ISSN:2790-167X Volume-5-(2023)

# Research on Corpus-based Foreign Language Teaching in China: A Bibliometric Analysis

Meixia Chen

School of Foreign Studies, North Minzu University, Yinchuan, Ningxia;

cmx@nmu.edu.cn

**Abstract.** Based on the data published in major journals, this study examines the research articles on China's corpus-based foreign language teaching published from 2000 to 2023. In order to thoroughly explore the development status, the number of articles published, the journals involved, keywords, authors, institutions, and research trends are analyzed. It is suggested that corpus-based foreign language teaching research in China has yielded fruitful results, supported by diverse research perspectives, micro and macro research focus, and virtual foreign language teaching practice. At the same time, the study also reveals weaknesses. These include less cooperation among authors and institutions, uneven research resources, and unbalanced development of teachers' information literacy.

**Keywords:** corpus; foreign language teaching; CiteSpace; visual analysis.

## 1. Introduction

Baker [1] defines corpus as any collection of running texts ... stored in electronic form and analyzable automatically or semi-automatically (rather than manually). Compared with small corpus relying on language learners' intuition and introspection, large corpus provides rich and authentic language materials that teachers can use to describe language features and generalize language patterns [2]. In this way, corpus is a useful tool to improve foreign language teaching.

Since the 1980s, corpus linguistics has developed rapidly overseas, and corpus linguists thus take an empirical approach to describing language: they insist on the primacy of authentic, attested instances of use [3]. Following this trend, domestic scholars began to introduce corpus and corpus-based research to China and sought to apply this new perspective to foreign language teaching. In 1986, Professor Yang Huizhong published an article on computer corpus and foreign language teaching, marking the beginning of studies on corpus-based foreign language teaching in China. Since then, many studies on corpus-based foreign language teaching have been conducted from different perspectives. Some scholars examine the application of corpus in foreign language teaching from a macro point of view [4, 5, 6], and explore how corpus can facilitate foreign language teaching [7, 8, 9, 10]. As research grows, the need to construct learner-based corpus and platforms calls for attention [11, 12, 13, 14]. In recent years, the rapid development of multimedia has urged teachers to adopt multi-modal materials in their teaching to capture students' interest, thus led to the construction of multi-modal corpus [15, 16]. Studies in recent years have focused on how corpus assist translation teaching [17, 18, 19, 20, 21]. These studies have greatly enriched foreign language teaching and promoted its diversity.

In 2019, the Ministry of Education issued a blueprint for China's education modernization, proposing a guideline to accelerate education reform to meet the challenges of the information age. Foreign language teaching always involves software, labs, and tools, so it urgently needs reform. Teachers and scholars are actively exploring approaches to apply technology in foreign language teaching to maximize learning outcomes. Zhong [22] suggests that the integration of information technology in foreign language teaching is an attempt to study the psychology of learners through human-machine synergy, natural language, corpus and other technologies to simulate human language. Against this background, it seems necessary to look back at the development course of foreign language teaching in China and to get an insight into the research focus and trends in this field. The present study uses the bibliometric technology CiteSpace to delineate the main research

topics and focus of corpus-based foreign language teaching in the China National Knowledge Infrastructure (CNKI) from 2000 to 2022.

## 2. Research design

### 2.1 Data collection

In this study, relevant academic literature is retrieved mainly from two comprehensive databases that cover a wide range of interdisciplinary subjects. One database is Chinese Social Sciences Citation Index (CSSCI), a database created in 1998 and A Guide to the Core Journals of China. The two databases are the most significant and reliable knowledge-based information resources in China. To get an exhaustive list of relevant literature from the databases, key words are specifically designed. “Corpus plus foreign language teaching” are used simultaneously as the key words with a time span of 22 years. Altogether, 1,675 articles were retrieved. Book comments, conference reviews, prefaces and literature which are not directly related to the present research are excluded in order to get a valid number of 1,032 articles for bibliometric analysis.

### 2.2 Research method

Mapping knowledge domains were first introduced to China by Chen in 2005, bringing a new way of bibliometric analysis for carrying out research. Mapping knowledge domains is a software tool that helps visualize the research evolution of a particular academic field. With the help of mapping knowledge domains, researchers can discern the implicit connections between knowledge units, discover research focuses and estimate future research trends. Nowadays, frequently-used bibliometric tools include BibExcel， CiteSpace， VOSviewer， CitNetExplorer, among which CiteSpace gains great popularity for its efficiency and simplicity. With adequate data, researchers can immediately generate visual information without much human intervention. In this research, CiteSpace is used to produce the publication of articles, journals, keyword co-occurrence, author and institution collaboration network and burst words to analyze the evolution of corpus-based foreign language teaching studies in China.

## 3. Descriptive analysis of corpus-based foreign language studies

Before applying CiteSpace to generate a knowledge map, the author firstly provides a descriptive analysis of corpus-based foreign language studies in China, including publication output and distribution of publications.

### 3.1 Publications output

As is shown in figure 1, studies on China’s corpus-based foreign language teaching generally shows two trends. Since the year 2000, scholars have begun to experiment with corpus tools for sentence segment, collocation, keywords and clusters to explore language patterns and provide authentic language materials for foreign language teaching. Their experiment has led to a surge in academic publication. Publication output reached nearly 200 in 2011, indicating a consistent growth of interest in corpus-based teaching. During this period, several large-scale learner corpus were built, such as the Chinese Learner English Corpus (CLEC), Spoken and Written English Corpus of Chinese Learners(SWECCL), Parallel Corpus of Chinese EFL Learners(PACCEL). Construction of these corpus greatly promoted the implementation of corpus in China’s foreign language teaching. After 2011, studies in the field showed a gradual decline with publication output decreasing year by year. It doesn’t mean that corpus-based teaching was abandoned by teachers and scholars alike. On the contrary, it was rather a shift from macro to micro perspective. Scholars began to view the function of corpus from a more rational and critical point of view.

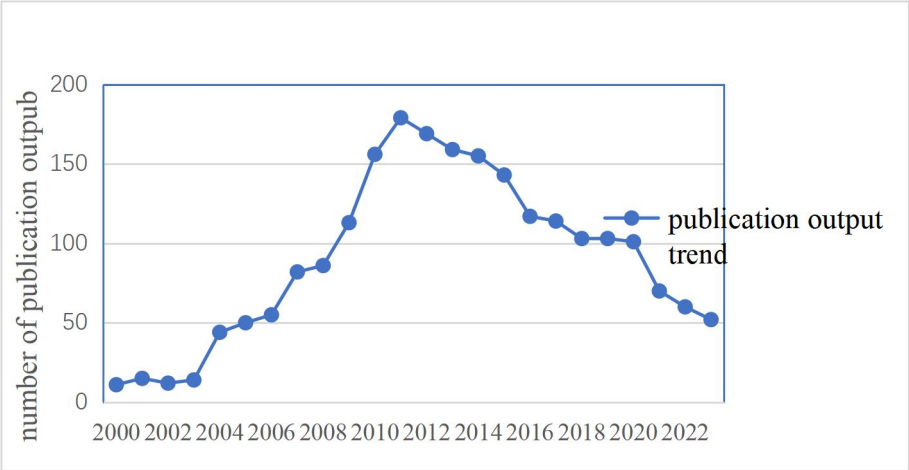


Fig. 1 Line graph of publication output from 2000-2022

### 3.2 Distribution of publications

Statistics show that research results related to corpus-based foreign language teaching have mainly been published on some core journals included in CSSCI (502 articles) and A Guide to the Core Journals of China (530 articles) published by Peking University Press. Core journals like Technology Enhanced Foreign languages, Foreign Language World, Journal of PLA University of Foreign Languages rank the top ten (Figure 2). Distribution of publications can indicate research priorities. Scholars believe that corpus-based approach is an effective way to modernize foreign language teaching, and have been dedicated to conducting high quality research to ensure reliable, feasible and sustainable results that can be directly applied to language teaching practice.

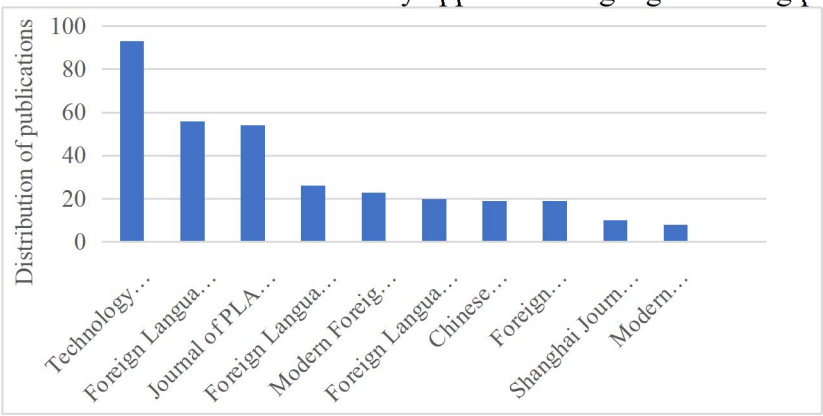


Fig. 2 Distribution of publications

## 4. A visual analysis of corpus-based foreign language teaching

CiteSpace offers visualized data in different forms, including keyword co-occurrence, author and institution collaboration network, country collaboration network, author co-citation and journal co-citation. As article data exported from CNKI cannot be used to generate author co-citation and journal co-citation, this paper just focuses on keyword co-occurrence, author and institution collaboration network and burst terms.

### 4.1 Keyword co-occurrence analysis

'The analysis of keyword co-occurrence is an effective way to show emerging trends and to track topics of research over time because keywords provide a concise and precise high-level summarisation of a document' [23]. In this research, papers from CNKI are imported into CiteSpace

(V6.1.R6) to visualize the status quo of corpus-based foreign language teaching in China. Timespan ranges from 2000 to 2023.Time slice is set as one year, node type as keyword, linkage intensity as Cosine. Then a keyword co-occurrence knowledge map of foreign language teaching is generated (Figure 3). CiteSpace provides two metrics, Modality value (Q value) and Weighted Mean Silhouette value ( S value), based on the clarity of the network structure and clustering, which can be used as a basis for us to judge the effectiveness of the map. According to the map, the clustering Modularity Q value is 0.6252 (>0.3) and Weighted Mean Silhouette value is 0.9228 (>0.5), which indicates that the clustering analysis is well structured and the modules in the network are correlated and are highly independent from each other.

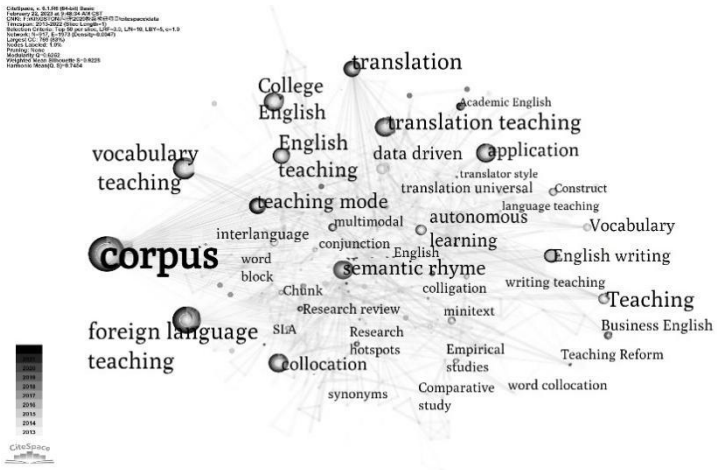


Fig. 3 Keyword co-occurrence map of CNKI（2000-2022）

The mapping presents circular nodes of different sizes, each node representing a keyword. Node size reflects the frequency of a keyword. Generally, the larger the node, the larger the font. The link between the nodes indicates the co-occurrence intensity of a keyword, reflecting the co-occurrence coefficient between keywords. As can be seen from Figure 3, 362 node words and 542 links are generated, with a density value of 0.0047. The keyword “corpus” is the largest node in the map. The nodes of foreign language teaching, vocabulary teaching, college English, translation teaching, English teaching, semantic rhyme, collocation, application, English writing, teaching mode, and independent learning are also very conspicuous, appearing in each time slice and with the darkest color in the outer circle, indicating that these node words are still the hot spots of research in the field. An analysis of co-occurrence of keywords reflects that corpus-based foreign language teaching research in China has formed a relatively mature academic research network.

In addition, the small nodes scattered around the mapping reveal the changing trends of hot spots in corpus-based foreign language teaching. Figure 3 shows that the nodes of interlanguage, data-driven, word block, translation universal, translator style, and connective are smaller and lighter in color, indicating that these node words used to be the focus of study, but now they receive less attention as researchers are getting a deeper insight into the implication of adopting corpus in teaching.

### 4.2 Institute and author collaboration network analysis

In generating the institute and author collaboration network, the paper sets time slice to one year, node type to institution, and the threshold to 3 (Figure 4). A total of 449 institutions are generated, with 79 inter-institution connection lines and a density of 0.0008. Figure 4 shows that the School of Foreign Languages of Shanghai Jiaotong University and Faculty of International Studies of Henan Normal University have the largest nodes, and they are the main research bases of corpus-based foreign language teaching in China. Other universities like Shanghai International Studies University, Beijing Foreign Studies University, Guangdong University of Foreign Studies,

Nanjing University, Liaoning Normal University, South China Normal University, Shandong University, Northeastern University and Yangtze University have also been actively involved in research on corpus-based foreign language teaching. Nevertheless, the number of linking lines suggests that cooperation between institutions and universities is relatively rare. It seems that most universities tend to carry out studies independently rather seek cooperation, resulting in a undeveloped collaboration network. Besides, it can also be seen in Figure 4 that institutions conducting studies on corpus-based foreign language teaching are mainly located in east China, while those in central and southwest of China are engaging less in this research field. Only one or two Institutions in northeast have done relevant research. This reflects an uneven distribution of research resources concerning corpus-based foreign language studies. Because of geographical advantages, institutions in east China have access to more resources and up-to-date information technologies.

The author collaboration network can be generated by setting the node type as author (Figure 5). The result shows that this network is composed of 427nodes and 67 linking lines, with a density of 0.0007. The nodes for authors like He An'ping, Li Guangwei, Wu Junli, Wang Lifei, Wang Kefei and Xie Jiacheng are larger than the rest, indicating that they are prolific scholars in this field. Most scholars have made some attempts but failed to explore further. It can be noticed that the linking lines between authors are rather weak, and only a few scholars have cooperated in their research. In general, academic cooperation concerning studies on corpus-based foreign language teaching is still among scholars.

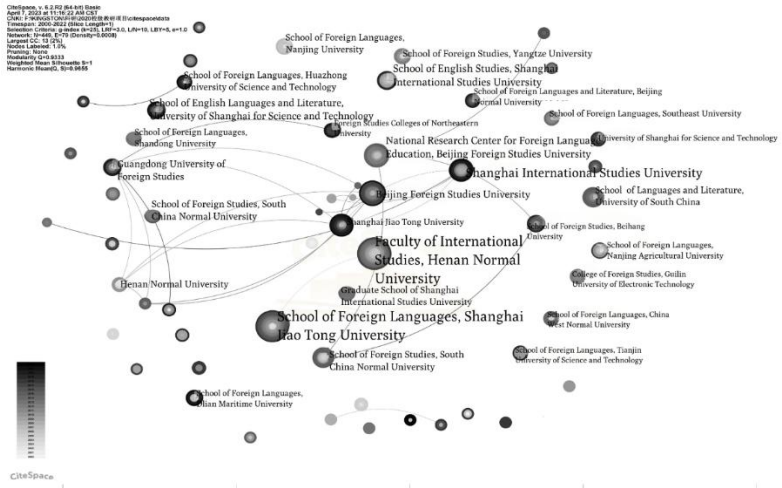

Fig. 4 Institution collaboration network

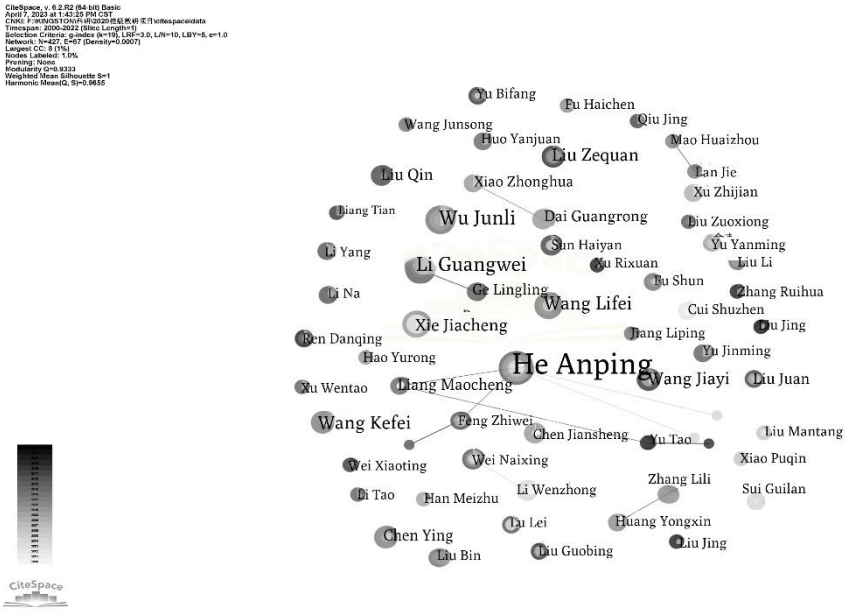


Fig. 5 Author collaboration network

### 4.3 Research trends analysis

Research frontiers as defined by Citespace emphasize the characteristics of new trends and sudden changes and are represented by emergent words obtained by the Kleinberg emergence detection algorithm [24]. The burst word is a term that appears more frequently at a certain time in the context of its subject knowledge. Analysis of burst words helps researchers keep track of frontiers of development and academic trends in the field, and thus better determine the future changes in a scientific field [25]. In order to study the evolution of hotspots in the corpus of foreign language teaching, this study analyzes the keyword mutation characteristics of 1,032 documents. Keyword is set as the Node Type and "Burst terms" as the Term Type. Table 1 shows the top 20 burst words.

Table 1. Burst terms

| No. | Keywords | Burst strength | Burst year | No. | Keywords | | Burst strength | Burst year |
|---|---|---|---|---|---|---|---|---|
| 1 | foreign language teaching | 9.28 | 2000-2006 | 11 | SLA | | 2.57 | 2010-2011 |
| 2 | index | 2.74 | 2001-2007 | 12 | data-driven | | 3.22 | 2012-2016 |
| 3 | language studies | 2.5 | 2002-2010 | 13 | mini-texts | | 3.19 | 2013-2017 |
| 4 | annotation | 2.73 | 2003-2009 | 14 | vocabulary | | 2.52 | 2013-2016 |
| 5 | vocabulary teaching | 5.55 | 2005-2010 | 15 | translation | | 3.11 | 2014-2022 |
| 6 | developing trend | 2.76 | 2005-2007 | 16 | Constructi-on | | 2.79 | 2014-2019 |
| 7 | index | 3.08 | 2009-2012 | 17 | review | | 2.31 | 2014-2015 |
| 8 | interlanguag e | 2.26 | 2009-2010 | 18 | semantic prosody | | 4.65 | 2015-2020 |
| 9 | English vocabulary | 3 | 2010-2011 | 19 | business English | | 2.98 | 2015-2020 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 10 | language teaching | 2.68 | 2010-2012 | 20 | translation teaching | | 2.66 | 2016-2022 |

The higher the burst value of the keyword and the newer the year of emergence, the more likely the keyword is a current research hot spot in the field and will become a future research hot spot. In Table 1, "foreign language teaching" has the highest burst value, which reflects the focus of research at the beginning of the introduction of the corpus into foreign language teaching. However, the topic of "foreign language teaching" is broad and unfocused, and only emerged until 2006. The emergence of keywords such as "retrieval, language research, vocabulary teaching, and mediated language" indicates that the academic community has a clearer understanding of the application of corpus in foreign language teaching, and the focus of research has shifted from the initial macroscopic research to the microscopic level, which is also the period when the research on foreign language teaching of corpus has made rapid development and promoted the next stage of research perspective. From 2012 onwards, the theme words "data-driven, microtext, translation, construction, semantic rhyme, business English, and translation teaching" have become hot topics, among which "translation" and "translation teaching" have emerged until now. "This reflects the consensus of academics and front-line teachers on the application area of corpus, that is, the large-scale natural language contained in the corpus forms a natural fit with translation teaching research, provides a large amount of authentic corpus for translation research, facilitates learners' access to authentic and natural language resources, and promotes data-driven teaching and learning. The corpus provides a large amount of authentic corpus for translation research, facilitates learners' access to authentic and natural language resources, and promotes data-driven foreign language teaching and learning.

Many scholars have published books and research papers on corpus-based translation studies during this period, including Liu Ping, Wang Kefei, Hu Kaibao and Dai Guangrong. They have expounded on how to collect corpus materials, construct a corpus, and apply it into teaching, which greatly promotes the empirical application of corpus in foreign language teaching.

## 5. Limitations of corpus-based foreign language teaching in China

As mentioned above, studies on corpus-based foreign language teaching have produced fruitful results. Researchers have conducted multidimensional research from diverse perspectives, which lays a solid foundation for future research. Nevertheless, there are still concerning research collaboration, scope, perspectives and resources. The present study proposes the following three suggestions for further improvement.

### 5.1 Coordinating research strength and build an academic community

In the 20th century, British scholar Brownie first introduced the concept of Academic Community (AC) in his article "The Autonomy of Science". An academic community is a collaborative group voluntarily formed by researchers and research institutions with the same or similar scientific interests, and the members of the group realize a common value vision through collaborative research around a certain topic [26]. The mapping of the cooperative network of institutions and authors in the previous section shows that domestic research on foreign language teaching and learning in corpus reflects the characteristics of independent research dominated by scattered research forces and lack of synergy, which to a certain extent restricts the development of research diversification in this field. Research institutions and authors should actively integrate their strengths, exchange and learn from each other, and realize an academic community with integrated thinking and shared resources. Especially, western universities should actively seek help from eastern universities in the absence of technical and research resources, and improve the state of academic development in western regions by establishing inter-university mutual aid networks and building mutual aid platforms.

### 5.2 Enrich resources and create diverse teaching mode

According to the results of CiteSpace spectrogram, the current domestic corpus foreign language teaching research is largely based on textual corpus with a single form of resources, which cannot meet the personalized needs of different majors. For example, the general corpus is not applicable to ESP teaching. Therefore, this study proposes that we should try to build a multimodal corpus containing four modalities: text, image, audio and video, and carry out diversified classroom teaching activities with the help of the rich corpus. In the present time of social and cultural pluralism and diversified cultural presentation, the use of multimodal English teaching mode and multimodal corpus for English education reform research in colleges and universities can help cultivate college students' independent learning ability in multimodal online classroom teaching [27]. Foreign language ability is a consideration of students' comprehensive ability, and the collection of multiple forms of teaching resources is helpful for teachers to create teaching situations, fully stimulate students' perceptual cognition, and input new knowledge from different levels. In addition, the use of multimodal corpora can help improve the ecological environment of foreign language education and achieve a virtuous cycle between teaching and learning.

### 5.3 Improve Teachers' technology literacy to enhance teaching effect

CiteSpace keyword co-occurrence mapping shows that most previous corpus-based foreign language teaching research has focused on the language itself, with little mention of teacher technology literacy research. In fact, teacher technology literacy is an essential element in the implementation of corpus-based foreign language teaching. At present, there are various corpus tools such as Paraconc, Antconc, Wordsmith Tools, Wmatrix, Sketch engine, etc., whose functions cover parallel text alignment, semantic annotation, lexical annotation, generation of word lists and subject terms, text retrieval, collocation, contextual keywords, word clusters, etc. The complexity of using them is daunting to many teachers and seriously hinders the practical activities of corpus application in foreign language teaching. Even in the United Kingdom and the United States, where corpus research started earlier and research results are more mature, the frequency of corpus use in university teaching and scientific research is low. Boulton [28] argues that the main reasons for this are firstly, the complexity of corpus use methods, which can easily lead to teachers' frustration, and secondly, teachers' concern about the technology's threat to their role as teachers. Therefore, it is possible to try to simplify the use of corpus technology by incorporating it into normal teaching activities and teaching materials. Therefore, to widely promote research on corpus foreign language teaching and learning in China, especially in the western region where resources are relatively backward, teachers' technological literacy should be given enough attention. Through regular training on corpus technology, teachers should be helped to master basic corpus production and the use of common corpus platforms.

## Conclusion

In this paper, CiteSpace is used to sort out the research results of domestic corpus foreign language teaching from 2000 to 2023, and outline the changes in research focus and research trends in the field over the past 23 years in terms of the number of publications, keyword co-occurrences, author and institutional cooperation networks, and emergent hotspots, respectively. In general, there are two main lines of development in the field of corpus foreign language teaching in China: macroscopic research until 2011, followed by a shift in focus to the integration of corpus technology in different disciplines as academics gain a deeper understanding of the nature of language learning and the scope of corpus technology applications, with the aim of enabling students to summarize language phenomena through observation, generalization, and induction. Despite the growing maturity of corpus foreign language teaching research, there are still some shortcomings, such as weak inter-institutional and inter-author cooperation networks, uneven

distribution of resources, and lagging information literacy of teachers, etc. It is expected that more attention will be paid to these aspects in future corpus foreign language teaching research.

## Acknowledgement

This research is funded by the Key Educational Reform Project of North Minzu University (2020ZDJY06) and Ningxia Higher Education Scientific Research Project (NYG2022054)

## References

[1] Baker, Mona. “Corpora in Translation Studies: An overview and Some Suggestions for Further Research.” Target 7, 2, 1995, 223-243.

[2] Xu Xiuling, Xu Jiajin. Looking back on the 40 years of corpus application in China’s foreign language teaching. Foreign Language Education in China, 2017, 10(04):62-68+88-89. Kenny, Dorothy. “Corpora in translation studies”, 1998.

[3] Huang Ping, Yan Chaoya. An analysis of current research status of corpus application in English Teaching in China, Shandong Foreign Language Teaching, 2010, 31(05):44-50.

[4] Gui Shichun, Feng Zhiwei, Yang Huizhong. Modern Foreign Languages, 2010, 33(04):419-426.

[5] Wu Juan. Discussion on the importance of corpus linguistics to foreign language teaching and research. Journal of Southwest Minzu University, 2011, 32(S2):254-257.

[6] An Xuehua. The application of corpus in L2 writing. Journal of Xi'an Foreign Languages University, 2013, 21(02):54-58.

[7] Zhao Lianzhen. Corpus-driven foreign vocabulary pedagogy. Journal of Southwest JiaoTong University, 2015, 16(04):44-51.

[8] Li Xiuwen. Corpus-driven foreign language pedagogy in local universities. Education Modernization, 2018, 5(18):97-98.

[9] Zhang Qi, Zou Yanli. On the application of corpus-based data driven learning in foreign language teaching. Journal of Changchun University of Science and Technology, 2022, 35(06):181-185.

[10] Liang Maocheng. Mini-texts and its application in foreign language teaching. Technology Enhanced Foreign Languages, 2009, No.127(03):8-12.

[11] Zhen Fengchao, Wang Hua. Application of learner corpora to foreign language pedagogy: ideas and methods. Foreign Language World, 2010, No.141(06):72-77+90.

[12] Ge Lingling, Li Guangwei, Liu Chaohui. Construction of corpus-based college English teaching platform and studies on its teaching mode. Foreign Language World, 2011, No.146(05):2-8.

[13] Peng Xinxia, Mark Davies. The internet-based English corpus IWeb and its application in teaching and learning English as a second language. Technology Enhanced Foreign Languages, 2020,No.194(04):73-81+12.

[14] Li Yuxiang. Constructing foreign language multimodal corpus of classroom teaching. Journal of University of Shanghai for Science and Technology, 2019, 41(01):1-11.

[15] Qi Taoyun, Yang Chengshu. Designing and building simultaneous interpreting corpus with multi-modalities: with ECTSIC-P as an example. Chinese Translators Journal, 2020, 41(03):126-135+189.

[16] Chai Mingjiong. Exploring solutions to the challenges facing translation training: an introduction to the corpus-based professional translation training platform. East Journal of Translation, 2019, No.58(02):4-7.

[17] Liu Xiaodong, Li Defeng. Using the monolingual English corpus COCA in Chinese-English business translation teaching. Chinese Science and Technology Translator’s Journal, 2020, 33(01):29-32+61.

[18] Zhao Zhenting, Chai Mingjiong. Toward a language service market-oriented and corpus-based translation model in the age of AI and Big Data. Technology Enhanced Foreign Languages,2021, No.201(05):88-95+13.

[19] Liu Bingdong, Cao Lingmei. On constructing a teaching model of corpus-based translation course for undergraduates. Technology Enhanced Foreign Languages, 2021, No.201(05):68-73+10.

[20] Dai Guangrong, Liu Siqi. Progress in corpus-assisted translation pedagogy research: A review of domestic and foreign journal paper from 2007 to 2022. Foreign Language World, 2023, No.214(01):40-48.

[21] Zhong Fuqiang. Construction of intelligent foreign language education system: an approach to teaching reform. Technology Enhanced Foreign Languages, 2021, No.197(01):85-91+14.

[22] Fang Yan, Yin Jie, Wu Bihu. "Climate Change and Tourism: A Scientometric Analysis Using CiteSpace." Journal of Sustainable Tourism, 2018, 26 (1): 108–126.

[23] Cai Jiandong, Ma Jing, Yuan Yuan. The research of the evolution, research fronts and focus of foreign CSCL: analysis based on CiteSpace. Modern Education Technology, 2012, 22(05):10-16.

[24] Tu Tao, Zhang Yuming. A research of "internet+ plus" based on knowledge mapping and co-word analysis. Journal of Southwest University (Natural Science), 2021, 43(01):1-11.

[25] Wang Lizhu, He Yunfeng. Visual analysis of curriculum ideological and political research of China based on CiteSpace. Education Theory and Practice, 2022, 42(24):27-31.

[26] Huang Lihe. Towards multimodal pragmatics: a study of illocutionary force in Chinese situated discourse. Shanghai Foreign Language Education Press, 2019.

[27] Boulton, Alex. Data-driven learning: On paper, in practice. Bern: Peter Lang, 2009.