Construction of Parallel Corpus for English Translation Teaching Based on Computer Aided Translation Software

Bingbing Pan[1] and Qiongfang Qin[1]

[1]School of Foreign Languages, Guilin Tourism University, Guilin, Guangxi, 541006, China, panbingbing0515@163.com

Corresponding Author: Qiongfang Qin, agathe_qin@163.com

Abstract. Computer-aided translation technology can greatly improve translation efficiency which also reduce translation costs, so it is gradually being favored by all walks of life. Based on this, the article analyzes the computer-aided translation technology in this era. The application value of parallel corpus in translation teaching and translation practice is widely recognized, and there is an increasing demand for the construction of parallel corpus. With the development of corpus technology, corpus is used as a new research paradigm to enter the field of translation teaching and practice. After reviewing the current status of translation teaching, discuss the application of parallel corpus in current translation teaching in terms of self-built bilingual parallel corpus, translation teaching application and translation teaching content, aiming to promote translation teaching reform in order to meet the requirement of regional social and economic development. This student-centered teaching philosophy can make up for the shortcomings of traditional teaching methods. Students summarize and summarize the characteristics and rules of language use through exploratory and discovery-based learning, which is conducive to cultivating students' autonomous learning ability. Based on the self-built parallel corpus and referring to the teaching theories of different schools of constructivism, the translation teaching model is constructed from the aspects of teaching objectives, procedures, content and evaluation.

Keywords: Artificial Intelligence; Machine Translation; Computer-aided Translation; Empirical Evidence; Parallel Corpus
DOI: https://doi.org/10.14733/cadaps.2022.S1.70-80

1 INTRODUCTION

With the development or concept of artificial intelligence, Internet and other modern information technologies, lives, work, and studies have become more and more integrated with modern information technology, and people are dependent on computer technology and artificial

intelligence technology [1]. Computer-aided translation technology (CAT) is a new type of language translation tool centered on the translator and assisted by machine translation and computer translation. The development time of computer-aided translation technology in our country is relatively short [2]. When it was first developed, due to the constraints of algorithms and other conditions, it could only provide translators with a certain degree of understanding and reference, and could not produce reasonable, fluent, and ideal translations. In the meantime, CAT technology will usher in a real development trend, and it will be possible to completely replace manual translation, Zhang's [3] research has achieved good results in this field. Therefore, this paper conducts this research based on the prospects of the field of artificial intelligence-assisted translation in the future, and builds a computer-assisted translation architecture based on hybrid strategies.

Umerova [4] points out that parallel corpus refers to a bilingual corpus that can retrieve and display the source text and its translated text. Its powerful search, comparison, storage and other functions, as well as the richness, authenticity of the corpus, the intuitiveness and accuracy of data statistics, greatly compensate for the shortcomings and deficiencies of traditional translation teaching methods, and the research of parallel corpus has also opened up a new way for translation teaching. This perspective provides a new theoretical framework for translation studies. Its application in translation teaching has been widely recognized by academia. Current research believes that the role of parallel corpora in translation teaching is mainly reflected in: (1) Compared with ordinary dictionaries and teaching materials, it can provide more abundant translations and contexts for search terms, especially for the translation of professional terms. (2) Parallel corpus can provide more abundant examples for students to learn translation skills, and effectively improve learners' translation awareness. (3) Parallel corpus helps to create a discovery-based learning environment. The parallel corpus retrieved by the corpus helps students discover or explore the similarities between languages, as well as the skills or rules of language conversion, and gradually form the ability to use translation strategies. It can be seen that parallel corpora have broad application prospects and value in translation teaching. In recent years, the construction of parallel corpora has achieved certain results. In the mainland of my country, the construction of multiple bilingual parallel corpora has been completed. For example, the parallel corpus built by the 863 projects undertaken by the Institute of Computational Linguistics, and the English-Chinese bilingual corpus of Harbin Institute of Technology (400,000-500,000 sentence pairs) [5]. In addition, there are some specialized corpora, such as the English-Chinese parallel corpus, the parallel corpus, and the Chinese-English parallel corpus of Zhang's novels [6]. In 2011, the University of International Business and Economics adopted a school-enterprise joint research and development method to establish a "computer-assisted business translation teaching system", which is also a corpus-based translation training teaching system. Some colleges and universities in China have used this system to offer the course of machine-assisted business translation [7]. Although parallel corpora have begun to be used in translation teaching, its promotion is still relatively lagging. The number of corpora that teachers can obtain for free is limited, and the content and scale are not fully suitable for classroom teaching, especially basic teaching at the undergraduate level. Moreover, English for special purpose contains a large number of terms, and the style is distinctive, which is different from ordinary English. However, with the development of computer-assisted translation technology, it is no longer a problem for teachers to build parallel corpora according to different teaching needs.

This article describes the process of building a corpus and the key algorithms used, and designs an auxiliary translation software to implement the key algorithms. Among them, the algorithm of Chinese text duplicate checking combines the research results of Chinese linguistics, and proposes a method to extract the feature string of Chinese text. In the design process of auxiliary translation software, the auxiliary translation engine based on translation memory technology is regarded as a search engine based on full-text indexing, so as to provide the translation of the same or similar sentences for the translator's reference. Expands the concept of word segmentation in Chinese processing, so that the index problem of Chinese and other

character system languages can be handled in a unified way, and the language scalability of auxiliary translation software is guaranteed.

2 CONSTRUCT A MULTILINGUAL PARALLEL CORPUS

2.1 Machine Assisted Translation System

The translation accuracy of the machine translation system (fully automatic) has been hovering around 70% for a long time. The readability of the translation, the coverage of the system to language phenomena, and the robustness of the system, especially the openness, are not satisfactory. Society urgently needs large-scale processing of real texts (especially massive online texts), and machine translation systems are far from the expectations of today's society for large-scale real text processing. The idea of Computer Aided Translation (CAT) was born under this background. Compared with the fully automatic machine translation system, the machine-assisted translation system is a human-machine interactive system [8]. In this translation mode, the computer is responsible for the task of assisting the translator. It not only provides the translator with some knowledge of vocabulary, terminology, and phrase translation, but also searches for the translation of the same or similar sentences from the translated text, so that the translator can avoid the problem. Necessary repetitive work and efficient translation work. The important idea of computer-assisted translation (including translation memory technology based on translation memory technology and translation technology based on example mode) is to search for the same or similar sentence or phrase in the translation memory (bilingual alignment library) and the example mode library, and give a reference translation. Translators make full use of existing translation resources and try to avoid duplication of work. This auxiliary translation mechanism is particularly suitable for the translation of texts such as scientific and technological monographs, scientific and technological literature, product manuals, manuals, United Nations documents, etc., which are long and frequently repeated language phenomena and need to focus on the translation of new content.

The machine-assisted translation software of machine translation memory technology is based on this simple fact: Because of the huge amount of translation materials involved in the professional translation field, but the scope is relatively narrow, it is concentrated in one or several majors, such as politics, economics, and military. Translation memory technology starts from here, first of all to eliminate the repetitive work of translators, thereby improving work efficiency. The technical principle is as follows: the user uses the original text and translation to establish one or more translation memories. During the translation process, the system will automatically search for the same or similar translation resources in the translation memory, give reference translations, so that users can avoid unnecessary repetitive work and only need to focus on the translation of new content. Its essence is a kind of translation strategy in which the translator uses computer programs to participate in the translation process [8].

2.2 Corpus and Corpus

The English word "corpus" of the corpus comes from the Latin "body" and is used to refer to any written or spoken text. However, in contemporary corpus linguistics, corpus refers to storing certain text in a machine-readable form. Texts for specific purposes or language usage are used to store language materials. The research based on corpus is corpus linguistics. Scholars have expounded the definition of corpus linguistics from different aspects. Alfuraih's [9] definition of corpus linguistics is: the study of language based on textual materials. Corpus language mainly studies the construction and compilation of corpus, the processing and management technology of corpus, the use of corpus in language research, and the application of corpus linguistics in computational linguistics. In order to conduct research, a powerful corpus is needed to support it.

The corpus contains a wide variety of content, which can include written or spoken language, and can also include ancient or modern texts from multiple languages. At the same time, these texts can be entire books, newspapers, magazines, and so on. The content included varies with the type of corpus. The universal corpus refers to the composition of universal texts, which do not belong to the same text format and the same field. The process of building a database is shown in Figure 1.

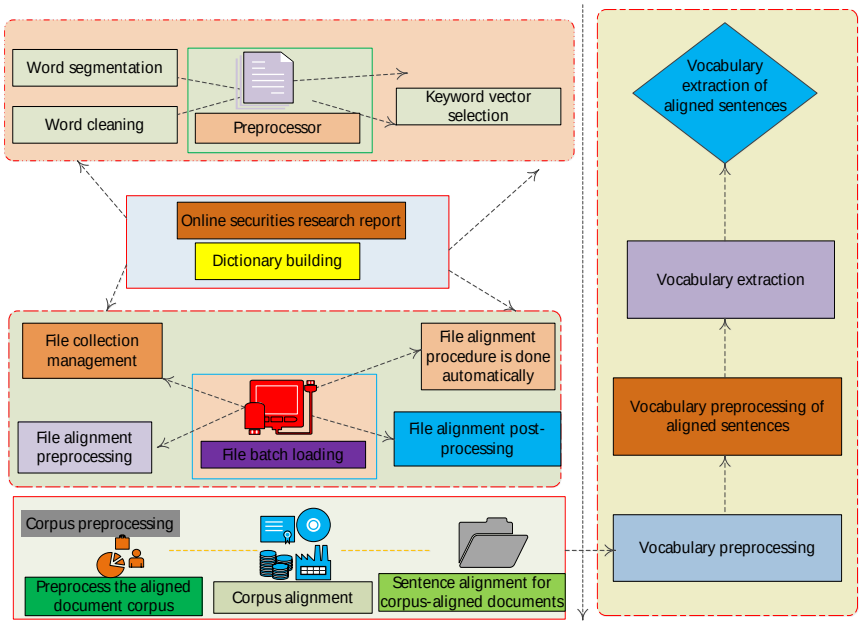

Figure 1: Database construction flow chart.

2.3 Semantic Similarity

The measurement of semantic similarity can refer to the vector model in information retrieval. The basic idea of the vector space model is to represent the text in a vector: $(J_1, J_2, J_3, ..., J_n)$, where J_i is the weight of the *i-th* feature item. You can select characters, words, or phrases as feature items. It is generally believed that selecting words as feature items is better than using words and phrases as feature items, and at the same time, the relative word frequency of words is used to represent the components of the vector. The most commonly used weight calculation method is the TF-IDF weighting method:

$$\exp(J_{ik} / f_{ik}) = N / n_i \tag{1}$$

where

$$f_{ik} = N_{ik} / \max_j(N_{jk}) \tag{2}$$

$$\exp(TF - IDF_i) = N / n_i \tag{3}$$

Among them, *N* is the total number of texts in the system, n_i represents the number of texts containing the word, f_{ik} represents the word *k*, the initial frequency in the text, and here refers to

the number of occurrences in the text. Then the normalized frequency of text and word *k* is f_{ik}, and the maximum value is obtained by calculating all words in text and *k*. $TF-IDF_i$ is the inverse document frequency of word *k*, J_{ik} is the weight of word *k* in the text . After obtaining the feature vector $(J_1, J_2, J_3, ..., J_n)$ of the article, the similarity between the articles can be calculated. Suppose the feature vector of document is $(J_1, J_2, J_3, ..., J_n)$. The feature vector of document is $(j_{1i}, j_{2i}, j_{3i}, ..., j_{ni})$, and the similarity is generally calculated by the cosine angle is the basis:

$$similarity(d_i, d_l) = 2(\sum_{k=1}^{n} j_{ki} j_{kl}) / (\sum_{k=1}^{n} j_{ki}^{\ 2} + \sum_{k=1}^{n} j_{kl}^{\ 2}) \quad (4)$$

3 BUILD A COMPUTER-AIDED TRANSLATION SYSTEM

After the library building tools are provided and used, the size of the corpus continues to increase, and these resources can be provided to auxiliary translation software.

3.1 Translation Process

The translation memory product will automatically "memorize" the translation of every sentence translated by the user. When translating a new sentence, search the translation memory, compare and match the paragraph with the unit of different translation in the memory, and select the translation unit that is closest to the original text, give a reference translation. Due to the relatively fixed vocabulary and sentence patterns in the professional field, when users have accumulated multiple memory banks of a certain size, they will encounter more and more repeated sentences, and the translation work will become easier and easier. The user admits the translation, or make some modifications, and the revised new translation will be automatically stored in the memory library for future use. Translation memory products also support network shared memory functions. In other words, when multiple people are translating at the same time, a translation memory can be shared through the local area network, and each online translator can call the work of others in real time.

3.2 Automatically Build a Library

For users who have accumulated a large amount of translation materials before using the translation memory product, the translation memory product will provide an automatic database building tool. The original text and the target text correspond to each other in sentence units. After the user has done some adjustments and proofreading, the tool will automatically generate a standard translation memory file. All the user's data can be recovered through this tool, so that the translation memory library can be established efficiently and quickly [10]. These libraries will be further supplemented and improved in the process of continuous use.

3.3 Terminology Management

Translation memory products also offer a very important function is terminology management. In the field of internet technology, each document contains a large number of professional terminologies, and the consistency of terminology translation is always one of the important contents of proofreading. The time-consuming and laborious work, and it is hard to guarantee that there will be omissions. Translation memory products use a term management tool (usually an electronic dictionary) to standardize all terms. The user only needs to create one or more standard term lists at a time (the list includes the original and translated terms). When using the translation memory system to translate, open the corresponding term list in the term management tool, and

it will automatically identify what are in the sentence are defined, so that the standard term translations can be given.

3.4 Automatic Typesetting

What people don't want to do is more than duplication of labor. The typesetting of electronic documents is also a headache for translators. Especially the localization industry has extremely strict requirements on the format of the translation, which must consist the format of the original document. In this regard, translation memory products are far ahead. Current translation memory products generally provide different formats [10]. The translation will automatically adopt the format of the original text, so translators don't need to bother with typesetting, as long as they concentrate on translation.

3.5 Overall Framework Diagram

According to the content described in Section 3.1, the framework of our auxiliary translation software is designed, as shown in Figure 2.

The design diagram consists of 4 parts, including user interface, auxiliary translation engine based on translation memory technology, multilingual parallel corpus, and import/export tools related to the resources in the corpus, and indexing tools. The corpus has been established, the other three parts belong to the content to be developed, and the instance-based auxiliary translation engine mentioned in the design drawing does not belong to the development content, so it will not be mentioned [11].

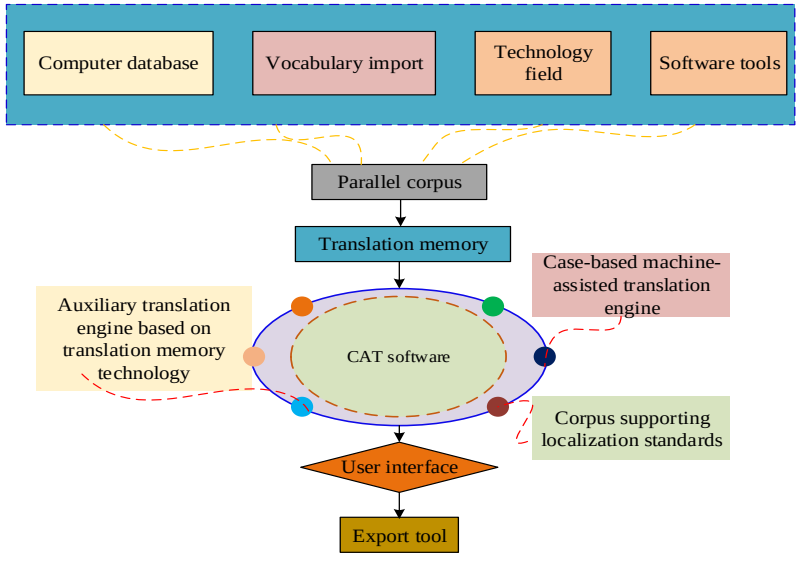

Figure 2: CAT software structure.

4 THE PERFORMANCE OF STUDENT TRANSLATORS' TRANSLATION ABILITY UNDER THE INTERVENTION OF PARALLEL CORPUS

This chapter mainly uses the student translator's percentile scores in each translation segmentation unit and the quantitative analysis of the software, and combines the student translators translations and the number of times they query the bilingual corpus during translation

to explore the actual translation process, the bilingual corpus Under the intervention, whether there is difference in the external performance level of the translators in the experimental group and the translators in the control group, and give a preliminary explanation for the reasons. The research findings in this chapter will answer the relationship between the use of bilingual corpus resources and the external performance of translation ability, translation direction, text type and language level.

4.1 The Overall Situation of the Intervention of the Bilingual Counterpart Corpus

This research mainly focuses on the difference between the subjective and objective performance of student translators' translation abilities after the intervention of the bilingual corpus. We did not allow student translators to perform simple single-sentence translation, and then instructed the student translators to find bilingual alignment corpora with a high degree of correlation with the single sentence to be translated, and then perform the translation. The research results of such an experiment can be said to be unnecessary to prove that the quality of the corpus determines the quality of the student translator's translation, and thus cannot provide us with the actual situation of the student translator's autonomous use of the bilingual corpus in text translation. Therefore, we provide student translators with relatively complete texts to be translated, pay attention to their self-inquiry of bilingual corpus in translation practice after they learn relevant corpus query knowledge, and study the changes in their translation ability under the intervention of bilingual corpus. In the four experiments, the number of times and the number of student translators inquiring into the bilingual corpus when translating each translation segmentation unit is shown in Figure 3.

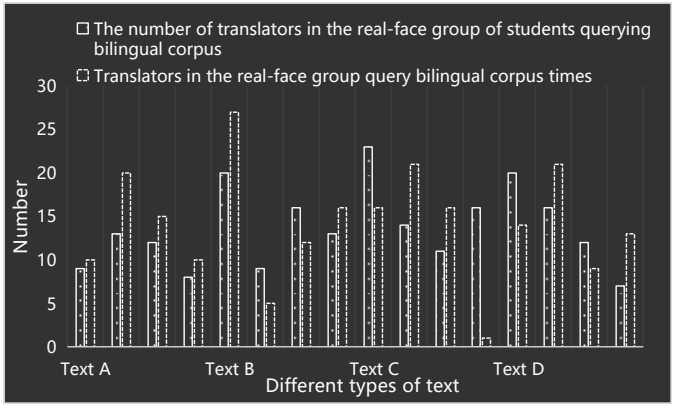

Figure 3: The overall situation of the bilingual corpus of translators in the experimental group.

We list the texts and reference translations of the four experiments and the corresponding average word counts of student translators. From Figure 4, it can be seen that the English translation of the Chinese text translated by the student translator is 30 to 40 English words different from the reference translation in the average number of words in each English translation. This shows that the student translator is in the process of Chinese-English translation. I tend to close the original text for word-to-word translation, and cannot translate as proficiently in foreign languages as professional translators.

In view of the large gap between the English translation of the student translator and the reference translation, we use the number of Chinese characters in the Chinese source text and the number of Chinese characters translated by the student translators in the experimental group as comparable data.

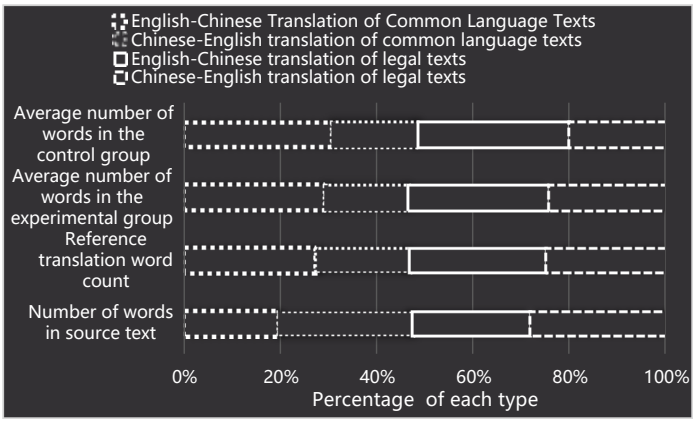

Figure 4: The number of words in the source text and the number of words in the translated text of the four experiments.

The average number of text queries to the bilingual corpus is standardized based on every 10,000 Chinese characters, and then a histogram of the number of times the experimental group student translators queries the bilingual corpus in four experiments is drawn, as shown in Figure 5.

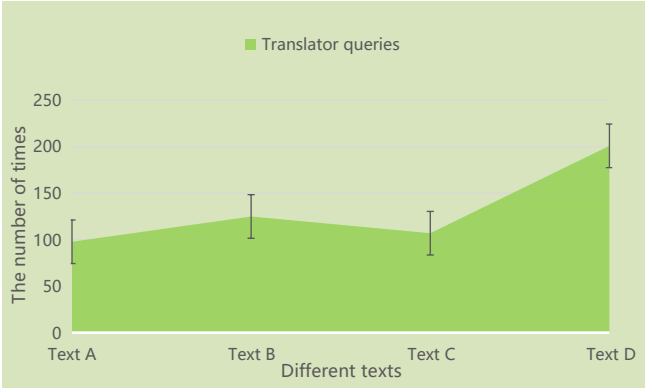

Figure 5: The number of times the student translator queries the bilingual corresponding corpus per 10,000 characters.

As can be seen from Figure 5, as far as the translation of texts in the register of special languages is concerned, when translating from Chinese to English, the average number of queries for the bilingual corpus of student translators is nearly greater than that of translations from English to Chinese. In the translation of common language register texts, the number of queries from Chinese to English is a bit more than the number of queries from English to Chinese, although it is not like in the special language register the gap in translation of similar texts is so obvious. Therefore, in terms of translation direction, student translators will make more use of bilingual corpus resources in the process of Chinese-English translation. This is because student translators cannot understand and express foreign languages as proficiently as native speakers. Big relationship. On the other hand, as far as the text category is concerned, the number of queries of the bilingual corpus of the student translators in the translation of the register text of the special language is higher than the number of queries in the translation of the corresponding direction of

the common language register text, which also means that for the translation of some special expressions (including professional terms and specific sentence patterns) in the register texts of special languages, student translators will rely more on the parallel alignment provided by the bilingual corpus translate the corpus.

4.2 The External Performance of Translation Ability in the Translation Process of Different Translation Directions and Text Types

This section mainly answers the first three questions in the first group of research questions. Specifically, it is to explore the translation practice of the experimental group student translators and the control group student translators in different translation directions and different text types under the intervention of bilingual corpus. Whether there are significant differences in the external performance of their translation ability, and give a preliminary explanation for the reasons.

We input the percentile scores of each translation segmentation unit of all 100 student translators who participated in the experiment in the four translation side tests for one academic year into the computer, and use the social science statistical software to perform independent sample t-tests. Then the average score of each group of each translation segmentation unit, the concomitant probability result of the parameter test of the difference between the groups, and the number of times and the number of people in the experimental group of student translators query the bilingual corpus are listed Figure 6. It can be seen from Figure 6 that among all the effective translation segmentation unit scores, only the fifth translation segmentation unit for the Chinese-English translation of common language texts and the seventh translation segmentation unit for the Chinese-English translation of special language registers Translation segmentation unit (the difference between the groups in the score reached a statistically significant level $p<0.05$). In addition, there are two translation segmentation unit scores that have differences between groups that are close to statistical significance, that is, the second translation segmentation unit for the English-Chinese translation of ordinary language texts and the English translation of special language domains. The second translation segmentation unit of the Chinese text, their concomitant probability, which is close to the set value of statistical significance level 0.05.

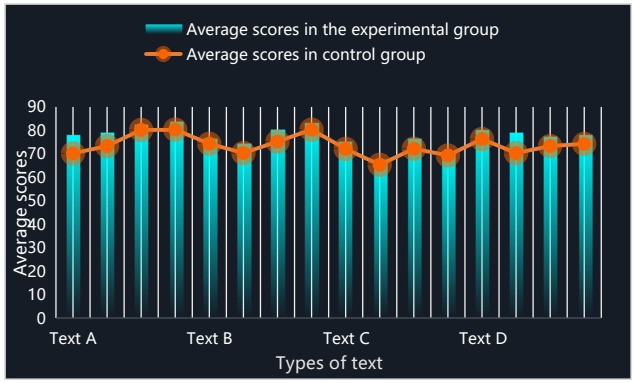

Figure 6: The external performance of the translators in the experimental group and the control group in different translation directions.

Based on the average scores of the two groups of student translators, as shown in Figure 7, among the above four translation division units, the other three are all students from other group. This shows that in the context of similar other influencing factors, whether or not the bilingual corpus is involved in the translation process of the above three translation segmentation units for

the student translators in the two groups is related to the quality of the translation. The average score of the translation of the student translator is higher than the average score of the student translator in the control group, and more than half of the students in the experimental group inquired into the bilingual corpus when translating the translation segmentation unit.The good or bad has a more obvious positive effect. In other words, the bilingual corpus has a more obvious impact on the external performance level of the translators in the experimental group. Therefore, we can make a preliminary inference that the bilingual corpus, as a reference resource for translators, can play an important role in certain parts of the text translation for student translators, and can improve the external performance level of their translation ability.

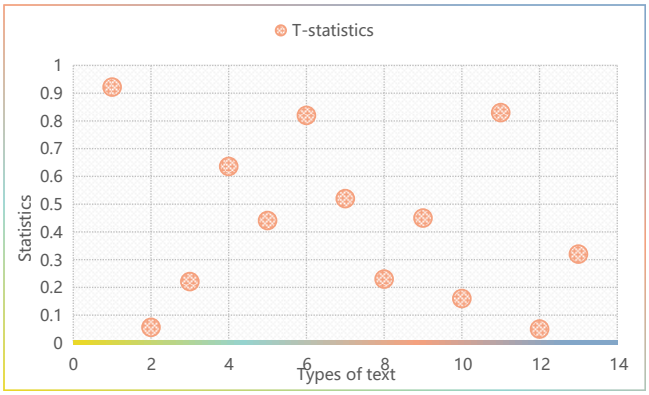

Figure 7: The T-test statistics of translation ability in the experimental group and the control group.

5 CONCLUSION

With the application and promotion of computer-assisted translation technology, the utilization rate of parallel corpus in translation teaching has been greatly improved. Modern education technology has promoted the pace of teaching method reform and also provided greater convenience for independent learning. However, this teaching mode has a certain dependence on hardware facilities such as computers and networks. If corpus retrieval is required in class, it is necessary to teach and study in the laboratory, which puts forward certain requirements on teaching conditions. Only by meeting the hardware conditions can the teaching model reform be effectively carried out. In addition, students are limited to language proficiency, and the analysis and use of parallel texts in the corpus still need teachers' guidance and help. Therefore, teachers should not rush to promote and use them in teaching, especially for undergraduates who have just gone through two years of language learning. Therefore, in the early stage of teaching, teachers should have an understanding of students' language level, translation ability and computer operation level, and should gradually guide students to use the corpus for learning after students can appreciate and judge the level of translation. The English-Chinese bilingual parallel corpus is widely used in English teaching. It can provide more English texts and examples for English teaching, and can provide teachers with more materials when teaching English. Through various methods, the parallel corpus can be effectively used in English teaching, which not only provides convenience for teachers, but also provides students with better English query materials. To a certain extent, it improves the quality of English teaching and promotes English. To sum up, the article firstly analyzes the software from the concept, main functions, and core technologies of computer-aided translation software; secondly, it looks forward to the technological change of computer-aided translation and the development direction of computer-aided translation. The construction of the system pointed out the direction; finally, the design and implementation of the English-Chinese auxiliary translation system based on the mixed strategy were analyzed.

Bingbing Pan, https://doi.org/0000-0001-9817-169X
Qiongfang Qin, https://doi.org/0000-0003-2433-2635

ACKNOWLEDGEMENTS

The phased achievement of the key project "Research on Translation Strategies towards the Tourism Culture of North Guangxi Ethnic Minorities" supported by Chinese Ministry of Education during the 13th Five-year plan (No. JKY10314).

REFERENCES

[1] Zhijie, Z.: A study on the computer aided English translation of local legal based on parallel corpus, In International Conference on Frontier Computing, 7(12), 2017, 241-256. https://doi.org/10.1007/978-981-10-7398-4_26
[2] Gao, T.; Wang, X.: The construction of computer-aided translation teaching platform based on borpus, In Journal of Physics: Conference Series, 1648(3), 2020, 32-59. https://doi.org/10.1088/1742-6596/1648/3/032059
[3] Zhang, J.: The influence of the computer-aided environment on the English translation teaching under the network teaching mode, In International Conference on Frontier Computing, 7(9), 2019, 1695-1700. https://doi.org/10.1088/1742-6596/1648/3/032060
[4] Umerova M.-V.: Parallel corpora in translation studies. Sciences of Europe. 1(2), 2018, 29-39. https://doi.org/10.2991/icemc-17.2017.71
[5] Lin, W.; Zhang, Y.; Liu, W.: A study on the foreign language translation ability and its training platform construction based on computer technology, In 6th International Conference on Education, Language, Art and Inter-cultural Communication, 1(1), 2020, 46-49. Atlantis Press. https://doi.org/10.2991/assehr.k.191217.076
[6] Zhang, B.: Construction and application of the English corpus based on the statistical language model, In International Conference on Frontier Computing, 7(1), 2018, 665-670. https://doi.org/10.1007/978-981-13-3648-5_82
[7] Lin, B.; Yi, P.-C.: On the construction and application of a platform-based corpus in tourism translation teaching, International Journal of Translation, Interpretation, and Applied Linguistics (IJTIAL), 2(2), 2020, 30-41. https://doi.org/10.4018/IJTIAL.20200701.oa3
[8] Frerot, C.: Corpora and corpus technology for translation purposes in professional and academic environments, Major achievements and new perspectives. Cadernos de Tradução, 36(5), 2016, 36-61. https://doi.org/10.5007/2175-7968.2016v36nesp1p36
[9] Alfuraih, R.-F.: The undergraduate learner translator corpus: a new resource for translation studies and computational linguistics, Language Resources and Evaluation, 54(3), 2020, 801-830. https://doi.org/10.1007/s10579-019-09472-6
[10] Tian, L.; Mu, Y.; Yang, W.: Designing a platform-facilitated and corpus-assisted translation class, In International Symposium on Emerging Technologies for Education, 8(1), 2018, 208-217. https://doi.org/10.1007/978-3-030-03580-8_22
[11] Hu, K.: Corpus-based study of translation teaching, In Introducing Corpus-based Translation Studies, 1(2), 2016, 177-191. https://doi.org/10.1007/978-3-662-48218-6_7