2018 4th International Conference on Education, Management and Information Technology (ICEMIT 2018)

Construction on Parallel Corpus System for English Translation of Liaoning Dialect

Jie Fu

College of Foreign Languages, Bohai University, Jinzhou, 121013, China

563714244@qq.com

Keywords: Liaoning dialect; English translation; parallel corpus; system construction

Abstract: Dialect is a unique national culture, each place has a unique dialect, with a rich cultural heritage. Dialects can represent both regional cultural characteristics and social phenomena. Aiming at the problems of dialects in Liaoning and the English translation of dialects, this paper follows the basic principles of parallel corpus construction and puts forward the construction plan of parallel translation corpus system of Liaoning dialect English translation. The specific construction process includes 8 steps such as "corpus collection, corpus processing, corpus classification, corpus markings, corpus tagging, corpus coding, corpus alignment, corpus warehousing". The research results of this paper can not only help students to learn dialects, but also assist in literary translation, and serve the local traditional culture of Liaoning.

1. Introduction

Dialects are also called vernacular, which is a variant of language. According to nature, dialects can be divided into regional dialects and social dialects. Regional dialects are variants of language that differ because of geographical differences. They are branches of the language of the whole people in different regions, and are a reflection of the imbalance in language development. Social dialects are members of society in the same region, and different social variants are formed because of social differences in occupation, class, age, gender, and cultural upbringing. The differences between modern Chinese dialects are manifested in various aspects of speech, vocabulary and grammar, especially in terms of speech. In the modern Chinese dialects of China, the northern dialect can be regarded as the development of ancient Chinese in the vast northern regions after thousands of years, while the other dialects are gradually formed by the northern residents in the history.

Liaoning Province has a long history and culture. The Hongshan Cultural Site of Niuheliang in Chaoyang County is about 5,000 years old. From the unearthed altars, temples, goddess painted figures, jade carving pigs and painted pottery, the following conclusions are drawn: A primitive civilized society that has begun to take shape in the country marks that Liaoning is one of the origins of Chinese civilization. From a historical point of view, the dialect vocabulary of Liaoning is mixed with some vocabulary of Manchu and Mongolian languages in history. The southern part of Liaoning also absorbs foreign vocabulary such as Japan and South Korea. Liaoning Province is located in the south of China's northeastern region, bordering on the Bohai Sea and the Yellow Sea in the south, echoing Shandong Province. Liaoning is the only coastal and border province in the Northeast, and the gateway to the northeast and the eastern part of Inner Mongolia Autonomous Region. Liaoning is an important part of the Northeast China Economic Zone and the Bohai Rim Economic Zone. It is also a foreign trade and international trade in Northeast China. An important channel for communication.

As a non-standard language, dialects are mainly spoken language-specific regional variants that only pass through a region. As a literary form, dialects have distinct regional characteristics, which can well create a local atmosphere, enhance the authenticity of the article, vividly shape the character and character of the characters, increase the sense of humor or satire of the works, and win many writers. Love has always been used frequently in ancient and modern Chinese and foreign literature. But for translators, the translation of dialects in literary works is the most headache. The differences in English dialects are mainly in pronunciation, and there are fewer differences in vocabulary. Because the same linguistic symbols are used, the author and the translator have less difficulty in the communication and rendering of the works. The dialects in Chinese not only have different pronunciations, but also different words. The words in many dialects cannot be transformed into words, which causes difficulties in communication between the author and the translator.

China's vast territory has created its own local characteristics, which in turn has formed its own dialect. Chinese dialects explain that there is quite a feeling of "only can be expected, can not be said." There are still such difficulties between Mandarin and dialect. Chinese dialects and foreign dialects belong to different language families and are more difficult to translate. Constructing the parallel corpus system of Liaoning dialect English translation can not only help students to learn dialects, but also assist in literary translation, and serve the local traditional culture of Liaoning.

2. Corpus Related Concepts

The corpus-related concepts related to the research content of this paper include the following three:

(1) Corpus. The corpus refers to a large-scale electronic text library that has been scientifically sampled and processed. It is the basic resource for corpus linguistic research and the main resource for empirical language research methods. Used in lexicography, language teaching, traditional language research, natural language processing, and statistical or case-based research. Since the emergence of computerized corpora, corpus linguistics has developed rapidly, and the use of corpus for language comparison research and language ontology research has achieved fruitful results. The corpus needs to clarify the following three basic understandings: the corpus stores the language materials that have actually appeared in the actual use of the language; the corpus is the basic resource for carrying the language knowledge by the electronic computer; the real corpus needs to be processed before it can become Useful resources.

(2) Corpus classification. There are many types of corpora, and the main basis for classification is the purpose and use of the research. Usually divided into four types: Heterogeneous, without specific corpus collection principles, widely collected and stored as many corpora; Homogeneous, only collect corpus of the same type of content; Systematic, according to predetermined principles and proportions to collect corpus, so that the corpus is balanced Sexual and systematic, able to represent a range of linguistic facts; Specialized, only collects corpus for a particular purpose. According to the language of the corpus, it can be divided into Monolingual, Bilingual and Multilingual.

(3) Parallel corpus. Parallel corpora is a bilingual or multilingual corpus composed of the original text and its parallel corresponding translated text. The degree of alignment can be word level, sentence level, paragraph level and stage level. The parallel corpus includes three forms: uni-directional parallel corpora, bi-directional parallel corpora and multidirectional parallel corpora.

3. Basic Research on Construction of Translation Parallel Corpus

The research content includes four aspects:

(1) Analysis of the problem of translation parallel corpus construction. The corpus construction is separate and lacks a large-scale, comprehensive and multi-purpose national parallel corpus. With a certain scale and a relatively complete corpus, the construction period is relatively long, and the repeated construction disperses the power, which limits the scale and processing depth of the corpus. The key point in the development of corpus translation studies is "the development of bilingual libraries, which are both technical means and infrastructure, as well as the research purposes of the developers." From the perspective of corpus classification, most corpora are still limited to literary and non-literary. The corpus of some specialized syllabuses is still relatively small.

(2) Basic principles for the translation of parallel corpora. Authenticity, corpus entry adopts the principle of “recording”, and the corpus should be marked with faithful originals to maximize the original corpus. No changes are made to the corpus at the time of entry; balance, different types of corpus should be as uniform as possible, and should be differentiated according to actual conditions; systematic, all corpora are complete, and can correspond one-to-one, related information Complete; dynamic, corpus needs to be constantly enriched and updated; convenience, clear corpus structure, simple interface, quick response, easy to use.

(3) The development trend of translation parallel corpus construction. Focus on the interdisciplinary integration of corpus research, and combine traditional linguistics, computational linguistics, computer science and lexicography to highlight the interdisciplinary nature of the corpus; put more energy into the study of multiple modal corpus, in the process of corpus processing More attention to non-linguistic factors; computer storage technology and data storage technology development to promote the corpus large-scale, the wide application of multimedia technology to promote multiple corpus mobilize, and the Internet, especially the wide application of mobile internet to promote corpus mobilization, with the field of research Continuously promote the specialization of corpus.

(4) Standardization for the translation of parallel corpus construction. The arbitrariness in the construction of translation parallel corpus has become a key issue that restricts the development of corpus construction. It is hoped that a set of scientific construction standards will be formed to promote the standardization and further development of corpus construction, and better serve the external communication and development of Liaoning dialect. The formulation of construction standards must have a sufficient realistic foundation. The design and planning of corpus construction, the collection and processing of corpus, and the maintenance of corpus management systems need to establish uniform standards to achieve a higher degree of sharing, and provide conditions for comparative study of different corpora.

4. Construction on Parallel Corpus System for English Translation of Liaoning Dialect

The construction of the parallel corpus system of Liaoning dialect English translation is divided into 8 steps in order, as shown in Fig. 1.

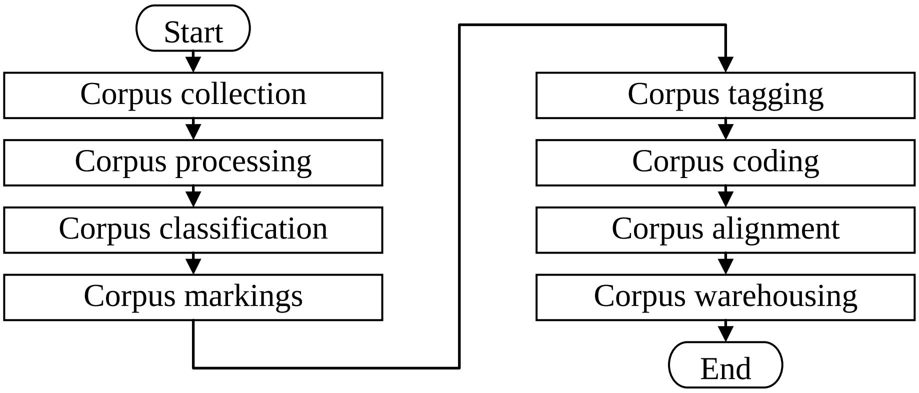

Fig. 1. Construction process on parallel corpus system for English translation of Liaoning dialect

(1) Corpus collection. Corpus collection is the most complicated task in the construction of corpus, which requires a lot of manpower and material resources. In order to ensure the authenticity and fresh activity of the corpus, the corpus of the Liaoning dialect corpus is a combination of words and voices. The corresponding English corpus is graduated from English majors and has been completed by experts who have been engaged in linguistic research for many years. In view of the scattered characteristics of Liaoning dialects, three collection groups were set up to go to Liaozhong, Liaonan and Liaoxi for collection work. To choose the ideal pronunciation partner. The pronunciation partner should be a local native, preferably a middle-aged and elderly person with

quick thinking, clear articulation, stable pronunciation and no influence on Mandarin. He can speak pure local dialect in Liaoning.

(2) Corpus processing. Use a text organizer to organize individual text into a standard format. Text Organizer is an easy-to-use and powerful text editing software with functions such as undo, redo, find, replace, code conversion and text editing. For files with irregular format, just a few buttons. It can be organized into a canonical format. The software covers a number of useful functions, including editing, removing spaces, removing paragraphs and leading spaces, paragraph finishing, punctuation, simplified text, correcting local garbled characters, etc., and supports batch processing. For a large number of dialect corpora collected on the spot, using a text organizer, it can be organized into a standardized format.

(3) Corpus classification. According to the different corpus requirements of each sub-library, the selected corpus is classified according to certain rules according to the register, style, difficulty and source. The existing corpus classification techniques are divided into two categories: expert knowledge methods and machine learning algorithms. Based on the existing expert knowledge, the expert knowledge method converts it into classification rules to classify the corpus, with emphasis on the establishment of classification rules. Different from the expert knowledge method, the classification rules of machine learning methods are not explicitly given before classification, but are automatically established according to the training corpus learning category attributes. The clustering algorithm in the machine learning field has been used as a corpus classification in recent years. Get better performance by training the data set.

(4) Corpus markings. The external information and structural features of the recorded text are automatically or semi-automatically performed by the computer. From the information type, all the information added to the corpus is meta information, that is, information about the information. The meta information is divided into four types: editorial information, analytical information, descriptive information, and management information. In the corpus database construction process, the tag information is generally encoded in the markup language, recorded in the file header, or managed with the database, and associated with the text. Although tagging information is important, inserting tags into text is an effective method. Saving the original format text allows for a more intuitive view of the corpus, but it is still a challenge for the computer to automatically identify and retrieve relevant structural features.

(5) Corpus tagging. Realizing the machine reading of corpus and improving the utilization value of corpus, the key lies in the corpus. The corpus annotation is to process the original corpus in the corpus, and mark the various linguistic features of the linguistic features on the corresponding linguistic components, including part-of-speech tagging, syntactic analysis, phonetic annotation, semantic annotation, pragmatic annotation, discourse annotation, Stylistic labeling and word labeling, etc., to facilitate computer reading. The end user of the corpus should be clear about the principles of labeling and the meaning of the code. The corpus annotation is not perfect, it is just a useful tool. Labeling tries to adopt the neutral mode that is generally accepted, and any annotation mode cannot be used as the first standard.

(6) Corpus coding. Including word class assignment and syntactic assignment, it creates conditions for the quantitative study of language and provides convenience for studying the probabilistic features of language. The word class assignment is to mark the word class attribute for each word in the text. This work is usually done on the basis of the traditional grammar on the word class, but the classification adaptation requirements are made finer. Syntactic assignment is a syntactic annotation of each sentence in the text. It is divided into three steps: the first step is to assign a syntax code to each word in the text; the second step is to find a special grammar code form and syntactic fragments, and the syntax The structure is modified as necessary; the third step is to complete the syntactic analysis and assign values one by one, from which the largest syntactic analysis is selected as the analysis result of each sentence.

(7) Corpus alignment. The original text and the translation establish a paragraph or sentence or even a word level correspondence. The purpose of corpus alignment is to build a corpus memory and apply it to computer-assisted translation, which can greatly improve translation efficiency,

especially for translation in specialized fields. The alignment of the semi-automatic English-Chinese bilingual parallel corpus is divided into two processes: the first process, which first divides the text of the two languages into sentences, each sentence occupies one line; the second process, based on the results of the first process, Manually align text in both languages at the sentence level. If the texts of the two languages differ in the segmentation of the sentence, try to keep the original sentence intact and adjust the translation to suit the original text.

(8) Corpus warehousing. Import the corpus into the database system and save the relevant information. The structure of the data table is shown in Tab. 1.

Table 1. Data storage table on parallel corpus system for English translation of Liaoning dialect

No	Fields interpretation	Fields Name	Type	Width	Remarks
1	Primary key	ZGJZ	varchar	20	
2	Dialect name	FYMC	varchar	100	
3	Dialect coding	FYBM	varchar	10	
4	Provider coding	TGZBM	varchar	10	
5	Provider name	TGZXM	varchar	30	
6	Dialect provider date	FYTGRQ	datatime	8	
7	Dialectal text content	FYWBNR	varchar	500	
8	Audio file format	YPWJGS	varchar	30	
9	Audio file storage	YPWJCC	varbinary	Max	
10	Video file format	SPWJGS	varchar	30	
11	Video file storage	SPWJCC	varbinary	Max	
12	Translation text content	FYWBNR	varchar	500	
13	Translation file format	FYWJGS	varchar	30	
14	Translation file storage	FYWJCB	varbinary	Max	
15	Note description	BZSM	varchar	Max	

Acknowledgment

This work is supported by social science fund project of Liaoning province (L15BYY007): Phonetics Characteristics of Liaoning dialect and its influence on English education in Liaoning.

References

[1] S. X. Xiang, "Bilingual Parallel Corpus-based Translation Teaching & the Improvement of Translation Competence," The Journal of Shandong Agriculture and Engineering University, vol. 34, no. 10, pp. 11-13, 2017.

[2] L. Liu, "Dialectal translation strategies in English literature," Journal of Suzhou University, vol. 31, no. 6, pp. 52-54, 2016.

[3] J. G. Bai, "A Preliminary Study of the Application of Parallel Corpus in Translation Teaching," Journal of Chifeng University (Philosophy and Social Science Chinese Edition), vol. 39, no. 5, pp. 161-164, 2018.

[4] N. Ding, "Liaoning dialect phonetic research overview," Journal of Jiamusi Vocational Institute, vol. 32, no. 11, pp. 322, 2015.

[5] G. L. OuYang, "Summary on the Research of Liaoning Dialect in Past 60 Years," Huazhong Normal University Journal of Postgraduates, vol. 17, no. 4, pp. 50-52, 2010.

[6] N. Ding, "The construction of Tianjin dialect in the view of corpus," Shanxi Archives, vol. 23, no. 6, pp. 87-89, 2017.