A Study on the Construction of the Small TCM English Corpus

WU Yanxia [1, a], Song Weicai [2,b]

[1] Institute of humanities, Jiangxi University of Traditional Chinese Medicine, Nan Chang, 330004, China

[2] Institute of Computer Science, Jiangxi University of Traditional Chinese Medicine, Nan Chang, 330004, China

[a]email: wuyanxia2000@sina.com

Corresponding Author: Song Weicai

Keywords: TCM English; Corpus

Abstract. TCM (Traditional Chinese Medicine) English is a special expression system generated by English language in the process of translation and exchange of TCM. The corpus now generally refers to the combination of a large number of language material and positioning retrieval software stored in a computer. At present, there is no ready-made TCM English Corpus. This study will suggest an operable way and explore the feasibility of building TCM English Corpus from the corpus collection, corpus annotation, corpus search and other aspects.

1. Introduction

TCM is our country's rare original culture and has a strong core competitiveness of the productive forces. TCM English is a special expression system generated by English language in the process of translation and exchange of TCM, a new member of ESP (English for Specific purposes) family (Li Zhaoguo, 1999). According to this definition, the main body of TCM description should be the theory and practice of TCM, and the language of description is English. The corpus now generally refers to the combination of a large number of language material and positioning retrieval software stored in a computer (Xie Jiacheng, 2003). Corpus function is very powerful, and the most prominent feature is the ability to quickly and accurately provide one or more keywords related to the bulk of the real corpus, thus revealing the nature of language and the use of law. However, in this era in which all fields of linguistics actively develop corpus, TCM corpus which shoulders the mission of promoting TCM to the world is not available. To establish a TCM corpus, it is necessary to define it according to the purpose of the building. According to different purposes, corpus can be divided into: general corpus, special corpus, monitoring corpus, colloquial corpus, student English corpus and parallel corpus (Yang Huizhong, 2002). TCM English, this definite theme and content determines that the TCM corpus is a special corpus that should contain written corpus in all fields related to TCM theory and practice as much as possible so as to establish a corpus that satisfies the research requirements.

At present, many TCM institutions and scientific research units in China have begun to study the construction of TCM corpus, and put forward a lot of reasonable opinions, ideas and suggestions. This paper searched the full-text database of the Chinese journal and obtained 80 articles containing "Chinese medicine", "English" and "corpus". After screening again, there were 27 articles related to TCM corpus from 1987 to 2016 Literature, the specific distribution is shown in Table 1.

Table 1 TCM English corpus articles

Years	2003	2004	2005	2006	2007	2009	2010	2012	2013	2014	2016
Number of papers	2	1	2	2	2	1	1	3	3	5	5

Most of the scholars above have put forward the theoretical principles from the macro, and at present there is no ready-made TCM English Corpus. Based on the research above, and after understanding the needs of teachers teaching and student learning by means of lectures, research, questionnaires and other means, this study will suggest an operable way and explore the feasibility of building TCM English Corpus from the corpus collection, corpus annotation, corpus search and other aspects.

2. Overall Framework

In order to make TCM English Corpus in the future construction be balanced, this study put forward the overall framework of TCM English Corpus construction based on the basic principles of corpus linguistics and combined with the author's long-term study.

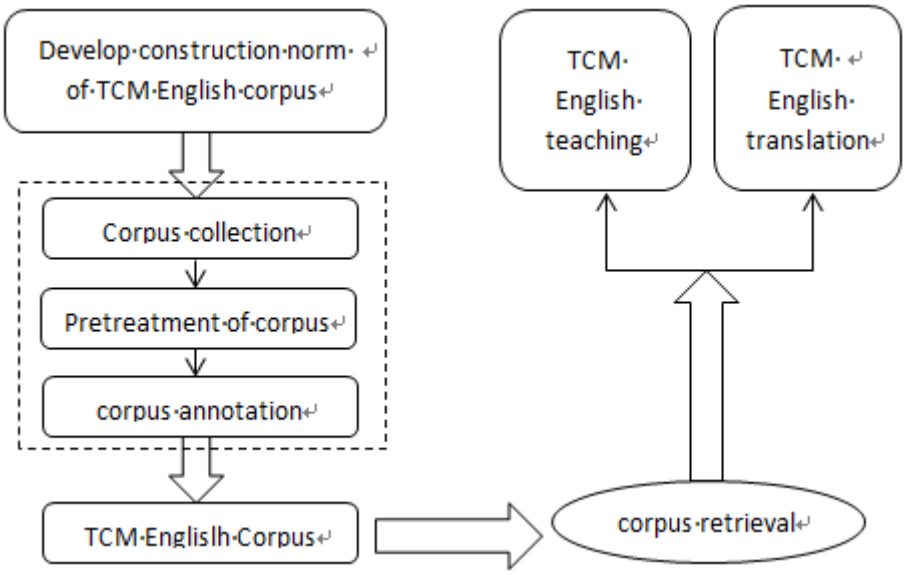

Fig. 1 The overall framework of TCM corpus

The initial design of the Chinese English corpus consists of eight sub-libraries: Yin-yang and Wuxing corpus, Zangxiang corpus, Qi,blood and body fluid corpus, Meridians and collaterals corpus, Causes of disease corpus, Pathogenesis corpus, Four diagnostic methods corpus and syndrome differentiation of eight principles corpus.

3. Research Ideas

3.1 Basic Ideas This paper studies three aspects of corpus building: corpus selection, corpus annotation and corpus retrieval.

2.1.1The Corpus Selection TCM corpus has the basic characteristics of special corpus, so we should pay attention to the representation and balance of corpus when collecting corpus. In the process of building TCM corpus, we should fully consider whether the collected corpus can truly represent TCM English. TCM English is a standardized English of descripting TCM theory and practice, the language material should include the theory and practice in all areas of TCM. Another problem to consider in establishing a corpus is equilibrium (Sinclair, 1991). At the beginning of constructing TCM corpus, we must first establish the scope of collecting language material, as well as the proportion of different language material. The collection of text should focus on TCM English teaching materials, Chinese translation of TCM English books and English version of the main TCM English papers; text selection criteria must be TCM English teaching materials and translation of the domestic and foreign authoritative publishing institutions, and the paper must be collected from the international popular first-class journals. The language material of TCM corpus is collected according to the discipline, the type of disease and so on.

3.1.2Corpus Annotation For the selected language material, we should mark the most basic background information, such as corpus of the author, translator, time, title, word, source, classification, style, style and other parameters. For corpus, markings and collations must be made for easy retrieval. We can use the standard method of Chinese TCM corpus, such as the labeling method ("Liu Yao, Zhou Yang, 2004") for a certain type of patient's condition, such as using "disease name identification + subject + specialist + disease name serial number + suffix". Although the raw language material can be multilingual, the annotated language material is easier to retrieve from many aspects.

The former is the premise and the foundation, and essential in the corpus construction process.

Taking the current world-wide labeling technology into account, TCM English corpus at this stage only label the part of speech. TCM English is a relatively standardized language, so in terms of the part of speech markings we can borrow other corpus labeling and retrieval tools to mark the TCM English corpus. The annotated TCM English corpus can provide high-frequency vocabulary, vocabulary collocation, word block, syntax, pragmatics, discourse and other aspects of the original data to achieve a more detailed description and more accurate understanding of TCM English.

3.1.3Corpus Retrieval In the analysis of TCM English corpus, we can use WordSmith Tools 5.0. WordSmith Tools 5.0 has three main functions: search, vocabulary and keywords. The primary role of the search is to query and count the frequency of a particular word or phrase in a particular text. TCM English has many distinctive features in terms of vocabulary, such as the specialization of ordinary English vocabulary, borrowing western terminology, generating new words by imitation, generating new vocabulary by word formation, borrowing Chinese language and so on (Li Zhaoguo, 1999). Through the search terms, you can provide a standardized study of the vocabulary characteristics of TCM and its translation. Using the search function, you can find the most commonly used expression method in the multiple meanings of the same meaning, and provide the de facto basis for the standardization of TCM English terminology. The vocabulary function reflects the frequency of vocabulary use in the creation of corpus and the overall characteristics of all texts. The lexicon function lists not only words but also lexical chunks. Through the statistics of high-frequency vocabulary and core vocabulary appearing in TCM English corpus, we can provide powerful material for compiling Chinese textbooks and dictionaries. The key word function is an important means to study the difference between the text content and the text language. The subject word refers to the vocabulary that the frequency is significantly higher than or less than the corresponding word frequency in the reference corpus. Through the keyword function, we can compare the word frequency in the corpus with the word frequency in the reference corpus to determine the difference between the corpus and the reference corpus, and provide good data for the study of the textual characteristics of the corpus.

4. Summary

A conclusion is gotten on the comparison above: small TCM English corpus can provide a more objective and comprehensive characteristics and internal laws of TCM English. Through a large number of examples, To study the language rules system in this field, make a qualitative analysis based on quantitative analysis and reveal the characteristics of Chinese language can provide a large number of language material which is objective, real and easy to retrieve to improve the quality of TCM English teaching and the standardization of Chinese translation of Chinese medicine English.

References

[1] Ji Zhe, Shi Yunzhong. A Comparative Study of Nigel Wiseman and Li Zhaoguo 's Academic Thoughts [J]. Study Journal of Traditional Chinese Medicine,2006,(1).

[2] Li Zhaoguo, Zhu Zhongbao. TCM English. Shanghai: Shanghai Science and Technology Press,2002.2.

[3] Ni Chuanbing. The Construction Principles of TCM Englislh Corpus [J]. Journal of Shanghai University of Traditional Chinese Medicine,2005,(3).

[4] Wen Yongyi, Fan Xinrong. Study on the Feasibility of Constructing TCM Corpus [J]. Shanghai Journal of Traditional Chinese Medicine,2003,(4).