Self-compiled On-line Parallel Corpus in Translation Teaching

Yushan Zhao
School of Foreign Languages, North China Electric Power University
Beijing, China
zhaoyushan1963@163.com

Juan Shi
School of Foreign Languages, Inner Mongolia University of Technology
Hohhot, China
shimaggie128@163.com

***Abstract*—Corpus has great potentials as a kind of advanced teaching. In China, corpus-aided translation teaching bears less fruit and there are few empirical studies of Chinese-English translation based on corpus. This paper tries to introduce the significance of self-compiled on-line corpus, the data processing, the structure of the self-compiled corpus system and its content and size to enable student translators to reduce the word stuffing, choose appropriate equivalents and appropriate collocations in the target language, and help them reduce the time spending on translating.**

Keywords—parallel corpus; translation; teaching

I. INTRODUCTION

Corpus is very popular in language teaching, but it is seldom used in translation teaching, because of many advantages, people apply it into the research of literature and translation studies as the fast development of corpus linguistics, concordance tools and software also flourish and are constantly being improved. However, in the area of translation teaching, corpus has not shown its values and advantages completely. Teachers pay little attention to its use in the classroom, especially in translation teaching. Therefore, this paper intends to construct a self-compiled online corpus to introduce a new approach to teaching by which learners enable to grip the different meanings of the same word in different contexts and gain the proper equivalent for translation and, at the same time, to improve teaching methods and teaching quality.

II. CORPORA AND TRANSLATION TEACHING

A. Corpus

What is corpus? Actually, a corpus can be defined in terms of both its form and its purpose. According to Crystal, a corpus is a collection of linguistic data, either written texts or a transcription of recorded speech, which can be used as a starting-point of linguistic description or as a means or verifying hypotheses about a language[1]. Sinclair also expounds it as a collection of naturally occurring language text, chosen to characterize a state or variety of a language[2].

On the whole, linguists have always used the word corpus to describe a collection of naturally occurring examples of language, consisting of anything from a few sentences to a set of written texts or tape recordings, which have been collected for linguistic study. More recently, the word has been reserved for collections of texts that are stored and accessed electronically. It is a new technology in the study of language, which depends greatly on the use of computer. It is a body of natural language material stored in computer-readable form. Programs can be written to manipulate the language material in various ways. It is a powerful resource for linguistic research.

Corpora are classified on the basis of varied criterions, because corpora can differ in both structure and content according to the purpose for which they were compiled.

According to the different compiling purposes, there are general corpora and specialized corpora. General corpora refer to the corpora which consist of a body of texts which linguists analyze to seek answers to particular questions about particular language. Specialized corpora are corpora with the language samples of a particular field or area.

According to the involved languages, corpora are divided into monolingual corpora and multilingual corpora. The former refer to those corpora composed of only one language while the latter refer to the corpora containing more than one language. Parallel corpora are those including parallel texts of two languages. Multilingual corpora, especially parallel corpora, are valuable resources in translation research and teaching.

B. Translation Teaching

Translation teaching is the cultivation of professional translators or interpreters. It gives translation lessons under the guidance of translation theories. It helps improve students' translation skills and strategies. It is

the higher stage of English teaching, which is closely linked with professional translation.

Translation teaching studies refer to the study on both teaching theory and practice. The former is generalized from the latter and can be the guidance for the latter. From the last decade of the twentieth century on, more and more teachers began to notice the significance of translation theory in translation teaching. Therefore, they advocated to add some courses about the history of translation development and to introduce some translation theories and basic translation skills or approaches to students. Some scholars think that basic knowledge of translation theory should be part of the course, such as definition, history, criterion and process of translation, translatability, different types of meaning, the consistency of form and content, context and culture. The past ten years witnessed the bloom of translation teaching in China. Many teachers have carried out research on translation teaching, translation teaching methodology and so on. This paper, from the practical perspective, discusses constructing self-compiled online parallel corpus in order to improve the translation teaching methods and help students improve the quality of translation version.

III. Self-compiled On-line Parallel Corpus

The idea of using small corpus in classroom ESL has been strongly supported by a number of Western researchers and applied linguistists, notably Stevens, Johns and Tribble over past twenty years. In China, however, small corpus is relatively new and study on small corpus construction over the past few years is regrettably insufficient in amount and meager in description, few people have actually set out to self-compiled small corpora and the intricate skills of using small corpus in EFL teaching were not applied and tested in a typical Chinese classroom setting. To do some complementary job, this paper endeavors to discuss on construction of small corpus and tries to describe in detail the self-compiled on-line corpus.

A. Significance of Self-compiled On-line Corpus

The self-compiled corpus for translation can meet the requirement in the following points: With the clearly defined objective, the author can collect pertinently the data which are fit for the students' competence and interests; The open end database ensure to enlarge the size by updating the data when necessary, so the corpus could be modern and practical.

B. The Data Processing

The construction of the self-compiled parallel corpus mainly contains two parts: processing data and developing a concordance web page. This section will focus on the processing data.

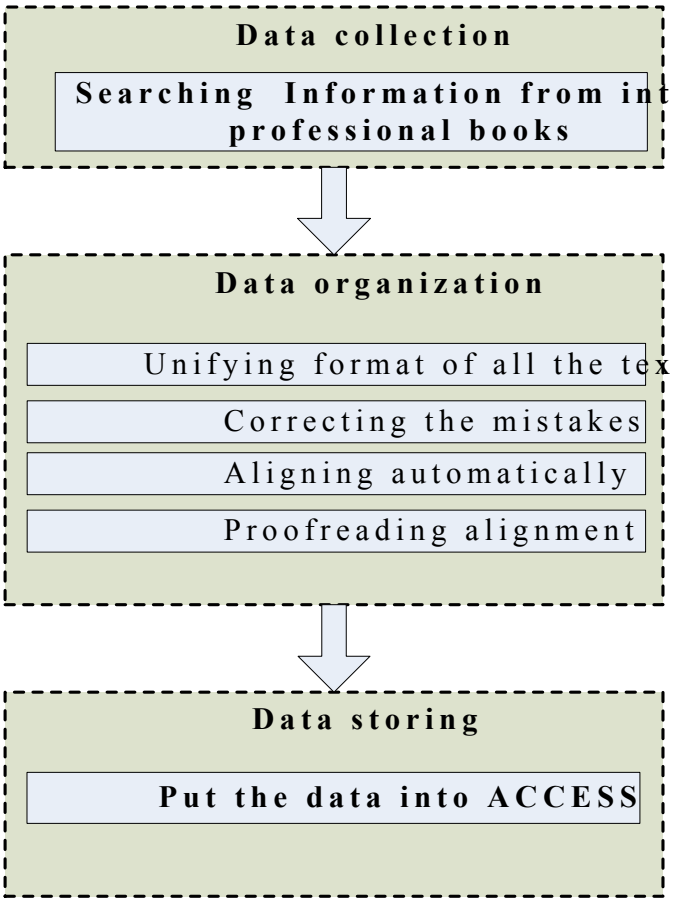

Fig.1 The Flow of Processing Data

The quantity and the quality of the corpus are determined by size and content of its data. Therefore the data collection is the first step of the corpus construction. The current self-compiled corpus will be used as a reference tool in teaching. The author collects articles about the politics and official documents. To ensure the quality of the corpus, the author chooses the materials from the official webs and reference books for CATTI2, CATTI3 and TEM8. The translations from these sources are completed or selected to publish by professional translators.

At first the format of all the texts should be unified and this step will convenient the job afterwards. Secondly, it is necessary to correct mistakes in the texts and make a quality as higher as possible. The third step is alignment. The precondition of this part is the articles and their translations are formal and the translations are relatively literally not too freely. Take the Chinese article as example. In the Microsoft Word, the author uses the wildcard ^p to replace the full stop as the tag of each sentences, and copy these sentences into Line A in Excel. Each sentence will occupy one block; and then using the same method to copy the translation sentences into Line

B. Next step—matching the sentence with its English translation—is a hard job because it must be finished one by one by hand. The last step is to store all the parallel sentences into the database; in the current corpus, we use ACCESS as database. By now, the procedure of data processing is finished.

C. The Structure of the Self-compiled Corpus System

Concordance is a main function of this corpus. To perform a concordance is to search and find all occurrences or instances of a particular word or phrase in a corpus and display them, usually in a certain format. A concordancer is certain software or computer program, often as a user interface or webpage available on line, designed to receive input from users and carry out concordances in the corpus. A concordancer is one of the most important tools in corpus-based research and usually is easily available. There are already many concordance tools available in the field, such as Wordsmith, ParaConc etc. But regrettably, these tools are not suitable for the research. Firstly, WordSmith is not free software the researcher cannot afford the cost to buy the registration codes for all subjects. Secondly, as in the case of the corpus in the experiment, which is designed to be accessed through the Internet, a user friendly concordance interface is of great importance to the project.

The current tools are not suitable considering the purpose of this study. Facing such difficulties, with ASP (Active Server Pages) Programming language and database software ACCESS, the author designs and develops a concordance program which can perform KWIC.

Since many teachers and researchers have made their corpus and receive good results, it is believed that a DIY corpus will meet the requirements of this study better.

Fig. 2 shows the structure of the program.

There are three layers in the corpus system: data layer, business layer and application layer, as shown in Fig. 3-2. In the data layer, the author uses ACCESS and IIS together as the foundation of the corpus. ACCESS as a database contains all the data- parallel sentences in this corpus. This software is user-friendly and easy to use. The state of the data in the system needs to be consistent across the lifetime of the system. Also, it comes as a component in the Microsoft office software. This enables the current study to lower the cost for the compilation of this corpus. IIS is the server and with it the corpus can be searched in the internet.

The business layer contains the code about recognition of Chinese and English, data binding, keywords highlight, data query and workflow. In this layer, the first step is to accept the data query from the interface. Then, after recognizing the Chinese and English, the program will decide the searching part in the database and bind the data. In the data binding, the

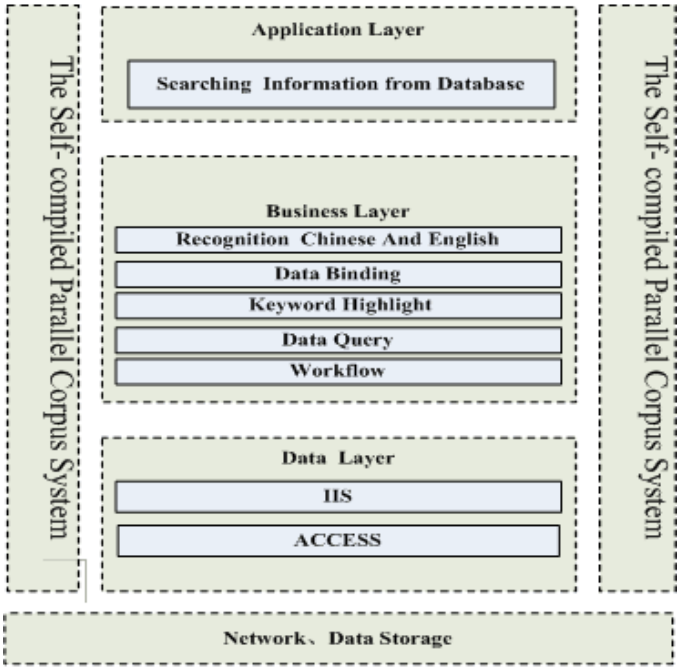

Fig. 2 The Structure of the Program

keywords which are queried from the interface would be highlight. When the program finishes all these steps, it will send these messages by workflow.

The application layer is the web page that we usually use. For the self-compiled corpus, we use ASP take text scripts in an HTML context and run them on the WEB server to create dynamic and interactive pages. A web-based interface is commonly provided to access the rest of the system—here refers to the parallel sentences in the database. In this interface, the users can entry the English words or Chinese phases in the searching block and then get the correspondent parallel translation—display window. The keyword will be highlighted in red to give the emphasis. Meanwhile the system will show a total number for clues which fit for the keywords. The following screenshot is the display of the concordance lines generated by the self-compiled parallel corpus about Chinese phrase “reunification”. Students can use this corpus concordance program to investigate the information contained within the corpus.

D. Content and Size

This self-compiled corpus is a small-sized one. To this issue, the “W3-Corpora”experts of Essex University explained in *World Wide Web Access to Corpus: Corpus Linguistics* [3], by now there is no definition of the size of corpus and no standard of the data amount to a certain research. It is important to get “enough” data, but how many is “enough” should judge all horses as individuals. Also, Carter and McCarthy argue that for the purposes of studying grammar in spoken language a relatively small corpus is sufficient[4]. Therefore, to certain specific

language research, small corpus would be better than the big one. And more important, big corpora are usually used for the authorized teaching plan and the textbooks compilation but not for teaching, especially for translation teaching therefore these big corpora cannot meet specific demands in the translation classroom.

A specific situation about the content of this self-compiled corpus are given: two white papers (*The Taiwan Question and Reunification of China,* and *The One-China Principle and the Taiwan Question*), three government leaders' (Deng Xiaoping, Jiang Zemin and Hu Jintao) important speeches on Taiwan question, *China's National Defense* (excerpt) from 2000 to 2008, and 56 articles about current affairs which are selected from reference books for CATTI2 (China Aptitude Test for Translators and Interpreters), CATTI3 and TEM8, and also 21 articles concerning other topics such as landscape and weather. It is necessary to mention that the China's National Defenses are huge documents, with the limited time the author just excerpt the parts related to Taiwan question from them and do the alignment. This self-compiled corpus contains 73,124 English words and the 102,515 Chinese characters, and it covers the 90% essential phases in the translation material.

Self-compiled online parallel corpus constructed can be used in translation teaching and learning. It offers a specific method to improve on the instruction manner of conventional teaching and conduct a constructivist translation teaching class. Because students are exposed to a large-scale of text and they learn through data, the results of present study suggest that, as the powerful tool for translation, corpus enables students to reduce the word-stuffing, choose the suitable equivalent and collocations and shorten the time in the translation.

ACKNOWLEDGMENTS

This paper are financially supported by North China Electric Power University--Teaching Reform Project: E-C Translation; Application of the Mini-corpus in the English Summary Writing of Academic Paper.

REFERENCES

[1] D.Crystal,Stylistic Profiling. In English Corpus Linguistics: Studies in Honour of Jan Swartvik. London: Longman. 1991, p.23.

[2] J. Sinclair, Corpus, Concordance, Collection. Shanghai: Shanghai Foreign Language Education Press. 1999, p.11.

[3] Hongzhan, Zhao, Self-compiled Small Translation Corpus. Beijing: The Chinese Translation of Science and Technology. vol. 5, 2007.

[4] R. Carter, & M. McCarthy, Grammar and the Spoken Language. Applied Linguistics. 1995.