ISSN 2616-5783 Vol.5, Issue 15: 139-144, DOI: 10.25236/AJHSS.2022.051519

Basic Research on Construction of Multimodal Parallel Corpus of Tourism Translation in New Media Era

Li Rui*, Gou Xiuli

College of Foreign Languages, Bohai University, Jinzhou, China
selinalr@163.com
**Corresponding author*

Abstract: *Parallel corpus are bilingual or multilingual corpus composed of the original text and its parallel counterparts in the target language. Bilingual parallel corpushas the closest relationship with translation, providing rich translated texts and convenient translation means. The new media era has put forward new requirements for tourism translation. As the number of foreign tourists coming to Liaoning increases year by year, the problems in tourism translation in Liaoning become more and more obvious. The basic research on Liaoning tourism translation multimodal parallel corpus in the new media era, including theoretical basis, technical support, development trend and standard formulation, will provide basic support for the construction of multimodal corpus and realize the integration of various media forms, such as words, images, voice and video, which can fully mobilize users' multiple sensory systems and provide comprehensive tourism translation services.*

Keywords: *New Media Era; Tourism Translation; Multimodal Parallel Corpus; Basic Research*

1. Introduction

Corpus refers to a large-scale electronic text base that has been scientifically sampled and processed. It is the basic resource of corpus linguistics research and the main resource of empirical language research methods. Since the emergence of computerized corpus, corpus linguistics has developed rapidly, so the contrastive study of language and the study of language ontology using corpus have achieved fruitful results. With the development of corpus, two research paradigms, corpus driven and corpus based, have emerged. With the extensive application of computer technology and the rapid development of Internet technology, corpus research has moved from "Corpus 3.0" to "Corpus 4.0", that is, multimodal corpus [1]. Corpus is no longer limited to text corpus, and a variety of corpus such as audio and video have been added, which can include various types of corpora, reflecting a new view of language.

At present, the study of tourism translation corpus is limited to the academic field, and no breakthrough has been made in its practical application. The new media era has put forward new requirements for tourism translation. With the increasing number of foreign tourists coming to Liaoning, the problems in tourism translation in Liaoning are becoming more and more obvious. The poor quality of publicity translation affects the development of tourism in Liaoning, the inaccurate translation of public signs causes misunderstanding to foreign tourists, tourism translation resources do not meet the needs of smart tourism. Most scenic spots have not built intelligent question answering systems, human-computer interaction systems and simultaneous interpretation systems, and it is difficult for foreign tourists to solve problems in time. With the revitalization of the old industrial base in Northeast China, and the construction of the "Belt and Road" and the implementation of various tourism policies under the construction of a new development pattern, more and more international tourists will travel in Liaoning in the future. In this context, it is of great significance to carry out research on the construction and application of a multimodal parallel corpus of tourism translation in Liaoning in the new media era to promote the development of the tourism industry in Liaoning.

2. Theoretical Basis

The theoretical basis is to study the general or main laws of social and economic movements, and to

ISSN 2616-5783 Vol.5, Issue 15: 139-144, DOI: 10.25236/AJHSS.2022.051519

provide a guiding common theoretical guidance for applied research. This research involves functional translation theory, information processing theory, corpus linguistics theory and structuralist linguistics theory.

2.1 Functional Translation Theory

Translation is to use the most appropriate, natural and equivalent language to reproduce the information of the source language from semantics to style. It does not seek rigid correspondence on the surface of the text, but to achieve functional equivalence between the two languages. Functional translation theory is a methodology system based on rhetorical functional equivalence, which is formed in the repeated process of practice and develops with practice, and has strong practicability and operability in translation practice. Translation makes the original text and the translated text equal in "rhetorical function". In addition to linguistic context, situational context, cultural context and pragmatic context of the original language must also be considered. It is very important to build a corpus with rich information materials. Excellent tourism translation needs to closely connect the scenic spot with foreign tourists, improve the tourism experience, and optimize the tourism experience [2].

2.2 Information Processing Theory

As the core of modern cognitive psychology theory, information processing theory points out that the psychological process of people's understanding of things should be regarded as the process of information processing. In this process, people are regarded as an information processing system that can receive, store, process and transmit information in turn. Information processing theory is based on behaviourism psychology and gestalt psychology, and regards complex learning cognitive process as information processing process [3].Information processing theory shows the steps of memory well, describes the process of memory, and explains the essence of memory. From the perspective of information processing theory of cognitive psychology, the application of multimodal parallel corpus to tourism translation has unique advantages, which provides a new way to improve the quality of tourism translation.

2.3 Corpus Linguistics Theory

Corpus linguistics uses a large number of naturally occurring language data, which puts forward new findings or theories based on linguistic facts, and provides philosophical ideas for linguistic research. A diachronic study of language phenomena in the term database can reveal the historical changes of a specific language phenomenon. Corpus linguistics is highly applicable and plays a more prominent role in the field of linguistics from the initial language analysis and theoretical research to today's lexicography, foreign language teaching and artificial intelligence [4].Corpus linguistics is very abstract. Whether it is traditional corpus linguistics research theory or modern corpus linguistics research methods, it is necessary to construct the application rules of language analysis through abstract research methods, so as to summarize the essential methods of language use.

2.4 Structuralist Linguistic Theory

Structuralist linguistics is any language study that regards language as an independent system with phonetic, lexical and grammatical characteristics. Structural linguistics believes that language is essentially a symbol system combining sound and meaning, and a structure. The nature of language elements determines the interrelationship between system elements [5].Structural linguistics not only affects all fields and schools of linguistic research, but also affects the development of other humanities and social sciences. Rigorous analytical methods have penetrated into philosophy, literary criticism, literary research and other fields, and have had an important impact on the development of humanities and social sciences. The current construction of multimodal corpus shows a trend of merging the two categories of "language structure" and "language function", paying attention to both language structure and language function.

3. Technical support

There are many technologies involved in the construction of multimodal parallel corpus of tourism

ISSN 2616-5783 Vol.5, Issue 15: 139-144, DOI: 10.25236/AJHSS.2022.051519

translation in the new media era. This paper mainly studies the technologies of corpus acquisition, corpus alignment, corpus tagging and corpus storage.

3.1 Corpus Acquisition Technology

There are two main methods to obtain parallel corpus [6]. One is to mine parallel corpus from various databases or literatures; Second, obtain corpus resources from bilingual Web sites, and generate parallel corpus after processing. With the development of the Internet, more and more websites provide bilingual information, which makes the Internet become a potential bilingual corpus information source containing rich bilingual resources, promoting the development of Web based parallel corpus acquisition methods. Web crawler is an active and specialized search technology. The search object is a web page. It can automatically grab data from the network according to the designer's requirements, download the web page according to the incoming URL, and convert the web page into a string [7]. Common web crawlers will adopt certain crawling strategies, mainly including depth first crawling strategy and breadth first crawling strategy.

3.2 Corpus Alignment Technology

Corpus alignment refers to establishing a corresponding relationship between different language units of two or more language texts, that is, determining which language unit of the source text and which language unit of the target text are mutually translated. According to the size of language units, corpus alignment can be divided into lexical level, sentence level and paragraph level alignment. Lexicon compilation requires lexical alignment, while translation requires sentence alignment. Sentence level alignment is a more commonly used alignment mode in translation practice. Its working principles mainly include three types: length-based alignment, vocabulary-based alignment, and hybrid alignment [8]. ABBYY Aligner is a translation alignment software, which can automatically realize bilingual alignment according to user settings, and can also edit the translated content on the software, supporting languages of multiple countries.

3.3 Corpus Tagging Technology

Corpus tagging is generally used to describe the relevant information such as the author and source of the corpus, as well as the linguistic features such as the part of speech and syntactic features of the corpus. The key to realizing the machine-readable corpus and improving the utilization value of the corpus is the effective tagging of the corpus [9]. After tagging, it is not only convenient for users to extract information from the corpus, but also can increase the reusability of the corpus. Doccano is a lightweight open-source data annotation platform, which is implemented by Django. It is very simple to deploy and use. Whether on Windows PC or Linux server, it only needs to be completed step by step according to the official guidance. It supports emotion analysis, named entity recognition, text summarization and other tasks. On a small corpus, all annotation can be completed in a few hours.

3.4 Corpus Storage Technology

The data in the database is organized, described and stored according to a certain data model, with less redundancy, higher data independence and scalability, and can be shared by various users. The corpus is mainly stored in a relational database. From the perspective of software development, the interface between users and relational database programming is flexible. Most RDDMS products use the standard query language SQL. Users can access the information of another product almost indiscriminately. The application software that interfaces with relational database has a similar program access mechanism, and provides a large number of standard data access methods. The database design and standardization process are also very simple, easy to implement and understand. The powerful function of relational databases has effectively promoted the development of multimodal parallel corpus.

4. Development Trend

Corpus linguistics is a marginal discipline developed in recent years, which is not only the historical accumulation of linguistic development, but also the product of contemporary linguistic research unique to the computer age [10]. Corpus linguistics integrates linguistic theory, mathematical thinking

ISSN 2616-5783 Vol.5, Issue 15: 139-144, DOI: 10.25236/AJHSS.2022.051519

mode and computer technology, and comprehensively and completely describes language in a scientific and objective way, so that people can re understand the nature of language.

4.1 Multimodality

Modern corpus can study not only language itself, but also many non-linguistic factors. Compared with the traditional unimodal corpus, the multimodal corpus not only collects spoken and written texts, but also collects many communication factors outside the text, including fonts, images, pronunciation, intonation and interpersonal interaction. These materials involve multiple modes, which are described and processed in advance by multimodal annotation tools. With the help of MCA multimodal corpus software similar to that developed by Professor Anthony Baldry's team, multimodal analysis can be carried out on the corpus, and various functions other than language research can be realized [11].

4.2 Multimedia

With the advent of multimedia and the Internet, audio and video materials have emerged in large numbers. It is a trend to establish multimedia discourse corpora that can contain audio and video materials. Corpus has developed from the stage of collecting text corpora to the stage of storing spoken language corpora. At present, it has gradually stepped into the development stage of multimedia discourse. Although there have been many researches on tourism translation corpus in the past, most of them focus on tourism publicity translation, tourism public signs translation and tourism text translation. The future tourism translation corpus is more in line with the needs of the new media era. The prominent feature is the multimodality of the corpus, the integration of various media forms such as text, image, voice and video, which fully mobilize the users' multiple sensory systems and can provide comprehensive tourism translation services.

4.3 Intersection

Corpus linguistics itself is interdisciplinary, combining linguistics and computer science to form a new language research method and field. As interdisciplinary corpora and cross register corpora have high requirements for researchers, they need to master knowledge of multiple disciplines. On this basis, they collect corpus from various disciplines or registers, label the corpus from the perspective of cross research, and discover new research concerns, research methods and research strategies. At present, there are few qualified researchers, and cross research needs to be developed. The future multimodal parallel corpus of tourism translation will be more and more widely studied, not limited to tourism discipline or tourism translation, but gradually extended to multiple disciplines and registers to promote multidisciplinary cooperation and communication and multi register comparison.

4.4 Intelligence

Driven by new theories and technologies such as mobile Internet, big data, supercomputing, sensor networks and brain science, as well as the strong demand for economic and social development, artificial intelligence has accelerated its development, showing new features such as human-computer collaboration, deep learning, cross-border integration, group intelligence openness and independent control, and promoting the accelerated leap from digitalization, networking to intelligence in all areas of the economy and society. Machine translation has entered the intelligent translation stage of neural network, and machine translation based on artificial intelligence has become a hot spot in recent years [12].Seizing the significant strategic opportunity of AI development, under the premise of applying AI and natural language processing technology, we invited computer and linguistic experts to classify and code the corpus and synthesize an advanced tourism translation corpus with advanced AI search and other functions [13].

5. Standard formulation

Standards are unified provisions on repetitive things and concepts, which are based on the combination of science, technology and practical experience, agreed by relevant parties, approved by the competent authority, and issued in a specific form as the guidelines and basis for common compliance. At present, there is no unified standard for corpus construction, and the construction practice is highly arbitrary. In order to improve the quality and efficiency of construction, standards

ISSN 2616-5783 Vol.5, Issue 15: 139-144, DOI: 10.25236/AJHSS.2022.051519

must be formulated.

5.1 Construction Process Standard

The standard process for the construction of multimodal parallel corpus of tourism translation in the new media era has been formulated, which provides a correct path for the work. First, the construction task is to define the specific objectives, explain what kind of corpus needs to be built, the reason, purpose and significance of the construction of the corpus, and solve the problem of the necessity of the construction of the corpus. The second is the overall design, which clarifies the characteristics of the corpus, determines the scale, structure, content, method, construction principle and use method of the corpus, and solves the feasibility problem of the corpus construction. Third, software development. While language researchers collect and label language materials, software designers and programmers simultaneously develop software systems. Fourth, release and opening. After the corpus has all the intended functions, it should be released and opened to the society through various ways to give play to its application value in the tourism service industry.

5.2 Corpus Collection Standard

Corpus collection is the most complicated work in corpus construction, which requires a lot of manpower and material resources. First, the authenticity and vividness of the corpus. The corpus of Liaoning tourism translation multimodal parallel corpus adopts a combination of words and sounds, and the corresponding English corpus is completed by experts engaged in tourism translation. In view of the scattered characteristics of tourist attractions in Liaoning, four collection groups were established to conduct collection work in northern, central, southern and western Liaoning, focusing on the tourist guides, brochures and public signs collected on the spot [14].Second, the balance and systematic standard of language materials. Balance means that different types of language materials should be distributed as evenly as possible, because too few language materials will not guarantee the objectivity of research conclusions, nor provide comprehensive support for translation practice. Systematicity refers to the fact that the corpus can reflect the features of Liaoning tourism translation and facilitate the observation and analysis of the corpus from all angles.

5.3 Corpus Entry Standard

According to different corpus requirements of each sub database, the standardized and sorted corpus shall be classified according to certain principles, and classification standards shall be established, which can be classified according to the source, collection method, processing degree, main purpose, corpus level, time attribute and language type [15]. In order to ensure the authenticity of the corpus, errors and writing formats of words, phrases, sentences, chapters and punctuation marks in the corpus shall be recorded as they are, without any change, and the original appearance of the corpus shall be maintained to the maximum extent. Wrong characters are difficult to enter. If there are no characters in the computer character library and it is impossible to enter directly, you can identify them with codes first, and then reflect the original appearance during later processing [16]. When the spoken language is transcribed into written form, the pause, repetition and phonetic errors should be truthfully reflected. For multimodal corpora, expressions and body movements associated with oral communication should also be depicted.

5.4 Corpus Tagging Standard

Corpus tagging refers to the processing of the original corpus in the corpus, and the tagging of various additional codes representing language features on the corresponding language components, including part of speech tagging, syntactic analysis, phonological tagging, semantic tagging, pragmatic tagging, discourse tagging, stylistic tagging and word tagging, so as to facilitate computer reading. First, it is comprehensive and relative. The more comprehensive the content marked in the corpus, the more it can meet the needs of various tourism translation. The construction of corpora is limited, and it is difficult to achieve fully "universal", so there is a problem of relativity. Second, scientific and universal. Scientific means that the tagging of the corpus should be correct, conform to the relevant norms of Chinese characters and words, and conform to general grammar rules. The labelling of similar language phenomena should be consistent. Universality refers to the standardization and generalization of corpus tagging codes, which is conducive to the sharing of corpus resources. Therefore, it is necessary to develop a generally accepted and willing to use corpus annotation specification as soon as possible.

ISSN 2616-5783 Vol.5, Issue 15: 139-144, DOI: 10.25236/AJHSS.2022.051519

6. Conclusions

Corpus, as a kind of real used language material, constitutes the basic resource of language knowledge and plays an important role in modern natural language processing research. The rapid development of computer technology has provided superior conditions for the development of corpus linguistics, which has gradually evolved into a multidisciplinary, multi-level, multi-functional and multi angle discipline to analyse language phenomena. The influence of corpus linguistics has gradually penetrated into all areas of language research, making corpus an indispensable tool for natural language researchers, lexicographers and ordinary language enthusiasts. The theoretical basis, technical support, development trend and standard formulation of this paper lay a foundation for the construction and application of multimodal parallel corpus of Liaoning tourism translation in the new media era.

Acknowledgements

This work is supported by 2022 annual social science planning fund project of Liaoning province (L22BYY004): Construction and Application of Liaoning Tourism Translation Multimodal Parallel Corpus in the New Media Era; Social science planning fund project of Liaoning province in 2021 (No: L21BYY002): Construction and application of bilingual parallel corpus for Translation for Chinese classics sub-confucian category.

References

[1] Y. Sun, "The application of multimodal Financial English Corpus in college English teaching," Foreign Languages and Translation, vol. 27, no. 2, pp. 70-76, 2020.
[2] Z. Z. Luan, "A study of tourism text translation from the perspective of functional translation theory: Taking the English translation of the Five Avenue Scenic Area in Tianjin as an example," English Square, vol. 12, no. 21, pp. 15-18, 2022.
[3] Q. Zhang, "A study of Higher Vocational English listening teaching under the guidance of Information processing theory," Journal of Kaifeng University, vol. 34, no. 4, pp. 69-71, 2020.
[4] J. H. Gong, "On the Practical Significance of Corpus Linguistics Theory in College English Teaching," Journal of Lanzhou Vocational Technical College, vol. 30, no. 10, pp. 130-131, 2014.
[5] J. Jiao, "A brief analysis of structuralist linguistic theory," Journal of Social Science of Jiamusi University, vol. 25, no. 3, pp. 50-51, 2007.
[6] J. Shao, C. Z. Zhang, "Automatic Acquisition of Domain Parallel Corpora from Internet," Data Analysis and Knowledge Discovery, vol. 35, no. 12, pp. 36-43, 2014.
[7] Y. S. Chi, "Research on Web Crawler Technology Based on Python," China Computer & Communication, vol. 33, no. 21, pp. 41-44, 2021.
[8] H. S. Wang, S. S. Wang, "On the Application of Corpus Alignment Technology in Translation," Chinese Science & Technology Translators, vol. 30, no. 4, pp. 16-19, 2017.
[9] A. H. Dong, "A Discussion to Annotation of the Corpora," Journal of Beijing Institute of Graphic, vol. 24, no. 5, pp. 67-70, 2016.
[10] Y. J. Zhou, "The Application of Corpus Linguistics and its Growing Trend in China," Journal of Qiqihar University(Philosophy & Social Science Edition), vol. 36, no. 3, pp. 138-140, 2007.
[11] Z. Y. Ye, "A Reappraisal of English Corpus Construction and Application," Journal of Ningbo Polytechnic, vol. 18, no. 1, pp. 78-82, 2014.
[12] Y. Lu, "Big data corpus construction under artificial intelligence translation," Gansu Science and Technology, vol. 35, no. 17, pp. 80-84, 2019.
[13] J. F. Gu, "Exploration on the significance and approach of constructing artificial intelligence corpus for international ommunication," International Communications, vol. 28, no. 1, pp. 40-43, 2021.
[14] S. E. Hu, "Design and construction of bilingual parallel corpus of tourist attractions: A case study of Zhoushan," Journal of Western, vol. 9, no. 8, pp. 66-68, 2021.
[15] Y. S. Luo, S. Fu, "Corpus classification system and college medical English teaching," Journal of Higher Education, vol. 3, no. 15, pp. 105-107, 2018.
[16] B. L. Zhang, X. L. Cui, "On the Standards of Building a Chinese Inter-Language Corpus," Applied Linguistics, vol. 24, no. 2, pp. 125-134, 2015.