Numerical Analysis and Optimization of Computer-Aided English Reading Corpus

Ke Wang(✉)

Institute of Foreign Languages, University of Jinan, Jinan 250024, China
13573550013@163.com

Abstract. In order to improve the scientificity and effectiveness of English reading teaching, a computer-aided numerical analysis and optimization method of English reading corpus is proposed. This paper introduces the relevant knowledge and theory of corpus, discusses the theoretical basis of computer-assisted English teaching based on corpus, and preliminarily discusses the classification of teaching materials and the application of different content corpora in English teaching. In particular, a pioneering attempt has been made to extract language materials from film and television subtitle files for auxiliary teaching, and a film and television corpus with a capacity of more than 1.1 million words, which contains more than 110 English subtitles of film and television works, has been constructed. The results show that computer-assisted corpus can fully improve the efficiency of English reading teaching in senior high schools.

Keywords: computer aided · English reading teaching · corpus

1 Introduction

With the development of humanity, changes are taking place in our curriculum. In order to improve teaching and promote the development of all aspects of student performance, the teaching concept of taking students as the main body has been gradually adopted. at various colleges and universities. The teaching of English reading in colleges and universities in China is characterized by large numbers of people, limited resources and the difference in student achievement. Teachers cannot control the learning of every student. The application of the corpus provides a foundation for students to learn easily [1]. With the development of information technology, corpus is used more and more widely. Corpora refers to language materials collected for language research and stored in the form of electronic data. It contains a lot of real and natural language data. It helps language researchers to observe and explain the complexity of natural language from a new perspective. Common English corpora are British National Corpus, which pays equal attention to both written and spoken English, and its vocabulary capacity exceeds 100 million. The largest English corpus in the world today is the American Contemporary English Corpus, with a potential vocabulary of 450 million. In the context of the Internet, the use of corpus as a new teaching method in the teaching of English reading in high schools can help students learn more about the text, understand the meaning, understand

D. Kumar et al. (Eds.): IEIT 2023, AHSSEH 10, pp. 849–855, 2023.
https://doi.org/10.2991/978-94-6463-230-9_102

its cultural purpose, improve their reading ability, improve their thinking, and improve the basic foundation of English language teaching [2, 3]. Computer-assisted language teaching uses modern teaching methods to turn the traditional classroom into a flexible and interactive classroom, and has the characteristics of both pictures and text. It tightly grasps the attention and interest of students, and selectively presents the course content to students. Especially, it has the characteristics of visualization, diversification and vividness, which is conducive to the multi-directional input of language teaching, stimulate students' innovative consciousness, and build their own knowledge system.

2 Research Methods

2.1 The Concept of Corpus

With the birth of the computer and the rapid development of information technology, the concept of corpus has also undergone many changes. The main form of the corpus has evolved from original text to electronic text. Now what we call corpus is inseparable from computer output. Now it is generally believed that the corpus is about a large electronic machine with the capacity, which is created based on some speech content, using test models, writing continuous words and use text or rhyming words. In summary, a corpus is an example of the use of natural language to represent some dimension of language structure in some subject. This topic focuses on the translation of corpora from the point of view of linguistic research. From the point of view of the organizational form of corpora, we can think that corpora is a set of work that contains a lot of words stored in computers and software. This good work can realize the repetition, statistics, analysis and other uses of the corpus [4–6].

2.2 Theoretical Basis of Computer-Assisted English Reading

Researchers who hold the theory of learning generally believe that learners should actively seek support and explore the environment. In the process of interacting with the environment, they gradually develop knowledge of the outside world, in order to develop their own intellectual structure. The students' experience does not come from the teacher's understanding, but is created by the students themselves. Through social, family, school experiences and various materials from the educational process, students gradually integrate the knowledge into their existing knowledge structure. And create new stable knowledge structures through their own creative process, so that their own understanding can reach new heights. Constructivism emphasizes the student as the center. It requires students to transition from external beneficiaries of external support and internal knowledge to subjects of information processing and knowledge producers of technical content. Know. He wants teachers to change from acquaintances and workers to helpers, supporters and guides of students' construction of meaning. Teachers should use new teaching methods New teaching methods and new instructional design strategies, change the teacher's foundation, in terms of knowledge transfer.

Traditional teaching methods refer to students as learning materials [7]. We know that the design of the learning environment is an important part of the design of teaching, which is an important role in the success or failure of teaching. The model consists of 3 parts, as shown in Fig. 1:

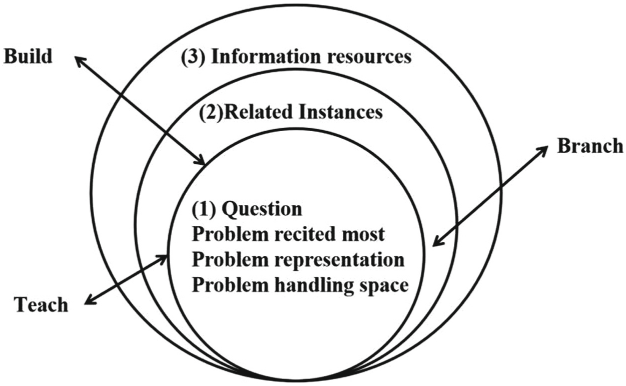

Fig. 1. Constructivism learning environment model

2.3 Application of Corpus in English Reading Teaching

For a long time, in some traditional application fields of corpus, little attention has been paid to the classification of corpus. The collection of corpus is either too comprehensive, taking into account various themes and styles, such as dictionary compilation; Or it is too specific, for example, some areas of linguistic research should even be specific to the study of the style of a writer's works. Of course, it is always reasonable to adopt the corresponding standards of corpus collection and classification in some application fields listed above, and it has been proved to be correct and effective by practice. However, the teaching corpus cannot completely copy the principles and standards of corpus collection and classification in other fields. It is very important to classify the teaching corpus. Because the classification of corpus is not arbitrary, but serves for specific teaching purposes and teaching applications. Should have its own characteristics and standards. Of course, the same category of corpus can have different specific application forms, but the category of corpus plays a considerable role in restricting and determining its application. At the same time, the proper classification of the corpus also provides a good idea for the collection of the corpus [8]. As shown in Fig. 2:

3 Result Analysis

3.1 Source of the Corpus

In the early stage, the source of the corpus was relatively single, mainly manual input of written printed materials. With the progress of information technology and the development of the Internet, the source of corpus has been quite rich. For the establishment of English teaching corpus, there are the following main sources of corpus: scanning written words, encyclopedias, English paper library, domestic and foreign news websites, English electronic publications, film and television subtitles, etc.

English news is closely related to current affairs and life, which can timely reflect the changes and development of contemporary English and represents the characteristics of contemporary English, so it is an important source of English teaching corpus. Collecting

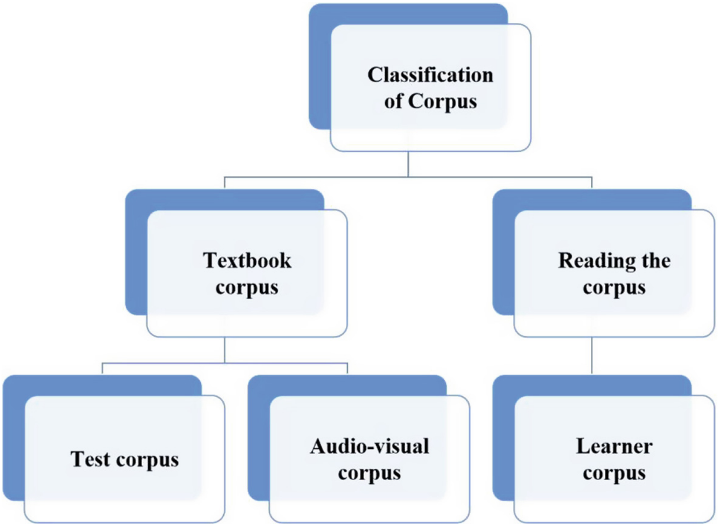

Fig. 2. Classification of corpus

Table 1. Famous English websites

New York Times	http://www.nytimes.com
Washington Post	http://www.washingtonpost.com
Los Angeles Times	http://www.latimes.com
Time	http://www.time.com
The Times	http://www.timesonline.co.uk
China Daily	http://www.chinadaily.com.cn
Pute English	http://www.putclub.com

English news text can use some well-known English news websites and English news learning websites at home and abroad. The website is shown in Table 1.

Using corpus indexing software and some other computer software tools combined with the above methods to collect and construct the teaching text corpus can realize many specific functions.

3.2 Use Statistical Information to Analyze the Difficulty and Discourse Style of English Reading Articles

Some measures of the corpus are correlated with many characteristics of the corpus, especially when many characteristics of the corpus are stable and consistent. Next, we analyze two small corpuses: the College Admissions English Comprehension Corpus and the National Postgraduate Admissions English Comprehension Corpus to illustrate this point. As shown in Table 2, Fig. 3:

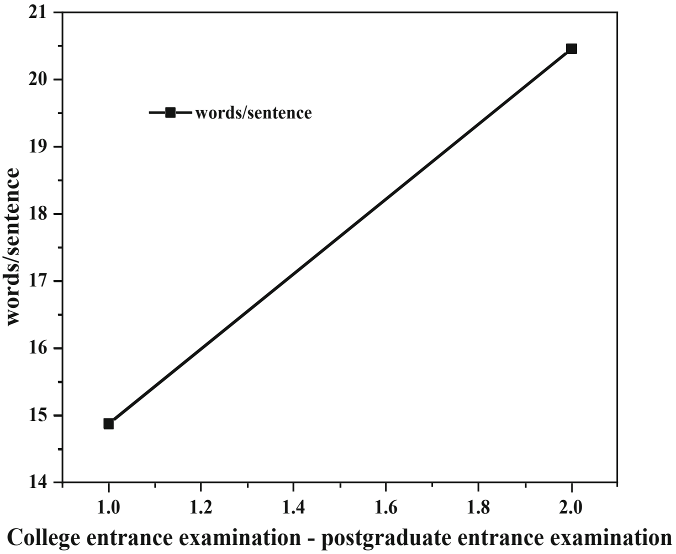

Fig. 3. A aph of average sentence length between two corpora

Table 2. Comparison of statistical parameters between two corpora

Parameter Kube	Average sentence length (words/sentence)	Number of symbols (tokens)
College entrance examination reading	14.8783	2260
Graduate Entrance Examination reading	20.4571	4285

From the analysis results in Table 2, qualitatively speaking, the average sentence length of postgraduate entrance examination reading is larger than that of college entrance examination English reading, the former involve more class characters than the latter. This is consistent with the predictions of most of us. From this example, we can see that some measures of the corpus can help teachers and learners to choose language learning materials for certain reasons. Although there is no strict and quantitative reference relationship between these statistical parameters and the text's difficulty, discourse style and other features, there is relative reference between some statistical parameters of different corpora, especially when the corpus is large enough.

3.3 English Film and Television Materials Combined with Film and Television Corpus to Assist English Reading Teaching

The combination of English film and TV materials and film and TV corpora to help teach English reading can create two important elements for learners, one is the audio

Table 3. Basic statistical parameters of self-built film and television corpus

Words (types)	31.311
Words (tokens)	1,150,070
Type/token ratio	36.73388
Characters	4396005

content-visual design of film and TV documents, and another is a summary of film and TV subtitles and corpora that are very easy to find and compare. This is consistent with instructional design that emphasizes the central role of content in construction. As an aid to English teaching, film and television materials should have two necessary conditions. First, there must be English dialogue; Second, there must be English subtitles. Many English movies in DVD format are provided with both Chinese and English subtitles. If you want to play movies and TV files in other formats, you can consider using MPC (Multi-Player Classic) with external subtitles to select Chinese and English subtitles. When using external captions, it must be noted that external captions should have the same name as the corresponding video files and be located in the same folder [9, 10].

We have collected the subtitles of 112 excellent English films and documentaries to build a film and television corpus with a wide range of topics, including classic business cards and some excellent films in recent years, so as to fully reflect the characteristics of contemporary English. The main parameters of this corpus are shown in Table 3:

4 Conclusion

This paper presents the knowledge and theory of corpus, and discusses the theoretical basis of corpus as computer-aided English reading, and preliminarily discusses the classification of teaching materials and the application of different content corpora in English teaching. In particular, a pioneering attempt has been made to extract language materials from film and television subtitle files for auxiliary teaching, and a film and television corpus with a capacity of more than 1.1 million words, which contains more than 110 English subtitles of film and television works, has been constructed. It is hoped that it can play a role in attracting jade, and arouse the interest and attention of educational technology circles in corpus-assisted English reading teaching, a field with great potential.

To sum up, with the development of information technology, corpus can fully improve the teaching efficiency of English reading in senior high schools, expand students' reading volume, broaden their reading horizons, improve students' ability to identify vocabulary usage, improve their thinking ability, and improve their ability guarantee for their lifelong development.

References

1. Harris, J. , & Rich, S. . (2021). Los Angeles Reading Corpus of Individual Differences: Pilot distribution and analysis. Proceedings of the Annual Meeting of the Cognitive Science Society, 84(2-3), 1-11.

2. Hu, Y. . (2021). Research on the Influence of Online Multimedia Corpus Indexing System on English Teaching in Colleges and Universities. CIPAE 2021: 2021 2nd International Conference on Computers, Information Processing and Advanced Education, 72(mar.), 316-322.
3. Mandelstam, S. , Knight, J. L. , Dean, R. J. , Richards, L. J. , & Barker, M. S. . (2021). Verbal adynamia and conceptualization in partial rhombencephalosynapsis and corpus callosum dysgenesis. Cognitive and Behavioral Neurology, 34(1), 38-52.
4. Wang, W. , Li, Y. A. , Ma, L. , & Qu, Q. . (2021). Research on Error Detection Technology of English Writing Based on Recurrent Neural Network. 2021 International Conference on Big Data Analysis and Computer Science (BDACS),22(31), 49-158.
5. Meng, Q. . (2021). The pedagogy of corpus-aided English-chinese translation from a critical & creative perspective. Theory and Practice in Language Studies, 11(1), 29.
6. Jacobs, A. M. , & Kinder, A. . (2022). Computational analyses of the topics, sentiments, literariness, creativity and beauty of texts in a large corpus of english literature.
7. Yvette, G. A. , Zhang, K. , & Jude, T. K. . (2021). Gelr: a bilingual ewe-english corpus building and evaluation. IJERT-International Journal of Engineering Research & Technology(8).
8. Zhang, W. , Wang, B. , & Zhou, L. . (2021). Analysis of text feature of english corpus with dynamic adaptive recommendation algorithm fused with multiple data source english language. Microprocessors and Microsystems(Issue: 6 special), 104075.
9. Collins, P. . (2022). Hypercorrection in english: an intervarietal corpus-based study. English Language and Linguistics, 26(2), 279-305.
10. Balabel, M. , Hamed, I. , Abdennadher, S. , Vu, N. T. , & Etinolu, Z. . (2020). Cairo Student Code-Switch (CSCS) Corpus: An Annotated Egyptian Arabic-English Corpus. The 12th Conference on Language Resources and Evaluation, 100(1), 1-4.