ISSN: 2688-8653

Study on Construction of a Bilingual Parallel Corpus of The Classic of Mountains and Seas

Yiming Wang *, Kerui Su and Yajun Pi

School of Languages and Literature, University of South China, Hengyang 421001, China.

* Corresponding Author

Abstract

The book, The Classic of Mountains and Seas is a comprehensive encyclopedia of ancient China, (Wang Hong, 2010) and the translation and its research of this book are also one of the hits in translation studies. This paper introduces the construction of a bilingual parallel corpus of Chinese-English translations based on the original text of The Classic of Mountains and Seas and the translations by Anne Birrell and by Wang Hong, mainly including the design of the corpus, the construction route of the corpus, the problems of and solutions to building the corpus. The construction of this corpus as a result would serve as a platform for the translation research of The Classic of Mountains and Seas, such as the study of translator's style.

Keywords

English translation of the Classic of Mountains and Seas; Construction of Corpus; Corpus Translation; Translation Research.

1. Introduction

In the preface of the Chinese-English bilingual edition of The Classic of Mountains and Seas in the series of library of Chinese Classics,[1] mentioned that this book is a content-rich encyclopedia of ancient China, because it records the countries, places, mountains, waterways, and mythological figures from ancient times, covering a wide range of topics and kinds in nature and in ancient China. Therefore, scholars at home and abroad have been researching on this book constantly, with its influence extending to various disciplines such as science and technology, geography, mineralogy, medicine, mythology, and religion. There are also various foreign language translations of The Classic of Mountains and Seas. However, when it comes to completeness among English versions, the main ones are the translations by American scholar [2] and that by Wang Hong and his team. Based on the original text of The Classic of Mountains and Seas and two target texts translated respectively by Birrell and Wang, this article explores how to design and construct a Chinese-English bilingual parallel corpus with translations of The Classic of Mountains and Seas. It also elaborately explains the purpose of building the corpus, the selection of corpus texts, methods of text collection, tools for corpus construction, the route of corpus construction, and the main problems of and solutions to building the corpus. The construction of this corpus provides an analytical platform for studying the translator’s style and other issues in translation studies between the two versions by Birrell and Wang, laying a solid foundation for further research.

2. The Research Status of Corpus-Based Translation Studies

[3]Introduced corpus into translation studies, a group of researchers in China, represented by Wang Kefei and Hu Kaibao, have made persevere efforts and gradually established a certain theoretical and disciplinary system. Corpus-based translation studies mainly focus on the following aspects:

Theoretical developments related to corpus. Researchers such as [4], [5]have summarized the development of corpus-based translation studies over the past decade or two decades (1999-2018); [6][7]and others have introduced an overview of translator style research based on corpus from 2002-2017; [8]introduced three types of translation corpus.

Construction of corpus. The construction of corpus is the foundation for conducting corpus-based translation studies and an essential component of it. [9]proposed the design and construction of bilingual corpus as early as 2004; [10]proposed a plan for the construction of interpreting corpus. [11][12] and others introduced the construction of Chinese-English bilingual parallel corpus for novels.

Corpus-based translator's style research. Where there are translations, there are translators, and the identity of translators leaves traces, or styles, in their translated works. Therefore, corpus-based translator's style research is a hot topic in corpus-based translation studies. [13] explored the translators' styles in the translation of "Dream of the Red Chamber"; [14] discussed the translators' styles in the three English translations of "Rickshaw Boy."

3. The Design of the Corpus

The design of a corpus is a comprehensive plan that mainly includes the purpose of the corpus construction, selection of texts for the corpus, methods of text collection, and the selection of tools for corpus construction.

The purpose of this corpus construction is to establish a Chinese-English bilingual parallel corpus for The Classic of Mountains and Seas, providing an excellent platform for the study of translator's style in the translation of this text. The quality and authority of the data entered into the corpus are essential, so the choice of texts is crucial. The selected Chinese source text of The Classic of Mountains and Seas and the first English translation come from the "Library of Chinese Classics" series published by Hunan People's Publishing House, which is the English translation by Wang Hong and others; the second English translation is "The Classic of Mountains and Seas" published by Penguin Books, with Anne Birrell as the translator. Both English translations are scholarly works of high quality and complete, offering significant value for corpus construction. The collection of texts is primarily based on the two printed copies of books mentioned above, by photographing and using photos to generate images, and then employing optical character or word recognition software to digitize the texts and save them as documents. Before the construction of the corpus, manual proofreading of those raw texts is necessary to ensure the accuracy of the digitalized texts. The main tools for constructing the corpus include a range of software and they are: EmEditor, mainly used for cleaning the Chinese language texts, such as, noise reduction; ICTCLAS, mainly used for segmentation and annotation of Chinese; CLAWS, used for annotating English materials; ParaConc, AntConc, and Wordsmith, used for aligning, searching, and statistical analysis.

4. The Construction Route of the Corpus

The construction route for building a Chinese-English bilingual parallel corpus of The Classic of Mountains and Seas can be divided into several modules, like, text cleaning, word segmentation and annotation, alignment, and retrieval. First, the collected raw materials undergo cleaning to remove typos, gibberish, impurities, redundant information, and EmEditor is used to remove noise information, such as extra spaces. Second, for the Chinese language texts, specialized word segmentation which identifies words out of individual Chinese characters, is required to complete the annotation. English materials do not require word segmentation and are directly annotated using tagging tools. Third, the processed texts are added into software for alignment. Since the current retrieval tools support sentence-level alignment, alignment is generally based on sentences. The original text of The Classic of Mountains and Seas is in Chinese, and during the translation process, the English translation may have been merged or split compared to the original text. To ensure the accuracy of alignment, manual proofreading is essential. Finally, the aligned texts can be used for retrieval and statistical work. Each module in the corpus construction process will encounter related problems, which are worth paying attention to. Analyzing these problems and proposing solutions are of certain reference value for the construction of bilingual corpus of Chinese classics. Next, this paper will discuss the main problems encountered during the corpus construction process and their respective solutions.

5. The Main Problems and Their Solutions

The main problems in corpus construction can be divided into two categories: data-processing problems and software problems. data-processing problems range from the original text to the two translations. Software ones mainly relate to text cleaning and annotation.

Data-Processing problems can be analyzed into three categories: the original text and the two translations. The data-processing problem of the original text of The Classic of Mountains and Seas are concentrated on the digital recognition of ancient Chinese characters. Since the original text is a classic, some ancient character text documents cannot be recognized. Moreover, document software cannot recognize images either, so simply using picturing-like software to piece together ancient characters, which seems obscure, although it achieves a visual effect to some extent, does not allow for text retrieval. Therefore, to quickly build the corpus and realize its functions such as retrieval and statistics to explore and compare the translators' styles of the two versions, this corpus construction uses a character substitution method. Obscure ancient characters are replaced with characters that can be generated in text documents. This corpus uses a substitution method for 160 ancient characters, covering 286 items, involving the following situations: first is the replacement of simplified with traditional characters. The source text uses simplified characters, while document software still retains traditional forms, such as replacing the obscure character "朱鸟" with "鴸", "鱼需" with "鱬", and the character composed of "此" above and "鱼" below with "鮆". Second is taking radicals. Since some ancient characters cannot be generated in text documents, their radicals are used instead, such as replacing "尚鸟" with "尚" and "付鸟" with "付". Third is the use of variant characters. Obscure characters in the original texts can be replaced by synonymous variant characters, such as replacing "月暴" with "皺". Through the substitution of obscure characters, text documents can correctly recognize them, thus smoothly proceeding with later cleaning, denoising, and other work.

Data-Processing problems related to the texts of Wang's version are involved in multiple modules of building the corpus, including spelling mistakes, homophones generated by transliteration, and difficulties in alignment caused by translation techniques.

Firstly, spelling mistakes. During the text cleaning process, there were 4 items of English spelling errors in Wang's version: the word "Rive" is a spelling mistake in “The Jianshui Rive originates from its northern slope and runs to the northwest before it empties itself into the Gushui River which is rich in kohl stone and dark cinnabar”; so is "shuch" in “It makes a sound like a crying baby and it may eat humans and reptiles shuch as snakes”. Such spelling mistakes should be rectified in the text cleaning process.

Besides, there were 30 items of homonyms generated by transliteration: in the two sentences below, “The Sishui River（泗水）flows out of the northeast of Lu and runs south and then southwest before it passes Huling in the west” , and “The Sishui River（肄水）flows out of the southwest of Linjin and runs southeast before it empties itself into the sea west of Panyu” , the rivers referred are different. The first one refers to "泗水" while the other refers to "肄水". Due to their same pronunciation, however, spellings of the two Chinese words after transliteration are the same, which causes certain obstacles to the retrieval of corpus.

Apart from this, problems happened in the module of alignment as well. Due to the adoption of some translation methods, there are some difficulties in the alignment of the Wang's version with the original text. (shown in Table 1) For example, there are 64 cases in Wang's translation where sentence splitting leads to misalignment. As can be seen from the following example sentences: “英水出焉，南流注于即翼之泽” was split into two sentences and required to be merged together to align with the original text. In other 18 instances in total, when a volume of the book shares the same title with a chapter, omitting is applied in Wang’s translation to wipe out one of the repeated titles (usually the title of a chapter) to avoid lexical overlapping. As shown in the table, the 15th volume of the book has the same title with the 15th chapter and Wang Hong chose to omit one of them. As a result, the chapter title has no translation to align with. Moreover, there are 10 cases of annotated translation or amplification in the translation. Relevant cases in the table appears at the end of the chapter, The Classic of Areas Within the Seas: the East, so there is no counterpart. However, if amplification was employed at the level of sentence without adding extra sentences, it won't affect alignment. According to the chapter, The Classic of Areas Overseas:the West, “服常树，其上有三头人，伺琅玕树” was translated as “On the top of fuchang, a legendary tree, there are people with three heads. They are guarding langgan, another legendary tree, nearby”, among which “a legendary tree” and “another legendary tree” were amplified. Since they were not single sentences, they won't affect alignment. Above alignment problems caused by translation techniques need manual proofread when building a corpus so that the texts can be aligned for better retrieval.

Table 1 examples of difficulties in the alignment of Wang's version

Translation Techniques	Original Texts	Translations
Sentence Detachment	英水出焉，南流注于即翼之泽。其中多赤鱬，......	There is also a river which flows out of this mountain and is called the Yingshui River.With its water flowing swiftly to the south,it finally empties itself into the Jiyi Lake where there are many red giant salamanders.
Omission	卷十五大荒南经	Volume Fifteen The Classic of The Great Wilderness:the South
	大荒南经	无译文对齐
Amplificatio n	无原文对齐	The above is collated by Ding Wang,Grand Master of Ceremonies and Expectant Appointee,and jointly edited by Wang Gong,...

Birrell's translation also has problems deserving discussion. First of all, there were omissions (6 cases) and errors (4 cases) in text cleaning process. For example, there is a sentence in The Classic of Regions Beyond the Seas: The South: “三首国在其东，其为人一身三首。一曰在凿齿东” was translated as “Threehead Country lies to its east. Its people have a single body but three heads”, among which “一曰在凿齿东” was omitted. As a result, the omitted one had no counterpart to align with. Meanwhile, “......是多白玉。其中多鮨鱼，其状如蛇而四足，是食鱼” in The Classic of the Western Mountains Chapter 3 was translated as “There are quantities of white jade here and in these waters there are numerous flashwing fish which look like a snake but have four feet” , among which “是食鱼” was omitted. However, since it doesn't form a single sentence, it has no impacts on alignment. In cases of The Classic of the Southern Mountains

Chapter 2, “又东四百里，曰洵山，其阳多金，其阴多玉” was translated as “Five hundred leagues further east is a mountain called Mount Weep”, among which “四百里” was translated as “Five hundred leagues” by mistake; in The Classic of the Southern Mountains Chapter 3, “又东五百里，曰发爽之山，无草木，多水，多白猿” was translated as “Four hundred leagues further east is a mountain called Mount Showlively”. Unlike the sentence before, “五百里” in original text was translated as “Four hundred leagues” mistakenly.

Besides, Birrell's version has 13 cases of amplification and 4 cases of annoted translation. Like in The Classic of Regions Within the Seas: The South, the sentence “枭阳国......，左手操管” was translated as “People hold a bamboo tube in their left hand [to trap them]”, among which “to trap them” is amplified and annotated in the text. Since this kind of amplification just amplified sentence constituents rather than complete sentences, it has little impacts on alignment. In cases of The Classic of Regions Within the Seas, “稷之孙曰叔均，是始作牛耕” was translated as “The divine Millet's grandson was Reap Even,and Reap Even was the first to invent the oxdrawn plough. Reap Even gave birth to Big Alike and Scarlet Shade”, in which the annotation displayed as a complete sentence. As a result, the annotation should be merged with previous sentence to meet alignment requirements.

Moreover, from the aspect of sentence, there are 114 cases of sentence combination which affect alignment. For example, in The Classic of the Western Mountains, “其上有兽焉，其状如牛，白身四角，其豪如披蓑......” was translated as “There is an animal on its summit which looks like an ox; it has a white body and four horns; its long hair is like a reed raincoat”. Obviously, the translation split sentences in original text into several clauses and link them with semicolons. However, semicolon is a symbol of segmenting sentences, so to align texts, clauses in the translation which are segmented must be emerged.

During the process of building a corpus, problems of applying software also worth considering. First of all, when collecting texts, there are some errors in TXT format when they are converted from PDF and Word documents; then when it comes to denoising, Emeditor wouldn't clean the text completely and the rest part like extra space characters, null strings and carriage returns must be deleted through manual proofreading. At the same time, Emeditor might have errors when identifying marks of alignment. And hence, manual proofreading is required to rectify them; moreover, in the process of word segmentation and annotation, ICTCLAS only recognize texts in ANSI form.

6. Conclusion

The construction of parallel corpus is an essential part of corpus-based translation studies and also one of the hits in this field. This paper discusses the construction of a Chinese-English parallel corpus of the book, the Classic of Mountains and Seas from several aspects, including corpus design, construction route, problems encountered during the process, and their solutions. By constructing the corpus in a standardized and effective manner, we provide a solid platform for further exploration into translation study issues such as the translator's style in the dual translations of the Classic of Mountains and Seas, and offer a modest reference for the construction of Chinese classic literature corpus.

Acknowledgements

Funded by Hunan Social Science Foundation program, “ A Corpus-Based Research on Translator’s Style in Translations of The Classic of Mountains and Seas”(20WLH11) .

References

[1] Wang, H. & Zhao Z. trans. 2010. The Classic of Mountains and Seas[M]. Changsha: Hunan People's Publishing House

[2] Birrell, A. trans. 1999. The Classic of Mountains and Seas[M]. London: Penguin Books

[3] Baker, M.1995. Corpora in Translation Studies: An Overview and Some Suggestions for Future Research [J]. Target, 7(2):223-243.

[4] Wang Kefei, Huang Libo. Corpus-based Translation Studies: Progress in Recent 15 Years[J]. Foreign Languages in China, 2008, (6)9-14.

[5] Li Yan, Feng Huali. The 20-Year Development of Corpus Translation Studies in China (1999-2018) [J]. Journal of Yanshan University(Philosophy and Social Science Edition) , 2020, 28 (01): 105-110.

[6] Li Keli, A Review of Corpus-Based Studies on Translator's Style in China (2007-2017) [J]. The Science Education Article Collects, 2018, (31): 178-179.

[7] Lv Qi, Wang Shuhuai. A Visualized Bibliometric Analysis of Corpus-assisted Translators Style Studies in China (2002-2016) [J]. Journal of Yanshan University(Philosophy and Social Science Edition), 2019, 20(01):42-49.

[8] Liao Qiyi. A Research on Corpus and Translation [J]. Foreign Language Teaching and Research, 2000 (9): 380-384.

[9] Wang Kefei. Design and Construction of a New Bilingual Correspondence Corpus [J]. Chinese Translators Journal, 2004,(6):75-77.

[10] Zhang Wei. Development and Construction of Interpretation Corpus: Some Issues in Theory and Practice [J]. Chinese Translators Journal, 2009, (3): 54-59.

[11] Huang Libo. Chinese-English Parallel Corpus of Contemporary Chinese Novels: Research and Application [J]. Foreign Language Education, 2013,(6):104-109.

[12] Liu Zequan, Yan Jimiao. Choice and Style: On the English Translation of the Reporting Verbs Headed by Dao in Hong Lou Meng [J]. Journal of PLA University of Foreign Languages, 2010, (4): 87-92, 128.

[13] Liu Zeqian, Liu Chaopeng, Zhu Hong. A study of the Translators' Styles in Four English Translations of "The Dream of the Red Mansions" - Based on Corpus Statistics and Analysis [J]. Chinese Translators Journal, 2011, 32(01): 60-64.

[14] Huang Libo. A Corpus-Based Study of Translators' Styles: The Translation of Discourse Presentation in Three English Translations of "The Rickshaw Boy" [J]. Journal of the PLA University of Foreign Languages, 2014, (1): 72-80, 99.