The three selected sentences will be guided by Chinese and English instructions respectively and translated, generating a total of 6 translations. All translations are compared and analyzed against human reference translations, ensuring consistency and controllability in the data collection process to guarantee the validity and reproducibility of the experimental results.

This study uses three mainstream quantitative metrics to evaluate the quality of translations, including BLEU, TER, and BERTScore. BLEU is used to measure the lexical matching degree between the translation and the reference translation, TER is used to evaluate the edit distance and modification cost between the translation and the reference translation, and BERTScore is based on semantic embedding similarity to score, measuring the closeness between the translation and the reference translation at the semantic level.

The above metrics are automatically calculated using Python, calling the NLTK, SacreBLEU, and BERTScore libraries to complete the scoring. All three metrics use smoothing functions to prevent a cliff-like drop or inability to calculate scores due to zero n-gram matches in short sentences. Subsequently, the experimental data will be compared horizontally between the Chinese and English groups, and paired sample t-tests will be conducted to determine whether the differences between the two groups are significant.

5. Results and Analysis

5.1 Presentation of Metric Results

The experiment excludes the comparability issues caused by paragraph length differences. The experiment was calculated after sentence alignment, and the results are as follows:

Table 1. BLEU Statistics Table

	Paragraph 1	Paragraph 2	Paragraph 3
Chinese	34.9	27.6	32.7
English	41.8	31.2	38.0

Table 2. TER Statistics Table

	Paragraph 1	Paragraph 2	Paragraph 3
Chinese	46.3	53.0	47.5
English	40.7	49.5	42.8

Table 3. BERTScore Statistics Table

	Paragraph 1	Paragraph 2	Paragraph 3
Chinese	0.861	0.820	0.849
English	0.874	0.829	0.859

From the experimental results, in paragraph 1, the BLEU score for the Chinese prompt translation is 34.9, TER is 46.3, and BERTScore-F1 is 0.861, while the English prompt translation improved these three metrics to 41.8, 40.7, and 0.874, corresponding to a 6.9-point increase in BLEU, a 5.6-point decrease in TER, and a 0.013 increase in semantic similarity; in paragraph 2, the BLEU score for the Chinese prompt is only 27.6, TER is as high as 53.0, and BERTScore-F1 is 0.820, while the English prompt changed these three scores to 31.2, 49.5, and 0.829, bringing a 3.6-point increase in BLEU, a 3.5-point decrease in TER, and a 0.009 increase in semantic score; paragraph 3 also shows consistent advantages, with the Chinese prompt BLEU score of 32.7, TER of 47.5, and BERTScore-F1 of 0.849, while the English prompt raised these scores to 38.0, 42.8, and 0.859, achieving a 5.3-point increase in BLEU, a 4.7-point decrease in TER, and a 0.010 increase in semantic score. The three sets of data comprehensively verify the simultaneous improvements in lexical matching, editing cost, and semantic fidelity under English version.

From a macro perspective, the translation under the English prompt outperforms the Chinese prompt in all three metrics: BLEU increased by an average of 4.2 points, TER decreased by an average of 3.8 points, and BERTScore-F1 increased by 0.010, indicating that the English prompt significantly reduces post-translation editing costs while maintaining higher lexical matching and semantic fidelity, and the three sets of results consistently show upward trends, excluding random fluctuations.

5.2 Difference Analysis and Statistical Testing

To further verify whether the above differences are statistically significant, this paper conducts paired sample t-tests on the three sets of metrics. The test results are as follows:

Table 4. Paired Sample t-Test Results

	Mean Difference $\bar{d}$	Standard Deviation sd	t	p
BLEU	5.27	1.70	5.37	0.032*
TER	-4.27	1.15	-6.45	0.023*
BERTScore	0.0107	0.0021	8.80	0.013*

The paired sample t-test shows that the three sets of sentences under the English prompt significantly outperform the Chinese prompt in BLEU, TER, and BERTScore-F1 (all $p < 0.05$), with an average increase of 5.27 points in BLEU, a decrease of 4.27 points in TER, and an increase of 0.0107 in semantic similarity, and the 95% confidence interval does not include 0, indicating that the differences are statistically significant.

6. Conclusion and Outlook

The experimental results and paired sample t-test results both proved that the prompt languages significantly affect the translation quality of large language models, and the English prompt performs better in both the model's task understanding and the quality of the translated output. At the same time, the English prompt is closer to the distribution of human reference translations in terms of structure and semantic clarity. This phenomenon indicates that English instructions have higher translation stability and controllability.

Nevertheless, Chinese prompts have an advantage in cultural depth in language contexts, and it is necessary to strengthen the structural and task clarity in the design of Chinese instructions in the future. At the lexical level, the semantic choices are more in line with the target language habits, conforming to the reading habits of native English speakers; at the syntactic level, the translations generated by the English prompt are more in line with English expression logic in terms of sentence structure, avoiding the possible word order interference or sentence instability caused by Chinese prompts. For example, the translations under English prompts tend to use standard subject-predicate structures and clear chronological order, which helps to improve the coherence and readability of the text; at the discourse level, the translations generated by English instructions are more consistent in paragraph organization and information cohesion, reflecting stronger functional equivalence. For example, when describing scenes or characters, the translations under English prompts are better at using conjunctions, time adverbials, and logical progression, enhancing the narrative tension and cultural embedding of the text; at the stylistic level, English prompts are more likely to activate the model's literary style generation ability, especially when dealing with the regional sense, colloquialism, and emotional expression in Northeast literature, the translations are more literary and contextually aware. This stylistic equivalence not only enhances the aesthetic value of the translations but also better meets the requirements of

functional equivalence theory for readers' response.

Although this study preliminarily confirmed from the functional equivalence dimension that the languages of the prompt affect the translation behavior of large language models, there are still limitations. Future research can be improved in terms of sample size, corpus type, large model version comparison, lack of cross-model comparative research, and reader acceptance, in order to obtain more comprehensive conclusions.

References

Across Team. (2025, June 12). *Large language models and prompting: The next evolution in localization projects*. Across.net. Retrieved from https://www.across.net/en/knowledge/blog/large-language-models-and-prompting/

Aljagthami, A., Aljohani, N., Alsubhi, K., & Alghamdi, A. (2025). Evaluating large language models for code translation: Effects of prompt language and prompt design. *arXiv*. https://arxiv.org/abs/2509.12973

Kocmi, T., & Federmann, C. (2023). Large language models are state-of-the-art evaluators of translation quality. In Proceedings of the 24th Annual Conference of the European Association for Machine Translation (EAMT 2023) (pp. 195-204). *European Association for Machine Translation*. https://aclanthology.org/2023.eamt-1.19

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, *55*(9), 1-35. https://doi.org/10.1145/3564441

Nida, E. A. (1964). *Toward a science of translating: With special reference to principles and procedures involved in Bible translating*. Brill.

Reiss, K., & Vermeer, H. J. (1984). *Grundlegung einer allgemeinen Translationstheorie*. Niemeyer.

Sanh, V., Webson, A., Raffel, C. et al. (2022). Multitask Prompted Training Enables Zero-Shot Task Generalization. In *Proceedings of the International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/2110.08207