Research on Corpus-based Cohesion: A Case Study of AI's Chinese-to-English Translation of *I and the Temple of Earth*

Yang,Xiaowei

Southwest Jiaotong University, Chengdu, Sichuan, 610031, China

Abstract: With the widespread use of this AI chatbot's translation capability, there has been a surge of interest in exploring its translation abilities. This study aligns with the research trend in AI translation and investigates the cohesion features of ChatGPT's Chinese-to-English translation of the text *I and the Temple of Earth* from the perspective of systemic functional grammar. The research reveals that the ChatGPT translation version of *I and the Temple of Earth* effectively employs repetition, personal reference, and conjunctions to establish discourse coherence, meeting the requirements for cohesive devices typically used in general texts. However, there are still some limitations in the discourse coherence of the ChatGPT translation, mainly manifested in the monotonous form of cohesion and the use of simplistic cohesive devices.

Keywords: AI translation; Cohesion; Discourse coherence; Corpus linguistics

DOI: 10.62639/sspjiss07.20250207

1. Introduction

Language is a system of social symbols that expresses meaning in a certain social context (Halliday, 2001). As the highest level of language structure, a discourse is a linguistic unit that expresses complete meaning in a certain context. Language communication is always in the form of discourse (Wilss, 1982). The basis of discourse coherence is the communication of cognitive and knowledge structures of the message sender and the message receiver. On the other hand, discourse coherence is realized through various means of coherence and any ideographic segment or discourse is not a random stack of individual sentences, but an organic aggregation of individual sentences within the whole segment and the whole discourse through different means of coherence. Therefore, the process of discourse translation can be said to be the process of identifying and reconstructing coherence (Wang Dongfeng, 1998).

In order to follow the current trend of AI translation, this paper takes the Chinese-to-English version of ChatGPT's *I and the Temple of Earth* as a corpus to explore the discourse coherence features in the translated text. *I and the Temple of Earth* is one of the most well-known prose texts by Shi Tiesheng, a famous contemporary Chinese writer. This essay is regarded as a concentrated manifestation of Shi Tiesheng's humanistic and philosophical writing. The beautiful and down-to-earth language of the prose can serve as a representative of contemporary Chinese literature, and can provide a textual example for the study of AI's ability to translate literary texts.

2. Literature Review

Discourse cohesion refers to the use of lexical and grammatical devices—such as reference, substitution, ellipsis, and conjunctions—to create a sense of "textual texture." These cohesive devices work by building inter-sentential

(Manuscript NO.: JISS-25-7-62015)

About the Author

Yang,Xiaowei (2001-), Femal, Miao ethnic group, Lichuan, Hubei Province, China, Master's Student, Master degree, Southwest Jiaotong University, Language Teaching and Second Language Acquisition.

links and establishing dependency relationships within the text. In doing so, they connect different parts of a text to form a coherent and unified whole. Cohesion is foundational to textual coherence. It ensures that the meanings expressed in different parts of the text are logically and semantically interrelated, enabling the text to function effectively as an integrated unit.

The importance of discourse cohesion lies in its significant impact on text comprehension and translation quality. A cohesive text facilitates the accurate transmission of information and enhances the ease with which readers can interpret and accept its content. Particularly in translation, maintaining the cohesion of the source text is a fundamental requirement. It directly influences the fluency, naturalness, and fidelity of the target text. When cohesive ties are missing or used inappropriately, the resulting translation may appear disjointed or awkward. Even if the translated sentences are grammatically correct in isolation, a lack of cohesion at the discourse level may cause the translation to sound unnatural, confuse the reader, or even distort the intended meaning of the original text.

With the rapid advancement of artificial intelligence (AI), and especially the growing use of machine translation (MT) tools such as Google Translate and ChatGPT, the definition of translation "quality" is being re-examined. Traditionally, translation quality was evaluated primarily based on word-level accuracy, i.e., how faithfully the translation reproduced the literal meaning of the original text. However, in today's increasingly complex multilingual and multicultural communication contexts, quality is now more broadly understood to include the ability to preserve cohesion and meaning across the whole text. A good translation should not only be accurate at the sentence level but also effectively convey the structure, flow, and nuanced relationships within the discourse of the original.

From the perspective of Systemic Functional Linguistics (SFL)—a theoretical framework that views language as a social semiotic system—machine translation can be evaluated in terms of how well it handles discourse cohesion. SFL identifies various types of cohesive relationships, including lexical cohesion, reference, and conjunction, among others. Research has shown that machine translations often fail to maintain these cohesive relations adequately. For instance, lexical cohesion errors, especially those related to terminology, are among the most common. These often arise because the translation system fails to interpret the meaning of words in light of their context, especially in cases where cultural or situational nuances are involved. SFL, with its focus on register variation, offers a useful lens for understanding such errors. It can also help guide translators in selecting more accurate and contextually appropriate equivalents, thus improving lexical cohesion.

Moreover, semantic coherence errors, sometimes known as "literalization" errors, indicate a failure to grasp the intended or pragmatic meaning of a word or phrase. This often results from an over-reliance on word-to-word translation strategies that ignore how meaning functions in context. For example, a machine may translate an idiom or metaphor too literally, creating a result that is semantically awkward or even nonsensical. SFL can help prevent such errors by emphasizing the importance of interpreting deeper-level meanings within the overall discourse structure. Maintaining semantic coherence is vital for producing translations that are not only accurate but also meaningful and appropriate in the target language and culture.

Another common issue in machine translation involves the inappropriate or insufficient use of reference and conjunctions. Studies have found that machine-translated texts tend to underuse cohesive devices such as reference, substitution, and ellipsis, in comparison to human translations. For example, in many machine-generated texts, substitution and ellipsis occur rarely, if at all. While connectives may appear frequently, they often lack variety and flexibility. They are typically applied mechanically, without full consideration of the discourse context, which can lead to a rigid and unnatural flow of ideas. Such patterns indicate that current machine translation systems are still limited in their ability to apply grammatical cohesion strategies dynamically and contextually.

These various error types—lexical, semantic, and grammatical—collectively highlight the deeper limitations of machine translation in handling discourse-level meaning. They are not just minor word or grammar mistakes; rather,

they reflect a failure to realize the cohesive structure that underpins a readable and coherent text. Without effective cohesion, the translated output may lack overall unity and fail to meet reader expectations for naturalness and clarity.

3. Research Methodology

(1) Research questions

This study mainly explores the following three questions: (1) What are the features of the English text of *I and the Temple of Earth* translated by ChatGPT in terms of lexical cohesion? (2) What are the features of the English text of *I and the Temple of Earth* translated by ChatGPT in terms of reference? (3) What are the features of the English text of *I and the Temple of Earth* translated by ChatGPT in terms of the use of conjunctions?

(2) Research instruments

In this study, quantitative research method was used to collect research data by utilizing corpus search capability. Firstly, the Chinese text of Shi Tiesheng's prose *I and the Temple of Earth* was translated into English text in ChatGPT. Then the translated text was loaded into AntConc, a corpus analysis tool, to retrieve different data according to the research needs. In addition, the software Wordnet, a large lexical database of English, was used to find out the lexical ensembles of the words.

4. Results and Discussion

(1) Lexical cohesion

Lexical cohesion is the semantic connection of a discourse established through lexical selection (Hong Liu, 1996). Lexical cohesion can be divided into two categories repetition and co-occurrence. In this study, taking into account the corpus search characteristics and limitations (not being able to recognize precisely the meaning associations between words and words as in manual search), we mainly discuss the aspects of repetition, i.e., repetition of the same word, synonyms, antonyms, etc.

After manually eliminating the keywords that have nothing to do with lexical cohesion (e.g., prepositions, pronouns, proper nouns, etc.), the list of keywords in the keywords is obtained, according to the criteria of the frequency of occurrence greater than or equal to 10 and the keyness greater than or equal to 1.

Then, with the help of the software Wordnet, this study established the lexical ensembles of the words in the list respectively, containing those words' repetitions, synonyms, antonyms. Take the word "ancient" as an example, the lexical ensemble of this word is: *Synonym (past, old, ancient); Antonym (present, future, young, immature).* What's more, the repetitions in the lexical ensembles contain different inflected forms of verbs. After putting the established vocabulary ensembles into AntConc to retrieve, it generates a table about the frequency of occurrence of the words therein (the table not shown in the text due to space constraints).

The retrieving results indicate that the total frequency of occurrence of all words is 621. The highest frequency in the vocabulary ensembles is repetition, with a total of 375 times, which is much higher than that of synonyms and antonyms; followed by synonyms with a total of 190 times, and antonyms with the lowest frequency of 128 times. That is to say, in terms of lexical cohesion, ChatGPT's Chinese-to-English version of *I and the Temple of Earth* uses more repetitive vocabulary to articulate the whole discourse and less use of synonym and antonym. The frequent appearance and use of repetition indicates that the vocabulary in the text is more monotonous, further revealing that the AI lacks a certain degree of flexibility in the use of lexical cohesion means.

From the perspective of their lexical properties, these words are mainly verbs, nouns and adjectives. Verb words are often repeated in the discourse in the form of synonyms and repetitions; nouns are often repeated in the form of synonyms; adjectives are often repeated in the form of synonyms and antonyms. That is to say, in order to articulate the discourse, this AI translations have a specific tendency to reproduce different lexical words in a specific way.

Then, the total category of the lexical ensemble is used as the classification criterion, and it is further analyzed from the three perspectives of repetition, synonym and antonym. In this study, the repetition of *saw*, the synonym of *desire*, and the antonym of *ancient* were selected, according to the criteria of the frequency of occurrences and the representativeness, and they were put into AntConc to retrieve the frequency of their left and right 1 word occurrences.

It can be observed from retrieving results that in the repetition ensemble of *saw*, *saw* itself occurs more frequently at 18 than the other words. This further indicates that the text uses the past tense more often than other tenses, and the text tends to talk about past events. The large number of repetitions of the same verb in the same tense form in the text constantly reminds the reader that the text describes an event that happened in the past. Its presence in the text as an identifier allows the reader to easily recognize whether or not the fragment being read is taking place in the present, in other words it makes the switch between the past and the present in the text more natural and easier to understand.

The retrieving results shows the frequency of the words appearing left and right of it shows that the participants associated with the verb *saw* are: *I, you, she, they, he,* and these participants are also the main participants in the text. The high frequency of the appearance of *saw* with these participants suggests that some of the characters' dialogues or character descriptions are interspersed in the text. Descriptions are interspersed throughout the text. These discourses are organically articulated throughout the discourse through the cooperation between *saw* and the participants as a kind of identification.

Desire and its synonyms were put into AntConc to retrieve the frequently occurring one word on the left and right of these words. The results show that the frequency of the word *still* on the left side of the *desire* and its synonyms is 4. Further searching within the text shows that *still* only collocates with *want* in the lexical composite. The frequency of *to* on the right side of the lexical items is 11, and it is most frequently collocated with "want" in the lexical items, and also collocated with "desire".

The word *ancient* and its anonyms were put into AntConc to retrieve the one word occurring on their left and right, and the results show that the frequency of the *word man* is 9, and an in-text search of the word in AntConc shows that the word man only collocates with *ancient* and its antonym *young*. The word *garden* occurs at a frequency of 4. An in-text search of this word found that it was only paired with *ancient* in the lexical collection.

The purpose of the search for synonym and antonym collocation vocabulary is to further observe the depth of the text's use of recurrent articulation devices. Through the above analysis, it can be seen that the ChatGPT translated text utilizes more recurrent articulation devices, especially the use of repetition and synonyms. However, when analyzing the left and right word collocation of synonyms and antonyms in the text, it can be seen that the antonyms and synonyms used in the text are only collocated with fixed words, and a large number of repetitive collocations appear in the text. That is to say, although the text uses the means of articulation of the same present, but the mode of application is too simple and boring, which can only satisfy the general standard of articulation of the content, and cannot satisfy the demand for the presentation of the beauty of the literary text.

(2) Reference

Grammatical cohesion refers to reference, substitution, ellipsis, and conjunction among the articulatory devices proposed by Halliday and Hasan, which are categorized in detail. Since corpus research mainly deals with data on a large scale, substitution and ellipsis are difficult to retrieve and process in corpus processors, so this study only

explores the two devices: reference and conjunction.

First of all, reference is the relationship between one word and another. Depending on the direction and location of the reference to the word, reference is categorized into external and internal reference. External reference means that "the identity presumed by the referent can be reproduced from the context of the text, and the referent links the text to its environment; but it does not contribute to the articulation of the text, except indirectly by repeating the same referent in a chain" (Halliday, 2004). As the definition explains, external reference refers to a reference to something outside the text itself, i.e. a reference to the environment, situational context or socio-cultural context, so finding external relations through the corpus is not possible because consistency in the corpus is a way of referencing that is limited in the text. Internal reference implies that the identity presumed by the reference item can be reproduced from the text itself or, more precisely, from the immediate system of meanings produced as the text unfolds. As the text unfolds, the speaker and the audience establish a system of meaning, and once a new meaning is introduced, it becomes part of that system, the correct category of things to be inferred by internal horizontal reference. There are actually two possibilities here. The internal reference can point "backwards" to the history of the unfolding text, i.e. to a reference that has already been introduced and is part of the "system of meaning of the text". This type of internal reference is called cataphoric reference, and the element to which the cataphoric reference points is called the antecedent. This kind of reference is very common. Internal reference can point to the future of the unfolding text, i.e., to a reference that has not yet been introduced, and this type of internal reference is called anaphoric reference (Hong Liu, 1996).

It is well known that whether it is anaphoric reference or cataphoric reference, they include both personal reference and demonstrative reference. Personal pronouns can be further divided into personal pronouns and personal possessive pronouns. Both anaphoric reference or cataphoric reference can find their respective objects of reference in the discourse, and thus can be detected by consistency in the corpus. Therefore, by bringing together the reference-related words in the Systemic Functional Grammar to form a vocabulary set, and then combining this vocabulary set with the person in the corpus text, the use of reference in the text can be observed by searching in the corpus. (Liu Jianpeng, 2020:162)

This study focuses on analyzing the text from the point of view of personal reference. The list of words and the list of keywords of the texts were observed through AntConc and an attempt was made to select words from them that could be used as indicators. It is observed that there are few specific personal nouns in the studied texts, and personal pronouns such as me, I, and she appear frequently in the word lists and keyword lists. If these personal pronouns in the text are taken as the indicators of personalization, then the retrieval of personalization in the corpus is not very meaningful. Therefore, in order to better fit the actual situation of the text, we directly put the personal reference words into the corpus for retrieval, observe the distribution and frequency of these words in the text, and present the results in the following chart and table.

Figure 1 Map of the distribution of restricted personal pronouns in the text

It can be seen that the word I has the highest frequency of occurrence (260) among the personal reference appearing in the text, accounting for 32% of the total frequency of occurrence (811). The personal pronouns me and mine corresponding to I occur 52 and 1 times respectively, accounting for about 6% and 0.1% of the total frequency; The word *it* has the second highest frequency (106), accounting for about 13% of the total occurrences; *you* has the third highest frequency (81), accounting for about 10% of the total occurrences; The frequency of *he* (76) ranks fourth, accounting for about 9% of the total frequency of occurrence, and the frequency of its counterparts, *him*

and *his*, are 13 and 36 respectively, accounting for about 2% and 4% of the total frequency of occurrence; *she* ranks fourth, accounting for about 9% of the total frequency of occurrence, and its counterparts, *him* and *his*, are 13 and 36 respectively, accounting for about 2% and 4% of the total frequency of occurrence. *she* ranks fifth in frequency (63), accounting for about 8% of the total occurrences, and its counterpart *her* occurs 54 times, accounting for about 7% of the total occurrences; The frequency of *they* is 34, accounting for about 4% of the total occurrences, and the frequency of its counterpart *them* is 10, accounting for about 1% of the total occurrences; *we* is 23, accounting for about 3% of the total occurrences, and its counterpart *us* is 3% of the total occurrences. The frequency of *us* is 2, accounting for 0.2% of the total occurrences.

From the data obtained from the above analysis, it can be seen that although the pronouns *it* and *you* have a high frequency of occurrence, they lack corresponding referents, so it can be assumed that they appear in the text as general personal words. However, *I, he* and *she* not only have a high frequency of occurrence, but also their corresponding referents appear more frequently in the text. Therefore, I, he, she and their corresponding referents are the most frequently used referent pronouns in the text.

As can be seen from Figure 1, personal reference words appear 811 times in the text, which is about 2% of the total number of words in the text 45802. Considering the characteristics of personal pronouns as meaningless grammatical vocabulary, the high frequency of occurrence of personal reference words in the text indicates that the means of personal care is fully utilized in the text. From the distribution map, the personal care words appear in the front, middle and back of the text, but the distribution is more uneven, and the distribution image is characterized by discontinuity. Through further search, it was found that a large number of keywords and high-frequency words appeared in the text paragraphs where the personal care words frequently appeared, and these keywords and high-frequency words were all personal pronouns. In the passages where personal pronouns appeared to be missing, the text lacked the referents of personal care pronouns, so there was no need to use personal care pronouns. That is to say, the frequent appearance of these personal care pronouns does not represent the characteristics of the Chinese-to-English text of ChatGPT, but is merely a natural presentation of the presentation required by the characteristics of the text's content.

(3) Conjunction

Halliday (2004:538) defines a conjunction as "a connected system of articulation that has evolved as a supplementary resource for creating and interpreting text" (Halliday 2004:542-544). Connectives are directly presented as words in the chapter and are more direct and obvious compared to their other connective means. Therefore, in the corpus, the use of the connective in the text can be observed by directly searching the connective you want to study in the document.

In this paper, we refer to the vocabulary ensemble of commonly used connectives summarized by Liu Jianpeng (2020: 177), and put the ensemble into AntConc for retrieval to observe the distribution and frequency of occurrence of these connectives in the text. After manually eliminating words with the same morphology but not belonging to the category of connectives, the search results are organized into the following figure and table.

Figure 2 The distribution of conjunctions in the text

Looking at Figure 1, we can see that there are a total of 398 conjunctions in the text (one word with the same form but not belonging to the category of conjunctions is manually excluded), accounting for less than one percent of the total vocabulary. The distribution of connectives in the whole text is relatively even, and they are detected in the front, middle and back of the text, with a slightly higher frequency in the middle of the text. Further analysis

shows that connectives, as grammatical words with no additional lexical meaning, account for a smaller proportion of the text than real words, so the result of less than one percent of connectives in the studied text is in line with the general discourse situation. The fact that connectives are evenly distributed throughout the text suggests that they are widely used in the text and that the text relies on them to some extent for discourse articulation. Tracking the middle part of the text in AntConc, where conjunctions appear more frequently, reveals that the middle part of the text has a large number of characters' dialogues compared to the front and back parts of the text. The text tends to use simple connectives such as and, but, etc. when articulating the dialogues between characters.

The retrieving results in corpus show that the frequency of the connective *and* is the highest, accounting for 59% of the total frequency of the connective, followed by *but, still,* and *then, then,* accounting for 17%, 7% and 6% of the total frequency of connectives respectively. The frequency of other connectives is less than 5%. In terms of the frequency and complexity of connectives in the text, the text tends to use simpler and more common connectives, such as *and*, *but* and so on. In terms of the classification of connectives, additional connectives, such as *and, but, however,* etc., appear most frequently in the text, as well as a small number of conditional connectives and explicative connectives. In general, the conjunctions appearing in the text are relatively simple and of a single category, which also indicates a single and simple pattern of conjunctions in the text.

5. Conclusion

This study, based on Systemic Functional Grammar, examines cohesion features in ChatGPT's Chinese-to-English translation of *I and the Temple of Earth* using corpus-based quantitative methods.

In terms of lexical cohesion, the translation frequently uses repetition, along with some synonyms and antonyms. Verbs often appear repeatedly as synonyms or exact repetitions, nouns mainly as synonyms, and adjectives as synonyms or antonyms. However, the overall use of lexical cohesion is somewhat monotonous and rigid. To further understand repetition, the study also looked at collocations of recurring words. These repeated words function as discourse markers, helping to connect ideas and signal tense shifts. Yet, the use of repetition remains simple and lacks literary depth.

Regarding reference cohesion, personal reference is commonly used to describe the main characters. However, this reflects the original text's needs rather than a distinctive strategy of the ChatGPT translation. In terms of conjunctions, the translation uses them consistently and appropriately to link ideas, following general discourse logic. But similar to lexical cohesion, the use of connectives is too uniform, lacking variety.

Overall, ChatGPT's translation makes effective use of basic cohesive devices—repetition, reference, and conjunctions—to ensure general coherence. Nonetheless, the translation still shows limited flexibility and superficial cohesion, which hinders its ability to convey the literary beauty of the original text.

References

[1] Halliday, M. A. K. (2004). An introduction to functional grammar (3rd ed.). London: Arnold.
[2] Qi, Y. (2014). Corpus-based functional discourse analysis: President Obama's 2013 inaugural speech. Journal of Language and Literature Studies, 10, 20–22.
[3] Gao, X. (2022). A study on complex noun phrases in English academic writing: Based on a comparable corpus. Corpus Linguistics in China, 9(2), 47–60. [in Chinese].
[4] Hong, L. (1996). Mechanisms of cohesion and coherence in English discourse. Shandong Foreign Language Teaching Journal, 1, 11–14. [in Chinese].
[5] Hu, Z. (1989). An introduction to systemic functional grammar. Changsha: Hunan Education Press. [in Chinese].