The Effectiveness of Using Corpus Technology in College English Teaching

Shan Xia(✉)

Chengdu Polytechnic, Chengdu, China
279499548@qq.com

Abstract. Corpus analysis adopts computer assisted method to process electronic language database to reveal language operational models with statistic methods, which usually centered on phenomenon of probability. The research tries to explore ways to integrate corpus technology in college English and to investigate if corpus-based teaching is effective. It adopted constructivism approach to design the lessons with corpus technology. Students were asked to translate corpus produced basic word list in their filed, learn vocabulary with corpus's processed results and retell text with high-frequency word cloud. A quasi-experiment with 81 non-English major colleges students was carried out for 6 weeks. The results showed students' post test scores were significantly higher than the pretest scores. Therefore, it drew the conclusion that corpus-based college English was effective. This research shed lights on the possible use of corpus technology in college English to promote the vocabulary learning outcome of college students.

Keywords: corpus · effectiveness · word frequency

1 Introduction and Literature Review

With the development of technology, more technologies were introduced to the classroom to help facilitate the learning process and the integration of corpus technology in language classroom also starts to get attention from EFL (English as a Foreign Language) teachers.

Corpus means electronic language database. Corpus analysis adopts computer assisted method to process electronic language database to reveal language operational models with statistic methods which usually centered on phenomenon of probability. [12] Corpus analysis is often used in language research filed to determine the significant differences between two corpora or language features relation within one corpora with statistical methods like chi-square test, log-likelihood rate and correlation analysis etc., Since Tim Johns [3] and others put forward the data-driven foreign language learning method, corpus has become an important method of modern foreign language teaching. This method accords with the learning concept of constructivism, and emphasizes that students should "learning by doing". [8] This "learning by doing" approach, originally put forward by John Dewey in his work "My Pedagogical Creed [2], goes well with college students who have fairly good operational ability in the learning process. Therefore, EFL teachers need to lead students to work with both their mind and hand.

V. Balakrishnan et al. (Eds.): ICMEIM 2022, AHSSEH 6, pp. 539–547, 2022.
https://doi.org/10.2991/978-94-6463-044-2_68

In fact, Constructivism theory also holds that students are the main body of learning, and teachers should not only impart knowledge to students, but also help students cultivate themselves as active builders of knowledge. Corpus-assisted teaching attaches great importance to the cultivation of students' autonomous learning ability and tends to guide students to build their own knowledge system. In corpus-assisted teaching, students can think independently under the guidance of teachers, actively experience the original language materials, and consciously explore and discover the inherent laws of language [13].

In the previous research of corpus-based study of language learning, some researchers used corpus technology to do synonyms comparison work. For instance, by comparing the usage of synonyms of the word common and ordinary in LOCNESS and CLEC's St5 and St6 sub-corpora, it was found that there were many misuses by Chinese English learners, which could be attributed to improper teaching methods, lack of learning means, interference of mother tongue, cultural differences and insufficient knowledge of cultural background. [9] Corpus was also used to help translation students and professional translators to check more appropriate ways of rendition so as not to cause cultural translation problems. [5] To solve the problem of lack of real context in classroom foreign language teaching, video corpus-based teaching was adopted. The results of teaching experiments showed that the video corpus-based teaching significantly improved the learning effect. It could help students memorize vocabularies and understand the meaning of new vocabularies in a better way. [4] Corpus technology can further be explored to apply to the following fields in the future more often: compiling teaching materials with corpus; broadening the channels of learning resources with corpus; establishing relevant small-scale corpora; and constructing data-driven learning mode with corpus. [7] For data-driven learning mode of vocabulary explanation, Chinese teachers often focus on explaining words word collocation, semantic tendency and col ligation. [13]

For vocational college students, whose priority is to prepare for the future career, they need to collect the basic career skills and professional knowledge in their three college years. Consequently, college English should aim to allow students to learn some career English related to their filed and develop independent and cooperative learning spirit. Specialized vocabulary collection helps students to combine their profession with EFL learning, thus providing students impetus for language learning. The concept of lexical syllabus was first put forward by Sinclair & Renouf. They maintained that language study needs to lay emphasis on the most common words, the most important forms and the typical collocations, they claimed the essence of lexical syllabus was to extract high-frequency words from corpus. West [1] developed GSL or, General Service List, which were thought to be the 2000 most frequently-used word families to learners of English. Coxhead [1] developed AWL, or Academic Word List, which contains 570 word families for academics texts but not included in GSL. Experts and scholars from 15 Chinese universities nationwide have been organized to create 19 professional English corpora, with the size of each corpus ranging from 200,000 to 1,000,000 words and have been trying to build professional English word lists for college students [11].

Based on the literature above, the research aims to improve students' vocabulary learning, adopts the corpus-based technology to design college English and explores if students' vocabulary skills develop through the process. Therefore, the research questions are: how to design college English vocabulary with corpus technology? What effects does corpus technology have on students' vocabulary test scores?

2 Research Design

This research used purposive sampling to explore the corpus technology effectiveness among 81 first-year non-English major students in a stated-owned college in Chengdu. These students are from the three classes of tourism department school, majoring in tourism management, hotel management and civil aviation service.

To measure the effectiveness of the use of technology, this research tries to adopt the quantitative method and carry out a quasi-experiment to compare the students' scores before and after the treatment which lasts about 6 weeks, during which each class will have 4 sessions of classes in a week. The research chose two units to carry out the research, namely exposition unit and travel unit, the teaching of each of these two units lasts three weeks.

The integration of corpus technology is designed in three aspects. First, while the teacher explains new words to students, he tries to include word collocations based on frequency, typical samples sentences, video clip with the word and then asks students to make up sentences. In comparison, in the pretest unit of traditional class, the teacher used the existing material on ready-made PPT to explain. Second, the teacher would incorporate a word cloud of the text and a mind map to help students practice summarizing the key points of the text. Thirdly, for the treatment unit of travel, the teacher used corpus technology to extract two word list for the three classes to translate and understand (Table 1).

The two word list derived from two textbooks, one is Tourism English, which were used as textbook for tourism major students, and another is the English-Speaking Guides in Sichuan, which were used as a reference book for provincial Tour Guide Qualification English exams. The researcher extracted the text from each of the book, excluding the exercise and translation part. For the book of Tourism English, around 5000 words corpus were extracted to make 100-word lists and for the book of English-Speaking Guides in Sichuan, around 70,000 words were extracted to make the 100 word lists based on

Table 1. Lesson Design [drawn by the author].

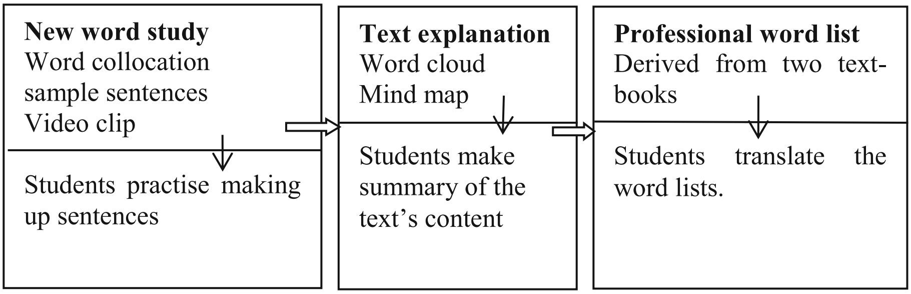
New word study
Word collocation
sample sentences
Video clip
Students practise making up sentences
Text explanation
Word cloud
Mind map
Students make summary of the text's content
Professional word list
Derived from two textbooks
Students translate the word lists.

Table 2. Two word lists based on frequency (the top five words processed from Antconc. The former table shows results from Tourism English; the latter table shows results from English-Speaking Guide in Sichuan) [from Antconc3.5]

Corpus Files	Lemma Types: 410		Lemma Tokens: 710	
Tourism English.txt	**Rank**	**Frequency**	**Lemma**	**Lemma word form(s)**
	1	24	china	china 24
	2	22	tourism	tourism 22
	3	13	airport	airport 13
	4	13	chinese	chinese 13
	5	12	beijing	beijing 12
Corpus Files	Lemma Types: 2537		Lemma Tokens: 6737	
Sichuan Tourism.txt	**Rank**	**Frequency**	**Lemma**	**Lemma word form(s)**
	1	120	dynasty	dynasties 14 dynasty 106
	2	106	chinese	chinese 106
	3	90	china	china 90
	4	81	sichuan	sichuan 81
	5	68	meter	meter 4 m 64

word frequency. In the “top preference”a stop lists of the 2000 most often used general words(GSL) and a lemma list were used to get a more accurate result. The corpus tool used in this part is Antconc, which is popular corpus tool (Table 2).

With these two-word lists in hand, the 81 students were asked to cooperate with each other to do the translation of these 200 words. They were also given the original text to confirm their correct understanding. So, class of tourism management was asked to translate word list 1 about tourism English. Class of hotel management were assigned to translated word list 2 about English Speaking guide in Sichuan. Class of civil aviation were given the task to check the translation by the former two classes. Overall, all the students put their effort into the translation or checking process. Finally, after the two word list with translation returned to the teacher for a final check, they would be distributed back to all the students for further memorizing work. In this kind of cooperative work, students get familiar with high-frequency words in their filed and can be more confident when they speak, listen, read or write English in their field.

The research instrument to collect data is 100-point vocabulary tests. Both the pretest and post test were designed in the same way. The tests are composed of three parts: 20 words dictation from the unit’s vocabulary lists (2 points for 1 words), 6 items of vocabulary blank filling exercise (6 points for each) chosen from the corresponding unit’s textbook exercise, and one reading passage from China daily website with four items of True or False questions (6 points for each).

Clearly, the word dictation and textbook exercise are of the same difficulty. To guarantee the difficulty of reading passage, corpus technology was used to verify it. A 2022 Beijing Winter Olympic Games report from China daily, labeled as passage 1 were

Table 3. Difficulty level of the two reading passages. [From Antword Profiler1.4]

Passage	LEVEL	FILE	TOKEN	TOKEN%	TYPE	TYPE%
1	1	The first 1000 words in GSL	205	67.88	73	56.15
	2	The last 1000 words in GSL	21	6.95	10	7.69
	3	The 570 words in AWL	33	10.93	22	16.92
	0	Words not contained in the above three lists	43	14.24	25	19.23
	TOTAL:	302				
Passage	LEVEL	FILE	TOKEN	TOKEN%	TYPE	TYPE%
2	1	The first 1000 words in GSL	272	84.74	107	76.43
	2	The last 1000 words in GSL	12	3.74	11	7.86
	3	The 570 words in AWL	1	0.31	1	0.71
	0	Words not contained in the above three lists	36	11.21	21	15
	TOTAL:	321				

chosen for the unit of exposition, as the one of the textbook materials is about winter Olympic Games. Meanwhile, an introduction about Jiuzhaigou Valley, a famous local scenic spot in Sichuan province, labeled as passage 2 was selected from China Daily website for the unit of travel, as one the text in the travel unit was related to scenic spot introduction. The two passages are of 302 and 321 words respectively, indicating a similar length. Then, Antword Profiler corpus tool was used to assess the difficulty level of the two passages. For passage 1, the type excluding from GSL and AWL is 25 (19.23%), denoting to the difficult word number. Meanwhile, for passage 2 the number of words excluding from GSL and AWL is 21(15.00%) (Table 3).

To balance the difficulty of the two passages and lower the comprehension burden of both the passages, some difficult words highlighted by Antquick Tool (an online corpus annotation) were annotated with Chinese translations manually by the researcher. Consequently, 11 English words from passage 1 were annotated, while 5 English words from passage 2 were annotated, reducing unfamiliar words to 14 and 16 respectively.

3 Data Analysis

The research adopted paired samples t-test to evaluate if the integration of corpus technology in teaching was effective for college students' language learning. It tried to assess students' memory of new words, their understanding for word use in context and their reading skills with these newly-learned words. SPSS25 were used to process the data collected before and after the treatment.

The research used the paired data of test scores from 81 students. Pretest scores were collected at the end the first three weeks when students completed their exposition unit. Post test were assigned to students at the end of the last three week after students finished

Table 4. of comparison for students' test scores (n = 81) [Self-made].

	Pretest on vocabulary		Post test on vocabulary			
	M	SD	M	SD	MD	T(80)
Test scores	49.96	17.98	66.67	78.25	-17.71	-2.113*

*p < 0.05.

their travel unit. They were given 25 min each time to finish the tests and were told the scores will be part of final exam scores, so everyone tries very hard to present a good score.

Table 4 shows that students pretest scores were significantly different from the post test scores (t (80) = −2.11, p < 0.05.). Inspections of the two groups means indicate that the average scores of pretests (49.96) are significantly lower than that of post test (66.67). The difference of means is −17.71 points on a 100-point test.

4 Results and Discussion

The overall results of corpus technology integration into the language classroom seems to be effective, as the post test vocabulary scores is significantly higher than the pretest scores, indicating students' progress in vocabulary study. This empirical study supports the previous scholars' view point that corpus-based teaching activate students' autonomous learning capability and independent thinking with the guidance of teachers and it provide students with original language material with which, they can make exploration and discoveries of the language [13].

In part of class, the teacher invited students to cooperate with each other to translate two word list and also visualized the word frequency with a word cloud picture, showing students a profile of the word list. Through the cooperation of each class in translating basic word list of tourism English, everyone feels that they are part of the teams for the whole project, which is closely related to their professional filed, so they are all highly engaged in their teamwork. Class monitor divided their class work to everyone and each of them were responsible for the translation of five words on average. The translation work returned to the teacher within one week and were passed on to the third class for a check and everyone was expecting for the final edition of their project. These word list involve the effort of all the participate with both hand and mind and make students to feel more confidence about language use in their field. It was an application of Dewey educational philosophy "learning by doing" [2] and as students construct their language skills and knowledge, they make progress.

The class extract two word lists from two textbook for students to carry out the translation project and students were satisfied to see the project outcome and keep the lists for further study in their spare time. According to the Education ministry of China (2021), for college students Emphasis should be placed on cultivating students' ability to communicate effectively in English in the workplace. This workplace English is difficult for college students in terms of professional vocabulary, followed by grammar and discourse. Each professional filed has its own specialized words and terminology, which lays the main barrier to understand. [11] Therefore, the two word list reflects students demand and proves to be effective for the rise of vocabulary tests scores on the unit of travel.

To help students to better understand the new words from the textbook, the teacher adopted corpus technology to make detailed study of the words before it is explained to students. The frequency of collocates, word cloud of collocates, semantic preference, col ligation as well as video clips were presented to students [10] providing students with rich context and innovative ideas. Through the process, students are more patient to learn and excited to see how different words are understood with big-data language technology. Based on the content, students are more actively in making up sentences after the teachers' explanation and produce more correct English sentence. This is probably because the collocation and structures for using the word were given. Which serve as guidelines to elicit more sentences.

To help students to retell the passage, they were given both idea mind map and high-frequency word cloud. Students prefer to use idea mind map than the word cloud produced by corpus technology. According to cognitive theory, this is probably because in language communication, people prefer to rely on meaning than on individual words. However, corpus technology can still be used by the teacher to control the difficulty level of passage for students to read. [7]

5 Conclusion

1) How to design college English vocabulary with corpus technology?

The lesson plan tried to integrate corpus technology in the EFL teaching. It incorporates the educational philosophy of constructivism approach and borrows the "learning by doing" concept from Dewey. The idea is that students construct their knowledge through experience, as active builders of language. They need to be both individual learners who develop autonomous learning skills as well as cooperative and interactive participants who contribute to the whole project of knowledge. Therefore, the treatment in lesson plan asked students to cooperatively translate the two high-frequency word list. One is taken from the tourism English textbook for college students, and another is derived from the English-Speaking guide in Sichuan. The Two 100-word lists were extracted from the textbook corpus by the teacher with corpus technology. The word list were distributed to three classes of students as cooperative project together with the original textbook corpus, with each student got 5 words to translate or check on average. The two words listed were submitted to the teacher for a second check and returned to students for further study.

During the class time, when explaining the vocabulary, the teacher tried to use corpus technology to produce word collocates, collocates word cloud, col ligation, semantic preference, video clip and invites students to make up or translate sentences into English. According to students' response, they felt a lot easier to produce sentences with the new word and made less mistakes compared with traditional teaching.

After the text study, the teacher used corpus technology to produce high-frequency words cloud to visualize the important words along with a mind map for students do the retelling work of the text. Due to students' habits, they prefer to use the mid map more than the word cloud, as they focus a lot on text meaning than the lexical items. However, the corpus technology help teachers to prepare appropriate reading material for students. If the reading material is too difficult, he may annotate difficult expressions with Chinese.

2) What effects does corpus technology have on students' vocabulary test scores?

As is in the previous data analysis part, the integration of corpus technology in college English teaching seem to produce a positive result for students' vocabulary learning, as the post test scores on vocabulary are significantly higher than the pretest cores. According to the tests design, students spelling memorization, words understanding and use, reading comprehension level improve to some extent. In fact, as has been mentioned, professional vocabulary are great obstacles for students to learn English in a specialized filed. [11] In practice students are rather weak to remember new words and use them in practice as is often reflected in classroom practice and final examination paper. So, the new way of teaching may shed lights on students' vocabulary learning.

3) Significance and limitations

The study is beneficial to classroom language learning. It proves preliminary evidence for the possible effectiveness of corpus technology in vocabulary teaching and how much progress language learners have made through the teaching reform. It also drew attention to student cooperative learning and specialized vocabulary for their future career.

The research also has its limitations. The 81 samples were chosen from only one college in Chengdu with the topic of exposition and travel. In the future, further study can explore if the corpus-based teaching can also be applied to other topics with students of different majors or ages to prove the effectiveness of corpus-based teaching.

Acknowledgements. This project is supported by 2021 Teaching Reform Project of Chengdu Polytechnic (JG-Y2148), Big Data Application Research Center for Higher Vocational Foreign Language Education of Chengdu Polytechnic(19KYPT06) and 2022 Scientific Research Project of Chengdu Polytechnic(22CZYR008).

References

1. Coxhead, A. (2000). A new academic word list. *TESOL quarterly*, *34*(2), 213–238.
2. Dewey, J. (1984). My pedagogical creed, 1897. *The philosophy of John Dewe*, 442–454.
3. Johns, T. (1991). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. *ELR journal*, *4*, 27–45.
4. Liu, Y., Han, L., Jiang, B., & Su, X. (2018). The application and teaching evaluation of Japanese films and TV series corpus in JFL classroom. *The Electronic Library*, *36*(4), 721–732. https://doi.org/https://doi.org/10.1108/el-09-2017-0193

5. Olalla-Soler, C. (2018). Using electronic information resources to solve cultural translation problems: Differences between students and professional translators. *Journal of Documentation*, *74*(6), 1293-1317. https://doi.org/10.1108/jd-02-2018-0033
6. West, M. (1953). A general service list of English words: with semantic frequencies and a supplementary word-list for the writing of popular science and technology. Longman.
7. Chang, H. (2012). Corpu-based English teaching resource analysis and suggeustions. *Heilongjiang Researches on Higher Education*(08 vo 30), 195–198.
8. Liang, M. (2009). Mini-test and its application in foreign language teaching. *CAFLEC* (03), 8–12.
9. Ren, P. (2008). Teaching and learning of synonyms: with common and ordinary as example. *Journal of PLA University of Foreign Languages* (04), 57–60.
10. Xu, X. & Xu, J. (2017). Four decades of corpus application to foreign language teaching in China. *Foreign Language Education in China (Quarterly)*(04 vo 10), 62–68+88–89.
11. Xu, J. (2022). Lexical syllabus design of higher vocational English and its instructional implications informed by the Principles of Four Usages. *Foreign Language Education in China(*01 vo 5),43–49+90.
12. Xu, J.(2019).*Corpora and discourse studies.* Foreign Language Teaching and Research Press.
13. Yu, X. (2015). Analysis of application prospect of language database in improving undergraduate English output capacity. *Heilongjiang Researches on Higher Education*(09), 166–168.