International Journal of Advanced Engineering, Management and Science (IJAEMS)
Peer-Reviewed Journal
ISSN: 2454-1311 | Vol-10, Issue-7; Nov-Dec, 2024
Journal Home Page: https://ijaems.com/
DOI: https://dx.doi.org/10.22161/ijaems.107.14

The Application of Corpus Data-driven Learning Model in College English Language Teaching in China

Li Keli

School of Foreign Languages, Zhejiang University of Finance & Economics Dongfang College, Haining, 314408 PRC China
Email: lkljasmine@126.com

Received: 15 Nov 2024; Received in revised form: 12 Dec 2024; Accepted: 18 Dec 2024; Available online: 24 Dec 2024

Abstract— *In recent decades, corpus linguistics has developed rapidly, and corpus data-driven learning has provided new ideas for English language teaching research and practice. The concept of data-driven learning highlights the advantages of corpus in assisting language teaching and cultivating students' autonomous learning ability, research ability, and ability to use modern information technology. China's college language English teaching is still confined to the traditional teaching model, and corpora are rarely used for teaching innovation. Starting from the concept of corpus and data-driven learning and the actual teaching needs, this paper briefly introduces the necessity of implementing this learning concept in college English language teaching of vocabulary, grammar, translation and writing, in order to promote corpus-assisted English language learning in college English classrooms and improve the advanced and innovative nature of Chinese college English language teaching.*

Keywords— ***data-driven learning; corpus; college English language teaching; China***

I. INTRODUCTION

Corpus linguistics is the use of computer technology to systematically describe and explore a large amount of real language or language variants, including its linguistic features, contextual meaning, pragmatic functions, etc. With the help of computer retrieval advantages, intuitive observation interface and huge amount of language information, corpus linguistics has become one of the main means of language research and teaching (Chen & He, 2017). Data Driven Learning is a language learning concept proposed by Tim Johns, which is based on the research results of the Collins COBUILD research project led by John Sinclair, the founder of corpus linguistics at the University of Birmingham in the UK. The core of the concept is to use information technology and a large amount of corpus data to allow students to observe and summarize language usage phenomena (Xu & Zhang, 2019).

The concept of data-driven learning highlights the advantages of corpora in assisting language teaching and cultivating students' autonomous learning ability, research ability and ability to use modern information technology. In language teaching, there are two main purposes for using this learning method: first, corpus linguistics emphasizes that meaning and function exist in context. Therefore, from the perspective of learning content, through index lines, students can more intuitively explore and discover the form, meaning and function of language through context (Lin & He, 2019); second, from the perspective of learning and teaching methods, this learning method emphasizes student-centeredness and challenges the traditional foreign language teaching concept centered on teachers and textbooks. In other words, data-driven learning requires teachers not to directly and explicitly convey information and knowledge to students, but to act as guides and collaborators, control the teaching process, and cooperate with students' knowledge input and output in a research and exploration manner. This language learning method is in line with the student-centered and teacher-led education policy, which is conducive to training students' ability to discover problems and think independently. When the concept of data-driven learning was first proposed, it was not widely recognized. However, with the application and popularization of computer technology and corpus technology in foreign language teaching, the importance of data-driven learning has become increasingly apparent (Talai & Fotovatnia, 2012).

II. CORPUS DATA-DRIVEN LEARNING CONCEPT AND ITS CHARACTERISTICS

The characteristic of corpus linguistics is to observe language features from a huge amount of real texts through frequency, collocation and index lines. Constructivist learning theory believes that the acquisition of language knowledge is not a simple process from teachers to students, but a process of discovery and exploration by students themselves (Zhou, 2009). Teachers need to build connections between new and old knowledge, and guide students to make positive transfers based on existing knowledge, experience and cognitive structure, so as to promote the formation of students' independent exploration motivation and master new knowledge. The data-driven learning concept based on corpus has the above characteristics. First, this learning concept uses real corpus as language input through corpus index lines. Secondly, the data-driven learning concept requires learners to be centered and autonomous learning to be the main focus. Students should self-manage, self-monitor and self-evaluate in the learning process. Finally, this is a bottom-up inductive learning method. Students need to find rules from a large amount of language data. This process can not only deepen their understanding of language knowledge, but also train their ability to explore the essence from the phenomenon (Sinclair, 2004).

The lexical grammar theory proposed by Sinclair (2004) includes five elements: node words, collocation, colligation, semantic tendency, and semantic prosody, which are the basic components of corpus linguistics research. Therefore, the implementation method of data-driven learning also corresponds to these five hierarchical systems. Specifically, it mainly includes the following methods and steps.

2.1 Searching for node words and observing concordance

This method mainly involves importing the target corpus file into the corpus analysis software, searching for node words, and presenting an index line on the interface. Students can intuitively observe the frequency of occurrence and context of the node words through the search results, and deduce their contextual meaning and function.

2.2 Analysis of collocation and colligation

The study of collocation is mainly based on the phrase concept of corpus linguistics, and the co-occurrence characteristics of language are understood through the analysis of frequency and contextual meaning. Sinclair (1991) defined collocation as "the co-occurrence of two or more words within a short distance in a text", that is, the words that co-occur at the same time have contextual meaning; the collocation concept proposed by Wei (2002) emphasizes the frequency feature, "a sequence of words that realize certain non-idiomatic meanings in the text and are used in a certain grammatical form in a certain way, and the words that constitute the sequence are mutually expected and co-occur with a probability greater than chance", that is, collocation in the statistical sense. As an important concept of word collocation in corpus linguistics, colligation refers to the combination of language between lexical and grammatical categories, that is, the co-selection relationship between words and grammar. Corpus linguistics research uses colligation to describe the grammatical structure at the horizontal combination level. "Class connection is not an abstraction parallel to word collocation, but a higher level of abstraction" (Wei, 2002; Zhen & Li, 2017). For example, "turn a blind eye to" reflects the lexical grammatical structure of "turn + a + ADJ + NOUN + to", that is, the colligation. The corpus gives the vocabulary collocation suitable for the real context according to the frequency, presenting its contextual meaning. This learning method can enable students to learn more authentic language expressions.

2.3 Understanding semantic preference and semantic prosody

Sinclair (2004) proposed that a meaning unit is composed of a main word and a class of words that often co-occur with this word, and has relatively stable semantic characteristics, which forms semantic tendency and semantic prosody. The speaker's point of view and attitude can be expressed through semantic rhyme. Therefore, recognizing and learning semantic tendency and semantic prosody can help us master accurate and appropriate language forms for communication while avoiding making up collocations according to grammatical rules (Gao & Wei, 2020).

The above series of operations based on data-driven learning allow students to see a large number of real language samples while being immersed in the situation. Classroom teaching and language learning planned in this way reflect the information and high-tech nature of modern education. In theory, it will improve students' ability to explore, summarize and learn independently. However, applying them to the practice of foreign language teaching has both gains and difficulties.

III. CORPUS DATA-DRIVEN LEANING AND ENGLISH VOCABULARY TEACHING

Since the concept of corpus was created in the 1960s, linguists have been exploring the use of electronic corpora to analyze language and gradually applying them to classroom teaching. Johns (1991) coined the term data-driven learning (DDL), and advocated corpus-based inductive "data-driven learning", which can help teachers guide students to use the index line function in the corpus to conduct independent research, thereby enriching students' vocabulary. By searching for synonyms of keywords that students use a lot, teachers can present vocabulary in the same semantic field to help students master the diversity of the vocabulary based on specific contexts. In addition, teachers use corpus-based data-driven learning methods to compare the definitions and collocations of easily confused vocabulary, deepen students' understanding of the meaning and usage of the vocabulary in multiple contexts, and thus help them distinguish and understand the meaning of words.

Later, many scholars use corpora to analyze vocabulary association and collocation, and explore their application in vocabulary teaching (Cobb, 1997; McEnery & Wilson, 1997; Nesselhauf, 2006; Barfield & Gyllstad, 2009). Cobb (1997) explores whether corpus-based classroom vocabulary teaching can produce obvious learning effects. This study first proposed the vocabulary learning hypothesis, which is to use corpus vocabulary indexes to learn vocabulary usage presented in different contexts. Studies have shown that when learners use vocabulary indexes, they will improve the accuracy of their vocabulary usage, which verifies that applying corpora to vocabulary teaching will produce positive learning effects. Nesselhauf (2006) mainly studies the vocabulary collocation in the English corpus of second language learners, explores the mistakes that learners are prone to make when using vocabulary collocation, and analyzes which factors affect the mastery of vocabulary collocation and ultimately affect language learning. In addition, related studies use corpora to strengthen vocabulary learning, explore the learning strategies and effects of vocabulary collocation and association, and provide practical suggestions based on this (Chen, 2013; Ground, 2019; Ghalebie et al., 2020). Reka & Eniko (2022) focused on the application of specific-purpose corpora in the research of English major vocabulary teaching. The study found that specific-purpose corpora can provide vocabulary learning support for learners and teachers.

Research shows that Chinese English learners have far more misused parts of speech than foreign learners, both in terms of the total number of misused parts of speech and the types of misused parts of speech involved. In college English language classes, how to help students explore and summarize language usage rules from real language data is a major challenge faced by Chinese college English teachers in their teaching. Traditional auxiliary English vocabulary teaching tools, such as textbooks, dictionaries, reference books, and English teachers' sense of language are all lacking in helping students fully understand and correctly apply English vocabulary. The examples in textbooks and dictionaries are relatively limited, so they can only provide limited meanings, usages, and grammatical collocations of target vocabulary. The comprehensiveness of the corpus with its huge amount of data and the convenience of using technology can help solve this problem.

In order to enable students to express their ideas better, teachers can use the DDL teaching model in teaching, find specific collocations from the BNC corpus, and help students understand and master more skills in using the language. The word "problem" is an important part of "solve the problem". When the information of "problem" is extracted from the BNC corpus, 568 index lines are extracted after screening. Among them, the most common verb collocations of "problem" are: tackle, unlock, remove and conquer. After careful comparison and study, students find that the words on the index line are used in daily life, and they also realize that there are some shortcomings when using these daily words. The collocation mentioned in the above example is a language technique that can help students better understand the words in the text and combine them in a more effective way. This technique makes the relationship between words that is not accidental become closer. To master this language technique, we must start with the formal characteristics of word collocation, such as the similarity between phrases, mutual expectations and attraction, and the similarity of collocation components, because these are important aspects of word collocation research. The corpus can be used to more effectively guide students to use common phrases for learning. For example, teachers can guide students to study the collocation of "absolutely" with other common adjectives in the BNC corpus and find out the common points between them: indicating the characteristics of a certain thing to a certain extent, such as "sure" and "certain"; referring to the judgment of the characteristics of a thing, and other terms that help describe people, things, and objects.

The collocation function in the corpus is usually displayed in the form of spans. The target vocabulary is used as a node word (or central word), and the number of words around the central word is set. The sum of the number of words on the left and right is the span. The words that appear at each position in the span belong to the collocation words of the target vocabulary. By observing the frequency of the target vocabulary and these collocation words, we can grasp the collocation rules, collocation patterns or habitual collocation usage of the target vocabulary, so as to understand the semantic characteristics (i.e. semantic prosody) or semantic tendency of the vocabulary. The submodule “collocation” in the multilingual database comprehensive platform can realize this function and can be directly applied to the collocation teaching of college English classroom.

Taking the central word “rule” as an example, in the British National Corpus (hereinafter referred to as BNC), the language is selected as English, the number of words in the left and right intervals of “rule” is set to 5, that is, the span is 10, and punctuation is enabled as stop words to reduce interference with the results. According to the query results, the maximum frequency value of the words in each span position is sorted out, and the following results are obtained: the central word “rule” appears 336 times in the platform database, and the top three collocations with it are right 1 “of” (305 times), right 2 “law” (288 times), left 1 “the” (223 times). From this observation, it can be identified that “rule of law” is a fixed collocation usage, meaning “法制 (fǎ zhì)”. Teachers can also show students specific examples of this collocation and its usage in different contexts through the KWIC (Key Word in Context) or index line function. In addition, the system can automatically generate cloud maps for vocabulary collocation, presenting the vocabulary collocation to users in a visual form. The closer the vocabulary is to the central word, the closer the relationship between the vocabulary and the central word is, and the higher the frequency of collocation in specific applications.

The above discussion shows that when teachers apply real examples in the database in college English language classes, it not only provides a scientific basis for teachers’ teaching, but also allows students to learn more authentic and customary expressions based on examples and context, ultimately improving teaching and learning efficiency.

IV. CORPUS DATA-DRIVEN LEARNING AND ENGLISH GRAMMAR TEACHING

When people learn a second language, they can significantly improve the effect of language learning by mastering certain grammatical rules. Therefore, grammar teaching is very important in English language teaching. According to relevant surveys and studies, although most Chinese college students have a certain concept of English grammar, less than one-fifth of students have a clear concept of English grammar, and some students even have no concept of grammar at all. In the process of learning a language, they rely entirely on their sense of language or reading Chinese translations. At the same time, the main purpose of most students learning English grammar is to take exams, and only about one-fifth of students learn grammar to improve their English literacy. At present, most students say that they have great difficulties in learning grammar, but they all want to improve their grammar level. The main way for students to improve their grammar level is still to rely on teachers' teaching and reading grammar books, or a large number of grammar exercises. The learning effect of this kind is often not good. Students' grammar learning is prone to fall into the misunderstanding of taking exams, and it is difficult to achieve positive transfer for the use of language in real scenarios. Therefore, in order to improve the effect of English grammar teaching, improve the accuracy of language learning, and realize the combination of language form, meaning and pragmatics, it is necessary to introduce corpus into English grammar teaching.

When teaching sentence patterns, Chinese college English language teachers are accustomed to conducting a large number of sentence pattern conversion exercises of affirmative sentences, negative sentences and interrogative sentences in the teaching process to strengthen students' proficiency in a certain sentence pattern. For example, when students learn the *there be* sentence pattern, they will do a lot of negative sentences *there be not* exercises according to the teacher's requirements. However, after searching various types of existential sentences (*there be* sentence patterns) on BNC, it was found that among the 186,030 examples of existential sentences (*there be* sentence patterns) retrieved, negative sentences only accounted for about 1.1% of the total. It can be seen that native English speakers generally use the *there be* structure in affirmative sentences and interrogative sentences in actual applications, and rarely use its negative form. In addition, the frequency of *there be* sentence patterns in academic English is extremely low. Language teaching separated from the corpus is inevitably lacking in accuracy and may mislead students to a certain extent. Kennedy (2000) believes that the frequency of a language phenomenon in the corpus should be the only criterion that can affect language teaching. In addition, according to the description of traditional English grammar, some is used in affirmative sentences, and any is used in negative sentences or interrogative sentences. Tognini-Bonelli used the Birmingham English Corpus to search and count, and extracted 35 of the 21,636 index lines of *any*, and found that a total of 19 *any* were used in negative and interrogative sentences, and the remaining 16 *any* were used in affirmative sentences, accounting for 46% of the total (Tognini-Bonelli, 2001: 15-17). It can be seen that the description of the use of *any* in traditional grammar is not accurate or complete. The use of *any* in affirmative sentences is an objective grammatical phenomenon, and English learners can use it boldly without hesitation because of concerns about the grammatical rules taught by teachers in class.

For another example, there is a rule in traditional grammar: in a clause introduced by the conjunction *that*, if the clause introduced by *that* is a direct object or complement, *that* is often omitted in informal usage. However, it is difficult for learners to grasp what frequency “often” refers to and how *that* is omitted in formal language based on the above general description. Corpora can provide learners with comprehensive and reliable statistics. The Longman Corpus mainly includes several domains such as conversation, news and academic articles. The search results of *that* clauses in the Longman Corpus show that in conversation, the proportion of that omitted in *that* clauses is about 85%; in news reports, *that* is omitted in about 25%; and in academic articles, it is very rare to omit *that*. In addition, regarding the omission of relative words, traditional textbooks often give a general rule explanation, that is, relative words are often omitted in informal styles such as spoken language. However, neither teachers nor scholars can give a clear explanation of what frequency "often" refers to. The search results of the Longman Corpus show that in conversations, about 25% of relative clauses omit relative words; in academic articles, only about 10% of restrictive relative clauses omit relative words. Of course, there are also omissions in the four domains of conversation, novels, news, and academic articles. When the subject of the relative clause is a pronoun, 60% to 70% of the relative clauses omit relative words; when the subject of the relative clause is a noun phrase, 80% to 95% of the relative clauses retain relative pronouns. After understanding the specific distribution and frequency of these grammatical rules in different domains, learners may have a more comprehensive, accurate, and in-depth grasp of the usage of that clauses and relative word omission.

It can be seen that compared with traditional grammar teaching, corpus-based grammar research observes the language actually used by native English speakers in different registers, and the units of description are grammatical meaning units including vocabulary and structure, which provides a more accurate and intuitive basis for language learners.

V. CORPUS DATA-DRIVEN LEARNING AND TRANSLATION TEACHING

With the popularization of machine translation, the statistical suggestions provided by corpora have begun to be valued. Parallel corpora used in translation can provide detailed language examples for translation teaching and improve the accuracy of translation (Liu & Li, 2020). Mona Baker (1995) was the first to combine corpora with translation teaching. Subsequently, foreign scholars such as Olohan (2004) and Gallego-Hernández (2016) discussed the index line resources, co-existing examples and their significance provided by corpora. Chinese scholars have also put forward improvement suggestions from the aspects of translation teaching design and teacher-student interaction mode, and have given an implementation plan for corpus-assisted translation teaching (Xiao, 2005; Qin & Wang, 2007). As a scientific and technological product in the era of information intelligence, corpus translation technology is an efficient translation tool that helps improve translation efficiency and empower cross-cultural communication. There is a broad space for exploring the integration of corpus translation technology into translation teaching. In the process of translation classroom teaching, by introducing professional field corpora, students can better master the correct expression of professional terms and effectively improve their professional translation ability and cross-cultural communication ability.

In traditional translation teaching, it is difficult for students to fully understand the original text, especially the original text with very new content, only relying on textbooks, dictionaries or translators' intuition. The powerful search function of monolingual corpora provides a good solution to this problem.

After being elected as the mayor of Boston, Chinese American Wu Mi said in her speech: We're ready to be a Boston where all can afford to stay and to thrive. And yes, Boston is ready to become a *Green New Deal* city.

Many students cannot understand *Green New Deal* in the above example. Through Baidu Encyclopedia, we can find that its corresponding Chinese expression is "绿色新政 (lǜ sè xīn zhèng)", but the original English background information of the system is difficult to obtain, and students still have a vague understanding of the connotation of this concept. If we use the search function of the corpus, this problem will be solved. Take the Corpus of Contemporary American English (hereinafter referred to as COCA) as an example. Enter *Green New Deal* in the search box on the homepage of the corpus, and soon get 246 search results arranged in reverse chronological order (2019-2008). The relevant materials come from speeches, blogs, magazines, news, etc. The first three search results appeared in 2008, two of which came from *The Washington Post* and the other from the *Christian Science Monitor* (CS Monitor). The former is the most influential daily newspaper in the United States, and the latter is an internationally renowned news organization. After these two media reported information related to the *Green New Deal*, other magazines, blogs, and speeches also appeared with related content. Using the context function of COCA, translators can click on each search result as needed to view the complete relevant news reports and obtain the original English information about the *Green New Deal* to have a more comprehensive understanding of this concept. Taking the 246th collection result with the earliest publication time as an example, the context function of COCA shows that the concept of *Green New Deal* was first published in the article "World wants green action, despite costs" on page 25 of the Christian Science Monitor on November 20, 2008. The article focused on the importance of energy conservation and renewable energy to revitalizing the economy and emphasized the inevitable connection between green development and national sustainable development.

It can be seen that by analyzing the data in the corpus, we can obtain information about the frequency of use, collocation rules and context of various words, phrases and sentences, find common expressions under specific topics, and learn from the words and sentence patterns of native speakers, which can better help translators understand the usage and habits of the target language, so as to choose the most appropriate expression and improve the accuracy and fluency of translation. At the same time, compared with the traditional teaching model, the corpus-assisted translation teaching model can allow students to have a more intuitive feeling and deeper thinking about the correct selection and use of corpus, and improve their independent learning ability, collaborative communication ability and inquiry learning ability in the process of continuous discovery and exploration.

VI. CORPUS DARA-DRIVEN LEARNING AND ENGLISH WRITING TEACHING

There are a large number of empirical studies on the application of data-driven learning methods to improve the writing ability of second language learners abroad. The research is relatively in-depth, the specific content of the research is more detailed, and there are relatively more quantitative analyses of experimental results. Vyatkina (2016) studied the effect of data-driven learning of collocation on the expressiveness, proficiency and perception of second language learners' writing, combined a variety of outcome measurement methods, and measured delayed learning outcomes. Walker (2017) proved through empirical research that students' direct exposure to the corpus MICUSP (Michigan Corpus of Upper-Level Student Papers) can help reduce errors in the use of English conjunctions. Mao et al. (2018) let students directly contact corpora such as BNC, COCA and English-Chinese Parallel Corpus in writing classes to examine the validity of data-driven learning methods in college English writing classes, and used questionnaires to understand students' attitudes towards this new learning method. Li (2017) explored the impact of direct use of corpora on learners' collocation ability in academic writing. In recent years, although there have been some studies on the application of data-driven learning methods in English writing teaching in China, the overall number is small and the influence is relatively small.

Many Chinese universities still use traditional methods to teach writing, such as lecturing on theories, analyzing model essays, centralized review, and teacher face-to-face review. In such traditional writing teaching, the interaction between teachers and students is not smooth, the student participation is not high, the classroom atmosphere is not good, and it is difficult to stimulate students' interest in writing. The introduction of corpus can greatly improve the effect of writing teaching. In the corpus, learners can retrieve specific words, phrases, sentences and other real language examples to understand their language rules and expression patterns, and enhance language acquisition. The corpus breaks the rigid pattern of traditional writing teaching, and encourages students to find language examples in the corpus, verify their language expression methods, and drive students to write independently.

In English writing, students can learn, draw lessons from and apply verbal expressions with the help of corpora. For example, for the sentence "As time went by, my confidence increased slowly", when we cannot confirm whether the expression is authentic, we implant it into the COCA free English corpus for verification. Enter "my confidence increased" in the result query area and the result is "zero". This means that there is no similar verbal expression example in the entire corpus, that is, the expression is unacceptable. We can enter different words or phrases in the query area, click on the index bar information respectively, and the related examples will be displayed. Different example results can help us correct the above hypothetical statement: "As time went by, my confidence grew slowly".

The extensive content and feature-rich search tools of corpora such as BNC provide valuable resources and support for professional English writing teaching. First, corpora such as BNC contain samples from a variety of text types, and students can find collocations of target vocabulary in different contexts and themes. This helps students to understand the usage of words in a comprehensive way, rather than just being limited to a specific field or theme. Second, the keyword search function of BNC enables students to easily find collocations of target vocabulary. This greatly improves their writing efficiency because they can quickly find the information they need without flipping through a large amount of text or relying on traditional paper dictionaries. This real-time, customized collocation search tool can meet the specific needs of students in the writing process. Third, corpora such as BNC provide examples of whole sentences using the target vocabulary collocation. By viewing the complete sentences, students can better understand the grammar and context of the collocation. This helps students to organically integrate these collocations into their own writing to ensure that the sentences are smooth and grammatically correct. In addition, word frequency information in corpora such as BNC provides students with insights into the frequency of using a certain collocation, which can help students determine which collocations are most common in their writing, so that their articles are more natural and in line with the actual usage of English. This data-driven approach helps students avoid using overly stiff or uncommon word combinations and improves the quality of their writing. Finally, actual grammatical examples from corpora such as BNC can help students better understand the grammatical rules and context of word collocation. This enables them to more confidently choose appropriate vocabulary collocations and incorporate them into their own writing. By learning from actual examples, students can improve their language skills in practice, which is essential for professional English writing.

In short, using corpus to teach English writing is a relatively novel and effective teaching method. This model can solve some of the problems that plague English writing teaching to a certain extent. For teachers, using corpus to teach writing can well make up for the defects in language textbooks and achieve flexible use of language; for students, corpus writing can improve writing enthusiasm, and through self-searching, they can also better master the use of target vocabulary.

VII. CORPUS DARA-DRIVEN LEARNING AND ORAL ENGLISH TEACHING

The use of online multimedia corpus indexing system in oral English teaching is to use online graphics, animation, sound and video and other information integration methods to form document materials to assist teaching and improve the efficiency of students' oral learning. This feature is not available in traditional pure text corpora. Online corpus materials are vivid, rich and full of content. Teachers and students can more easily obtain targeted corpus to assist teaching in the process of teaching and learning, making what is taught and learned more effective, and fundamentally promoting the improvement of students' oral ability.

Phonetic problems are a major problem in oral teaching. In teaching practice, many students do not have enough recognition of phonetic symbols and have no awareness of connected reading and meaning groups. They ignore the intonation problem in oral expression and tend to say words one by one and stiffly, while skipping the meaning groups and connected reading parts composed of words. This not only makes the expresser lose confidence and gradually no longer want to speak English, but also makes the audience lose patience, directly or indirectly causing various misunderstandings in communication. To address this problem, teachers can use the convenience of the online system to insert various videos or audio clips that are suitable for students' levels and that they like into oral teaching for students to learn and imitate. For individual difficult sentences, the stressed and connected parts should be marked so that students can practice repeatedly with key points. This can not only improve students' pronunciation and intonation, but also enhance their understanding of English culture and improve students' oral literacy as a whole.

To achieve effective communication, the accumulation of vocabulary and the correct use of grammar are indispensable elements. In oral communication, students often realize that their vocabulary is extremely limited. Even if they have their own opinions on certain topics, they cannot communicate fluently due to limited vocabulary. Insufficient vocabulary and outdated vocabulary greatly hinder students' enthusiasm and initiative in free communication. Frequent grammatical errors are also a major difficulty in students' oral learning, which seriously affects the communication effect. In teaching, teachers should obtain daily high-frequency words and grammatical points that are suitable for students' acceptance ability through the online corpus retrieval system, and insert relevant videos, animations and other courseware in multimedia corpora into teaching in a targeted manner to present knowledge in a multimodal manner. In this way, key knowledge is more intuitive and visual, which promotes students' comprehensible input, and further guides and encourages students to fully and diversely apply daily high-frequency vocabulary and oral sentence patterns in actual oral communication, and constantly achieve internalization and appropriate use.

Flexible and diverse use of discourse markers can ensure the natural fluency of language. However, most Chinese English learners use such language markers too much or too little, and even use them indiscriminately, such as you know, well, so, I think, etc., which affects the meaning of the text to a certain extent. In daily oral learning and training, students do not pay attention to the acquisition of pragmatic knowledge of markers, and their awareness of the use of marker language blocks is not strong enough, and their output is insufficient, resulting in stiff language expression, inappropriate communication, and inappropriate consequences. In this regard, teachers should strengthen the training of college students in the use of markers, collect the language phenomena of various online markers and the markers that appear in the text of student conversation recordings, and form a teaching corpus of oral markers. In daily oral teaching, teachers can use typical examples to explain the pragmatic functions of markers to students, gradually cultivate students' awareness of marker use and strengthen students' ability to use markers correctly and appropriately in discourse communication.

Traditional English teaching overemphasizes exams and does not pay enough attention to oral practice, resulting in the fact that most Chinese students have the ability to "talk on paper". Once they actually use English to communicate with others, they will be timid. Sometimes they are so nervous that their minds go blank and they can't speak. Even if some students with strong psychological qualities can communicate with others orally, their English oral practice is too written, and they pay too much attention to details such as whether the structure is rigorous, whether the language is accurate, and whether the use is standardized, resulting in insufficient timeliness in English speaking. Once the other party's words exceed the scope of their own knowledge, it is difficult to continue the communication. The key to oral communication in English is fluency, not fancy words. When teaching oral English, teachers should pay attention to the accumulation of students' daily language, not just the accumulation of written language, to avoid students developing a mode of thinking of using written language for dialogue. By increasing "interesting knowledge" through online multimedia corpus, various topics of oral knowledge can be pushed to students in a regular manner in the form of small modules, which can accumulate little by little and continuously enhance students' awareness of the use of oral words and sentences. Online indexing of various types of oral learning resources can create a personalized learning atmosphere for students, experience authentic foreign languages, stimulate interest in oral learning, master common language blocks, and lay a solid foundation for good communication.

Obviously, network technology can not only be applied to teachers' classroom teaching, but also help students to study independently after class. As far as college English oral teaching is concerned, the multimedia corpus indexing system provides college students with equal opportunities to listen to classes to a certain extent, making teachers' teaching more flexible and the quality of teaching higher. Therefore, teachers should make full use of information technology, develop educational and teaching resources, help students expand learning channels, improve learning methods, and improve learning efficiency.

VIII. CONCLUSION

In summary, scholars and educators in China and abroad have conducted in-depth and extensive explorations on the concept and practice of data-driven learning, and on this basis have drawn similar conclusions and teaching inspirations. For example, by comparing the differences in language use between students and native speakers, searching for high-frequency collocations, and helping teachers and students master authentic word collocation behaviors; at the same time, in vocabulary teaching, we should not ignore the contextual meaning of vocabulary due to excessive attention to the size of students' vocabulary, but actively use the corpus to search for conventional word blocks, and learn to memorize words efficiently in context . The reasonable use of data-driven learning methods can fully mobilize students' enthusiasm, give full play to their initiative, give students enough space for thinking and expression, cultivate divergent thinking, improve students' ability to use their hands and brains, and achieve the established teaching goals.

In language teaching, any method must overcome some obstacles when applied in practice, and data-driven learning is no exception. Many teaching research practices have found that the acceptance of corpus technology by students at different stages will affect its role in the learning process. The higher the language ability and level of students, the better the mastery of data-driven learning methods, while learners with lower language levels do not accept this method well. In response to this problem, Sinclair (2003) suggested bringing texts retrieved and screened from the corpus to the classroom for discussion with students, and gradually operating them in person, so that students can more easily master data-driven learning methods, become familiar with the principles and functions of corpora, and accelerate their learning progress. Therefore, in the classroom of data-driven learning, teachers should not only have basic knowledge of corpus linguistics theory and the use of corpus tools, but also guide students to reasonably arrange the use of corpora for independent exploration when using data-driven learning teaching methods, and explain the functional rationales behind different language examples in a simple and easy-to-understand way, so that students can be familiar with this learning method, improve their interest in learning and their ability to use language comprehensively. This is undoubtedly a challenge for the teachers' prior knowledge reserves, lesson preparation, and classroom control and coordination abilities.

Driven by science and technology, corpus linguistics research has been developing rapidly, and foreign language teaching research and practice based on corpus linguistics have also been widely valued and applied. In the era of rapid development of information technology, data-driven learning methods are expected to become mainstream. For students, the use of corpus tools can solve many uncertain vocabulary usage, sentence expression and other problems. For teachers, the teaching model based on data-driven learning helps to achieve advanced, innovative and challenging teaching goals. It should be pointed out that although data-driven learning is a teaching method advocated by many experts and scholars, it is still in the exploration and trial stage. It still requires teachers and students to boldly try, make full use of all corpus resources, and conduct in-depth data-driven learning and teaching research to broaden the road for English language teaching.

REFERENCES

[1] Chen, H. Y. & He, An. P. (2017). Research Progress on English Chunks for Academic Purposes Abroad. *Journal of Jiangxi Normal University (Philosophy and Social Sciences Edition)*, 1:138-144.

[2] Xu, L. & Zhang, Y. (2019) Research on English Vocabulary Teaching from the Perspective of Data-driven Learning. *Contemporary Education Theory and Practice*, 5:142-147.

[3] Lin, W. Y. & He, An. P. (2019). A Study on English Metadiscourse Ability of College Students from Four Countries Based on Machine-cut Spoken Language Chunks. *Foreign Languages in China*, 1:71-78.

[4] Talai, T. & Fotovatnia, Z. (2012). Data-driven Learning: A Student-centered Technique for Language Learning. *Theory & Practice in Language Studies*, 2 (7).

[5] Zhou, X. (2009). The Guiding Role of Constructivist Learning Theory in Foreign Language Teaching. *China Adult Education*, 24:135-136.

[6] Sinclair, J. (2004). Trust the Text: Language, Corpus and Discourse. Nottingham: University of Nottingham.

[7] Sinclair, J. (1991). Corpus, Concordance, Collocation. London: Oxford University Press, pp170-75.

[8] Wei, N. X. (2002). Definition and Research System of Word Collocation. Shanghai: Shanghai Jiaotong University Press, pp100-67.

[9] Zhen, F. Ch. & Li, W. Zh. (2017). Firth's Theory of Meaning and Corpus Linguistics. *Foreign Languages*, 4:15-24.

[10] Gao, G. & Wei, N. X. (2020). The Study of Meaning in the British Linguistic Tradition: from Firth, Halliday to Sinclair. *Academic Exploration*, 93:25-34.

[11] Johns, T. (1991). From Printout to Handout: Grammar and Vocabulary Teaching in the Context of Data-driven Learning. *CALL Austria*, 1:10-34.

[12] Cobb, T. (1997). Is There Any Measurable Learning from Hands-on Concordancing? *System*, 3:301-315.

[13] McEnery, T. & Wilson, A. (1997). Teaching and Language Corpora (TALC). *ReCALL*, 1:5-14.

[14] Nesselhauf, N. (2006). Collocations in a Learner Corpus. Amsterdam: John Benjamins Publishing.

[15] Barfield, A. & H. Gyllstad. (2009). Researching Vocabulary Through a Word Knowledge Framework: Word Associations and Word Collocations, *System*, 1:121-135.

[16] Chen, Y. (2013). Corpus-based Vocabulary Learning: A Study of College English Learners' Use of Corpora. *ReCALL*, 3:320-336.

[17] Ground, P. R. (2019). Vocabulary Learning Strategies (VLSs) Employed by Learners of English as a Foreign Language (EFL). English Language Teaching, 5: 177-189.

[18] Ghalebi, R., F. Sadighi & S. M. Bagheri. (2020). Vocabulary Learning Strategies: A Comparative Study of EFL Learners. 7: 1-12.

[19] Reka, R. & C. Eniko. (2022). The Routledge Handbook of Corpora and English Language Teaching and Learning. London: Routledge.

[20] Kennedy, G. (2000). An Introduction to Corpus Linguistics. Beijing: Foreign Language Teaching and Research Press.

[21] Tognini-Bonelli, E. (2001). Corpus Linguistics at Work. Amsterdam: John Benjamins Publishing Company.

[22] Liu, X. D. & Li, D. F. (2020). Application of COCA English Corpus in English-Chinese Business Translation Teaching. *Chinese Science and Technology Translation*, 1: 29-32.

[23] Baker, M. (1995). Corpora in Translation Studies. *Target. International Journal of Translation Studies*, 7(2).

[24] Olohan, M. (2004). Introducing Corpora in Translating Studies. London and New York: Routledge, 2004.

[25] Gallego, H. D. (2016). New Insights into Corpora and Translation. Cambridge: Cambridge Scholars Publishing.

[26] Xiao, H. (2005). Application of Translation Workshop in Translation Teaching. *Journal of Sichuan International Studies University*, 1:139-142.

[27] Qin, H. W. & Wang, K. F. (2007). Application of Corresponding Corpus in Translation Teaching: Theoretical Basis and Implementation Principles. *Chinese Translation*, 5: 49-96.

[28] Vyatkina, N. (2016). Data-driven Learning of Collocation: Learner Performance, Proficiency, and Perceptions. *Language Learning & Technology*, 3:159-179.

[29] Larsen-Walker, M. (2017). Can Data Driven Learning Address L2 Writers' Habitual Errors with English Linking Adverbials? *System*, 1:26-37.

[30] Mao, L., Liu, Y. & Zhang, M. (2018). Research on the Effectiveness of College Student English Writing Teaching Based on Data-driven Learning. *Educational sciences: Theory & Practice*, 5:1160-1169.

[31] Li, S. (2017). Using Corpora to Develop Learners' Collocational Competence. *Language learning & Technology*, 3:153-171.

[32] Sinclair, J. (2003). Reading Concordance. London: Pearson Education Limited.