**Construction of Jiangxi Tourism Resource E-C/C-E Parallel Corpus**

Liyong Zhu\(^1\)*, Zhijuan Zhu

**Abstract**

This paper discusses the construction of Jiangxi Tourism Resource E-C Parallel Corpus on the basis of description of the necessity in its construction. It first describes its framework, then its construction. In construction, it includes data collection, the standard for the sorted data, data selection, data digitalization, data tagging and data aligning. With the introduction of its construction, it realizes its purposes and functions of the corpus, but also provides others with ways or means to use the corpus and to establish such kind of corpus.

**Keywords:** Jiangxi tourism resource; parallel corpus; E-C; construction and application

I. Introduction

Jiangxi is rich in tourism resources, for its natural generosity as well as its acquirement from Chinese history. Its beautiful natural mountains such as Mount Lu, Mount Sanqing, etc., which provide its fame in tourism resources. The red tourism resources also greatly enhance its beauty. For it has contributed a lot during the revolution against Japanese army and the Kuo Min Tang Army to the foundation of new China. In recent years, these areas that the red army had battled were developed to be the red tourism areas to make more Chinese people remember the past and cherish the present. Among the visitors there are foreigners who do not know much about China and Chinese. When they are there, they may not quite understand why those areas are so attractive to Chinese. Thus it is quite necessary to give the information in English to make foreign visitors understand the situation and the scenery and corpus is a way to realize that. Furthermore, the historical buildings and relics also increase its tourism resources.

Crystal (1991) tends to take corpus as a collection of linguistic-data, either written or spoken speech, that can be used as a means or starting-point of linguistic description about a language. Sinclair explains it as a collection of naturally-occurring language text, chosen to characterize a state or variety of a language. Kennedy defines it as a body of written text or transcribed speech which serve as a basis for linguistic analysis and description. Feng Zhiwei, a Chinese professor, believes it is a collection of certain scale language data with structural property which is representative and can be retrieved. According to the above four definitions, we can conclude that a corpus is a collection of chosen language materials that are stored in computer or electronic database and can be retrieved and analyzed for different purposes of language description or research. If the corpus is with language samples of a particular area or

\(^1\)Jiangxi Science & Technology Normal University, Nanchang, 330013, China

*Corresponding author email: johnzly@163.com
field, it is the special corpus. And if each of the text in a corpus is translated into one or more than the original and each of them has its paralleling text, the corpus is a parallel corpus. This research will construct a bilingual parallel corpus with Jiangxi Red Tourism related resources both with Chinese and English version.

A corpus can offer large database of naturally-occurring and authentic discourse, so researchers, language teachers and learners can analyze the language on the basis of authentic structures and patterns rather than intuitions and perceptions. Thus the construction of Jingxi Tourism Recourse bilingual parallel Corpus will be necessary and useful which will offer natural and authentic language data both in English and Chinese to language researchers, teachers, students, tourist guide, foreign travelers, translators and dictionary compilers. It can also provide researchers with natural and standard translation to develop intelligent bilingual tourist guide.

II. The general design of Jiangxi tourism resource E-C/C-E parallel corpus system

2.1 The principles

This corpus is designed with the principles of high practicability, easy management and maintenance, expandability, and security.

For high practicability, the design of the corpus takes consideration of the requirement of managers and users. It has analyzed the idea of the other design, tried to construct a high usability corpus with the perfect or mature function.

For easy management and maintenance, the managers can log in the system with browser to manage and maintain the corpus to guarantee its efficient performance.

For expandability, it means the corpus is not a closed one, rather, it is open to more related bilingual texts to be added and more functions to be developed in future. That is to say, its storage can be enlarged and more functions could be developed, the corpus has reserved space for these two aspects.

For security, the design will try to realize the distribution of access rights in order to ensure the security access to the system.

This corpus will have two main functions which are the management of the corpus and the concordance of the corpus. The management is realized by the insertion, delete, and the renew functions of the corpus. Software and manpower are combining used to segment text into sentences and store in the corpus. For the concordance, it uses phrases and sentences as the key words to realize the classification retrieval and the general retrieval. The corpus is designed mainly based on these two functions. The corpus will be designed for three purposes for its concordance. These purposes are research, teaching and learning as well as practicability.

2.2 Framework of the corpus

Normal users can demand and search language data in the corpus. The managers mainly administrate classification, language data, users, system configuration, query searches, etc.

Classification refers to the first class and second class classification for the language sources, which include the functions of insertion, deletion and edition, etc.
Administration of users includes deleting, editing, examining and verifying users, checking the number of corpus users.

Language data administration means that the managers can revise, add and delete the data in the corpus. When adding new data, the system can take advantage of the concordance function of the regular expression to segment text into E-C/C-E parallel sentences, and store them into the corpus.

Management of the system configuration offers the functions of check and revision of the manager manual, revising users’ help, and so on.

For concordance, the system offers not only the full corpus search, but also the classification search. It also includes the exact search and fuzzy search. The keywords are the English or Chinese character string, between each string, there is a blank space. The system is added with automatic distinguishing and phrase filtering mechanism.

III. The construction of the bilingual parallel corpus

After the system design of the bilingual parallel corpus, it is the time to construct it. The construction mainly includes, corpus design, data collection, standards for the chosen data, selecting data, data digitalization, data tagging and paralleling and concordance. It can be described in the following chart.
3.1 Design of the corpus

We analyzed the features of Jiangxi tourism resources and divided them into three areas. They are the natural tourism resources where famous natural beauty exists; The red tourism resources significant events, meetings or battles took place, the former offices of the important institutions, the former residences or memorials of the great figures of the time, cemeteries of revolutionary martyrs, memorials at all level and the historical buildings and relics. They could be shown in the following chart.

Figure 1. Construction of the bilingual parallel corpus.
When the subjects and areas of the corpus have been decided, the purpose of the corpus should be designed, for any corpus should be constructed according to its functions and purposes, if the corpus is of no use, it is not necessary to construct it. Jiangxi Tourism Resource E-C/C-E Bilingual Parallel Corpus is constructed for three purposes. Firstly, it is constructed for research. Secondly, it is used for teaching and learning. Thirdly, it is for practicability.

For research, the corpus will be used to study the features of language in tourism and its translation. For the features of language in tourism, the corpus can be used to study features of text, rhetorical devices, and lexicology in tourism and acquire the language difference in tourism and general language. For the research of translation, it can be used to study the difference in tourism translation and general translation. The questions may be whether there are common features in these two kinds of translation or whether there are any features of its own in translation and whether the features are based on the translation norms during certain time or beyond the limitation of time and space.

For teaching and learning purpose, this corpus could offer authentic and live teaching materials in teaching Chinese-English tourism translation for the teachers. It can also provide
teachers with real and diversified materials in data-driven learning (DDL) teaching. Data-driven learning means that making use of computer concordance help students find out the principles in target language and develop various learning and practices. For this corpus, we decide to develop various DDL teaching methods in tourism translation teaching with the help of the concordance provided by corpus. Students could acquire the translation skills and strategies with the help of the corpus and they could get them by themselves actively, not taught by teachers and accepted passively, thus they could learn better in tourism translation and improve their initiative in learning.

For practicability, this corpus provides tourism practitioners in Jiangxi with convenient reference translation or materials for self teaching and for improving their proficiency. It can also give a comprehensive description to the whole Jiangxi red tourism resources, which does favor to the foreigners and the guide. It also provides companies with language data to develop intelligent bilingual guide to spread red culture, improve the image of Jiangxi and China and save labor of the guide.

3.2 Data collection

After design of the general framework, the research group members went to different tourism spots to carry out the filed research and took photos for all those listed areas to collect language data. The bilingual introduction materials about the red tourism are bought, such as brochures of scenery spots and travels, maps and other printed materials or books. Other research group members carry out the online material collection including the introduction textual materials on the websites of the National Tourism Agency, Jingxi Provincial Travel and Tourism Administration, Nanchang Travel and Tourism Administration, other famous travel agencies, etc.

3.3 Standard of the chosen data and data selection

When language data had been collected and sorted, it’s time to make standard to decide what kind of data could be stored in the corpus. Three standards are made for the language data as the following. Firstly, no mistakes in the material whether they are in Chinese version or English version. Secondly, the unnatural translated text could not be selected and stored in the corpus. Thirdly, the Chinese data and English should be checked, reviewed and approved by experts and guides in tourism to ensure there is no single mistake in corpus. These standards are made and followed to ensure the quality of the language data and corpus.

With the standard, the workgroup members read all the language materials collected from various places and acquired with various ways. After reading and discussion among the members, language data that may be stored are selected. These language materials will be read and checked again by experts and guides again to ensure that the standards are strictly followed. With all these steps and measures, the data that could be stored in corpus are selected or chosen.

3.4 Data digitalization

When language data have been chosen, the next step is to digitalize them and transfer
them into word format. During the transformation, we use scanner to scan the printed text PDF format and then use OCR to change PDF into Word format. After that we read and compare these two formats to check and revise if there are any mistakes. We must ensure there is no mistake in the text. For the pictures or photos, we type the words by hand. For the e-text, it is easy to change the form and check whether there are any mistakes.

3.5 Data tagging

For data tagging, we use human-computer interaction to tag the data. That is to say, computer is used to classify and segment the text first, and then human finally decide whether the tagging is correct or not. CLAWS is used to tag the English text. It can classify and tag text, which is called part of speech tagging, also POS. For Chinese text, we use ICICLAS which is developed by Chinese Academy of Sciences. It can determine the classification of Chinese text and tag the words according to requirement. For the purpose of the corpus, we will also tag the rhetorical devices in language data. There is no such code in CLAWS software. It has to be decided and tagged by human beings. We use RHE to represent the whole rhetorical devices. For each rhetorical device, we use RHE and the first letter or word of the rhetorical device to represent the rhetorical device. For example, for metaphor, we use RHEM to tag it. In this way, the corpus can be used to compare the difference of Chinese and English tourism text to realize its function of research and also do favor to practitioners including teachers, learners and translators and guide.

3.6 Corpus paralleling

Corpus paralleling means the source texts and its translated texts are stored in equivalent relationship. There are four kinds of paralleling which are chapter paralleling, paragraph paralleling, sentence paralleling and word paralleling. English and Chinese belong to different language system, so it is hard to realize word paralleling. For the purposes of the corpus, this corpus only parallel chapters, paragraphs and sentences. In Chinese-English translation, there exist the phenomena of omission, expansion and selected translation, so sentence paralleling could not be fully realized. During paralleling, we use human-computer interactive ways. The software used is Paraconc which is used firstly. Paraconc cannot reach 100% right paralleling, so after that we have to check and revise the alignment by handwork to ensure there is no mistake in paralleling. After the step of paralleling, the users can do the work of concordance. That is to say, users can search corpus to get information they want.

IV. Conclusion

After the above steps, the construction of the corpus has been finished, but it is not the end of the construction. It could be added more language data and be expanded continuously. The more important is to apply the corpus. Besides the functions and purposes discussed in the above, it may be applied more widely. The following work for the group members is to study the application and acquire the effects of its application and research how to make full use of the corpus and achieve efficiency.
Acknowledgment
This work is supported by Jiangxi Social Science Planning Funding (No:15WX308).

References