ISSN 2522-6398 Vol. 4, Issue 9: 53-58, DOI: 10.25236/FER.2021.040910

Constructing an ESP Bilingual Parallel Corpus Based on AntConc: Application and Assessment

Xi Lu[1], Kai Jiang[2,*]

[1]*Department of Common Required Courses, Hubei Institute of Fine Arts, Wuhan, China*
[2]*College of Foreign Languages, Huazhong Agricultural University, Wuhan, China*
[*]*Corresponding author: jiangkai@mail.hzau.edu.cn*

Abstract: *AntConc is a free and green corpus tool developed by Japanese scholar Laurence Anthony featured by three main functions: concordance, wordlist and keywords. The author first describes the principles and process of constructing a web-based bilingual parallel corpus for the purpose of translation studies and ESP research based on AntConc. Constructing principles include strict linguistic standards, balance of corpus, and appropriate size. Based on the principles, language data is accumulated, processed and entered. After that, language processing software CLAWS part-of-speech tagger and Wmatrix are respectively used for text marking, annotating, high-frequency vocabulary extracting and corpus distribution balancing. In the end, texts, paragraphs and sentences are aligned by the corpus tool ParaConc. After the construction, the author uses the retrieval software Wordsmith 4.0 and statistical software SPSS 11.5 to prove the feasibility and effectiveness of the corpus with a six-month experiment that covers 16 translators, 25 teachers and 285 students.*

Keywords: *Computer Aided Translation, corpus, AntConc, Wmatrix, computational linguistics*

1. Research background

The construction of modern corpus began in the 1960s. In the recent decade, the scale of corpus construction has been continuously expanded, and has realized the transition from single language to multiple languages. At the same time, because of the rapid development of storage media and information technology, and the upgrading of labeling and retrieval software, corpus has been widely used. Corpus for specific purposes has attracted soaring attention in modern language teaching and research due to its large capacity, corpus based on authentic environment and fast retrieval.

Bilingual or multilingual corpus is a new and practical research method and practical training mode in the field of translation due to its innovative, data-oriented and student-centered characteristics. The data in bilingual corpus is not only an important resource of vocabulary and terms, but also a source of information where relative text, discourse structure and text type can be located, as well as a practical database for professional translation. One of the important uses of electronic corpus is language retrieval, which is also a way to realize the application of corpus in translation. Retrieval software plays an important role in translation. By retrieving the specific words and expressions in the corpus, the translator can effectively strengthen the understanding and control of the text, and ensure the accuracy and consistency of the main terms. This is of great help to translation, especially to the translation of texts in special fields. At present, there are many mature retrieval software in the world. The more practical monolingual retrieval software includes WordSmith and WordCruncher, Bilingual retrieval software is represented by MULTICONC, ParaConc and Bilingual Corpus System.

2. Purpose and Significance

In terms of fully automatic machine translation and real-time voice translation and other related technologies, many domestic research institutions and companies have realized them at present. However, such technologies have not yet been effectively applied in art translation and the teaching aimed for art students. The main reason is the lack of high-quality parallel corpus in related fields. There are few professional English corpus of art in China, and the capacity of the only art and design corpus is only around 300,000 words. Moreover, most of the corpora simply accumulate corpus without deep processing according to the characteristics of corpus, resulting in difficulties in corpus retrieval and low efficiency in corpus operation.

ISSN 2522-6398 Vol. 4, Issue 9: 53-58, DOI: 10.25236/FER.2021.040910

The English-Chinese & Chinese-English parallel corpus of fine arts provides professional services for fine arts translation and English teaching aimed at art students. The content covers painting, sculpture, design and architecture, and it is also a Chinese-English bilingual corpus with English as the source language, Chinese as the target language, or Chinese as the source language and English as the target language. The corpus can be used for English-Chinese Chinese-English machine-assisted translation, English-Chinese Chinese-English translation teaching and training, and provides a wealth of authentic corpus for English textbook compilation in the field of art (Zhao Yong, Zheng Shutang 2003), English-Chinese Chinese-English bilingual lexicography, natural language processing research and other work. Most importantly, the corpus of fine arts can greatly improve the efficiency of translation work, and also provide empirical research basis for the development of bilingual parallel corpus or multilingual parallel corpus of other language pairs, bilingual comparison and research, and corpus research in other fields.

3. Corpus Construction

3.1 Principles

The construction of fine arts English corpus should be based on strict linguistic standards.

3.1.1 Strict linguistic standards

The bilingual parallel corpus for fine arts is constructed according to the following principles. Strict sampling principles such as word frequency, word frequency and word class are established at the beginning of the design to ensure the randomness of sampling and the balance and representativeness of corpus information. The corpus should be able to comprehensively cover common English expressions in the field of fine arts and take the social application of English language in the field of fine arts as the main purpose, so as to meet the needs of teaching and translation.

3.1.2 The balance of corpus

Corpus collection is an important part of corpus construction, especially for fine arts English corpus. In the process of collection and selection, researchers should fully ensure the diversity, coordination and consistency of the sampled corpus. At the same time, corpus selection should cover materials that span a certain period, and the sampling range shall be reasonable.

3.1.3 Appropiate size

The size of fine arts English corpus can be divided into several categories according to different needs. Some large corpora involve many aspects, such as vocabulary, grammar structure, language structure and language behavior, etc., which can be called giant databases. Taking into consideration such factors as the word order, the number of texts and the sample size of the selected materials, we designed the corpus capacity to be 500,000 words. This can not only meet the translator's translation needs, but also provide a reasonable corpus capacity for the future expansion of corpus.

3.2 Data accumulation, processing and entry

The construction consists of three steps as shown in Fig. 1.

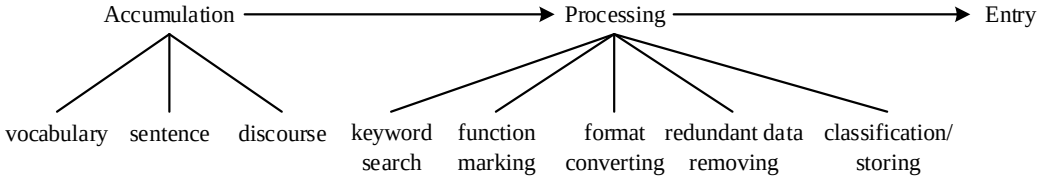

Figure 1: Data accumulation, processing and entry.

3.2.1 Accumulation

Corpus accumulation is the most important work in corpus construction. Since its open to the public, Trados has been favored by a large number of translation professionals, with more than 40,000 users at the global enterprise level. Trados is featured with powerful functions of storage and retrieval, and its memory bank is the core. Thus, the software is used to create corpus (memory bank) in specific fields of fine arts.

ISSN 2522-6398 Vol. 4, Issue 9: 53-58, DOI: 10.25236/FER.2021.040910

Corpus sampling must comprehensively consider the particularity of texts in the field of fine arts, the diversity of topics covered and the authority of text sources. Corpus content should include multiple forms of corpus such as vocabulary, sentences and discourse. When the author builds the database, the corpus is divided into four sub-libraries: painting, sculpture, design and architecture. Taking the compilation of painting as an example, the corpus is derived from various professional literature collections, art professional journals and art websites, such as art professional English textbooks, relevant materials, reference books and dictionaries, and all kinds of domestic and international representative art design academic journals, with topics related to Traditional Chinese painting, oil painting and printmaking. The key to corpus articles selection are materials with appropriate length, being concise and comprehensive, and finished in standardized language. The topics are highly related to the four fields, with strong professional representation and normative language.

3.2.2 Processing

Firstly, the special software technology and KWIC index technology are used to conduct keyword search for the corpus electronic text, and meanwhile the functions of the keyword such as part-of-speech, significance statistics and collocation statistics are marked out. Some documents in the form of PDF and Word retain the original document attributes, so there might be some redundant spaces, scrambled codes and invalid characters when the above text materials are converted to plain text format. In this case, text processing software such as EditPadPro, PowerGrep are used to remove redundant data and get clean plain text.

Then, the processed text data is classified and stored according to its attributes, such as content, attribute, meaning, date and theme, so as to build the basic prototype of the corpus. In the classification and storage of corpus, the data should be classified in a scientific naming way to reflect the basic information of corpus. For example, the file name "DA12EAD16 text" refers to "written language + paper + the 12th journal + from English for Fine Arts + the 16th text". Text stored in this way can be directly transformed into the corpus.

3.2.3 Entry

In order to make the corpus created open and shareable, and to facilitate other teachers and translation enthusiasts to flexibly use the corpus, the author first uses Heartsome TMX Editor Software to convert the Excel file that has been edited into TMX memory file. Then, the author imports the TMX text into Trados 2017 to form a corpus: open Trados 2017, click the "Translation Memory Bank" in the lower left corner, and then click the "Import" button in the top column. After obtaining the aforementioned TMX file, corpus users can import the TMX file into their own Trados according to the operating process described by the author, so as to retrieve and utilize it in translation.

4. Research Methodology

4.1 Marking and annotation of corpus

CLAWS part-of-speech tagger and Wmatrix are used for marking and annotation. The reference specification for aligning parallel corpora is the Chinese-English bilingual Corpus Markup Specification of Peking University (Chang Baobao, Bai Xiaojing 2003), and the markup can be set uniformly in XML extensible Markup Language. Specific markup work can be designed in a software platform to realize automatic markup setting.

In addition, the marking and labeling of corpus should not only meet the current needs, but also consider the possibility of continuous expansion of corpus and cooperation and sharing of multi-unit corpus in the future. Therefore, at least the markup and annotation of each unit corpus should be unified as far as possible. GALE corpus contains the initial manual annotation and the subsequent automatic machine annotation.(Olive et al.2011:1) the sentence alignment English-Chinese Chinese-English parallel corpus mainly serve the auxiliary machine translation, English teaching aimed at fine arts students and English-Chinese & Chinese-English art dictionary compilation, but in the future, it is likely to be used in fully automated machine translation and cross-language information retrieval, therefore, random marking or labeling is not recommended, otherwise a series of unwanted trouble may occur in subsequent corpus expansion and possible future cooperation.

ISSN 2522-6398 Vol. 4, Issue 9: 53-58, DOI: 10.25236/FER.2021.040910

4.2 Extraction of high-frequency vocabulary and corpus distribution

The purpose of corpus is to represent the whole linguistic facts with limited corpus. Firstly, according to statistical principles and common analysis software such as WELCTOOLSV, corpus was analyzed and selected, and the frequency standard of vocabulary selection was set to form a new word list. Software was used to analyze corpus, extract various statistical data, and corpus retrieval function was added at the same time. The preliminary selection of corpus requires the researchers to make further manual selection based on the following criteria, namely, meaning, etymology, ambiguity comparison, knowledge aggregation, difficulty, and frequency of use, in order to make it easy to tag or parse, and thus to make the choice of corpus highly professional and instructional.

4.3 Parallel alignment of corpus

Using the software ParaConc, parallel corpus can realize parallel alignment at multiple levels, such as text, paragraph, sentence, phrase and vocabulary. Text alignment is be achieved while collecting and organizing data.

But when parallel corpus paragraphs are not aligned, it is more difficult to align paragraphs manually. However, using the length relationship of bilingual paragraphs and computer aided proofreading can achieve 100% paragraph alignment faster. The author has tried the method of paragraph recombination alignment, but it doesn't prove successful. After all, paragraph is an important language unit (Liang Maocheng, Xu Jiajin 2012: 37). In order to maximize the benefit of corpus construction, it is suggested to keep paragraph marks as far as possible in parallel corpus.

Sentence alignment is also worth studying. Sentence alignment is not always one-to-one in parallel corpus, so it is not easy work. There are three main sentence alignment methods: 1) Sentence length-based method (Brown et al.1991; Gale & Church 1991); 2) Methods based on bilingual vocabulary translation information (Kay & Roscheisen 1993); 3) The method of mixing sentence length method and bilingual vocabulary translation. Due to the huge difference between Western and Eastern languages, the sentence alignment method based on sentence length is not suitable for the Chinese-English sentence alignment processing. Practice has also proved that the length ratio of English texts in the field of fine arts to Chinese texts is too long, usually between 0.8 and 7.5, thus sentence alignment based on sentence length is not effective enough.

The sentence alignment method based on bilingual vocabulary translation information is the better choice. It can be subdivided into the sentence alignment method based on bilingual dictionary and the sentence alignment method based on corpus vocabulary statistics. In order to test the difference between the sentence alignment methods based on statistics and dictionary when corpus is insufficient, we used Hunalign based on corpus vocabulary statistics and Champollion based on dictionary to test the sentence alignment of a section of Chinese-English art parallel corpus with 210 sentences and 4,040 words in English. The Chinese section has 161 sentences and 6 613 Chinese characters. The result of sentence alignment is shown in Table 1.

Table 1: Comparison of sentence alignment between Chinese and English parallel corpora in the case of insufficient corpus

Actual sentence pairs	Alignment method	number of pairs identified by system	number of pairs correctly aligned by system	accuracy	Recall rate	F-value
139	Hunalign(chinese words not divided)	170	13	7.6%	9.2%	8.8%
139	Hunalign(chinese words divided)	170	38	22.4%	27%	25.2%
139	Champollion	149	125	83.9%	90.5%	86.9%

The statistical data in Table 1 shows that, in the case of insufficient corpus, the alignment effect of Hunalign sentences based on corpus vocabulary statistics is significantly better than that of Champollion based on dictionary, indicating that the lexicography-based sentence alignment method is effective.

5. Application and Assessment

The analysis of corpus after running for a period of time is an important guarantee for the continuous improvement of corpus. There are two main levels of corpus evaluation. First, researchers shall retrieve common words and judge whether there is language error in the corpus according to the corpus labeling

ISSN 2522-6398 Vol. 4, Issue 9: 53-58, DOI: 10.25236/FER.2021.040910

provided by the corpus. If there are, errors shall be fixed immediately. Secondly, the developer shall evaluate and improve the search tools currently used to ensure that the search tools can instantly retrieve commonly used words in the industry and provide researchers and learners with information such as frequency of use, syntactic structure and collocation methods. By using the retrieval software Wordsmith 4.0 and statistical software SPSS 11.5, the feasibility and effectiveness of the corpus is proved.

In this stage, the corpus is accessible for six months and the operating effect of the corpus is evaluated and analyzed. Questionnaire survey and semi-open interview were used to conduct qualitative research on database users. A total of 16 translators, 25 teachers giving ESP (English for Specific Purposes) courses aimed at art majors and 285 art students participated in the survey.

Observations and on-the-spot notes are used as the basis for question formation, and then semi-structured interviews are designed. For more accurate information, with the consent of the interviewee, the interview is conducted in the form of notes and recordings. After the interview, the researcher transcoded the interview recordings and wrote memos. After the data were sorted out, they were analyzed.

The qualitative research results show that this corpus can effectively provide accurate bilingual parallel texts for translation and ESP courses aimed at art majors. For translators, corpus labeling can provide accurate language reference most of the time. Corpus can provide the retrieval results of common words in the field and provide specific information of relevant results, such as usage frequency, syntactic structure and collocation methods. In terms of ESP teaching for art students, the corpus can ease the burden on the part of teachers in the process of vocabulary teaching. The parallel texts and annotations can help students acquire knowledge in language and in specific fields, broaden professional horizon, and cultivate students' independence in learning and the strength the ability to solve problems they meet. However, some students respond that as the corpus offers large amount of information, students at lower language level or those who lack generalizing and summarizing ability sometimes have difficulties in understanding the retrieval results provided by the corpus. Therefore, teachers should gradually require students to practice and enhance the habit of independent learning supported by corpus based on their practical ability.

6. Conclusion

Practice has proved that the resources provided by the corpus for fine arts have high authenticity, diversity and practicality. The massive application information and data in the database provide professional translators, translation researchers and people interested in translation with rich language data in the field of fine arts, offering them comprehensive and reliable high-frequency vocabulary and collocation information in translation research and practical translation. The corpus can assist professional translators to translate more quickly and accurately in translation practice. It can also be widely used in translator training and ESP courses to help students understand texts in fine arts in the context of real corpus, so as to improve their professional English level and increase their understanding of relevant professional knowledge.

For ESP teaching, the corpus does provide a large amount of authentic material to teachers and can facilitate the teaching process, but the following questions are also worth pondering: how to get the most appropriate information from the huge amount of data? How to distinguish the rules of language application behind the retrieval results? How to make students recognize and accept the necessity of an ESP classroom aided by corpus? To answer these questions, teachers in charge of the classroom shall have a proactive and comprehensive understanding of the corpus. Being able to use the corpus effectively means that one can successfully locate corpus data needed, select effective, direct and reasonable examples to prove the common rule or trait of a certain language point. But due to the complexity and vast amount of data contained in the corpus, the requirements of different information retrieval for different teaching goals cannot be met directly and accordingly, it is suggested that teachers acquire necessary computer skills such as accurate data classification, analogy and contrast.

Acknowledgments

The paper is part of the research findings of *Research on College English Curriculum Reform in Art Colleges based on SPOC Model*, the Teaching Research Project of 2020 supported by Teaching Affairs Department, Hubei Institute of Fine Arts (Project No. 2020019), and the research findings of *A Corpus-based Study on the Translation and Dissemination of the Political Discourse "A Community of Shared Future for Mankind"*, the Philosophy and Social Sciences Research Project of 2020 supported by the D epartment of Education of Hubei Province (Project No. 20G030).

ISSN 2522-6398 Vol. 4, Issue 9: 53-58, DOI: 10.25236/FER.2021.040910

References

[1] P. F. Brown, J. C. Lai, R. L. Mercer. "Aligning sentences in parallel corpora". Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, pp.169-176, 1991.
[2] W.A. Gale, K.W. Church. "A program for aligning sentences in bilingual corpora". Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, pp.177-184, 1991.
[3] M. Kay, M. Roscheisen. "Text-translation alignment", Computational Linguistics, vol.19, issue 1, pp.121-142, 1993.
[4] J. Olive , C. Christianson, J. McCary. Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation. New York: Springer, 2011.
[5] B.B. Chang, X.J. Bai, "Beijing University Chinese-English Bilingual Corpus Markup Specification", Journal of Chinese Language and Computing, vol.02, pp.195-214, 2003.
[6] M.C. Liang, J.J. Xu. "The addition of meta-information and two-level alignment of paragraphs and sentences in bilingual corpus construction", Chinese Foreign Language, vol.06, pp.37-42, 2012.
[7] Y. Zhao, S.T Zheng, "Focus on the core vocabulary in college English textbooks", Foreign Languages and Foreign Language Teaching, vol. 06, pp.21-24, 2003.
[8] W.L. Liu, "Application of electronic corpus in translation teaching", Shanghai Translation, vol. 04, pp. 67-72, 2013.
[9] K.F. Wang, H.W. Qin, "The use of parallel corpora in the teaching of translation", Foreign Language Teaching and Research, vol. 47, issue 5, pp.763-772, 2015.
[10] K.F. Wang, "On the design and construction of the super-large-scale China English-Chinese parallel corpus (CECPC)", Foreign Languages in China, vol. 9, issue 6, pp. 23-27, 2012.