



A Study on the Application of a Corpus-Based Data-Driven Learning Method Utilizing an Online and Offline Blended Teaching Model in a College English Reading Course

A Case Study of “A Very Big Bang” in *Liberal Education Advanced English*

Ling Liang¹(✉), Kai-ying Chen¹, Shu-yi Huang¹, and Zhong-zheng Guo²

¹ School of Foreign Languages, Nanfang College, Guangzhou, Guangdong, China
16450818@qq.com

² School of Public Theoretical Courses, Guangzhou Donghua Vocational College, Guangzhou, Guangdong, China

Abstract. Data-driven learning, DDL, is a method of learning a foreign language based on corpus data which provides new ideas for the reform of foreign language teaching methods. This paper applies DDL corpus technologies when teaching a college English reading course. In this paper an online and offline blended college English course is modeled based on corpus DDL and it illustrates this process with a case study of “A Very Big Bang”, an article found in *Liberal Education Advanced English*. The tools used under the corpus based DDL online and offline blended teaching model are corpus retrieval and analysis tools, such as Ant Word Profiler, AntConc, Sketch Engine, as well as CAT (Computer Aided Translation) such as Tmxmall and SDL Trados. The former is used to analyze vocabulary difficulty, language characteristics, keywords, main events, dispersion plots, and their development. The latter can help students create their own translation memory by adopting machine translation plug-in components and translation skills. After applying this teaching method for three semesters the method has proven itself to be both positive and effective on students’ overall learning ability, learning satisfaction, innovative thinking, and learning initiative. The model’s effectiveness can be demonstrated from data collected and analyzed by SPSS.

Keywords: Corpus · DDL · Ant Word Profiler · AntConc · Sketch Engine · CAT · Online and Offline Blended Teaching · College English Reading · SPSS

1 Introduction

China’s foreign language teaching reform is currently facing two major trends. On the one hand, in 2020, the covid-19 virus caused a worldwide pandemic which compelled the Education Bureau to call for a “suspending classes without stopping school”. Since the impact of covid-19, the trend of online and offline hybrid teaching reform has become

increasingly common. Mainstream online teaching platforms (such as Chaoxing) have the advantages of massive free educational resources, data synchronization and powerful social functions. Therefore, how to effectively use various online platforms, learn from each other's strengths and how to complement one's weaknesses, and assist foreign language teaching is an important issue in teaching reform. On the other hand, big data technology such as the cloud computing and artificial intelligence in today's globalization and information age have also gradually affected the teaching of foreign language courses. Since 2018, the Ministry of Education launched six outstanding talent training programs and one basic discipline student training program. To implement plan 2.0 on the "Construction of First-class Course", that is, to create a new an interactive liberal arts program, we should integrate teaching with modern technology, break the closed technical knowledge structure, adhere to the ontology of the humanities, aim at cultivating cross-language and cross-cultural talents to improve the language service industry. In view of the aforementioned teaching methodologies, this paper, from the perspective of corpus-based data-driven learning model, tries to explore the integration of educational technology while teaching an English reading course. This paper demonstrates the use of long-term and effective online and offline teaching model to help students establish an efficient and effective learning model. The online-offline teaching method provides a paradigm for current college level English instruction.

2 Theoretical Foundation

2.1 Concepts and Features of DDL

Tim Johns proposed "Data-driven Learning" (DDL) in the early 1990s. DDL is a foreign language learning and teaching method based on corpus data. Students acting as researchers, observe and summarize language use phenomena based on a large number of corpus data, and self-learn word collocations, grammatical rules and pragmatic features. Teachers acting as guides, provide students with learning resources and guidance. This learning process of exploration and discovery fosters the ability of students' autonomous learning [3]. Johns and Rings (1991) believed that teachers should guide students to apply the concordance function of computer technology to retrieve the target language in classroom teaching, that is, to present all the contexts of a word or a string in the corpus, and design various class activities and exercises based on the index. ZHEN Feng-chao. Comments on four features of data-driven learning in "Corpus Data-Driven Foreign Language Learning: Ideas, Methods and Techniques", those being: student-centered autonomous learning, providing students with real data and an authentic learning environment, focusing on the process of self-discovery, and being a bottom-up interactive learning method [8].

2.2 The Integration of Corpus Research Method (CRM) and DDL

With the advancement of computer technology, the application of CRM is becoming more and more popular. CRM research methods have played an important role in linguistics since the 1980s. CRM in linguistics adopts the observation and processing of data,

which is embodied in the research process which comprises: extraction, observation, generalization, and interpretation [5]. In practice, researchers first use CRM to automatically process textual raw data and quantify language information. Next, researchers observe, describe and analyze the contextual information, meaning, and functional characteristics of specific languages according to the overall characteristics and trends of the data set. For example, analysis on the form, meaning and function of high-frequency words can directly reveal the characteristics of the language as used in a text. This method analyzes typical characteristics, idioms, grammar, slang, and other characteristics. Since word frequency or word probability distributions is an important property of a language's systems, this process of research is called frequency-based study or probability-driven study.

A corpus based DDL language learning model means that students comprehensively apply various proven corpus retrieval platforms to collaborate with teachers and peers while conducting data analysis on text and observing language phenomena. Through learning strategies, knowledge and acquired information, students summarize and analyze the language rules of a text, and finally internalize the gained knowledge.

2.3 Corpus-Based DDL Online and Offline Blended Teaching Model (CBM)

In order to better apply the online-and-offline blended teaching model to educators the three-stage theory of data-driven learning proposed by Tim Johns was utilized. Educators ask students questions to identify, generalize, and classify materials [4]. The author of this paper designed an online and offline blended teaching model based on the Chaoxing auxiliary teaching platform. The three stages of the DDL teaching method, those being the pre-class, in-class and post-class are integrated corresponding to teaching activities which are carried out online and offline. During class, the author refined DDL into three steps, which are pre-reading, active-reading and post-reading. The three-step method requires students to use corpus retrieval tools to observe and analyze class reading texts and extracurricular texts to achieve an automatic-inquiry learning method. (See Fig. 1).

In this paper an online and offline blended teaching model based on the Chaoxing teaching platform was applied in a classroom setting. This blended teaching model aims to provide knowledge dissemination and time management through a micro-service architecture. Teachers and students can find pictures, audio material, videos, forums, lectures and other multimedia and apply them to the blended teaching model. During the pre-class stage, teachers deconstruct the overall narrative structure of the preassigned text, design modular teaching content, and build a multi-dimensional evaluation system. Students form project teams for online learning during this phase of the blended teaching model. During the in-class stage, teachers design teaching activities utilizing Chaoxing. Attendance, video exercises, quizzes, discussions, group tasks, polls, and questionnaires are utilized on Chaoxing. Afterwards, students while interacting with teachers, learn thematic cultural knowledge, language knowledge, and reading skills. Teachers and students have team discussions about the overall learning process while carrying out projects. Teachers give feedback on the above-mentioned exercises and activities. During the post-class stage, according to the project evaluation criteria and weights, students perform self-evaluations in-groups and afterwards they also perform

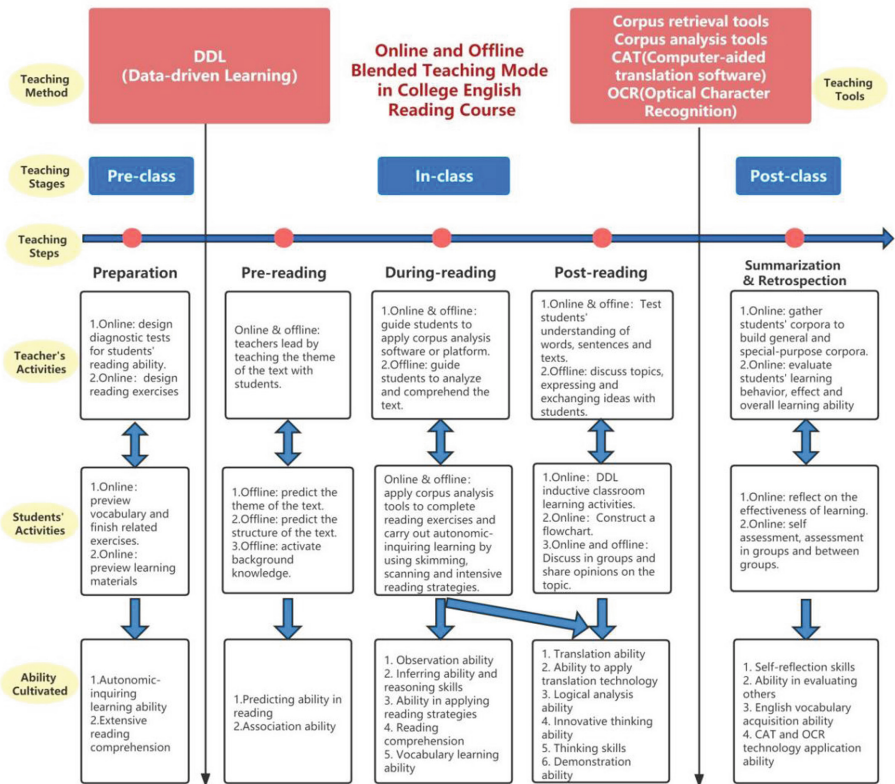


Fig. 1. Corpus-Based Data-Driven Learning in Online and Offline Blended Teaching Models in A College English Reading Course

group evaluations. Students submit questions while utilizing the online platform while teachers give feedback both online and offline.

The tools used under the corpus based DDL online and offline blended teaching model are corpus retrieval and analysis tools, such as Ant Word Profiler, AntConc, Sketch Engine, as well as CAT (Computer Aided Translation) such as Tmxmall and SDL Trados. The former is equipped with its own reference corpus, which compares and automatically analyzes natural language usage. The former also provides information on various aspects of vocabulary usage, such as word difficulty, index, collocation, word meaning, and synonyms. The latter is also equipped to improve the efficiency and quality of translation by using translation memory, term base and fuzzy matching to ensure the fast speed and consistency of bilingual translation. The former can help students skim, scan, speedread and perform intensive reading so as to gradually explore and learn independently. For example, students can grasp the main idea of an article by retrieving the high-frequency words of the text by searching for positioning words. Students can further query the context information provided in the index row by retrieving and querying grammatical construction and gapped construction. During this process,

students' cognitive ability, including reading comprehension, interpretation, vocabulary analysis, evaluation, reasoning, and interpretation have all been trained and improved. The latter can help students create their own translation memory by adopting machine translation plug-in components and translation skills so they can further discuss what they have learned about the text in the article.

3 A Paradigm Design Under the Corpus-Based DDL Online and Offline Blended Teaching Model

3.1 Pre-class Stage - Preparations

3.1.1 Preparation of CBM Resources

- Teachers create a text specific CBM for students to read before class. The difficult English text “A Very Big Bang” was selected from the textbook *Liberal Reading-Advanced English* as the starting point for the CBM. Students convert the text into TXT format while utilizing English Web 2020 (enTenTen20) as a reference starting point for the CBM.
- Students measure the readability of English text by Ant Word Profiler 1.5.1w (Windows) (2021 version);
- Students apply the retrieval software AntConc to analyze the distribution of high-frequency words in the full text, visualize the data and to understand the description of scientific and technological terminology intuitively.

3.1.2 Training Tools for the Use of CBM

Teachers upload CBM retrieval tools, analytic tools, computer-aided translation software, video explanations, software downloads, and other instructions from the Chaoxing platform. Students are encouraged to attempt to master various electronic tools before class to facilitate the teacher's guidance while participating successfully during classroom activities.

3.1.3 Design of a Corpus-Based DDL Classroom Model

- Teachers are responsible for designing reading ability diagnostic tests and creating online exercises while instructing students to use DDL analytics software or platforms testing students' comprehension of words, sentences, and articles. Teachers are also responsible for summarizing students' language skills while building general and specifically tailored course material. Teachers also interact with students offline while guiding students to apply corpus analysis software to analyze and understand articles.
- Student are responsible for forming a 5–7 person team to preview new words online, complete relevant pre-class exercises, self-learn relevant materials, analyze articles, complete reading exercises with the help of DDL analysis tools, learn new reading strategies, complete various DDL inductive classroom activities, draw flow charts, discuss questions in groups while expressing opinions, review and reflect on learning effectiveness, performing self mutually and inter-group evaluations, discussing the topic of the article and exchanging opinions, and to present the project results.

3.2 The In-Class Stage: The Classroom Practice Process Under the Corpus-Based DDL Model

During this process, teachers guide students to observe the corpus retrieval results, discuss and analyze the lexical difficulty and language characteristics, plot distribution and develop a context of the main events and the language used in the selected scientific and technological articles. The in-class stage comprises:

- Applying the word frequency table (Wordlist) and the text readability measurement software Ant Word Profiler, to analyze the lexical difficulty and language characteristics, while browsing the index lines by the search function (Concordance) to understand the language characteristics of the selected articles.
- Comparing the reference corpus, while making a keyword list by selecting the most important words in accordance with a key degree ranking mechanism while inferring the main events of the scientific and technological texts combined with the classification of a words meaning within a text.
- Finding the approximate position of search terms in the text via a Dispersion Plot while analyzing the text such that the plot is correlated by the function of indexing the original text file. This is done so as to aid the student's understanding of the developmental context of events while applying various reading strategies.
- While utilizing the online alignment function of the translation memory retrieval and exchange platform, a corpus translation memory record will be established to ascertain students' knowledge of the vocabulary used within the context of the text.

During the whole process, students need to simultaneously complete the reading exercises designed by the teacher on the Chaoxing platform (This article omits the discussion of reading exercises).

3.2.1 Analysis of Vocabulary Difficulty and Language Characteristics

In order to visually observe the lexical difficulty of a text, teachers should guide students to apply the English text readability measurement software called, "Ant Word Profiler" to rank the vocabulary in the selected texts. Ant Word Profiler's file viewer and editor tools allow users to view individual files and to highlight different ranking of words within the file by color coding. Ant Word Profiler also shows the student's overall knowledge across different vocabulary levels. By default, three levels of basic vocabulary lists are included in the program. Each table corresponds to each level of vocabulary. The first document corresponds to the first-level vocabulary, which refers to the 1,000 most frequently used words in the English language. The second document corresponds to the second-level vocabulary, and the second-level vocabulary comprises the second most frequently used words in the English language, that being, the 1001st to 2000th words. The coverage rate of first- and second-level words in general language discourse is 81.3% [6]. The third document corresponds to the third-level vocabulary, which consists of 570 academic words, which are the most difficult to master and this comprises 85% of the technical English words [1]. The degree of lexical difficulty in "A Very Big Bang" is shown in Table 1.

Table 1. Lexical Difficulty in “A Very Big Bang”

Word	Token	Level Coverage%
Level 1	1946	74.36%
Level 2	164	6.27%
Level 3	111	4.24%
Level 0	396	15.13%
Token Coverage	2617	84.9%

As can be seen from Table 1, the coverage rate of the first- and second-level vocabulary in “A Very Big Bang” is 80.63%. Compared with the coverage rate of 81.3% of general texts, students can be guided to draw the following two conclusions: First, 3/4 of the vocabulary used in this technical text are high-frequency words, so the reading difficulty is slightly lower than that of general English texts. Second, it is easy to read which is typical with the reading level of a second-year college junior. In order to verify the students’ comprehension of the text, teachers apply the Wordlist function in the AntConc retrieval software to generate a table of the word frequency. The table helps students to understand the nuances of frequency distribution of the words used in the text such as density, diversity, recurrence rate, and high-frequency semantic words [7]. The retrieval results show that there are 2620 tokens in this text, including 887 types. Table 2 shows the top 30 word frequency after articles and prepositions were removed.

In the classroom, teachers can guide students to observe and analyze Table 2 in groups while discussing and summarizing the distribution within the group. For example, relative conjunctions such as “and”, “that”, “as”, “but” in the list means this text uses: more complex sentences, attributive clauses, object clauses, and adverbial clauses.

Teachers will further guide students to browse the index lines of the original text of these four words (Fig. 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 and 12) to understand the language characteristics of the article by concordance. For example, with the knowledge of grammatical rules and sentence meaning, students can infer that “and” and “but” are mainly used to connect several natural phenomena that occur one after another. Within the meaning of a progressive or changing state of the events “that” is mainly used as attributive clauses and object clauses. With words that express past tense, past perfect tense and past future perfect tense, words such as “followed”, “would have done”, “had done” are used. Students can conclude that the point-in-time the author chooses to describe an event such as at the moment the asteroid hit the earth, follows as time goes backward, as is shown in the author’s use of words indicating perfect or past tense, such as “have been”, “would have been”, “had”, “was”, “were” and the noun “years” which appears mostly in the form of “millions of years ago”. In summary, the occurrence of an event can be traced back in time. “As” in the text is used to express time and effect by indicating that the author tends to describe the reason why a certain natural phenomenon happened. The author also uses metaphors and examples to describe the event more vividly. The verbs “was” and “were” are usually followed by past participle, indicating that many

Table 2. The Top 30 Words Most Frequently Used in “A Very Big Bang”

Rank	Frequency	Word	Rank	Frequency	Word
1	67	and	16	15	but
2	51	to	17	14	million
3	40	have	18	14	species
4	37	that	19	14	this
5	26	it	20	13	had
6	26	would	21	13	they
7	22	as	22	12	atmosphere
8	22	impact	23	10	any
9	17	earth	24	10	began
10	17	was	25	9	ago
11	17	were	26	9	ground
12	17	years	27	9	rock
13	16	km	28	8	creatures
14	15	all	29	8	dinosaurs
15	15	been	30	8	like

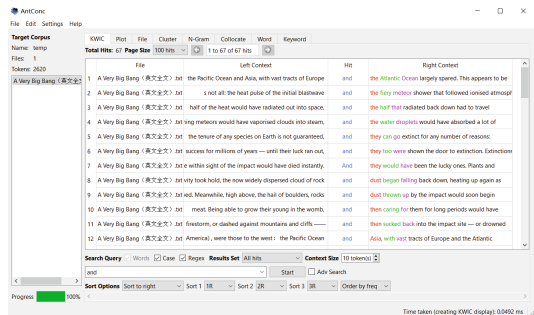


Fig. 2. The Original Index Row of “And” Used in “A Very Big Bang”

sentences are used in passive voice, which is typical of most scientific and technological articles.

3.2.2 The Analysis of Keywords and Main Events

A Keywords list refers to a list of words of significant differences generated by comparing a complete and continuous text within a larger reference corpus. The difference between the observation corpus and the reference corpus can be found from the key sources of the keywords. The keywords list can objectively reveal the general content and development

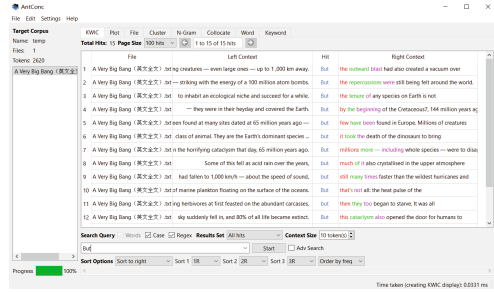


Fig. 3. The Original Index Row of “But” Used in “A Very Big Bang”

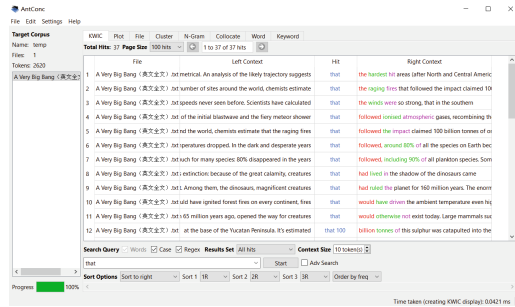


Fig. 4. The Original Index Row of “That” Used in “A Very Big Bang”

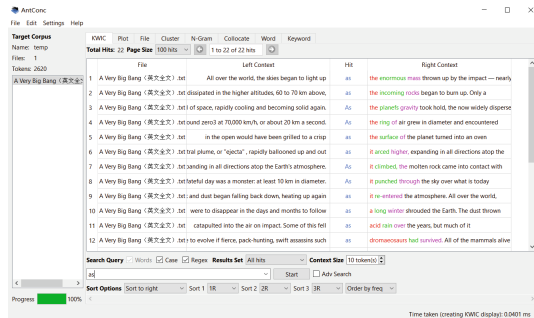


Fig. 5. The Original Index Row of “As” Used in “A Very Big Bang”

of the events within the text [2]. The author uses “A Very Big Bang” as the observation corpus, and English Web 2020 (enTenTen20) as the reference corpus, which has a total of: 44,968,996,152 tokens, 2,099,033,556 sentences, 789,418,319 paragraphs, and 81,323,314 documents. The author uses the Keywords function in Sketch Engine, a corpus search software, which can list the key sources compared with the relevance of reference corpus to facilitate the extraction of the top 50 keywords, as shown in Table 3.

Teachers first guide students to observe the top 5 keywords “blast wave”, “vaporize”, “cataclysm”, “planet”, “shockwave” and various other keywords indicating bad results,

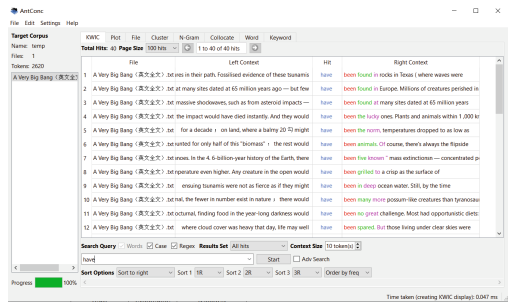


Fig. 6. The Original Index Row of “Have” Used in “A Very Big Bang”

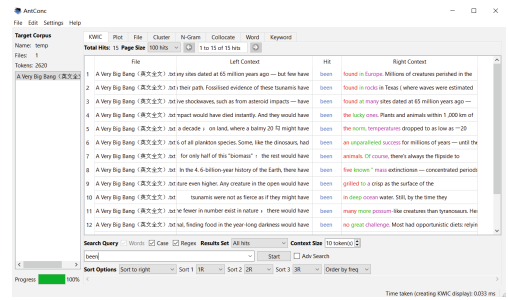


Fig. 7. The Original Index Row of “Been” Used in “A Very Big Bang”

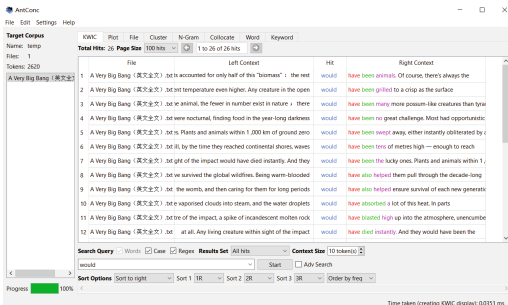


Fig. 8. The Original Index Row of “Would” Used in “A Very Big Bang”

such as “cataclysm”, “cataclysmic”, “hellish”, and “extinction”, in order to grasp the core events described in the article, such as the shockwave caused by the asteroid hitting the earth brought great disasters to the earth.

Next teachers guide students to observe other special keywords, categorize them according to part of speech or meaning, and speculate on several important events that will follow in time. For example, the geographical location of the asteroid impact can be inferred from the location named Yucatan. The words like “Triassic”, “tyrannosaur”, “dromaeosaur”, and the related words like “mammal”, “wombat”, “walrus”, “herbivore”,

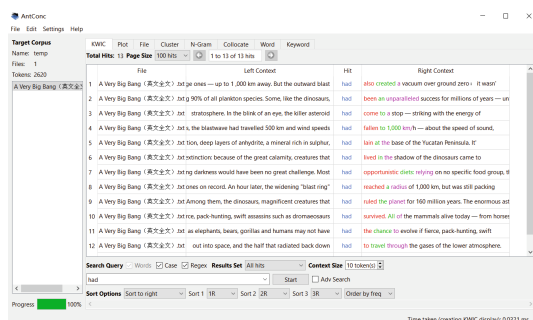


Fig. 9. The Original Index Row of “Had” Used in “A Very Big Bang”

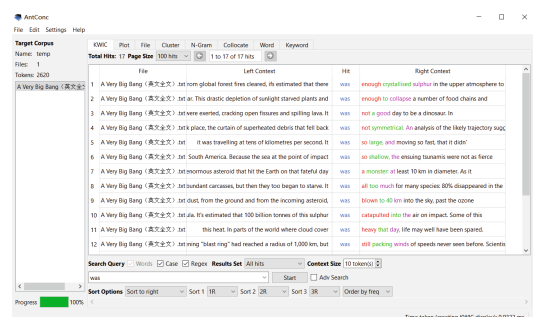


Fig. 10. The Original Index Row of “Was” Used in “A Very Big Bang”

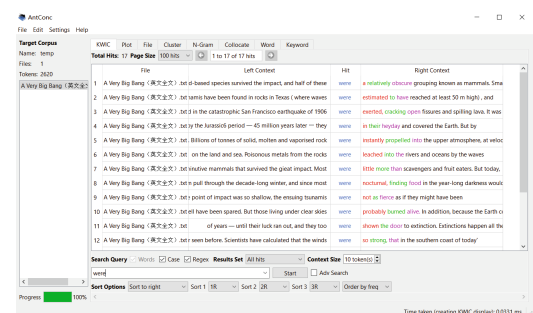


Fig. 11. The Original Index Row of “Were” Used in “A Very Big Bang”

“possum-like”, “warm-blooded”, all indicate that this event will bring a devastating blow to dinosaurs, and the fate of warm-blooded mammals will forever change. The result that an asteroid hitting the earth will bring a series of natural disasters, such as volcanic eruptions, seas of fire, hurricanes, heat waves, hurricanes and tsunamis which will devour all life on the planet, can be inferred from the following: chemical-related words such as “Sulphur”, “anhydrite”, “nitric”, “photosynthesis”; physical related words such as “molten”, “vaporize”, “crystallize”, “pulverize”, “fossilize”, “ionize”,

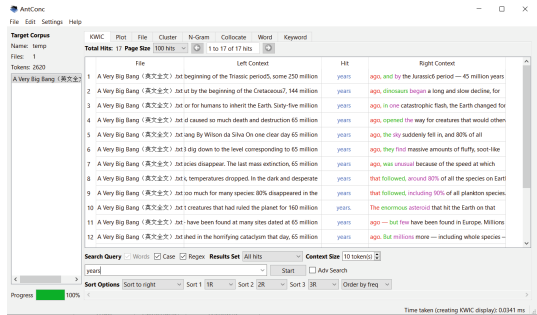


Fig. 12. The Original Index Row of “Years” Used in “A Very Big Bang”

“radiate”, and words related to weather such as “superheat”, “incandescent”, “inferno”, “super-volcano”, “ejecta”, “firestorm”, “cyclonic”, “tsunami”.

3.2.3 Analysis of the Dispersion Plot and Its Development

The Dispersion Plot refers to the content of the story in accordance with a certain developmental criterion described by narrative language. Since this scientific article is a description and explanation of the Big Bang and the occurrence of various natural phenomena, how will students quickly understand how these events developed is of great importance.

Teachers can guide students to list the top 30 high-frequency nouns, and then allow students categorize them according to their meaning. As shown in Table 4, the top 30 high-frequency nouns can be classified into five categories of high frequency nouns. The first category is about the when and where natural phenomena can occur, such as “km”, “earth”, “atmosphere”, “ground”, “sky”, “world”, “air”, “planet”, and “ocean”. The destructive effects of natural phenomena belong to the second category, such as “tsunamis”, “extinction”, “impact”, “blast wave”, “shockwave”, “cataclysm”. The third category is about the objects that the events happen to, such as “rock”, “creature”, “dinosaurs”, “species”, “asteroid”, “animal”, “mammals”, “forest”, “plant”, “dust”, “heat”, and “wind”. The fourth category is about the impact energy when natural phenomena occur, such as “impact”, “heat”, “wind”. The fifth category is about expressions of time, such as “years”, “time”, “today”, and “day”. Based on these five types of nouns, students can predict that this text mainly describes the time, place, process and destructive effects of natural phenomena when an asteroid hit the earth.

Finally, in order to analyze the order of occurrence of various phenomena more clearly, teachers instruct students to construct a Dispersion Plot analysis on these categories of nouns. Therefore, the approximate distribution position of specific words in the whole article can be visualized by the student. According to Table 4, the frequency distribution is as follows: “impact” (22 times), “earth” (17 times), “planet” (9 times). These words are written from beginning to the end in the text. We see from the following frequency distribution, “Km” (17 times), “ground” (9 times), “air” (7 times), and “tsunamis” (6 times) these words are in the first half of the article. These words are located in the original text by the File program. After analyzing 69 paragraphs, it can be

Table 3. Top 50 Keywords in “A Very Big Bang”

Rank	Frequency (focus)	Key Score	Keyword	Rank	Frequency (focus)	Key Score	Keyword
1	5	1637.723	blast wave	26	2	295.085	nitric
2	3	887.546	vaporize	27	1	294.914	pulverize
3	4	798.342	cataclysm	28	1	287.092	repopulate
4	9	661.865	planet	29	1	286.8	warm-blooded
5	4	632.499	shockwave	30	1	284.327	fossilize
6	5	552.808	molten	31	1	283.64	paleontologist
7	2	488.901	superheat	32	2	264.593	incandescent
8	2	486.321	crystallize	33	1	259.079	ionize
9	2	401.596	Yucatan	34	1	258.544	cyclonic
10	7	372.642	asteroid	35	2	257.153	inferno
11	2	359.163	seafloor	36	1	256.317	ejecta
12	9	352.583	dinosaur	37	1	239.463	flipside
13	2	336.777	plankton	38	1	227.643	wombat
14	1	331.263	tyrannosaur	39	1	220.507	decade-long
15	1	330.829	possum-like	40	1	220.191	crisscross
16	1	330.8	soot-like	41	6	219.954	mammal
17	1	328.808	mountain-sized	42	1	217.559	sixty-five
18	6	328.026	tsunami	43	1	212.928	cataclysmic
19	1	327.119	super-volcano	44	1	210.653	Triassic
20	1	325.54	dromaeosaur	45	1	194.634	walrus
21	3	323.091	Sulphur	46	1	186.884	hellish
22	1	321.961	photosynthesize	47	3	185.862	radiate
23	7	311.205	extinction	48	1	184.631	firestorm
24	1	307.147	recombining	49	1	176.584	herbivore
25	1	298.171	anhydrite	50	1	175.808	stratosphere

seen from paragraphs 1–30 that the first half of the article explains in detail the strong shaking of the impacted surface, sky and ocean after the asteroid hit the earth. The plots that mention “rock” (13 times), “atmosphere” (12 times), and “heat” (6 times) are distributed in the middle of the text, i.e. paragraphs 24–45. It can be seen that the asteroid impact caused the rocks on the ground to rush up into the sky and changed from solid to gas as the temperature changed over the atmosphere, and finally returned back to the ground to attack the life on it. The plots that mention “species” (14 times), “dinosaurs” (9 times), “animals” (7 times), “extinctions” (6 times), “mammals” (6 times), “food” (6 times), “plants” (5 times) are distributed in the last third of the text, i.e. paragraphs

Table 4. Dispersion Plot of Top 30 High-frequency Nouns in “A Very Big Bang”

Rank	Frequency	Noun	Dispersion Plot	
1	22	impact		
2	18	year(s)	Dispersion 0.579	Plot
3	17	earth	Dispersion 0.736	Plot
4	16	km	Dispersion 0.505	Plot
5	14	species	Dispersion 0.498	Plot
6	13	rock(s)	Dispersion 0.579	Plot
7	12	atmosphere	Dispersion 0.539	Plot
8	11	creature(s)	Dispersion 0.592	Plot
9	10	time	Dispersion 0.592	Plot
10	9	ground	Dispersion 0.613	Plot
11	9	sky	Dispersion 0.396	Plot
12	9	dinosaur(s)	Dispersion 0.480	Plot
13	9	planet(s)	Dispersion 0.688	Plot
14	8	today	Dispersion 0.551	Plot
15	8	world	Dispersion 0.750	Plot
16	7	day(s)	Dispersion 0.491	Plot
17	7	asteroid	Dispersion 0.571	Plot
18	7	animal(s)	Dispersion 0.289	Plot
19	7	air	Dispersion 0.628	Plot
20	6	extinction(s)	Dispersion 0.317	Plot
21	6	mammals	Dispersion 0.000	Plot
22	6	forest(s)	Dispersion 0.447	Plot
23	6	heat	Dispersion 0.396	Plot
24	6	number	Dispersion 0.385	Plot

(continued)

Table 4. (continued)

25	6	ocean(s)	Dispersion	Plot
			0.333	
26	6	food	Dispersion	Plot
			0.556	
27	6	tsunamis	Dispersion	Plot
			0.491	
28	5	blastwave	Dispersion	Plot
			0.553	
29	5	plants	Dispersion	Plot
			0.553	
30	5	dust	Dispersion	Plot
			0.667	

39–68. It can be seen that after the asteroid hit the earth, animals and plants on land suffered unprecedented destruction and the herbivores could not survive, which directly led to the rapid demise of the carnivorous dinosaurs. However, since mammals are trivial animals, they can forage and survive in harsh environments. Finally, trivial animals such as mammals have risen and dominated the earth since the impact.

3.3 Post-class Stage - Summarization and Retrospection

3.3.1 The Building of Common and Special Corpora

Teachers require students in groups to translate the text into Chinese, and to apply OCR technology to scan the original English text so that a Chinese-English translation is generated online. By utilizing the translation memory retrieval and online alignment functions (such as Tmxmall) or computer translation tools (such as SDL Trados), original terms and translation terms are collected to build a learning corpus Translation Memory, as shown in Fig. 13 and 14. This activity can not only improve students’ English vocabulary acquisition ability, but it also allows students to directly practice their ability in applying technology. This allows students to meet the requirements of a new liberal arts education that is the combination of technology and foreign language learning. Teachers aggregate and classify students’ corpora to build a general-purpose corpus and a special-purpose corpus that are applicable to the whole class, so as to help students review and articulate learning materials.

In practice, students use Tmxmall and Sketch Engine to extract terminology from the reading texts in TEM4 and TEM8 from the past ten years to build up a general corpus for learning. To date, 5380 words of Chinese and English terms have been extracted from the 2008–2019 TEM 4 reading program; 9604 Chinese and English terms have been extracted from the 2010–2021 TEM 8 reading program; and 38,386 Chinese and English terms have been extracted from the 2011–2021 CET 4 reading program. In total, 53370 vocabularies have been extracted offering rich learning materials and teaching resources. The corpus is systematically categorized and organized according to the exam type, difficulty, and completion time. In addition, a student team applied SDL Trados to

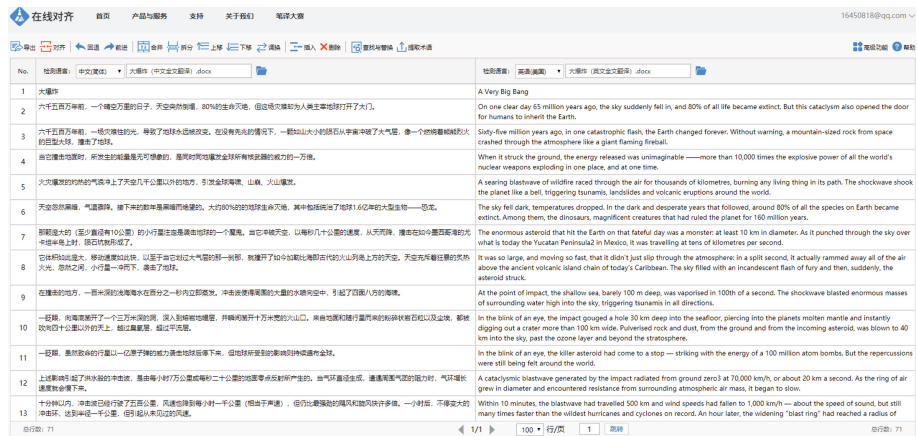


Fig. 13. Online Alignment Editing Interface of “A Very Big Bang”

序号	术语原文	术语译文	词频
1	酸雨	acid rain	1
2	气温	temperatures	1
3	熔岩	the molten rock	1
4	大型动物	large ones	1
5	全球所有核武器	explosive power of all the world's nuclear weapons	1
6	分布很广	a wider distribution	1
7	持续不断的流星	the incoming meteors	1
8	灭绝的边缘	the edge of extinction	1
9	特定的食物	specific food group	1
10	地球	the planet	3
11	尤卡坦半岛	the Yucatan Peninsula	1
12	如今四处分散的云状般的岩石与灰尘	the now widely dispersed cloud of rock and dust	1
13	有足够的阳光进行光合作用	enough light to photosynthesise	1
14	行星撞击地球后所产生的烈火	the raging fires that followed the impact	1
15	数月	several months	1
16	保证	guarantee	1
17	海底	the seafloor	1
18	雨水	the water droplets	1
19	那些生活在恐龙阴影下的生物	creatures that had lived in the shadow of the dinosaurs	1

Fig. 14. Corrected Translation of Terminology Extraction of “A Very Big Bang”

extract terms from three books reflecting Chinese culture, including *Historical Records*, *The Analects of Confucius* and *Silk Road*, and they established a special-purpose corpus, with a total of 12,216 extracted terms.

3.3.2 Evaluation and Reflection

Students can review their learning process through a personalized historical record of their past submissions to reflect on correcting their problems. When leaning the English language, students are required to learn about the evaluation standards on Chaoxing platform, and conduct self-evaluations, intra-group evaluations, and inter-group evaluation by scoring and texting. This method helps students not only to enhance self-regulation or reflection skills such as self-assessment and self-correction, but it also helps develop

Table 5. Comparison of the Effectiveness of CBM Model And Traditional Teaching

Average	CBM Model	Traditional Teaching Class 1	Traditional Teaching Class 2
Overall evaluation score	84.94	83.87	81.35
Self-evaluation	96.58	90.87	92.57

their ability to evaluate other students' strengths and weaknesses. During this process, teachers comprehensively evaluate students' reading abilities by analyzing their learning behaviors, online data, self-evaluations, as well as post-class questionnaires and interviews.

After teaching 567 junior college students who were English majors, for three semesters, the author concludes that a corpus based DDL utilizing an online and offline blended teaching model has four positive aspects as compared to traditional classroom teaching methodologies: 1. The former is more conducive to improving students' learning which is reflected in the student's overall improvement. (See Table 5) 2. The former is more conducive to improving the student's satisfaction when learning. For example, the self-evaluation scores among students increased significantly by 6 points, and the comments were more positive than their past comments. (See Table 5) 3. The former is more conducive to improving the students' innovative thinking ability. This is reflected in the measurement of fluency and flexibility in college juniors' creative thinking ability. Their scores for originality and creativity (total score of the three factors) were significantly higher than those who received traditional teaching methods. (See Fig. 15) These metrics were taken from "Torrance Test of Creative Thinking" (TTCT) and "Test of Young People's Scientific Creative Thinking Ability" developed by Hu Wei-ping and were applied to the corpus-based DDL teaching method. 4. The former is more conducive to enhance students' overall learning. For example, 96.88% of students agreed they had improved in their overall language ability. (See Fig. 16).

In terms of actual scores, Fig. 15 shows that there is no significant difference between grades when comparing originality and creativity ($p > 0.05$) but some differences in fluency and flexibility exist ($p < 0.05$). Students in different grades showed a significant level of fluency 0.05 ($F = 2.811$, $P = 0.040$), and by comparison, the average scores of the columns with a significant difference are "1.0 > 2.0; 1.0 > 4.0" respectively. Meanwhile, students of different grades showed a significant level of flexibility 0.05 ($F = 2.650$, $P = 0.049$), and by comparison, the average scores of the columns with a significant difference are "3.0 > 1.0; 3.0 > 2.0" respectively.

The author surveys 128 students enrolled in 2019 from 12 aspects and conducts follow-up interviews for special data after the survey. According to the five-level scale, the questionnaire gives options of "very satisfied", "relatively satisfied", "generally satisfied", "not very satisfied", "very dissatisfied", and assigned 5.4.3.2.1 points respectively. The average score comparison is shown in Table 6. A positive change in learning attitude is presented according to the proportion of the number of respondents. Specifically, the rate at which students learn is reflected in 12 areas which can be visualized (see Table 6). According to the table, after utilizing the t-test (Independent Sample t-Test), a total of 11

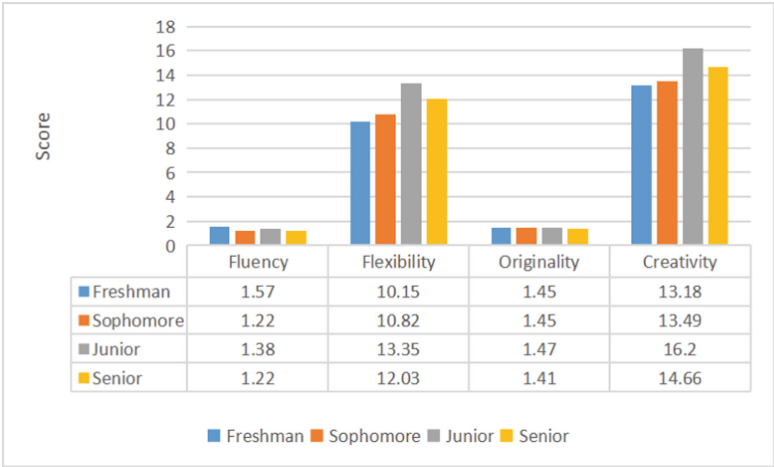


Fig. 15. Total Scores and Factor Scores of Innovative Thinking Ability in Different Grades

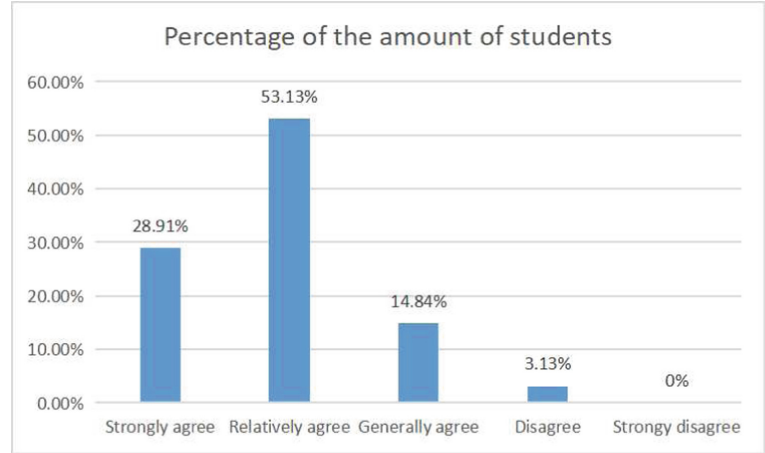


Fig. 16. Current Status of Student Learning Initiatives in the CBM Model

items have no significance ($p > 0.05$), and this shows consistency and no differences. In addition, male and female sample sets show a significant difference 0.01 ($p < 0.05$) for “my ability to extract terms using CAT”, which means that there are differences in the ability of male and female students in this aspect. Specific analysis shows that gender has a 0.01 level of significance for this item ($t = 2.677$, $p = 0.008$), and the average of males (2.57) is significantly higher than that of females (1.97).

Table 6. t-Test Analysis of Learning Effect of English majors under CBM Model

	Genders (Mean \pm SD)		<i>t</i>	<i>p</i>	Average
	Male (n = 14)	Female (n = 114)			
The degree students are satisfied with the course schedule	1.79 \pm 0.58	1.83 \pm 0.62	-0.272	0.786	4.17
The degree students are satisfied with the Internet and related learning resources obtained in the Advanced English Reading course	1.86 \pm 0.53	1.74 \pm 0.61	0.704	0.482	4.25
The degree that CBM Model has effectively improved my learning autonomy	2.07 \pm 0.73	1.90 \pm 0.75	0.791	0.431	4.08
The degree that students can learn more from Advanced English Reading course under CBM Model	2.00 \pm 0.55	1.81 \pm 0.62	1.107	0.27	4.17
The degree of overall satisfaction with Advanced English Reading course under CBM Model	1.86 \pm 0.53	1.83 \pm 0.62	0.137	0.891	4.16
Students' degree of willingness to utilize the CBM Model in the future	2.14 \pm 0.53	1.86 \pm 0.70	1.456	0.148	4.11
The degree of satisfaction and convenience of using Chaoxing operation functions	2.14 \pm 0.86	2.03 \pm 0.76	0.535	0.594	3.96
Students' ability to extract terms using CAT has improved	2.57 \pm 0.94	1.97 \pm 0.77	2.677	0.008**	3.96
The use of corpus retrieval analysis tools and CAT to organize vocabulary study files	2.43 \pm 0.85	2.03 \pm 0.78	1.801	0.074	3.93
The use of corpus retrieval, analysis tools, and CAT to assist reading comprehension and grasp the main idea of the reading materials	2.50 \pm 0.85	2.12 \pm 0.78	1.695	0.093	3.84
The use of corpus retrieval, analysis tools, and CAT to assist in the discussion of text topics	2.64 \pm 0.93	2.25 \pm 0.78	1.713	0.089	3.7
Students' learning attitudes have positively changed under the CBM Model	3.64 \pm 1.55	3.18 \pm 1.48	1.09	0.278	39.06%

* $p < 0.05$ ** $p < 0.01$

4 Conclusions

Corpus retrieval tools can provide rational data analysis and visual models for language learning. Corpus-based DDL when used in an online and offline blended teaching model has broken the traditional teaching and learning models that teachers normally rely on. Corpus-based DDL helps to create an objective analysis of articles based on the collected corpus data and students' reading strategies. Furthermore, the corpus based DDL learning activities are integrated into online and offline blended teaching methods. Students' application of educational technology has been strengthened such that foreign language students can overcome four major learning problems, which are: 1) thinking and theory over practice; 2) singularity over pluralism; 3) humanity over technology, and 4) input over output. Students therefore explore information technology while expanding inter-professional knowledge. Students also learn to utilize learning resources which helps to improve comprehensive learning skills and cultivate thinking while promoting efficiency. This teaching model transforms teachers from classroom authorities to knowledge guides, as well as collaborative learning guides, promotes the interaction between teachers and students, and is in accordance with the new direction of teaching innovation.

References

1. COXHEAD. A New Academic Word List [J]. *Tesol Quarterly*, 2000, 34(2):213–238.
2. JIANG Xiao-yan. A Corpus-based Representative Analysis of Language and Keywords in Jane Eyre [J]. *Journal of Jiangsu University of Science and Technology (Social Science Edition)*. 2016, 6:77–84.
3. Johns T. Should you be persuaded: Two examples of data-driven learning [J]. *English Language Research Journal*, 1991, 4: 1–16.
4. JOHNST, KING P. 1991. Classroom Concordancing. *English Language Research Journal (New Series)* 4. Special Issue [C]. Birmingham: University of Birmingham.
5. WEI Nai-xing. The Methodology and Related Concepts of Corpus Linguistics [J]. *Foreign Languages Research*, 2009, 5:36–42.
6. WESTMA. General Service list of English Words [M]. Cambridge: Cambridge, 1953.
7. ZHAO Qiong, XIE JING-jing. A Study on the Application of Corpus-Based Data-Driven Learning in the Teaching of College English—A Case Study of Maugham's Novel *A Friend in Need* [J]. *Journal of Chongqing Electric Power College*. 2019, 8:43–47.
8. ZHEN Feng-chao. Corpus Data-Driven Foreign Language Learning: Ideas, Methods and Techniques [J]. *Foreign Language World*, 2005, 4:20–27

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

