Trends in Social Sciences and Humanities Research
Print ISSN: 2959-992X
Online ISSN: 2959-9938
DOI: https://doi.org/10.61784/tsshr3278

AI AND HUMAN TRANSLATION OF CONCEPTUAL METAPHORS BETWEEN CHINESE AND JAPANESE: EVIDENCE FROM AN INSTITUTIONAL REPORT

YuQian Xin
School of Foreign Studies, Nanchang University, Nanchang 330031, Jiangxi, China.

Abstract: This study compares the human Japanese translation of conceptual metaphors in a Chinese institutional report with the Japanese translation generated by artificial intelligence (AI). The study selected 10 cases containing metaphorical expressions through the Metaphor Identification Procedure and sorted them out into coding cases (C1–C10). The study compares the translations generated by GPT-4 multiple times and the resulting back-translations and subsequent Japanese translations with the reference Japanese translations, focusing on the dimensions of metaphor recognition, translation strategy, cultural adaptation, style and grammatical naturalness. The results show that when the metaphor has a common conceptual basis, the performance of the model is more reliable; while when the metaphor interpretation relies on specific cultural mapping, context inference or formal institutional style, its performance declines. Human translation can more flexibly choose to retain, neutralize or reconstruct metaphorical mappings, while models tend to use literal conversion or direct interpretation. Research shows that although generative artificial intelligence can improve the efficiency of professional translation, it still needs to be supplemented by expert proofreading when dealing with metaphors with cultural connotations or texts with high style requirements. This study offers a qualitative comparison of metaphor translation across the selected cases.
Keywords: Cognitive linguistics; Conceptual metaphor; Generative artificial intelligence; Human translation; Translation strategy

1 INTRODUCTION

Conceptual Metaphor Theory believes that metaphor is not only a rhetorical means, but also a cognitive mechanism rooted in physical experience [1]. From this perspective, expressions such as "path", "obstacle", "light", "tool" and "barrier" can activate the structured mapping relationship between the specific source domain and the more abstract target domain. Therefore, translation is by no means a simple word replacement. The translator must judge whether this mapping relationship is common among different languages; whether the image is a customary expression in the target culture; and whether the retention of this image fits the genre and register of the translation.

Such issues are particularly prominent in institutional reports. Metaphorical expressions in such reports often combine condensed images with formal communication functions. If literal translation is used, although the image at the lexical level can be retained, it may cover up the connotation of its original meaning; if the interpretation is used, although its connotation can be clarified, it weakens the original image. In addition, metaphors can be recast as similes, or replaced with another image in the target language. Which strategy to adopt depends on the specific semantic unit, the context in which it is located, and the expected requirements of the corresponding style in the target language.

Generative AI expands the scope of translation tools used by researchers and practitioners and improves the translation speed. Recent evaluation results show that generative artificial intelligence performs well in high-resource language pairs; however, the study also found that its performance will vary depending on the field, prompts, language pairs and context levels [2-6]. Research at the document level further shows that even if the translation at the sentence level is accurate, it cannot ensure the consistency of the long text or the effective control of the style [3, 5]. These findings support a case-based comparison, that is, recording the translation choices actually observed, rather than generally asserting that artificial intelligence translation or human translation is better overall.

The existing research on artificial intelligence-assisted translation is mostly focused on texts with English as the target language, while there are relatively few studies on Japanese [7-8]. The study of the selected source text also focuses on its English translation [9]. This study aims to fill this gap. By comparing the Chinese and Japanese translations, it focuses on the "conceptual metaphor unit". This study raises three research questions: how consistent is GPT-4 in rendering metaphorical expressions; what translation strategies are used in artificial intelligence translations and reference human translations respectively; and what impact will back translation and repeated translation have on metaphors, register and grammatical naturalness?

2 RESEARCH DESIGN

2.1 Corpus and Sample Selection

© By the Author(s) 2026, under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0).

The Chinese source excerpts and their reference Japanese translations were retrieved from the multilingual parallel-query platform maintained by Shanghai International Studies University (https://document.shisu.edu.cn/home20/). In order to facilitate reproduction and maintain consistency, the original document will be marked as "Source Report" (SR for short) below. All institutional and policy-related expressions extracted from SR are only treated as language data and semantic units; their inclusion in this article does not constitute any evaluative claim, nor does it represent the author's own position. This study uses Metaphor Identification Procedure (MIP) to identify metaphorical units [10]. The study selected 10 cases containing metaphorical expressions and numbered them from C1 to C10; among them, the source data set contains 360 Chinese characters, and the reference data set contains 629 Japanese characters. Table 1 adopts the English coding scheme instead of directly listing the Chinese sentences in the original text. During the analysis process, in order to maintain consistency, Japanese example sentences are presented in Roman characters with English interpretations.

Table 1 Coding Scheme for the Selected Conceptual-Metaphor Units

Case	Metaphor Category	English Gloss	Analytical Function
C1	Journey/path	Maintaining a chosen route and rejecting alternative routes	Retention of the path mapping
C2	Journey/obstacle	Entering deep water, tackling hard bones, and crossing dangerous shoals	Retention, neutralization, and simile
C3	Light/visibility	An institutional power operating in sunlight	Retention versus explicitation
C4	Embodied strength	Ambition, courage, and confidence represented as bodily resources	Embodied semantic transfer
C5	Land-people identity	The country and its people represented as mutually constitutive	Metaphor neutralization
C6	Animal/action	Coordinated actions represented through tiger, fly, and fox images	Back-translation of parallel images
C7	Journey/examination	Institutional renewal represented as a new examination route	Genre-sensitive mapping
C8	Equipment	An innovative theory represented as equipment for thought and action	Source-domain retention
C9	Instrument/resource	A coordinating framework represented as an effective instrument	Target-culture substitution
C10	Barrier/chain	Restrictions represented through walls, barriers, and broken chains	Register and lexical choice

2.2 Translation and Back-Translation Procedures

This experiment adopts a three-stage translation and back-translation process. The labels used in this article are as follows: Cn indicates the Chinese source unit; TrJ indicates the published Japanese reference translation; GPT1 indicates the initial Japanese translation generated by GPT-4 from Cn; GPT2 indicates the Chinese back-translation generated from TrJ; GPT3 indicates the final Japanese translation generated from GPT2.

Step 1. Input 10 Cn units into GPT-4 respectively, and attach the fixed instruction "translate into Japanese". The resulting output is recorded as GPT1 (without any post-editing).

Step 2. Input each TrJ unit into GPT-4, with the fixed instruction "translated into Chinese". The result of the obtained Chinese back translation is recorded as GPT2.

Step 3. Input each GPT2 unit into GPT-4 with the instruction "translated into Japanese". The obtained Japanese output is recorded as GPT3.

Step 4. One day later, repeat the above process under the same prompt. Retaining these repeated outputs, it aims to observe whether its main strategy and style selection remain stable, rather than treating a single response as a conclusion about the behavior of the model.

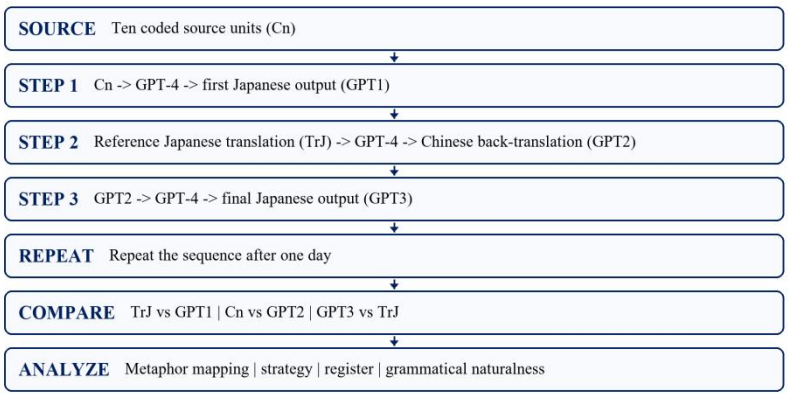


Figure 1 Research Procedure for the Three Translation Comparisons

Subsequently, three groups of studies were compared: TrJ and GPT1, Cn and GPT2, and GPT3 and TrJ. The analysis covers the following dimensions: metaphor recognition, retention or neutralization of source domain mapping, simile-based reformulation, target culture replacement, grammatical naturalness and register. Among them, the three categories

of "retention", "metaphor removal" and "simile-based reformulation" follow the strategic framework adopted in Japanese translation research in the past. The summary of the research process is shown in Figure 1.

3 RESULTS

3.1 Comparison of TrJ and GPT1

3.1.1 Metaphor identification

In coding cases, differences are most common in situations that need to be interpreted by context inference, specific cultural mapping or non-literal meaning. The following analysis only describes the observed output results, and does not infer the unobservable cognitive state inside the model. The difference between GPT1 and TrJ is only seen as evidence for specific sampling output and prompt configuration, not a universal feature shared by all large-scale language models or all translation situations.

C2 source unit. Its key expressions - "deep water", "hard bone" and "dangerous shoals" - conceive the difficult and uncertain task as a physical obstacle on the journey. In this context, the source domain unit of "Dangerous Shoals" does not refer to the specific geographical location, but constitutes an obstacle mapping.

TrJ. The phrase "kiken na hayase no yo na shiren" (meaning "a trial like dangerous rapids") in the translation constitutes the last difficulty. Among them, the comparison particle "yo na" clarifies this non-literal relationship and retains the original image through a simile.

GPT1. The sampling output adopts a more general expression "kiken na chiiki" (i.e. "dangerous area"). The translation retains the broad meaning of "risk", but does not reflect the obstacle mapping relationship presented by the reference translation (TrJ). Therefore, in TrJ, C2 is classified as "simile-based reformulation"; in GPT1, it is classified as "generalized".

This case shows that metaphor recognition cannot be evaluated by word matching alone. The output may retain metaphorical images, but leave implicit inferences outside the words; it may also omit images, but still convey the corresponding discourse function. Therefore, the appropriate judgment depends on the purpose of the analysis. This study records semantic transfer, image retention, naturalness of target language naturalness and register as independent dimensions, rather than combining them into a single "correct" or "wrong" label.

3.1.2 Translation-strategy differences

The reference translation presents three recurring strategies: retaining similar metaphors, neutralizing metaphors through direct expression in the target language, and recasting metaphors as similes. In the extracted cases, GPT1 used the first two strategies, and there was no clear simile-based reformulation. The result is only for the selected 10 sample units and should not be interpreted as evidence that "the model cannot produce similes under other prompts or conditions".

C1 embodies metaphorical retention. The original text uses expressions related to "route" ("lao lu" and "xie lu" respectively), which means to compare one established route with another unacceptable route. The TrJ translation uses "katsute no michi" and "jado", while the GPT1 translation uses "furui michi" and "jado". Although the vocabulary selection is slightly different, both versions retain the mapping relationship of "journey". Since the "path" experience provides a well-known conceptual structure in both languages, this image can be retained without much interpretation.

C5 embodies the "neutralization" strategy. The source word "Jiangshan" historically refers to land or territory, but in specific contexts, it refers to "country". TrJ translates it as "kuni", and GPT1 translates it as "kokudo". Both translations tend to directly express the concept of "country" or "territory", thus weakening the specific cultural image contained in the original word. Although this strategy improves the instant understandability, it narrows the semantic scope of the source word. Therefore, the case can be classified as: while successfully conveying the meaning of the proposition, accompanied by the partial loss of metaphorical resonance.

In contrast, C2 embodies the value of simile-based reformulation. The translation retains the image as an obstacle in the original text and adds a clear comparison, so as to guide readers to associate with the challenges referred to from the specific image. GPT1 transforms "hard bones" into "difficult problems" and "dangerous shoals" into "dangerous areas". This expression is intuitive and easy to understand, but the strategy weakens the parallel image that originally connected these three obstacles. The difference is a choice at the strategic level, not a simple translation of right and wrong.

In the selected case, the reference translation will flexibly adjust the strategy according to the semantic unit; in contrast, GPT1 alternates between "direct retention" and "neutralization processing". Therefore, the sampled AI outputs show a narrower range of strategies, especially in the case that similes can be used to preserve images and take into account clarity. Many tests have shown that although the specific wording may be different, from the level of macro strategy classification, its performance is more stable than the choice of specific words.

3.2 Comparison of Cn and GPT2

The "Cn-GPT2" comparison aims to examine whether the Chinese back translation generated by TrJ can restore the mapping relationship of the source text. When TrJ retains highly similar metaphors in cross-language correspondence, the GPT2 translation is usually closer to Cn (original Chinese). C6 is a typical example. The source text uses three juxtaposed "animal + action" images, namely "striking tigers", "swatting flies" and "hunting foxes". TrJ retains these parallel animal images, and GPT2 also restores the same three units. Although the expression that originally meant "coordinated action" has been simplified to the expression of "simultaneous action", the parallel structure and its core image are still clearly recognizable for Chinese readers.

When the reference translation (TrJ) neutralizes the metaphor or replaces it with the target cultural image, GPT2 may not be able to restore the original source domain. In Case C9, the word "fabao" refers to a particularly effective tool or method at the metaphorical level. The translator used "kirifuda", which is an image in the target language, which means "decisive card" or "decisive resource". The result given by GPT2 is "wangpai"; the word is semantically close to "kirifuda", but different from the original image. This is not a random mistake, but reflects the metaphorical expression that actually exists in the current Japanese input text. The case is classified as "target culture replacement" and is accompanied by "back-translation with a changed source-domain image".

The back-translation results show that semantic proximity can remain high even when the original image is not recovered. They also demonstrate a limitation of using back-translation as evidence of source fidelity. A fluent back-translation may conceal a shift introduced at an earlier stage, while a lexically different back-translation may still preserve the relevant function. For this reason, the present study compares each stage directly instead of treating GPT2 alone as a quality score.

3.3 Comparison of TrJ and GPT3

GPT3 and TrJ are largely similar in propositional content because GPT3 is generated from a back-translation that already reflects many choices in TrJ. Differences remain visible in lexical specificity and register. C10 illustrates this pattern. TrJ uses the formal institutional expression shoheki no setchi for the construction of barriers, whereas GPT3 uses the more general expression kabe o kizuku koto, literally building a wall. Both versions convey the barrier image, but the first is more nominal and institutional, while the second is more concrete and conversational.

In the coding scheme adopted in this paper, C10 is classified as a change from "formal institutional register" to "more general register". This observation is only for the current sample and does not involve a broader value judgment. The output of GPT3 is still readable, and it retains the core contrast of "barrier", "decoupling" and "severed connection"; however, its lexical selection makes the expression traces generated by the machine in the intermediate stage more obvious.

This comparison also shows that repeated translation can not only make the grammatical structure smoother, but also retain the previous semantic shift. Once the metaphor is "neutralized" (that is, the metaphorical color is lost) in GPT2, it is almost impossible to find clues to restore the original image in the final Japanese stage. On the contrary, if the source domain is retained in the back-translation process, GPT3 can often generate expressions similar to "Japanese reference translation" (TrJ). Therefore, this translation sequence forms a chain, in which each stage restricts the option of the next stage.

4 CONCLUSION

In the 10 cases analyzed, GPT-4 can usually retain the meaning of the proposition better; however, there are differences in its output in terms of metaphor recognition, cultural adaptation and register. The human translation as a reference shows richer translation strategies, including retention of metaphors, neutralization, target culture replacement and simile-based reformulation. When the compact image in the source language not only needs to be interpreted in context, but also requires the expression of the target language to meet the specific style norms, the difference between the two is the most significant.

The results of the study also show that "back translation" should not be regarded as a single independent indicator of fidelity. Although GPT2 can generate content that is fluent in language and semantically similar to the following Japanese input, it does not necessarily recover the original source image. Similarly, the content generated by GPT3 may be semantically similar to "Japanese reference translation" (TrJ), but there are differences in institutional register. Phased comparative analysis can provide more valuable evidence, because it can not only reveal where a shift in imagery or meaning first occurred, but also show whether this shift is retained, amplified or offset in the subsequent stage.

The above results are limited to the selected corpus, prompts, GPT-4 configuration and repeated tests, and should not be extended to all Chinese-Japanese translation tasks. In this sample, for translations involving metaphors and register, generative artificial intelligence has great practical value in auxiliary drafting and comparison, but it still cannot replace expert judgment. Future research should expand the scale of the corpus, complete the report model and prompt setting, introduce multiple human evaluators, and adopt explicit measures of inter-rater agreement to examine whether the observed strategy differences remain stable in different styles and model versions.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Lakoff G, Johnson M. Metaphors We Live By. Chicago: University of Chicago Press, 1980.
[2] Jiao W X, Wang W X, Huang J T, et al. Is ChatGPT a good translator? Yes with GPT-4 as the engine. arXiv, 2023. DOI: 10.48550/arXiv.2301.08745.
[3] Karpinska M, Iyyer M. Large language models effectively leverage document-level context for literary translation, but critical errors persist. Proceedings of the Eighth Conference on Machine Translation, 2023: 419-451.

[4] Kocmi T, Federmann C. Large language models are state-of-the-art evaluators of translation quality. Proceedings of the 24th Annual Conference of the European Association for Machine Translation, 2023: 193-203.
[5] Läubli S, Sennrich R, Volk M. Has machine translation achieved human parity? A case for document-level evaluation. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018: 4791-4796.
[6] Moslem Y, Haque R, Kelleher J D, et al. Adaptive machine translation with large language models. Proceedings of the 24th Annual Conference of the European Association for Machine Translation, 2023: 227-237.
[7] He X T. A study of Japanese translation in the context of artificial intelligence. International Public Relations, 2024(5): 23-25.
[8] Wen X, Tian Y L. A study on the effectiveness of ChatGPT in the translation of discourse with Chinese characteristics. Shanghai Journal of Translators, 2024(2).
[9] Wang L L. On the principle of fidelity in the translation of Party and government documents: A case study of the English translation of the 20th CPC Report. Chinese Translators Journal, 2023, 44(6): 116-123.
[10] Pragglejaz Group. MIP: A method for identifying metaphorically used words in discourse. Metaphor and Symbol, 2007, 22(1): 1-39.