Application and Sharing of Corpus in College English Teaching System Under the Internet Environment

Ning Wang(✉)

Heilongjiang Polytechnic, Harbin 150070, China
18645070503@163.com

Abstract. In recent years, with the rapid development of computer technology, corpora have played an important role in promoting the research of Chinese, English, and languages around the world. The construction of corpora has also attracted widespread attention at home and abroad. Corpus takes conversational language as the research object and establishes relevant discourse corpora, which helps people express the structural rules of language more formally and computationally.This paper introduces the corpus teaching system based on Internet technology to non English majors to verify the specific effectiveness of data-driven learning method for English vocabulary learning. This article takes 100 students from two classes in a certain university as the subjects. Through experiments, it is found that under the corpus teaching method of the college English teaching system, students spend 34% more time memorizing 10 English words on the same day than ordinary teaching methods. One week later, the number of students who can remember the same words using the corpus teaching method is 42 more than that using ordinary teaching methods. And the corpus data-driven learning method in teaching takes shorter learning time. The experimental results show that the corpus based teaching method of college English teaching system under the Internet environment has a good role in promoting English vocabulary learning.

Keywords: Vocabulary Corpus · Teaching System · Language Teaching · Vocabulary Memory

1 Introduction

The English course is a compulsory subject for Chinese university students, covering topics such as listening, speaking, reading, writing, vocabulary, grammar, etc. Such a complex curriculum structure results in students spending a lot of time learning English. To address this issue, teachers will integrate multiple course types into a comprehensive teaching approach, which not only improves class efficiency but also saves students' learning time. But this approach can also lead to blurring the boundaries of various course types, with vocabulary classes being the most typical.

From the perspective of students, due to the huge amount of English vocabulary knowledge and mechanical vocabulary learning methods, students often spend too much

J. C. Hung et al. (Eds.): FC 2023, LNEE 1131, pp. 82–88, 2024.
https://doi.org/10.1007/978-981-99-9299-7_11

time on learning [1]. In class, students mainly rely on the teacher's unilateral teaching, repeated copying outside of class, rote memorization, and mechanical memorization of vocabulary. This learning method is usually passive and receptive, lacking opportunities for active exploration and application of knowledge. Not only does it require students to spend a lot of spare time learning vocabulary, but it also weakens their productive abilities in speaking or writing [2]. Applying corpora to teaching is in line with the development of educational trends, using data-driven theory to guide students to explore independently, in line with the student-centered teaching philosophy, and to promote the integration of discovery based, autonomous, and exploratory learning concepts with big data information technology. As early as the 20th century, linguistic researchers have proposed presenting language data in the form of corpora [3]. The traditional corpus stage and the modern corpus stage are two stages in the development of corpora. The modern corpus stage is machine readable corpus, and the earliest machine readable corpus in the world is the Brown corpus, which belongs to the first generation corpus. The first generation corpus usually converts the corpus into electronic symbols and stores them in electronic computers [4]. Currently, researchers have conducted many related studies on the application of corpora in foreign language teaching, especially in the field of vocabulary teaching. However, through literature analysis, it has been found that the application of corpora in vocabulary teaching is mostly theoretical research, and the research subjects are mainly college students. In view of this, this article uses a corpus based vocabulary teaching method to conduct an empirical study on the ability of high school students to produce vocabulary, in order to provide data reference and theoretical basis for future students to improve their vocabulary production level [5].

This paper first combs the relevant research on English grammar and corpus assisted English teaching through literature, understands the direction and methods of existing research, on this basis, defines the connotation of English grammar and corpus that this research focuses on, and puts forward the innovation of this research. Secondly, the study went deep into the English grammar teaching classroom and conducted a teaching experiment with a period of two weeks. After the experiment, a questionnaire survey was conducted on the students, and the results of the questionnaire and the test were analyzed to test whether the corpus can promote English vocabulary memory and whether it can improve students' academic performance and the aspects it reflects.

2 Overview of Relevant Concepts

2.1 Classification of Corpus

(1) According to style [6]: corpus can be divided into written corpus and spoken corpus. Compared with spoken language, written language corpus is easier to collect and has a larger relative capacity. Like Brown Corpus. Oral corpora typically include transcribed text and audio files. The creation of the spoken language corpus is more than that of the written language corpus in the process of rewriting, and in this process, whether to transcribe pauses and whether to mark the length of discourse need to be discussed, which is more difficult than the establishment of the written language corpus. A typical example is the Cambridge version of the Wall Street Journal colloquial corpus with a British accent [7, 8].

(2) According to the corresponding method, a comparative corpus is one or more corpora composed of similar text content or different language texts similar in content, register, communication environment, etc. in a language comparison environment for comparative research. The Collins Birmingham University International Language Corpus led by Sinclair is one of the largest applied contrastive research corpora in the world today [9].

(3) Divided by time: The corpus can be divided into synchronic and diachronic corpora. A synchronic corpus is a corpus collected within a specific time range, used to horizontally compare certain language patterns within that specified time frame [10]. A diachronic corpus is a corpus collected over a long period of time, used to study the changes in certain language modules over a longitudinal period of time. The Contemporary English Corpus of the United States is a standard diachronic corpus that collects corpus from multiple fields within the United States over the past 30 years and is updated at least twice a year [11].

(4) According to the content and attributes of the collected corpus, it can be divided into heterogeneous, homogeneous, systematic, and specialized types. Heterogeneity refers to the lack of a fixed principle for collecting corpora, widely collecting and storing various corpora as is, with the most representative being the UK National Corpus [12]; Homogeneous type refers to collecting corpus of the same type of content. For example, Xinhua News Agency's news corpus; Systemic type refers to the corpus collected in advance based on predetermined principles and proportions to represent linguistic facts within a certain range [13].

2.2 Language Transfer

Transfer refers to the influence caused by the similarities and differences between the target language and any other acquired language. The language transfer theory attempts to explain what aspects of transfer occur in the process of second language acquisition, as well as the reasons for transfer. As for language transfer, comparative analysis hypothesis and connection theory have emerged in different periods to analyze it reasonably. Comparative analysis assumes that there are certain differences between the mother tongue and the target language, and indicates that these differences determine the difficulty level of learners. In contrast, the connectionist theory provides a more reasonable explanation for language transfer, which explains intralingual transfer and interlingual transfer [14].

3 Research Design

3.1 Test Method

Before and after the two-week corpus based vocabulary teaching in the experimental class, relevant data was collected and analyzed through test papers and writing essays. Testing is divided into pre testing and post testing. The test format, question setting method, and difficulty level of the pre test and post test are the same and both include two parts of the test content. The first is a test aimed at controlling students' vocabulary output. According to the controlled vocabulary production test designed by Laufer&Nation,

participants' controlled vocabulary size was measured [15]. The second is a test aimed at students' free vocabulary production. Mainly, students are asked to write essays on designated topics, and the vocabulary frequency profile test proposed by Laufer&Nation is used to input their essay texts into the RANGE32 software developed by Nation. The quality of vocabulary usage in the essay is analyzed by calculating vocabulary density, complexity, and diversity. After the pre test and post test, collect two test papers and compare the data using SPSS 22.0 to verify the changes in students' vocabulary production ability under the new teaching method intervention.

3.2 Questionnaire Investigation

The questionnaire survey method is mainly used to solve the third research question, to investigate students' attitudes towards the corpus based vocabulary teaching method. After the teaching is completed, questionnaires will be distributed to 50 students in the experimental class. The actual number of questionnaires distributed is 50, and 50 will be collected. After inspection and confirmation, 50 valid questionnaires will be collected, and the effective rate of paper collection is 100%. The filling out of the survey questionnaire should be done during class to avoid students being careless in filling out the questionnaire during breaks due to environmental or personal reasons. After filling out the test papers, collect and organize them, and then use SPSS 22.0 for analysis and processing to understand students' true attitudes towards corpus based vocabulary teaching method.

3.3 Corpus Analysis

The test content of this study is divided into two parts, including a controlled output vocabulary test and a free output vocabulary test, both of which include pre and post tests. For the controlled productive vocabulary test, according to the controlled productive vocabulary achievement test method, the test paper content needs to be prepared by selecting the test vocabulary from the designated scale. Since the subjects are college students, the vocabulary before and after the test is selected from the 2000 and 3000 scales; There are 40 questions in both the front and back test papers, and 20 words are selected from each scale for question writing. Each question is scored 1 point (the spelling and tense of the words are correct to score), a total of 40 points. For free productive vocabulary testing, it is mainly measured by the quantity and quality of relevant indicators of vocabulary used in students' compositions. The essay is presented in a propositional manner. The pre test questions are sourced from the first test after enrollment, and the post test questions are sourced from the final exam essay questions. The two questions are equally difficult and have the same amount of known information. After collecting essays, no correction is made, but the written text is converted into electronic version and input into RANGE32 software for vocabulary analysis and calculation.

3.4 Related Data Statistics Formula

For group spacing grouping data, first identify the group with the most frequent variable values, which is the group with the mode, and then calculate the approximate value of

Zhongshu according to the following formula. Lower bound formula:

$$M_0 = S + \frac{\Delta_1}{\Delta_1 + \Delta_2} * i \tag{1}$$

In the formula: represents the mode; S represents the lower bound of the mode; Represents the difference between the mode and other arrays; I represents the group spacing of the array.

Upper limit formula:

$$M_0 = U - \frac{\Delta_2}{\Delta_1 + \Delta_2} * i \tag{2}$$

4 Experimental Results

4.1 Frontal and Posterior Analysis

Table 1. Statistical analysis of pretest description

	N	Minimum	Maximum	Sum	Mean	Std. Deviation
Experimental group	50	15.00	35.00	1213.00	24.26	4.95206
Control group	50	13.00	34.00	1199.00	23.98	4.84237
Valid N	50					

Before the experiment began, the author conducted a controlled output vocabulary test in both the experimental and control classes. After the test was completed, the test scores were input into SPSS 22.0 for descriptive statistical data analysis, as shown in Table 1. All 50 test papers were valid. The minimum score for the experimental class is 15, the maximum score is 35, and the overall average score for the experimental class is 24.26. The lowest score of the control class is 13, the highest score is 34, and the overall average score is 23.98. The difference in overall average scores between the two classes is relatively small.

After the experiment, the control output vocabulary test scores of the experimental and control class students were input into SPSS 22.0. The descriptive statistical analysis results are shown in Table 2. The lowest score of the experimental class is 18 points, the highest score is 38 points, and the overall average score of the experimental class is 29.88. The minimum score of the control class is 16, the maximum score is 35, and the overall average score is 24.86. It can be seen that the average score of the experimental class after the experiment is higher than that of the control class.

Table 2. Post test description statistical test

	N	Minimum	Maximum	Mean	Std.Deviation
Experimental group	50	18.00	38.00	29.88	3.76146
Control group	50	16.00	35.00	24.86	4.78096
Valid N	50				

4.2 Survey Questionnaire Analysis

From Fig. 1, it can be seen that by conducting a survey on various data related to the application of corpora in English teaching among students in two classes, the results show that an average of about 72% of survey respondents believe that the corpus system has a certain degree of operability; About 86% of survey respondents believe that the corpus system still has a certain degree of practicality; About 93% of survey respondents believe that the application of corpus systems in universities is achievable and can meet daily evaluation needs. At the same time, an average of 96% of survey respondents believe that they are satisfied with the teaching results and process of this corpus system.

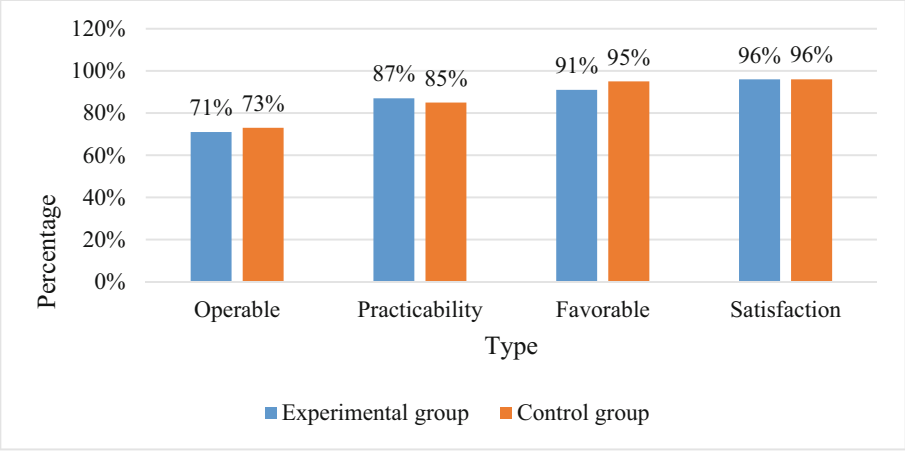

Fig. 1. Investigation Analysis Table

5 Conclusions

This study takes college students in a certain city as the research object, aiming to investigate the impact of using this vocabulary teaching method on students' vocabulary production ability. To ensure the reliability of the validation, another class using traditional vocabulary teaching methods was used as the control class. Through the analysis of student performance collected through the controlled output vocabulary test, it was

found that under the influence of corpus based vocabulary teaching, the controlled output vocabulary of the experimental class students was significantly improved, and both classes showed interest in the application of this teaching method, with a recognition rate of 96%.

References

1. Farahani, M.V., Khoshsaligheh, M.: Review of Liu (2020): corpus-assisted translation teaching: issues and challenges. Babel **67**(6), 845–848 (2021)
2. Maglie, R.B., Centonze, L.: Reframing language, disrupting aging: a corpus-assisted multimodal critical discourse study. Working Older People **25**(3), 253–264 (2021)
3. Meng, Q.: The pedagogy of corpus-aided English-Chinese translation from a critical & creative perspective. Theor. Pract. Lang. Stud. **11**(1), 29 (2021)
4. Merkulova, T.K., Galkina, A.A.: Personal approach with the usage of a corpus course schooling a foreign (ENGLISH) language as a method of evolving the cognitive curiosity of a high school scholar. Primo Aspectu **4**(44), 70–76 (2020)
5. Fernández, L.G.: Sources and steps of corpus lemmatization: Old English anomalous verbs. Revista Española de Lingüística Aplicada/Spanish (J. Appl. Linguist.) **33**(2), 416–442 (2020)
6. Khamis, N.: Corpus-based data for determining specialised language features. Int. J. Adv. Trends Comput. Sci. Eng. **9**(1), 36–41 (2020)
7. Curr, T.: Habeas Corpus, its versatility on both sides of the 'pond', and when right against remedy becomes quixotic. Global J. Comp. Law **9**(2), 220–244 (2020)
8. Levāne-Petrova, K., Auzia, I., Pokratniece, K.: Latvieu valodas apguvēju korpusa datu ieguves un apstrādes metodoloijas izstrāde. Valodu apguve problēmas un perspektīva zinātnisko rakstu krājums = Language Acquisition Problems and Perspective conference proceedings **16**, 299–309 (2020)
9. Noguera-Díaz, Y., Pérez-Paredes, P.: Teaching acronyms to the military: a paper- based DDL approach. Res. Corpus Linguist. **8**(2), 1–24 (2020)
10. Schäfer, L.: Topic drop in German: empirical support for an information-theoretic account to a long-known omission phenomenon. Zeitschrift für Sprachwissenschaft **40**(2), 161–197 (2021)
11. Sun, C., Wei, L., Young, R.F.: Measuring teacher cognition: comparing Chinese EFL teachers' implicit and explicit attitudes toward English language teaching methods. Lang. Teach. Res. **26**(3), 382–410 (2022)
12. Bose, P., Gao, X.: Cultural representations in Indian English language teaching textbooks. SAGE Open **12**(1), 517–542 (2022)
13. Mckinley, J., Rose, H.: English language teaching and English-medium instruction: putting research into practice. J. English-Medium Instr. **1**(1), 85–104 (2022)
14. Moodie, I., Meerhoff, L.A.: Using mock data to explore the relationship between commitment to English language teaching and student learning. Breast Cancer Online **53**(1), 121–123 (2020)
15. Khan, I.U., Rahman, G., Hamid, A.: Poststructuralist perspectives on language and identity: implications for English language teaching research in Pakistan. Sir Syed J. Educ. Soc. Res. (SJESR) **4**(1), 257–267 (2021)