Research on the Creation of English-Chinese Database in the Perspective of Big Data Driven Learning

Fei Yuan

Nanchang Institute of Technology, Nanchang, Jiangxi 330000, China

Keywords: Corpus, Data-driven learning, Creation

Abstract: The corpus is a computer-aided language learning tool that truly reflects the whole picture of language in terms of pronunciation, semantics and pragmatics, and has great potential for promoting college English teaching. The corpus is actually based on the effective use of discourse fragments and text, to build a rich electronic library of text. The specific application of corpus data-driven learning mode in college English vocabulary teaching includes: the student's self-learning as the main process feature, based on the corpus to define the key vocabulary phrases, the real language as the main language input, emphasizing the learning process of exploration and discovery, advocate the use of inductive bottom-up learning methods and take concrete measures in teaching.

1. Introduction

At present, in the actual foreign language teaching, corpus technology has been widely popularized and further applied, and the data-driven learning model has also attracted people's attention. The data-driven concept emerged in the 1990s as a corpus-based language learning model. In general, it does not directly teach a language feature, but indirectly gives students a linguistic fact. Let students obtain the corresponding language rules based on the observation and analysis of the index.[1] On the other hand, data-driven learning is mainly to further support students to actively and actively think, so that they can independently observe, generalize and generalize the laws of language by virtue of the real language environment. Since the 1990s, the research and construction of corpora has received increasing attention from language researchers and learners. As a relatively new teaching model, corpus provides a reliable and realistic data resource for language research and learning. It is mainly based on a large number of real and natural languages to construct a highly representative language. Compared with the traditional teaching mode, the corpus has incomparable advantages, which are both highly innovative and open, and also have certain authenticity and flexibility.

2. The Creation of the Corpus

The creation of a corpus is time consuming and laborious. If you want to create such a corpus, it is recommended to use simple paragraph alignment or sentence level alignment. Even the division of sentences is sometimes difficult to define; after the division is too fine, standards and sizes are often difficult to grasp. For example, the tagging of part of speech, although the English and Chinese related annotation software has, the success rate of marking is more than 90%, but the remaining 10% processing is still a problem. This 10% may be distributed in various places in the corpus.[2]The discovery and correction will eventually require 100% manual intervention. Moreover, the segmentation and labeling below the sentence level can be used for quantitative analysis; for translation teaching, the sentence-level division has already achieved the goal. The author participated in the creation of the “Chinese-English Material World”of the School of Foreign Languages. In the process of creation, he explored a more pragmatic creation procedure. In the translation of teaching, I collected materials, organized and purchased software, and created some personal parallel corpora. The specific creation steps are as follows:

  DOI: 10.25236/ISMHI.2019.012

2.1 Electronic Text Material

English-Chinese bilingual materials can come from books or from the Internet. Internet resources, if they are in PDF format uploaded by book scanning, can also be regarded as books; but if they are editable TXT,DOC, PDF files, they can only be used as a reference. Scan paper text into PDF or JPEG format first. Chinese and English language electronicization requires different processing software in order to achieve optimal results. After the Chinese scan, it is saved as a JPEG file, and processed by the software to obtain a TXT text file with a recognition rate of over 96%.[3] After scanning in English, save it as a PDF file and use the software to get a TXT text file or a DOC file of Word with a recognition rate of 99%.[3[If the page is both Chinese and English after scanning, it will be saved as an image file. Use the scanner's own image software to cut and get Chinese and English image files. Then use “Omnipage” (omnipage can convert JPG or PDF files into editable files) to get the text. Import the word document, use the find and replace function, use wildcards as needed, remove sentence breakpoints, add paragraph marks, and more. Finally, the book is electronically processed and two documents are obtained: one is the original Chinese text; the other is the English translation.

2.2 Corpus Alignment At the Sentence Level

Since the creation of a parallel corpus at the sentence level, there are many softwares for bilingual corpus level alignment at home and abroad. But neither “Paraconc”nor “Textprocessing” can achieve satisfactory reliability and validity. If the corpus is too large, the computer is likely to crash. It is recommended to first convert the two document files in English and Chinese into “text conversion to form”, and “text division position” use “other characters” to obtain two table files of Chinese original and English translation respectively, and copy the table with fewer lines to In the table with a large number of rows, the sentence level initial alignment is completed. If you want to achieve a certain degree of precision and reliability, complete sentence segmentation and manual operation are essential. Take the English-Chinese parallel corpus as an example. If the English sentence corresponds to a Chinese translation, then there is no problem.[2]

2.3 Classification and Retrieval of Parallel Corpora

Corpus resources can be stored in the teacher's computer in word or excel format, or backed up into CD-ROMs, mainly used as translation aids, especially for English-Chinese, Chinese-English comparison and translation and translation. Of course, it can also be saved as an XML document for future corpus tagging and retrieval, as well as online publishing. The corpus classification can be classified according to its own needs.[4] Create separate documents for different authors and different styles. For example, it can be classified into parallel corpora of literature, science and technology, newspapers, and press conferences according to the style. It can also be classified according to different authors.

3. The Specific Application of Corpus Data-Driven Learning Mode in College English Vocabulary Teaching

3.1 The Main Process Characteristics of Students' Self-Learning

Nowadays, in the actual university foreign language teaching process, teachers are still the protagonists of the whole teaching work. Whether in the teaching or learning process, teachers have always had unquestionable authority. In the traditional English teaching process of the university, the teacher always controls the teaching work, and the authority of the teacher has always been implemented in the related activities such as classroom organization and teaching content. At present, with the rise of the corpus-driven model, students have gradually become the absolute center of teaching work.[5] This type of teaching mode mainly emphasizes the self-study of students and makes the students' personality characteristics fully vivid. The corpus-driven model requires students to self-manage, self-monitor, and self-assessment in the actual learning process, and on this basis, it has a positive and effective impact on the individual factors of students.

3.2 Defining Key Vocabulary Phrases Based on Corpus

The current situation of college English teaching requires that the number of English words that students master is relatively large, and new words are constantly emerging due to the changing times. How to quickly and reasonably enable students to master the huge amount of words with the times has become the primary consideration for teachers in the actual teaching process. On the other hand, in the process of learning words, some learners focus on the vocabulary specified in an examination syllabus or simply recite the syllabus. Although this method of memorizing words simply can cope with the test in a short time, it is difficult to truly expand the vocabulary and use the words flexibly. This is mainly because when the new words are recited, the learner does not know which words are the key words, but gives each word the same attention. [5]

3.3 Enter the Real Language as the Main Language

As we all know, linguistic materials, as a natural language, are not produced solely for the purpose of teaching. Teachers can give rich language application data to learners to the greatest extent, so whether it is a complete text or some words that can be used frequently, the index context, its data is relatively large, which is also Other teaching methods in the traditional sense are unmatched. However, for various reasons, the main language information sources of today's foreign language students are mainly textbooks and teachers' language output, which limits the real language of students' contact. This gives the corresponding development space for the corpus data-driven learning model. The real communication environment is one of the main sources of linguistic data that the corpus data-driven learning model can give students. It can effectively simulate the actual communicative context and facilitate the actual communicative activities. On the other hand, it can also trigger the students' knowledge about the language system, the real world and the discourse, thus perfecting an environment worthy of in-depth study.

4. Implementation of Teaching Application

4.1 Teaching Preparation and Selection of Experimental Subjects:

In order to ensure the feasibility of the implementation of the corpus English teaching model, the teaching is carried out in two aspects. The first is to conduct research on the learning strategies and willingness to learn for the pre-teaching objects; the second is to introduce the corpus-based data-driven teaching model. The author has never heard of 149 of the 149 students surveyed about data-driven language learning and found that 23 people have heard of the concept but don't understand it. After the teacher's introduction, 46% (69 students) expressed their willingness to participate in the teaching mode, 32% (45 people) are very willing to participate in the teaching attempt, 6% (8 people).[6] I think that I am not comfortable with the teaching mode. In addition, influenced by the teaching conditions and myself, 16% of the students think that they are not suitable for the teaching mode for the time being. According to the results of the questionnaire, 60 students who scored 110 or more in the college entrance examination were randomly selected from the experimental group and the control group, and the students in the experimental group finally decided to accept the teaching concept as a premise.

4.2 Research Methods

Based on the requirements of the college English syllabus, this study uses vocabulary teaching as a starting point, introduces the concept of data-driven language learning, and conducts a new model of data-driven language learning in teaching, for data-driven language learning and student self-learning. The vocabulary situation is observed and studied. The comparative test of vocabulary teaching focuses on understanding the effect of learner learning strategies on vocabulary learning. The purpose of questionnaire survey is to focus on whether the new model can promote learner self-learning, and further demonstrate the application of this model in college English vocabulary teaching.

4.3 Teaching Evaluation and Results Analysis

Through vocabulary examination, relevant data were collected and compared with the experimental teaching group and the control group for statistical analysis. After one semester of study, we conducted a unified test of the subjects in the experimental group and the control group, and conducted related questions on the learning strategies adopted by the learners in the learning process and the learners' responses to the new teaching model. The investigation. The answers to the questions are all based on the five-level scale method commonly used in foreign language teaching research. After SPSS is used to process the statistical data of the questionnaire survey, we mainly analyze the following two problems:

(1) What is the relevance of vocabulary test results to learner learning strategies? (Results are shown in Table 1)

(2) What is the effect of the new model on the learner's own self-learning? (Results are shown in Table 2)

Table 1 Statistics of Vocabulary Examination and Learning Strategy Survey Results

	Group	N	Mean	Standard deviation	Mean standard error	Mean value equation test				
Vocabulary test	Experimental group	60	62.63	10.17	1.85	T	df	Sig.	F	Sig
	Control group	30	56.94	8.37	1.53	2.36	58	0.022	1.28	0.26
Data driven learning mode	Experimental group	30	60.13	5.39	1.00	2.87	58	0.006	1.73	0.19
	Control group	30	55.47	6.93	1.26					

It can be seen from Table 1 that in the vocabulary test and learning strategy survey results of the experimental group and the control group, the variance test results show that the significance probability (Sig.) of the two groups is 0.26 and 0.19, respectively, both greater than 0.05, indicating that the two The variance of the group on the two variables is equal. Therefore, look at the data of the Equal variances assumed line as the result of the t test; from the table, the significant probabilities of the vocabulary test and the learning strategy are 0.022 and 0.006, respectively. Less than 0.05 indicates that after one month of actual classroom operation and self-practice after class, the scores and learning strategies of the experimental group and the control group are significantly different.

Table 2 Data-Driven Learning Patterns and Vocabulary Test Description Statistics

	Group	N	Mean	Standard deviation	Mean value equation test		
Vocabulary test	Control group	30	62.63	10.17	T	df	Sig.
					2.17	28	0.583
Total score of strategy	Experimental group	30	4.10	0.81	1.96	28	0.001

Table 2 shows that most of the students participating in the experimental teaching can use the corpus of the teaching materials to conduct vocabulary independent learning according to their own learning requirements and purposes; the vocabulary test and learning strategy survey results of the experimental group and the control group are tested by Levene homogeneity variance test. The results showed that after one month of actual classroom operation and self-practice after class, the scores and learning strategies of the experimental group and the control group were significantly different. In addition, most students who participate in experimental teaching can use the corpus of teaching materials to conduct vocabulary independent learning according to their own learning requirements and purposes.[8]The use of corpus data-driven learning mode in college English teaching can promote learners in the process of language acquisition. Independence, with the help of existing language resources to improve the interest of learning and the ability to learn independently, so as to achieve the effect of improving the quality of learning.

5. Conclusion

On the one hand, data-driven can motivate learners to think and train them by observing the target language patterns and analyzing their typical combinations-collocation and words. The ability to characterize lexical chunks; on the other hand, a large number of corpora and related teaching tasks are designed to help students actively participate in language learning activities and help to form effective teaching outcomes. Data-driven language learning is positively correlated with vocabulary autonomous learning. The data-driven teaching model has certain help and effect on improving students' interest in learning and motivating learners to actively participate in teaching according to individual differences of learners. However, this model emphasizes the guiding role of teachers and the active participation of students. As the teaching organizer, the overall role of the teacher in the whole process of classroom teaching is from the original single knowledge disseminator to the coordinator, promoter and guide in the teaching process, and the student's autonomy is strengthened. Therefore, teachers must proceed from the actual situation of students, taking into account the characteristics of students' physical and mental development and the actual ability of students to accept. While actively learning through information technology, multiple learning strategies and various forms of activities, students continuously transmit feedback information to teachers. Based on the feedback of students, teachers continuously implement random control of the teaching process to achieve a harmonious resonance state.

Acknowledgements

The authors acknowledge the 2019 University Humanities Project of Jiangxi Province, No.656: The Research on Ocean Environmental E-C Parallel Corpus Based on The Translation of Ocean's Review.

References

[1] Engelseth P, Wang H. Big data and connectivity in long-linked supply chains[J]. Journal of Business & Industrial Marketing, 2018, 33(8): 1201-1208.

[2] Wu X H, Xiao Y, Li L S, et al. Review and prospect of the emergency management of urban rainstorm waterlogging based on big data fusion [J]. Chinese Science Bulletin, 2016, 62(9): 920-927.

[3] Fu Y, Hao J X, Li X, et al. Predictive Accuracy of Sentiment Analytics for Tourism: A Metalearning Perspective on Chinese Travel News[J]. Journal of Travel Research, 2019, 58(4): 666-679.

[4] Liu W Y, Jin J. Big data in outcome prediction of cancer: Current landscape and perspective[J]. Chinese Science Bulletin, 2015, 60(30): 2836-2844.

[5] Jiang H, Qiang M, Fan Q, et al. Scientific research driven by large-scale infrastructure projects: A case study of the Three Gorges Project in China[J]. Technological Forecasting and Social Change, 2018, 13(4): 61-71.

[6] Wang F, Humblé P. Corpus-Informed Translation Studies–Looking Backward and Forward: A Chinese Perspective[J]. Studia Neophilologica, 2019, 91(1): 119-125.

[7] Chen Y, Zhao C, Zhang L, et al. Toward Evidence-Based Chinese Medicine: Status Quo, Opportunities and Challenges[J]. Chinese journal of integrative medicine, 2018, 24(3): 163-170.

[8] Jeaco S. Helping Language Learners Put Concordance Data in Context: Concordance Cards in The Prime Machine[J]. International Journal of Computer-Assisted Language Learning and Teaching, 2017, 7(2): 22-39.