Education Insights

Article

Applying CQPweb in Academic Writing Course for ESL Biomedical PhD Students: An Exploratory Study

Lei Zhang [1,2,*]

[1] Journal Center, Chongqing University Cancer Hospital, Chongqing, 400030, China
[2] Department of Medical English, College of Basic Medical Sciences, Army Medical University, Chongqing, 400038, China
* Correspondence: Lei Zhang, Journal Center, Chongqing University Cancer Hospital, Chongqing, 400030, China

Abstract: CQPweb, a web-based fourth-generation corpus analysis tool, combines corpus resource sharing with online retrieval functions. This paper selects online corpora based on CQPweb and introduces data-driven learning (DDL) to a medical academic writing course for English as a second language (ESL) biomedical PhD candidates. The study indicates that query through CQPweb not only helps to solve lexico-grammatical problems encountered in PhD students' writing process such as word choice and collocation, but also discover and summarize syntactic discourse problems such as sentence form and text structure. Although the post-course questionnaire showed that learners faced difficulties such as "too many concordance lines" and "difficulty in formulating search formulas for syntax searches". However, overall, respondents provided positive feedback on the CQPweb-based corpus-assisted academic writing. The results show that CQPweb serves as an effective reference tool and reliable resource for academic writing, and the CQPweb-based DDL is conducive to promoting students' interest in learning, increasing the quality of academic writing, boosting their confidence, and cultivating independent learning capability. However, teachers should strengthen the training of corpus searching, especially syntax search, and also randomly thin the query results to reduce the cognitive load, protect students' learning motivation, and increase the use of corpus-assisted writing after class, in order to improve the quality of research paper writing.

Keywords: CQPweb; data-driven learning; medical research paper writing; corpus-assisted academic writing

Published: 20 January 2025

1. Introduction

English, as a lingua franca, plays a significant role in academic research and international scholarly communication, becoming an essential academic skill [1]. Recently, an increasing number of Chinese researchers have been publishing articles in English globally, with the number of publications ranking second in the world. However, the average citation rate remains relatively low, still far behind the traditional Western technological powers, which is, to some extent, related to the weak academic English writing skills of Chinese researchers [2]. Chinese scholars often find their research papers rejected by academic journals due to language issues. At the same time, more and more medical colleges and universities require the publication of research papers in international journals as a graduation requirement for PhD students. For researchers whose native language is not English, they are often encountered with the pressure of "publish or perish." In light of this, the necessity of offering academic English writing courses for non-native English as a Second Language (ESL) biomedical PhD students is self-evident. Such courses help them grasp the intrinsic characteristics and patterns of academic English, cultivate awareness of academic register and genres, and lay the foundation for future academic research and communication.

A corpus is a collection of published texts that are specifically gathered, structured, and searchable by computer programs. In recent years, with the rapid development of information technology, especially with the recent wave of artificial intelligence (AI) sweeping the globe, corpus construction and application have made considerable progress. A number of influential academic English corpora have been built both in China and internationally, including the British Academic Written English (BAWE), the Michigan Corpus of Upper-level Student Papers (MICUSP), and the Database of English for Academic Purposes (DEAP). The greatest advantage of a corpus is that it provides a large amount of authentic and reliable language materials, which have substantial teaching value and are widely applied in language teaching, especially in ESL writing courses [3-6]. However, there is less teaching research on academic English writing based on corpora in the biomedicine field, and currently, there is no research on the application of CQPweb-based corpus in ESL writing instruction among biomedical PhD students. Moreover, due to the technical barriers of corpus technology and copyright issues, academic English corpora have not yet played their role as powerful tools for language learning.

1.1. Features and Advantages of CQPweb

When dividing corpus analysis tools into generations, McEnery & Hardie [7] classified browser-server (B/S) model-based corpus retrieval tools as the fourth generation of corpus analysis tools. Among them, CQPweb is a representative of the fourth generation of corpus analysis tools, developed by Professor Andrew Hardie of Lancaster University in the UK. The web-based fourth-generation corpus analysis tool CQPweb makes corpus resource sharing and online retrieval possible, while users cannot copy or obtain corpus texts, protecting the copyright of the corpus [8]. Compared with the representative third-generation corpus analysis tools, such as Antconc and Wordsmith, CQPweb has two unique advantages: first, the integration of the corpus and analysis tool into one; and second, the use of indexing technology for fast response times [9].

The research selected is based on CQPweb-based DEAP Corpus (114.251.154.212/cqp/), which was created by Professor Xu's team at the Chinese Center for Language Education and Research. The current size of the corpus is 135 million word instances (as of March 2023) and will continue to expand. Unlike BAWE and MICUSP, which lack attention to ESL learners, the DEAP corpus is built entirely in line with the academic discipline settings in China, aiming to serve the teaching and research of ESL academic English in China, and has strong guiding implications for academic writing in China and other countries where English is not the native language. Currently, 29 academic English corpora have been established across various disciplines mostly by English teachers in Chinese universities and colleges, each with a size of about 5 million word instances, all of which are freely accessible using the user ID and password as "test". Moreover, each discipline corpus is built following a unified sampling plan and construction process. Based on the existing academic English corpora, the CQPweb-based DEAP corpus was employed in the teaching of the academic writing course for ESL biomedical PhD students at a University in Chongqing, China.

1.2. Data-driven Learning Concept

Data-driven learning (DDL) was proposed by Tim Johns in the 1990s, positioning learners as researchers who observe a large number of corpus concordance lines to explore and induce the linguistic characteristics and patterns of the target language. It is a student-centered, exploratory, and discovery-based learning model [10]. According to language acquisition theory, the memorization of linguistic knowledge does not guarantee the correct use of language; a substantial amount of language input is required for students to master the language they have learned [11]. Unlike traditional one-way teaching by teachers, DDL based on corpora presents a large number of published and authentic examples of language use, providing not only a large amount of language input but also input in

real contexts. Furthermore, unlike the direct instilling of language knowledge, the autonomous exploration and discovery-based learning of language use patterns is beneficial to enhance learners' initiative and interest, leading to a deeper understanding and more lasting memory of the learning content [12].

2. Application of CQPweb in Biomedical Academic Writing Course for ESL PhD Students

Based on classification criteria, corpora can be categorized in various ways. According to application orientation, they can be divided into general corpora and specialized corpora. Unlike general corpora, specialized corpora are built for specific research purposes, collecting texts from a particular field, such as collecting research articles from highly influential journals from the basic medical sciences field, for corpus construction to analyze the linguistic characteristics of specific fields and compare them with other fields. Considering the teaching target as doctoral students, who are advanced ESL learners with high initiative and strong self-learning abilities [13,14], it is decided to attempt to apply the DDL based on CQPweb to the compulsory writing course among 23 biomedical PhD students in a class for one semester.

DDL emphasizes students' bottom-up exploration, discovery, and summarization of language usage patterns, rather than conventionally instilling knowledge from teachers based on their own language accumulation. Teachers can fully leverage the students' primary role by creating teaching cases or lecture notes from corpus search results, designing relevant teaching activities centered on students, guiding them to discover language usage rules, and promoting their language sensitivity. Compared to the limited number of language examples in conventional teaching, DDL can provide a wealth of language usage examples, which is incomparable in terms of the quantity of language input provided to students. Moreover, since the academic corpora used in this study are collections of a large number of authentic language instances that have been edited and published, rather than language examples written by compilers based on their own experience and intuition in conventional textbooks, DDL is superior to conventional language teaching in both quantity and quality.

Considering the characteristics of students' major distribution, the Clinical Medicine Academic English Corpus (http://114.251.154.212/cqp/meddeap/) and the Life Sciences Academic English Corpus (http://114.251.154.212/cqp/biodeap/) are mainly adopted in class, which are related to clinical medicine and biomedical sciences. By introducing the basic usage of the CQPweb online corpus to PhD students and showing them how to use functions such as Frequency Breakdown, Distribution, and Collocation to correct language errors in writing, strengthen article readability, and increase English paper writing literacy, it is possible to better promote academic exchanges with international peers. Due to space limitations, this study mainly demonstrates the application of CQPweb in medical English paper writing from the levels of vocabulary, grammar, and sentence structure and discourse.

The online corpus tool CQPweb can quickly perform batch searches at the micro-level of vocabulary and grammar, and further analyze the search results through random sampling, frequency breakdown, collocation calculation, etc. At the vocabulary level, we mainly focus on word choice and collocation.

2.1. Word Choice and Collocation

Determining the right word is the most fundamental issue in writing. As second language learners, even if students know the meaning of a word, they may still wonder which word to use and its usage in the written language domain. For example, there are degree adverbs such as "essentially", "fundamentally", and "basically" to express the meaning of "in a very important or basic way". Which one is used more in the written language context? Students can search for answers in the corpus with questions. Using the Frequency

Breakdown function of CQPweb, the occurrence frequency and percentage of each adverb can be easily found in the corpus: "essentially" (91, 71.09%), "fundamentally" (23, 17.97%), "basically" (14, 10.94%). These data are self-evident in answering the students' confusion, suggesting that in written language, the first two words should be used as much as possible. Meanwhile, teachers can guide students to analyze the reasons behind the data, helping them recognize that the word "basically" is informal and should be avoided in academic texts to enhance their awareness of academic English register.

In addition, students can use the Collocation function to observe the left and right collocations of the search term, summarize and generalize higher-level colligation relationships, and further explore their semantic tendencies and semantic prosody. Below is a brief illustration of the word item analysis approach with the search term "insights" as an example.

By browsing the search lines, it is found that a typical collocation word for "insights" on the right is the preposition "into," and the search term often forms colligation with verbs (provide, yield, gain) on the left. To further explore the adjectives in the position one to the left of the search term, the CQPweb collocation calculation function can be used, with the position set to 1 to the left, and the statistical parameter selected as Log-likelihood, while restricting the part of speech to J.* (adjectives), the following results can be obtained (Figure 1). By observation, it is not difficult to see that the adjectives preceding "insights" all express some novel (new, novel) or important (important, valuable) meanings, demonstrating positive semantic prosody, which also highlights the purpose of medical research to explore new knowledge and the importance of research findings' implications and application.

There are 563 different words in your collocation database for "[word="insights"%c]". (Your query "insights" returned 149 matches in 95 different texts) [0.138 seconds - retrieved from cache]

No.	Word	Total no. in whole corpus	Expected collocate frequency	Observed collocate frequency	In no. of texts	Log-likelihood value
1	new	2,254	0.059	21	17	208.212
2	novel	587	0.015	12	10	137.216
3	important	2,921	0.076	11	11	88.372
4	further	3,028	0.079	8	7	58.464
5	valuable	177	0.005	3	3	33

Figure 1. Distribution of adjectives one position to the left of "insights into".

2.2. Phrase Discrimination

Through searching in CQPweb, not only can the usage of single words be explored, but also the usage of multi-word expressions (MWEs). In academic English, MWEs are an important component of academic discourse and play a crucial role in the teaching of academic English [15]. For non-native English learners, how to use MWEs accurately and idiomatically is a challenge. Compared with conventional dictionaries or machine translation engines, CQPweb provides language usage frequency and context, revealing the usage patterns of MWEs, which is helpful for synonym discrimination. For instance, when expressing that this study is different from previous studies for comparison, ESL learners can find phrases such as "in contrast," "by contrast," and "on the contrary" through dictionary lookup. So, which one is used in academic texts? By inputting the search formula "(in contrast|by contrast|on the contrary)" under simple query mode, the occurrence frequency of the above phrases can be searched in the corpus instantly. By clicking on the frequency breakdown function, it can be found that "on the contrary" has the lowest frequency, accounting for only 2.08%, which preliminarily indicates that "on the contrary" has a lower frequency in academic written language texts, and "in contrast" and "by contrast" should be preferred in writing. Teachers can guide students to further explore the

reasons behind the data, triggering students' realization that "on the contrary" expresses opposition to others' opinions or previous views and has a subjective nature, which should be avoided in medical academic papers to enhance their awareness of academic genre conventions.

No.	Search result	No. of occurrences	Percent
1	in contrast	1164	83.68%
2	by contrast	198	14.23%
3	On the contrary	29	2.08%

Figure 2. Distribution and proportion of phrases expressing "contrast" in the corpus.

2.3. Syntactic Structure Analysis

The advantage of online corpora lies in the rapid batch retrieval of micro-level vocabulary and grammatical features [16], but there is still a lack of exploration in the macro-level sentence structure and discourse language structure. This study selects the representative passive voice structure in academic papers for demonstration.

One of the characteristics of medical journal writing is the frequent use of the passive voice, which has become a convention [17,18]. Since the search formula for the passive voice requires the use of the more complex CQP syntax search, teachers need to write the search formula in advance ("[pos="VB.*"][pos="XX"]*[pos="R.*"]{0,4}[pos="VVN"]") or embed the search results in the courseware in advance in the form of hyperlinks. First, select the complex search (CQP syntax) mode, and then input the search formula. Taking the self-built HZAU CQPweb corpus as an example [8], the distribution of search items in different parts of academic articles can be further viewed, that is, the specific distribution of passive voice in the common structural moves of papers (Figure 3), and the results show that the passive voice is most frequently used in the methods section, allowing students to have an overall understanding and grasp of the distribution of passive voice in various parts of the paper.

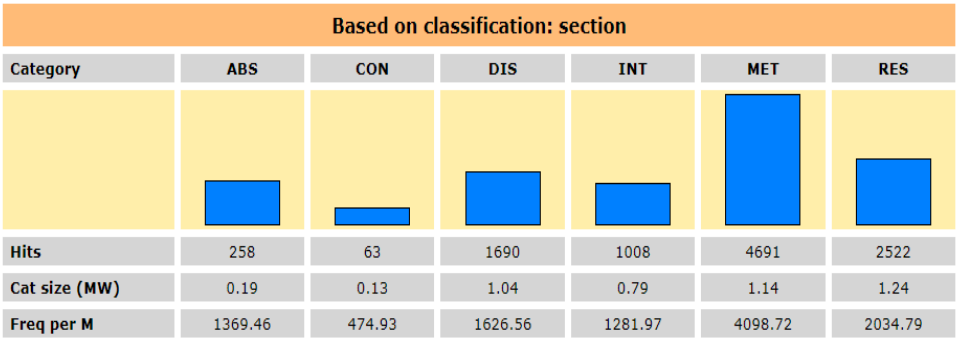

Figure 3. Distribution of passive voice in research papers based on different sections.

*Abbreviation: 1. ABS: Abstract 2. Con: Conclusion 3. DIS: Discussion 4. INT: Introduction 5. MET: Methods 6. RES: Results.

After students have a macro grasp of the passive voice, teachers can guide students to explore its characteristics and patterns from a micro perspective. Taking the Clinical Medicine Academic English Corpus as an example, the same search formula is input, resulting in a total of 90,054 search results, appearing in all 1,186 texts. Through frequency breakdown, the top ten most frequently used verbs in the passive voice in this Corpus and their distribution can be summarized and calculated (Table 1).

Table 1. The top 15 verbs in passive voice in the clinical medicine academic English corpus.

Verb	Observed Frequency	Dispersion (%)
use	4526	83.81
associate	3837	67.45
perform	2976	65.01
find	1908	56.24
report	1808	58.77
consider	1798	62.48
show	1744	57.17
observe	1683	48.06
assess	1205	46.96
include	1179	47.47
obtain	1144	47.98
calculate	1045	41.15
measure	970	36.42
define	968	39.97
identify	949	41.15

Furthermore, research has shown that both direct and indirect learning can play a role, and in addition to incidental learning, courses that encourage explicit attention to target vocabulary produce better results [19]. One application of keyword analysis in teaching is to generate a Word Cloud [20] to emphasize the relative importance of specific vocabulary in their disciplines to students. The word cloud shows the top 15 verbs used in the passive voice in the Clinical Medicine Academic English Corpus, with the size of each word reflecting its frequency of occurrence (Figure 4).

Figure 4. Word cloud of the top 15 verbs in the passive voice in the clinical medicine academic English corpus.

It is worth noting that most current corpus research is still at the micro-level of vocabulary and grammar, and there is less research on the macro-level of sentence structure and discourse, which is both the advantage and limitation of corpus methods. Follow-up studies can further explore sentence structures from the corpus, such as emphatic sentences, evaluative clauses, etc.

3. Post-Class Questionnaire Survey

After the course, a questionnaire survey on the effectiveness of corpus-assisted academic writing based on CQPweb was conducted, including 1 single-choice question, 4

multiple-choice questions, 2 subjective questions, and 1 question on attitude statements via a 6 Likert scale. The results showed that before taking the course, the vast majority of PhD students (95.65%) had not heard of corpora, which to some extent reflects the insufficient application of corpora in academic writing for postgraduate students and their potential as powerful tools for language learning has not yet been fully realized. Most students (69.57%) would use machine translation engines for reference when encountering vocabulary and grammar issues, even though the majority of respondents (60.87%) knew that machine translation was clumsy to some extent. Regarding the advantages of corpus resource retrieval, students held that the greatest advantage of using corpora was helping to learn the usage of fixed phrases and idioms (78.26%), followed by understanding the meaning and collocation of vocabulary (56.62%), and that analyzing search results helps to discover grammatical rules (56.52%). Meanwhile, respondents indicated that there were some challenges and difficulties in corpus-assisted academic writing based on CQPweb, including difficulty in screening and/or processing a large number of sentences retrieved (65.22%), difficulty in formulating correct search strategies/methods, and difficulty in writing regular expressions for complex searches (43.48%), indicating the need to develop strategies to address these challenges and issues. Overall, however, students gave positive feedback on corpus-assisted academic writing based on CQPweb, believing that the academic writing course should include modules on corpus-assisted writing and training on its search strategies, and expressed willingness to use corpora in future academic writing to increase their confidence in academic English writing (Table 2).

Table 2. Students' attitude scale towards corpus-assisted academic writing*.

Item	1(No.)	2(No.)	3(No.)	4(No.)	5(No.)	6(No.)	Mean	Standard Deviation
I am willing to use academic English corpora as reference resources in my future English writing	1	0	1	2	9	10	5.09	1.20
Corpus is more helpful for my English writing than dictionaries and translation websites	1	0	1	6	8	7	4.78	1.20
I believe the use of corpora is more helpful for English writing than reading.	1	0	3	4	8	7	4.7	1.29
The teaching content of academic writing courses should include corpus teaching.	1	0	0	4	9	9	5.04	1.15
The teaching of academic English courses should include the teaching of search strategies for mainstream academic English corpora.	1	0	1	1	11	9	5.09	1.16
Next time I write academically, I want to use a corpus.	1	0	1	3	9	9	5	1.20
I would recommend the use of corpora to friends who are learning English.	1	0	1	3	10	8	4.96	1.19
When I encounter problems in academic writing, I will seek help in corpora.	1	0	1	4	8	9	4.96	1.22

Learning corpora can increase my confidence in academic English writing.	1	0	2	2	7	11	5.04	1.30
Total	9	0	11	29	79	79	4.96	1.2

*Where 1-6 indicates a range from strongly disagree to strongly agree.

4. Discussion

4.1. Addressing DDL Challenges

Admittedly, facing the vast amount of data in corpus-assisted DDL, students may feel overwhelmed and discouraged, which can harm their enthusiasm for learning. To address this issue, on one hand, we can draw on Sinclair's [21] practice of randomly sampling 30 concordance lines of search results, and in teaching, we can use the CQPweb query result random Thin function to let students observe a certain number of concordance lines and summarize linguistic phenomena. On the other hand, we can also tailor language materials following the level of students, referring to the practice of mini-files proposed by Liang et al [16]. First, use the CQPweb download search results function to save the randomly sampled concordance lines as a separate small electronic text, i.e., mini-files, and then edit the mini-files as necessary according to the teaching objectives, removing unsuitable language materials, and using language materials that match the students' language level to achieve a customized, student-centered teaching effect. This also reflects the student-centered teaching philosophy, designing teaching activities around the teaching objectives. Concurrently, based on mini-files, supplemented with text processing tools such as PowerGrep, exercises can be compiled for students' in-class and after-class activities.

4.2. Corpus Advantages Over GenAI

With the rapid development of large language models (LLM) and generative artificial intelligence (GenAI), the explosive growth of applications led by ChatGPT has greatly changed our daily lives. Does this rise of LLM and GenAI mean the fall of corpus-assisted DDL? Not necessarily. It is important to note that corpora have the following advantages over GenAI in understanding language patterns and usage. First, the corpus provides clear and reliable data sources. For example, users can know the texts that make up of large general corpora and extract the full text of these corpora when needed. Considering the current "black box" attribute of GenAI, this is highly difficult to realize in the current LLM [22]. Second, the data in corpora are published texts, ensuring the authenticity and applicability of the content, which is particularly important for ESL learners, as they can rely on these published language data to verify language usage. Third, the research results of corpora are reproducible, and users can easily replicate given findings on the same dataset using the same queries, which is a major advantage in consistency. Fourth, corpora encourage exploratory, discovery-based learning, actively involving learners in the exploration and learning of data, rather than passively receiving information [23]. In summary, corpora still have irreplaceable value in language education and research, promoting deeper language learning and research. However, it is important to note that GenAI has its unique advantages in expanding our understanding of language usage, which corpora have not yet achieved. It is believed that by combining corpus and GenAI methods, ESL learners can gain a more comprehensive understanding of how language operates in different contexts than by using a single method alone.

4.3. Combine Corpus with Conventional and Emerging Teaching Tools

It is worth noting that although online corpora have obvious advantages, they cannot replace conventional language learning tools, including dictionaries and style manuals. To some extent, the CQPWeb-based corpus is a supplement to these reference books and

tools. For example, when students encounter tricky grammar or usage problems in academic writing and cannot find answers in reference books or dictionaries due to the continuous evolution of language and the not timely updated reference resources, they can search the corpus to obtain authentic and reliable evidence, exploring how a large number of authors or editors handle this problem during the paper publication process. Additionally, due to the limited size of corpora, it means that some questions cannot be answered through corpus query. Therefore, in addition to combining with conventional resources, we also need to combine with emerging transformative tools based on LLM and GenAI, such as ChatGPT, to realize complementary effects. Combining CQPweb online corpus with conventional and emerging resources and tools, and assessing their complementary roles and joint effects through qualitative and quantitative methods, will be a topic worth exploring in the future.

5. Conclusion

This study introduces the application of the fourth-generation online corpus retrieval platform CQPweb-based DDL in the teaching of English academic paper writing for ESL medical PhD students, which is a beneficial exploration and attempt to the existing academic English writing teaching model among biomedical postgraduates. By introducing the teaching concept of DDL, students engage in exploratory, discovery-based autonomous learning, obtaining a large amount of published, reliable language input from corpus searches, strengthening their awareness of academic language and genres in the related field, and helping them master the linguistic characteristics and patterns of medical academic English, laying the foundation for future international academic exchanges. However, there are still many limitations in the current teaching of corpus-assisted academic writing based on CQPweb, such as the difficulty in writing complex search strategies, too many concordance lines, the cost of time and energy, and the combination of corpora with conventional and emerging resource tools. Solutions to these challenges need to be further optimized to promote the application and promotion of DDL based on CQPweb in academic English writing courses among biomedical PhD students, assisting non-native researchers in academic English writing and enhancing their paper writing skills.

References

1. K. Hyland, Academic discourse: English in a global context, Continuum International Publishing Group, 2009, pp. 1–215.
2. Y. Chen and X. Xiang, "Study on the corpus-based teaching of English academic writing," Mod. Educ. Technol., vol. 12, pp. 84–89, 2015, doi: 10.3969/j.issn.1009-8097.2015.12.013.
3. M. Chen and J. Flowerdew, "Introducing data-driven learning to PhD students for research writing purposes: A territory-wide project in Hong Kong," English for Specific Purposes, vol. 50, pp. 97–112, 2018.
4. N. Dolgova and C. Mueller, "How useful are corpus tools for error correction? Insights from learner data," J. Engl. Acad. Purp., vol. 39, pp. 97–108, 2019.
5. P. Liu and P. Wang, "An exploration of corpus-based project-oriented innovative experimental teaching," in Proc. 2nd Annu. Int. Conf. Social Sci. Contemp. Humanity Dev., Dec. 2015, pp. 250–254, Atlantis Press.
6. Q. Luo and M. Shi, "On the application of CQPweb-assisted data-driven learning in English writing classes," J. Yangzhou Univ. (Higher Educ. Study), vol. 06, pp. 111–118, 2020, doi: 10.19411/j.cnki.1007-8606.2020.06.018.
7. A. Hardie, "CQPweb—combining power, flexibility and usability in a corpus analysis tool," Int. J. Corpus Linguist., vol. 17, no. 3, pp. 380–409, 2012.
8. P. Liu and L. Wu, "The construction and application of the online corpus analysis system CQPweb: A case study of HZAU CQPweb," China Univ. Teach., vol. 05, pp. 70–75, 2016, doi: 10.3969/j.issn.1005-0450.2016.05.015.
9. J. Xu and L. Wu, "Web-based fourth generation corpus analysis tools and the BFSU CQPweb case," Technol. Enhanced Foreign Lang. Educ., vol. 05, pp. 10–15+56, 2014.
10. T. Johns, "Should you be persuaded: Two samples of data-driven learning materials," vol. 4, pp. 1–16, 1991.
11. F. Jiang, Corpus and academic English research, Foreign Language Teaching and Research Press, 2019. ISBN: 9787513575164.
12. Z. Li, "Construction and application of a corpus-based teaching platform for English translation of Chinese classics," The Educ. Rev., USA, vol. 8, no. 10, pp. 1209–1216, 2024.

13. D. Lee and J. Swales, "A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora," English for Specific Purposes, vol. 25, no. 1, pp. 56–75, 2006.
14. L. Flowerdew, "Using corpus-based research and online academic corpora to inform writing of the discussion section of a thesis," J. Engl. Acad. Purp., vol. 20, pp. 58–68, 2015.
15. X. Lu, O. Kisselev, J. Yoon, and M. D. Amory, "Investigating effects of criterial consistency, the diversity dimension, and threshold variation in formulaic language research: Extending the methodological considerations of O'Donnell et al. (2013)," Int. J. Corpus Linguist., vol. 23, no. 2, pp. 158–182, 2018.
16. M. C. Liang, W. Z. Li, and J. J. Xu, Using corpora: A practical coursebook, Foreign Language Teaching and Research Press, Beijing, 2010, pp. 12–13.
17. D. Biber, "Intra-textual variation within medical research articles," in Variation in English, pp. 108–123, Routledge, 2014.
18. R. J. Amdur, J. Kirwan, and C. G. Morris, "Use of the passive voice in medical journal articles," AMWA J., vol. 25, no. 3, 2010.
19. R. Ellis, The Study of Second Language Acquisition, 2nd ed., Oxford University Press, 2008.
20. A. Gilmore and N. Millar, "The language of civil engineering research articles: A corpus-based approach," English for Specific Purposes, vol. 51, pp. 1–17, 2018.
21. J. Sinclair, Reading concordances: An introduction, 2003.
22. P. Crosthwaite and V. Baisa, "Generative AI and the end of corpus-assisted data-driven learning? Not so fast!," Appl. Corpus Linguist., vol. 3, no. 3, p. 100066, 2023.
23. P. Crosthwaite and V. Baisa, "A user-friendly corpus tool for disciplinary data-driven learning: Introducing CorpusMate," Int. J. Corpus Linguist., 2024.