Sinologia Hispanica, China Studies Review,
17, 2 (2023), pp. 139-158

Received: June 2023
Accepted: November 2023

Research Status and Current Problems of Corpus Linguistics in China

Estado de la investigación y problemas actuales de la lingüística de corpus en China

中国语料库语言学研究：现状与问题

廖悦 liaoy33@mail.sysu.edu.cn Liao Yue*

李坤瑜 lky1997tbc@gmail.com Li Kunyu**

School of International Sutdies
Sun Yat-sen University
Guangzhou, China 510000

Abstract: After more than 40 years of development, China has made significant achievements in corpus-based research, while problems still remain: corpus-based studies of linguistic phenomena are

* Dr. Liao Yue is the lecturer of the Department of Spanish, School of International Studies, Sun Yat-sen University. Her main research interest is linguistics.

0000-0002-8364-2097

** Li Kunyu is graduate student of School of International Studies, Sun Yat-sen University. Her research interest is translation studies.

0009-0000-1056-1061

not thorough enough, practice and research have not yet been bridged, and while the English corpus has made significant progress, the multilingual one lag far behind. It has become urgent to accelerate the construction of corpora in China to keep pace with international corpora. With the development of linguistics research on Chinese corpus and corpus construction in China as the research object, this study has adopted both diachronic and synchronic research methods, combing the history and current studies in corpus linguistics in China systematically. The paper is designed to summarize the current bottlenecks and problems of Chinese corpus construction with academic consensus, and to comprehensively and objectively analyze the problems in Chinese corpus linguistics research and the difficulties in solving these problems. Hopefully, it could draw the attention of academic circles in China and abroad, provide international experience in advanced corpus construction, solve the problems restricting the development of Chinese corpus, and promote corpus linguistics research and corpus construction in China.

Key Words: Corpus linguistics; Chinese corpora construction; research status; discipline development.

Resumen: A través de más de 40 años de desarrollo, China ha logrado avances significativos en investigaciones basadas en corpus, sin embargo, aún persisten algunos problemas: los estudios basados en corpus sobre fenómenos lingüísticos no son suficientemente exhaustivos; la práctica y la investigación no se han enlazado eficientemente; y, mientras que la investigación sobre el uso de corpus en inglés ha avanzado significativamente, el corpus multilingüe está aún notablemente rezagado. Es por eso que es imperativo acelerar la construcción y uso de corpus en China para alcanzar el paso de desarrollo de corpus de otros lenguajes y regiones. Con el desarrollo de la investigación lingüística sobre corpus chino y la construcción de corpus en China como objeto de estudio, este trabajo ha adoptado métodos de investigación diacrónicos y sincrónicos, analizando sistemáticamente la historia y los estudios actuales en lingüística de corpus en China. El artículo hace un especial énfasis distinguir y presentar de manera concreta los obstáculos y problemas actuales en la construcción de corpus chinos con consenso académico, y, al mismo tiempo, analiza de manera completa y objetiva los problemas en la investigación lingüística de corpus chino, así como las dificultades para resolver estos problemas. Por medio de este análisis se espera dirigir la atención de los círculos académicos en China y en el extranjero hacia esta área de investigación, de proporcionar experiencia internacional en la construcción avanzada de corpus, ayudar a resolver los problemas que limitan el desarrollo de los corpus chinos y promover la investigación en lingüística de corpus y la construcción de corpus en China.

Palabras clave: lingüística de corpus; construcción de corpus en China; estado de la investigación; tendencia del desarrollo.

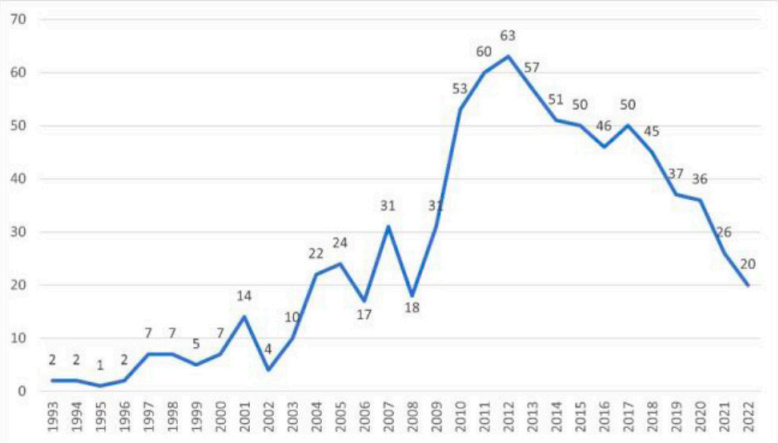
摘要：经过40多年的长足发展，中国在语料库语言学研究方面取得了不少成就，但也同时存在研究不够深入，学术研究成果与语料库建设实践脱节，非英语语种语料库建设及多语种平行语料库建设较为落后等问题。因此，加快中国语料库的建设，与国际语料库接轨，已成当务之急。本研究以中国语料库语言学研究的发展和中国语料库建设为研究对象，采用历时与共时相结合的研究方法，系统梳理了中国语料库语言学的历史发展和客观现状。本文旨在总结当前中国语料库建设的瓶颈和问题，并形成学术共识，全面客观地展现中国语料库语言学研究中存在的问题，同时尝试性提出解决这些问题的一些思路和想法。也希望能通过历史和现状的梳理和总结引起国内外学术界的关注，提供国际先进的语料库建设经验，帮助解决制约中国语料库发展的瓶颈问题，推动中国语料库语言学研究和语料库的发展。

［**关键词**］语料库语言学；中国语料库建设；研究现状；发展趋势

1. The Development of Corpus Linguistics Research in China

Corpus linguistics was introduced into China four decades ago, and its development has been basically synchronized with that of the world. Corpus linguistics research in China can be divided into three stages in general: theory introduction (1980-2002), corpus construction in the computer age (2003-2015), and large-scale corpus construction (2016-). In general, despite its relatively late introduction, corpus linguistics in China has enjoyed rapid development and yielded fruitful academic outcomes. With increasingly refined research areas, corpus linguistics in China has continuously received growing attention when gradually integrated with information technologies.

Figure 1: The trend of corpus linguistics publications in China, 1993-2022



Built in 1963-1964, the Brown corpus (Brown University Standard Corpus of Present-Day American English) is viewed as a pioneering landmark for global corpus construction, containing one million words of written English around 1961. Created by Francis and Kucera, the corpus has become a standard reference for more than 30 computerized English corpora constructed over the following two decades, among which the most influential and distinctive ones include the Lancaster-Oslo/Bergen Corpus (LOB Corpus), Collins Birmingham University International Language Database (COBUILD), the British National Corpus (BNC), the International Corpus of English (ICE), the Association for Computational Linguistics Data Collection Initiative (ACL/DCI), and the Network of European Reference Corpora (NERC). In the 1980s, corpus linguistics stepped out of the shadow

of transformational-generative grammar, and several second-generation corpora (linguistically annotated corpora) were built by dint of computer technology (Johansson, 2008, p. 33). Inspired by the globally burgeoning corpus linguistics, scholars in China started their research.

Professor Yang Huizhong is one of the first Chinese scholars to recognize the prospects of corpus linguistics, who has presided over the construction of the first English corpus in China, the Jiao Da English for Science and Technology (JDEST) Corpus, as early as 1982 (Zhen & Zhang, 2004). JDEST later became an important reference for syllabuses and vocabulary lists of College English, exerting a significant impact on English teaching in Chinese universities. The establishment of JDEST has also witnessed the beginning of corpus linguistics research in China. By 2003, seven English learner corpora, 16 parallel corpora, three special English corpora, and 13 Chinese corpora had been built. However, in the stage of theory introduction, due to the limitation of computer technology, China's corpus construction was confined to only a few universities equipped with outstanding language majors and mainframe computers (Gui, 2014). Meanwhile, the research outcomes were only limited to foreign language teaching, which was closely related to corpus linguistics in China, while the research gap in other fields in corpus linguistics remained unfilled. Also, confined by the technological limitation, corpus construction at this stage faced multiple problems: repeated construction of low-level corpora, lack of software development, deficiency of theoretical research, and the underdeveloped application of corpus in foreign language teaching (Zhen & Zhang, 2004).

After 2003, as global corpus linguistics ushered in the computer age, corpus linguistics in China also entered the primary stage of development. There has been a significant increase in the number of publications in Chinese Social Sciences Citation Index (CSSCI) journals since 2003, which peaked in 2010. The major features of this period can be summarized as follows.

Corpus studies have become a research focus in English studies when English corpus linguistics carved its niche (Wang, 2007). The construction and development of parallel corpus have become a new trend in corpus linguistics. In colleges and universities, various bilingual parallel corpora were under construction, such as the Chinese-English parallel language corpus (PCCE), constructed by the National Research Centre for Foreign Language Education of Beijing Foreign Studies University (BFSU), and the computational bilingual corpus by the Institute of Applied Linguistics (Li & Zhao, 2007).

After 2010, relevant literature on corpus linguistics in China National Knowledge Infrastructure (CNKI) decreased to a certain degree, while the overall quantity of publications still surpassed the theory-introduction stage. With the development of computer technology and the accumulation of academic achievements, the construction and research of Chinese corpus expanded in volume, depth and application. Hence, multilingual and self-built corpora emerged in this stage. Various new research focuses have been found, and corpus linguistics in China has shown interdisciplinary features in the combination with bibliometrics, sociolinguistics, and Chinese linguistics (Liu, Xu, & Liu, 2014). Foreign language teaching, a research direction closely related to corpus construction, has displayed various development trends. The development of corpus studies has drawn researchers' attention to the quantitative and qualitative analysis of language teaching (Liu, 2020). After a long term of corpus construction and research, distinguished scholars and research teams stood out and exchanged their outcomes at academic conferences during this period.

Since 2016, with the advent of the age of big data and the development of information technology, new demand for corpus linguistics has been put forward, and the discipline has entered another new stage in China in the information age when the use of big data led to new changes. China's corpus construction is about to be equipped with larger scales, diversified forms, refined classification of fields, deeper processing capability, and convenient operation. In this stage, corpus plays its part in education while also serving as a bridge of informatization in the age of big data. Moreover, it has made significant contribution to the critical discourse analysis of China's national image based on reports and news of foreign media (Chen, 2023). At the same time, the progress of application technologies and artificial intelligence (AI) based on language data is raising new topics and challenges for language research. The increasing volume of the corpus has become challenging for corpus linguistics in the selection and processing of corpus materials (Liang, 2021).

2. Milestones in Chinese Corpus Research

Up to now, Chinese and English corpora in China have been quite well-developed, and several high-quality corpora have emerged, with which various types of research have been carried out.

One of the representatives of Chinese corpora is the Center for Chinese Linguistics of Peking University (CCL corpus). Established in 2004, the corpus has undergone multiple expansions and upgrades. Till 2019, it has accumulated nearly 1.2 billion bytes of modern Chinese language material,

approximately 400 million bytes of ancient Chinese language material and about 71.6 million bytes of Chinese-English sentence alignment, and also developed various retrieval functions. The modern Chinese language material includes 12 categories, such as literature, drama, newspapers and periodicals, translation, TV series and movies, etc. (Zhan, Guo, Chang, Chen, & Chen, 2019)

Since its establishment, CCL Corpus mainly serves the study of Chinese linguistics. With this corpus, Jiao (2022) studied the differences among six Chinese hypothetical conjunctions (如果，要是，若，一旦，假如，万一) in grammar and pragmatics with various research methods (Jiao, 2022); Zheng (2020) demonstrates the subjective differences between "顶多" and "至多" in three aspects: syntactic forms, semantic features and pragmatic functions (Zheng, 2021); Tian leads qualitative and quantitative researches on the three most commonly used quantifiers (个，条，位) in modern Chinese (Tian, 2013). Besides, some comparative studies between Chinese and other languages have been conducted based on the combination of CCL corpus and corpora in other languages. Jin (2021) compares the negative functions of interrogative words in Chinese and Korean with CCL corpus, BCC corpus and Korean corpus (Jin, 2021); Under the guidance of the two-level word-class categorization theory, Wang (2020) investigates the usage patterns of "hypocrite" and "虚伪" respectively with Sketch Engine and CCL Corpus (Wang, 2020); Zha and Wang (2019) investigate and compare the grammaticalization processes of "超–" in Chinese and Japanese with CHJ corpus, BCCWJ corpus, BCC corpus and CCL corpus (Zha & Wang, 2019). In addition to research in linguistics, CCL also contributes to the study of applied linguistics. For example, Cheng and his colleagues (2019) revise their translation using Brigham Young University Corpus (BYU Corpus), CCL Corpus, BCC Corpus and CNKI Translation Assistant, thus compiling the Sanskrit-Latin-English-Chinese Quadrilingual Dictionary (Cheng, Wang, Wang, & Wang, 2019); Wang and Ji (2018) study book-jacket design with data analysis of the meaning of "情" in CCL corpus, and develop a design method based on abstract emotion (Wang & Ji, 2018).

The Chinese Learner English Corpus (CLEC) is one of the representatives of English corpora in China. Established by Yang Huizhong, who is the first-generation corpus linguist in China, it is supported by Gui Shichun, Li Wenzhong, Pu Jianzhong and other experts on platforms of Guangdong University of Foreign Studies and Shanghai Jiaotong University.

As a corpus for language learners, CLEC provides support for research on error correction in English teaching, one of which is to combine CLEC corpus with native English corpus to explore and correct vocabulary usage.

For example, Zang (2021), based on BAWE Corpus and CLEC Corpus, analyzes the similarities and differences between "improve" and "enhance" in English, and the similarities and differences in the use of these synonyms between Chinese learners and native speakers from the aspects of word frequency, colligation, collocation and semantic prosody (Zang, 2021); Yu and Zhao (2021) conduct a univariate analysis of the morphological syntax that affects the positional variation of particles in phrasal verbs based on CLEC Corpus and LOCNESS Corpus (Yu & Zhao, 2021); Zhou (2021), based on Cook's case grammar, compares the experiential usage of the verbs "concern" and "endure" in COCA corpus and CLEC corpus (Zhou, 2021). In addition to serving English teaching from the perspective of semantic prosody, the research based on this corpus also verifies the new model of information-based teaching. Wang and Mao (2019) look into the retrieved results of synonyms "chance" and "opportunity in CLEC corpus, and find that Data-driven Learning (DDL) can promote autonomous learning and discovery learning of in-depth vocabulary knowledge (Wang & Mao, 2019). In addition, research based on CLEC corpus also contributes to corpus construction. Huang and Gao (2019) analyze the grading standards on the language material of writing practices in CLEC corpus, suggesting that the existing grading standards based on learners' age and major are incapable of grading language ability precisely, and the construction of learner corpus should adopt more objective and comprehensive standards (Huang & Gao, 2019). Zhang (2016) studies English learners' acquisition of the Chinese negative structures at Tarim University, where CLEC corpus is used as a reference for Tarim Learner English Corpus (TLEC) (Zhang, 2015)

In addition to the Chinese and Chinese-English corpus, the corpus of other languages has also achieved success. In 2005, National Cheng Kung University (NCKU) began to build CEATE, a Spanish corpus for learners from China, which laid a foundation for an oral corpus called COATE in 2013. The two corpora eventually merged into CATE for Spanish learners in the Taiwan region (He & Liu, 2018). By 2006, three bilingual parallel corpora between Japanese and Chinese, French and Chinese, and Mongolian and Chinese had been developed respectively, when some scholars had begun the construction of a Russian corpus (Wang, 2007).

Apart from the large and systematic corpora, many small corpora have also facilitated academic research. These corpora reflect the slowdown in large-scale corpora construction in the late 1990s and the results of a rich supply of online texts and corpus materials (Zhen & Zhang, 2004). They were built by both research institutes in universities and scholars who apply self-built corpora for research, leading to fruitful outcomes in corpus-

based teaching about the self-built corpus. For example, based on a self-built parallel corpus, Bai and Yu built a process model and its corresponding computer-aided translation (CAT) system for Chinese scholarly monographs (Song & Wang, 2013). In addition to corpus construction, corpus software for special use also has yielded new outcomes, including PatCount developed by Liang and Xiong in 2008 (2008), and Colligator by Xu and Xiong in 2009 (Gui et al., 2010; 2009).

3.Leading Institutes and Scholars in Corpus Linguistics in China

At the beginning of the 1980s, the goal of corpus construction in China was mainly the statistical study of Chinese vocabulary. Since the 1990s, the corpus method has been widely used in the field of natural language processing, and the construction of corpus has also developed rapidly. The following basic information on main corpora in China is provided through investigation of corpus development organizations and relevant references.

3.1 Main Research and Development Teams of Chinese Corpus

At present, Chinese corpora in China are mainly constructed by the National Language Commission of China, Peking University, Tsinghua University, Beijing Language and Culture University (BLCU), Beijing Normal University (BNU), Central China Normal University (CCNU), Beihang University, Chinese Academy of Social Sciences (CASS), and Nanjing Agricultural University (NUA). Corpora of significance include the Chinese Corpora by National Language Commission, the CCL Corpus of Peking University, the Large-scale Contemporary Chinese Corpus (the annotated corpus of the People's Daily), BLCU Corpus Center (BCC) Corpus of Beijing Language and Culture University, Global Chinese Interlanguage Texts Corpus, HSK Dynamic Composition Corpus (no official English name found yet), Beijing Mandarin Spoken Corpus (BJKY), THCHS-30 Corpus and THU Open Chinese Lexicon of Tsinghua University, New Era People's Daily Segmented Corpus of Nanjing Agricultural University, Contemporary Chinese corpus of Baihang University, Chinese Interlanguage corpus of Central China Normal University, Mandarin Multimedia Child Speech Corpus of Chinese Academy of Social Sciences, and so on.

However, not all of these corpora are accessible to the public because of ongoing construction, upgrading, and accesses only for insiders. Seven of the above corpora currently open to the public with their information and access URLs are listed in the table below.

Table 1: Rsearch and development teams of Chinese corpus

Corpus	Website	Information
CCL Corpus (北京大学现代汉语(CCL)语料库)	http://ccl.pku.edu.cn:8080/ccl_corpus/index.jsp	The corpus was initially established in the end of 2004 and was expanded and updated three times in 2006, 2009, and 2014 respectively. It has nearly 1.2 billion bytes of modern Chinese language material, nearly 400 million bytes of ancient Chinese language material, and about 71.6 million bytes of Chinese-English sentence pairs, and has developed various retrieving functions.
BLCU Corpus Center (BCC) Corpus (北京语言大学语料库中心)	http://bcc.blcu.edu.cn/	The BCC corpus is a multilingual corpus with ancient and modern Chinese language materials as well as synchronic and diachronic discourse. It also provides cloud services.
THCHS-30 Corpus (清华大学中文语音识别数据(THCHS-30)语料库)	http://www.openslr.org/18	Completed in June 1994 and appraised by the State Education Commission, the corpus is a large general corpus with 770,000 words of language material.
THU Open Chinese Lexicon (THUOCL) (清华大学开放中文词库)	http://thuocl.thunlp.org/	THUOCL (THU Open Chinese Lexicon) is a high-quality Chinese thesaurus compiled and launched by the Natural Language Processing and Social Human Computing Laboratory of Tsinghua University, whose vocabulary list comes from mainstream websites' social tags, searching buzzwords, input-method thesaurus, etc. The thesaurus includes categories such as IT, finance, idioms, placenames, historical figures, poetry, medicine, food, law, cars, animals, and will continuously update the existing glossary with more categories. It provides free assess for universities, research institutes, enterprises, institutions and individuals at home and abroad.

New Era People's Daily Segmented Corpus (NEPD) (南京农业大学人民日报标注语料库)	http://corpus.njau.edu.cn/	New Era People's Daily Segmented Corpus of Nanjing Agricultural University has now exceeded 23 million words, all of which are manually segmented and labeled. It is free and assessable to the academic community, and is currently the world's largest general corpus of Chinese finishing.
Global Chinese Interlanguage Texts Corpus (北京语言大学全球汉语中介语语料库)	http://qqk.blcu.edu.cn/#/login	Currently, the corpus covers 3 million words of written language, 100 hours of spoken language, and 20 hours of video. When fully completed, its scale is expected to reach 50 million words, including 45 million words in written language; 450 hours of spoken language (150 words per minute, approximately 4 million words); and 110 hours of multimodal language material (150 words per minute, approximately 1 million words in total).
HSK Dynamic Composition Corpus (北京语言大学高等汉语水平考试(HSK)动态作文语料库)	http://hsk.blcu.edu.cn/Login	The construction began in July 2003 and completed in December 2006, put into trial operation on the Internet since then. After modifications, it is now officially accessible to the public. The original material is compositions by foreign students taking the Higher Chinese Language Test from 1992 to 2005. The first version of this dynamic corpus contains 11,569 articles in total (4.24 million words).

3.2 Main research and development teams of English corpus

Apart from Chinese corpora, researchers and scholars have also developed English corpora with the support of leading universities, among which Shanghai Jiao Tong University (SJTU) has made prominent contribution with its Institute of Corpora and Intercultural Studies from School of Foreign Languages. In the 1980s, SJTU has established the first self-developed corpus in China, the JDEST corpus (Jiao Da English for Science and Technology Corpus), which is expected to contain one million words with texts from industries such as computer, manufacturing, electrochemistry, and mathematical aviation, significantly improving the teaching efficiency of English for Science and Technology (Yang & Huang,

1982). As for spoken linguistic data, transcribed audio-visual recordings of the National College Test-Spoken English Test (the CET-SET) was collected by the institute for the construction of College Learners' Spoken English Corpus in China (COLSEC), providing researchers with recordings with improvisational conversation, which precisely reflect language proficiency of learners (Wei, Li & Pu, 2007). In 2006, Hu Kaibao and his team in SJTU built Chinese-English Conference Interpreting Corpus (CECIC) with three sub-databases: the Chinese-English parallel corpus for press conferences, the English original corpus for press conferences, and the Chinese-English parallel corpus for government work reports. CECIC makes it possible for researchers to conduct studies on Chinese-English conference interpreting and on its idiomatic and grammatic features. Recently, SJTU has opened its self-developed corpus retrieval and application platform to the public, in which researchers can register and then have access to its English-Chinese Parallel Corpus of Shakespeare's Plays and Chinese-English Conference Interpreting Corpus (CECIC). In addition to SJTU, Beijing Foreign Studies University (BFSU) built Super-large-scale China English-Chinese Parallel Corpus (CECPC) in a stand-alone version with bilingual search engine. It is applicable for research in natural language processing and language engineering, research of English-Chinese synchronic and diachronic comparison, and joint research of translation, translation teaching, and bilingual lexicography (Wang, 2012).

However, parts of the corpora mentioned above provide accesses only to school personnel. The table below has listed three of the corpora mentioned above and other corpora accessible with significance.

Table 2: Research and development teams of English corpus

Corpus	Website	Information
Chinese-English Conference Interpreting Corpus (CECIC) (上海交通大学汉英会议口译语料库)	https://instcorpus.com/ (please register before using the corpus)	CECIC comprises three sub-databases: the Chinese-English parallel corpus for press conferences, the English original corpus for press conferences, and the Chinese-English parallel corpus for government work reports. The existing storage capacity is 544,211 words. The corpus contains the original Chinese and English translations of press conferences held by the Chinese central government, relevant ministries and commissions of the State Council from 1988 to 2008.
English-Chinese Parallel Corpus of Shakespeare's Plays (莎士比亚戏剧英汉平行语料库)	https://instcorpus.com/ (please register before using the corpus)	With a capacity of approximately six million words, the corpus includes English source text of Shakespeare's works and three versions of translation. The corpus supports one-to-one and one-to-many alignment.
Super-large-scale China English-Chinese Parallel Corpus (CECPC) (大规模英汉平行语料库)	http://114.251.154.212/cqp/ (Both user ID and password are "test" for freely available corpora.)	The corpus includes a Japanese-Chinese translation corpus of 20 million characters and a general Chinese-English parallel corpus of 30 million words. At present, the 30-million-word vocabulary database has basically accomplished sentence-level matching, and the 20-million-word vocabulary database has been finalized, annotated, and bilingually connected.
THUMT: An Open-Source Toolkit for Neural Machine Translation (清华大学中英平行语料库)	http://thumt.thunlp.org/	Developed by the Natural Language Processing Group at Tsinghua University, it contains 2.85 million Chinese-English parallel sentence pairs.

The PKU Chinese-English Parallel Corpus (北京大学汉英语料库)	http://ccl.pku.edu.cn:8080/ccl_corpus/index_bi.jsp	The corpus is a large-scale Chinese-English and Chinese-Japanese bilingual corpus, including 200,000 Chinese-English aligned sentence pairs, 20,000 Chinese-Japanese aligned sentence pairs, and 10,000 pairs of Chinese-English aligned vocabularies. It is designed for the research and development of machine translation and other systems, providing basic resources and standardized evaluation.
Asian Corpus of English (ACE)	https://corpus.eduhk.hk/ace/	The corpus contains naturally-occurring spoken materials of interviews, press conferences, service encounters and seminar discussions for 1 million words in total. ACE is tagged manually using a set of conventions developed by the VOICE team.
English-Chinese Parallel Concordancer	https://corpus.eduhk.hk/paraconc/search	The English-Chinese Parallel Concordancer is a corpus project by the Education University of Hong Kong with a size of 4.5 million Chinese characters and 2.9 million English words. Users can search the English-Chinese parallel corpus online, and choose to do parallel concordancing or monolingual concordancing according to their needs.
Chinese Learner English Corpus (CLEC) (cooperative construction) (中国学习者英语语料库 [合作建设])	http://114.251.154.212/cqp/ (Both user ID and password are “test” for freely available corpora.)	The one million words in the corpus were retrieved from writings by high school students, answer sheets by non-English major students at the CET Band 4 and Band 6, and written homework of English major students of all grades.

Multilingual corpus is also important for corpus construction. Chinese academics have actively drawn on the advanced experience of the United Nations, the European Union, and developed countries such as the United States and Japan, introducing multilingual text-processing technologies to build multilingual parallel foreign language corpora. *Xi Jinping: The Governance of China* Multilingual Database Comprehensive Platform by

Corpus Research Institute of Shanghai International Studies University (SISU) is one of the accessible multilingual parallel corpora, which builds 28 sentence-level-aligned corpora in Chinese, English, Korean, German, Russian, French, Mongolian, Japanese, Thai, and Turkish with the source text from *Xi Jinping: The Governance of China* as the center language. The released Platform is currently in the on-campus trial stage, and the URL is http://202.121.96.180.

4. Problems in Corpus Linguistics in China

With the challenges in corpus construction and development, based on former research, it is suggested that we should promote corpus construction in China through cross-disciplinary communication, talent cultivation, multilingual technological cooperation and resource sharing.

(1) The cohesion between teaching practice and scientific research is far from satisfying. The primary problem for corpus linguistics in China is that most people perceive corpus as a research tool rather than a discipline. Scholars tend to apply corpus linguistics in research rather than focus on in-depth corpus study because of the easy access to corpus data. Therefore, scientific research cannot keep up with the demands of teaching and research practices due to the lack of investment and talents in corpus construction, technology that falls short of the international level, and shortage of interdisciplinary talents with both liberal arts and science knowledge.

Chinese corpus construction suffers from the confined size of existing corpora and the underdeveloped processing capability of corpus software. Also, Chinese scholars' research in corpus linguistics is relatively narrow and unbalanced, where descriptive analysis is attached with more attention than the in-depth study of linguistic phenomena (Song & Wang, 2013). At the same time, different types of corpora are imbalanced in proportion. Wang believes that there is a lack of communication in corpus construction, deep processing in corpus annotation and tagging, and exploration of research potential in corpus application (Feng, Wang, Wei, Pu, & Liang, 2012). When sharing his experience in corpus linguistics research, Wei mentions that, due to the limitation of knowledge, Chinese corpora in the early days are of small scale in volume, insufficiency in diversity, and inaccuracy in tagging (Feng et al., 2012). Xiao explains the problem of insufficient academic exchange, inadequate empirical studies in publication, and low academic profile. He also points out the deficiency of deep and systematic corpus analysis (2015). The problem of inefficient corpus construction has been noted by Yang in 2003 and remained unsolved till 2015 (Xiao, 2015).

(2) The lack of interdisciplinary communication and cooperation holds back corpus development. Corpus construction involves linguistics, computer science and other disciplines, where cooperation could greatly contribute to corpus quality: linguistic experts could introduce new theories into corpus tagging when computer experts improve the efficiency of corpus processing and point out the latest demand for industrial application. However, at present, due to the different research objectives, corpus experts in the two fields rarely conduct cooperative research (Xiao, 2015), hence corpus construction could not learn from the advantages of the two. On the contrary, it has caused the disconnection between linguistics and relevant industries (especially in the field of natural language processing), which constrains the development of corpus construction.

(3) The lack of interdisciplinary talents hinders the development of corpus linguistics. Insufficient talent cultivation in corpus construction could not be neglected. In addition to enhancing disciplinary cooperation, corpus construction is also in need of interdisciplinary talents who have received systematic education in linguistics, corpus and natural language processing. They are familiar with relevant knowledge and skills, and able to guide the construction of corpus, hence better communicating with linguistics and computer science experts. At present, with the lack of interdisciplinary talents, linguistic researchers are often short of the knowledge and skills of natural language processing, and computer researchers are not familiar with linguistic theory, which is one of the bottlenecks for corpus construction.

(4) With multiple achievements in English corpus, the development of non-English corpus lags behind, and most of the corpora do not have any open access. With the emphasis on corpus construction at the national level in China, projects of corpus construction have been increasing, while problems have emerged, such as repeated investment, lack of sharing and low utilization. Xiao (2015) points out that most government-funded corpora are only for internal use after completion, which leads to low utilization and repeated investment, especially the repeated construction of low-level corpus (Xiao, 2015). Therefore, scholars have begun to advocate corpus sharing. Tan (2014) suggests that maximizing resource sharing and the beneficiary groups of corpora is worth studying. Liu and Yang (2012) propose the idea of a library-led corpus resource-sharing platform. Xiao (2015) suggests that the relevant departments should issue regulations that the national, provincial and ministerial vertical projects must open the corpus to the public within a certain period after the completion of projects. Zhang and Cui (2015) put forward that data sharing should be regarded as the premise of public funding and the evaluation index of projects, and should

be included explicitly in the provisions of the national scientific research management department. Further research and implementation of these measures will significantly improve the utilization of corpus resources, and at the same time help to ensure the research quality of corpus projects.

5. Conclusions

After more than 40 years of development, corpus linguistics has achieved fruitful outcomes in various fields of language studies in China and has been proved to be an effective research method that has brought profound changes in language research. However, compared with advanced countries in corpus construction, we still fall far short. It is mainly manifested in low construction quality, unbalanced language development, lack of sharing, cooperation among related disciplines, interdisciplinary talents and shared corpus resources, and so on.

The government and the academic circles have gradually realized the above problems. Government agencies such as The National Social Science Fund of China have further strengthened corpus construction through national projects. It mainly supports corpus with various research and application demands, complicated construction process and limited number, and focuses on corpus construction with significance to discipline development. Also, they have supported outstanding research teams in corpus research nationwide, funded young scholars, and encouraged corpus study in linguistics courses for undergraduates and postgraduates. There are plans to cultivate young research talents at different levels, establish a corpus research echelon, and accumulate reserve forces for corpus construction in China. At the same time, it is necessary to strengthen the construction of computational linguistics major within the language discipline, and encourage joint research with computer discipline and interdisciplinary talents cultivation.

In the future, corpus capacity will continue to expand, and small corpus with professional discourse will become increasingly specialized and diversified, involving multilingual and multimodal data. With the instrumental nature of corpus, corpus linguistics will surely integrate with other disciplines in development. Its openness and interdisciplinarity make it a discipline brimming with vitality beyond linguistics. With technologies related to corpus linguistics growing in China and self-developed multilingual corpora opening to the public, corpus linguistics in China will receive more attention from the global linguistic community and play an important part in global corpus linguistic research.

REFERENCES

Chen, Z. & Cai, H. 2023. The National Image of China in Cuban Reports on COVID-19 Pandemic: A Study Based on the Linguistic Corpus and Critical Discourse Analysis. *Sinologia Hispanica*, 1: 103-124.

Cheng, S., Wang, S., Wang, Z, & Wang, H. 2019. On the translation of Ayurveda vocabulary through quantitative analysis and qualitative analysis within corpus in compiling Sanskrit-Latin-English-Chinese Dictionary of Ayurveda 基于语料库的定性定量分析在阿育吠陀词汇梵-拉-英-汉四语词典编纂中的翻译应用. *Journal of Panzhihua University*, 4: 66-74.

Feng, Z., Wang, K., Wei, N., Pu, J. & Liang, M. 2012. Excerpt of speeches at the expert forum of the first Corpus Linguistics in China Conference 语料库语言学在中国专家论坛发言摘登. *Foreign Language Teaching and Research*, 3: 371-375.

He, X. & Liu, Y. 2018. Building a corpus for Spanish learners in China: Plans and prospects 中国西班牙语学习者语料库(cace):规划与展望. *Corpus linguistics*, 2: 98-108.

Hu, K. & Tao, Q. 2010. The Compilation and Application of Chinese-English Conference Interpreting Corpus 汉英会议口译语料库的创建与应用研究. *Chinese Translation Journal*, 5: 49-56+95.

Johansson, S. 2008. Some aspects of the development of corpus linguistics in the 1970s and 1980s. In Corpus linguistics. *An international handbook*. Berlin: Walter de Gruyter, 33-53.

Li, J. & Zhao, X. 2007. A summary of the Symposium on Applied Linguistic Studies in Honor of Professor Yang Huizhong 庆贺杨惠中先生执教50周年暨应用语言学研讨会综述. *Foreign Language World*, 3: 75-79.

Liang, M. 2021. Corpus linguistic research in an era of big data[大数据时代的语料库语言学研究探索. Foreign Languages in China, 1: 13-14.

Liang, M. & Xiong, W. 2008. Applications of PatCount in foreign language teaching and research 文本分析工具patcount在外语教学与研究中的应用. *Technology Enhanced Foreign Language Education*, 5: 71-76.

Liu, R. & Yang, Z. 2012. An idea of the resource sharing platform of corpus 语料库资源共享平台建设构想. *Journal of Academic Library and Information Science*, 2: 4.

Liu, S. 2020. Corpus-based error analysis of written works in Chinese by Spanish students. *Sinologia Hispanica,* 2:101-128.

Liu, X., Xu, J. & Liu, L. 2014. An overview of corpus linguistics research in China (1998-2013): A CiteSpace based analysis 基于citespace的国内语料库语言学研究概述 (1998-2013). *Corpus linguistics*, 1: 69-77+112.

Song, H. & Wang, X. 2013. A review of corpus linguistics studies in China in recent ten years 近十年国内语料库语言学研究综述. *Shandong Foreign Language Teaching Journal*, 3: 41-47.

Tian, X. 2013. Case study on individual qualifiers "个", "条" and "位" "由" "个" "条" "位" 谈个体量词的泛化. *Journal of Ningxia University (Social Science Edition)*, 4: 38-42.

Wang, J. & Mao, S. 2019. English synonyms teaching and learning from the perspective of data-driven learning——take "chance" and "opportunity" as examples 数据驱动学习视角下的英语近义词词汇深度教学研究——以 chance, opportunity 词汇深度自主学习为例. *College English Teaching and Research*, 2: 73-80.

Wang, K. 2012. On the Design and Construction of the Super-large-scale China English-Chinese Parallel Corpus (CECPC) 中国英汉平行语料库的设计与研制 *Foreign Languages in China*, 6:23-27.

Wang, K. & Liu, D. 2017. Concordance and Application of the Super-Sized English-Chinese Parallel Corpus: A Big Data Perspective 大规模英汉平行语料库的检索与应用:大数据视角. *Computer-Assisted Foreign Language Education in China*, 6: 3-11.

Wang, L. & Ji, Y. 2018. On the clothing accessories design through corpus 基于语料库的服装配饰设计研究. *West Leather*, 21: 76-77.

Wang, Q. 2020. A contrastive study of the word class labeling between the Chinese word "Xuwei" and the French word "Hypocrite" 汉法"虚伪/hypocrite"词类标注对比研究. *Journal of Chongqing College of Electronic Engineering*, 1: 92-96.

Wang, Z. 2007. Russian corpus linguistics in China: Status and prospect 俄语语料库语言学研究现状与瞻望. *Russian in China*, 2: 44-47.

Wei, N. & Li, W.& Pu, J. 2007. Design principles and annotation methods of the COLSEC corpus COLSEC 语料库的设计原则与标注方法. *Contemporary Linguistics*, 3: 235-246+286.

Xiao, Z. 2015. Some reflections on Corpus Linguistics upon request 肖忠华语料库语言学答客问. *Corpus linguistics*, 2: 1-14+115.

Xu, J. & Xiong, W. 2009. Learner corpus based colligation research: Concepts, Methods and Sample Analyses 基于学习者英语语料的类联接研究：概念、方法及例析. *Technology Enhanced Foreign Language Education*, 3: 18-23.

Yang, H. & Huang, R. 1982. JDEST Corpus of English for Science and Technology JDEST科技英语计算机语料库. *Foreign Language Teaching and Research*, 4:60-62.

Yu, J. & Zhao, X. 2021. A corpus-based study on the particle displacement of Chinese English learners 基于语料库的中国英语学习者小品词位置变异研究. *Journal of Lvliang University*, 6: 21-26.

Zang, Y. 2021. A comparative study of synonyms in Chinese learners based on corpus: A case study of “improve” and “enhance” 基于语料库的中国学习者近义词对比研究——以improve和enhance为例. *Overseas English*, 12: 109-111.

Zha, Y. & Wang, X. 2019. A contrastive study on the grammaticalization of “Super ～” in Chinese and Japanese based on corpus 基于语料库的汉日语”超～”的语法化对比研究. *Journal of Hangzhou Normal University (Humanities and Social Sciences)*, 6: 101-109.

Zhan, W., Guo, R., Chang, B., Shen, Y. & Chen, L. 2019. The building of the CCL corpus: Its design and implementation 北京大学ccl语料库的研制. *Corpus linguistics*, 1: 71-86+116.

Zhang, B. & Cui, X. 2015. On the Standards of Building a Chinese Inter-Language Corpus谈汉语中介语语料库的建设标准. Applied Linguistics, 2: 10.

Zhen, F. 2004. The minutes of the 2003 International Conference on Corpus Linguistics 2003 语料库语言学国际会议纪要. *Foreign Language World*, 3: 80.

Zhen, F. & Zhang, X. 2004. Trends in the development of corpus linguistics – A summary of the “2003 International Conference on Corpus Linguistics 语料库语言学发展趋势瞻望——2003语料库语言学国际会议综述. *Foreign Language World*, 4: 74-77.

Zheng, H. 2020. Subjectivity differences and manifestations between supremum adverbs “dingduo” and “zhiduo”上确界副词”顶多” ”至多”的主观性差异及其表现. *Journal of Yanbian University (Social Sciences)*, 1: 63-70+141.

Zhou, Y. 2021. Corpus-based contrastive analysis of the experiential usage of verbs——with the verb “concern” and “endure” as examples 基于语料库的动词体验用法的对比分析——以动词concern和动词endure为例. *Journal of Wuyi University*, 10: 56-62.