The Design and Construction of the Corpus of China English

Lixin Xia
Laboratory of Language Engineering and Computing of Guangdong University of Foreign Studies
cdhuiyi@aliyun.com

Yun Xia*
Nanfang College of Sun Yat-Sen University/ Guangdong University of Foreign Studies
344192304@qq.com

ABSTRACT
The paper describes the design and construction of the Corpus of China English (CCE). With the emergence of China English as a developing variety in the family of the world Englishes, more and more research has been done to explore its use in China. In order to provide a reliable resource for researchers in the field of China English, CCE was built with due consideration given to its representativeness and authenticity. The general principles for the corpus were authentic, representative and manageable in size. It was composed of more than 13,962,102 tokens in 15,333 texts evenly divided between the following four genres: newspapers, magazines, fiction and academic writings. The texts cover a wide range of domains, such as news, financial, politics, environment, social, culture, technology, sports, education, philosophy, literary, etc. It will be a helpful resource for research on China English, computational linguistics, natural language processing and corpus linguistics. Moreover, it is also a valuable resource for English language teaching in the context of China English as a lingua franca (ELF) in which the theories and practice of English language teaching shift from teaching the "Standard" English to English as a lingua franca. With the corpus evidence, English teaching textbooks and materials can be developed with due consideration given to China English.

CCS CONCEPTS
• Applied computing; • Theory of computation; • Information systems;

KEYWORDS
Corpus, China English, Computational linguistics, Corpus linguistics, Natural Language Processing, English language teaching

1 INTRODUCTION
China claims to have the largest English learning population in the world. Some 30.721% of the national population, totaling 390.16 million, reported to have learned English at junior secondary schools or above in a national survey conducted in 2006 [1], which outnumber the population of the United States of America. The English they use develops distinct characteristics at phonological, lexical, semantic, syntactic and pragmatic levels, which is now known as China English. The emergence of China English as a variety of the English language in the family of World Englishes draws close attention of linguists who carry out research into its linguistic features. In the past, this was done primarily on the basis of a small amount of real or introspective data. With the rapid development of information technology and the Internet, the research on the linguistic features of native varieties of English has changed from the intuition-based one to the corpus-based one since the 1990’s. Then more corpora of non-native English varieties were built, among which the International Corpus of English (ICE) is the most influential.

However, corpus-based studies of China English have been scarce, and fall behind other non-native English varieties. One of the major reasons is the lack of a corpus of China English. Although some attempts have been made to build such a corpus, they are not freely-available for public use. In order to support the studies on the linguistic characteristics of China English, we are building a Corpus of China English (CCE), which is to be introduced in this paper.

2 A BRIEF REVIEW OF EXISTING CORPORA OF CHINA ENGLISH
Several corpora of China English have been built or are being built since the 2000s in China with different purposes and different sizes. In this section, they are to be introduced.

2.1 China English Corpus (CEC)
Li Wenzhong started to build the China English Corpus in 2001 [2]. It is a written corpus of China English, consisting of data from the following four genres: fiction, newspapers, magazines and academic texts. The text covers a wide range of domains: natural science, applied science, social science, world affairs, economics, arts, beliefs, leisure, literature [3, 4]. The corpus has a total of 17, 534 texts and 37, 470, 040 tokens.

The texts of the corpus came from the following sources: newspapers, magazines, academic journals, graduation theses, literary works, government documents, Internet and others. They cover a wide range of fields: natural science, applied science, social science,
Therefore, texts written by both Chinese speakers of English and this is a collection of written text which is about or in connection with China or Chinese culture by native or non-native speakers of English outside China. For the second type of text, the text is mainly from English media in China, English literary works by Chinese or overseas learners of English is excluded as their English is regarded as an interlanguage instead of a legitimate variety of English. The three corpora described above are all written corpora. CEC is said to be representative of China English and can be used to study the linguistic features of China English. However, the text is not evenly distributed in the four domains. For example, the number of tokens from social science is more than 16 million, which is twice as much as that from literature. CCCE is in fact an English media corpus, which is representative of English used in Chinese media. In terms of its size, CETC is the largest and most representative of them all. However, it is mainly built for the purpose of lexicography. Therefore, more corpora of China English are needed to facilitate research into this English variety.

### 3.1 General Principles

The primary objective of this corpus is to build a computerized resource for scholars and researchers to study the linguistic features of China English and those who are interested in China English. To achieve the objective, the corpus is designed and constructed in line with the following general principles: authentic, representative, and manageable.

As for the authenticity of CCE, only the text written by proficient Chinese speakers of English is collected in order to guarantee that the data represents the real use of the English language in Chinese context as a variety. Therefore, the text written by Chinese learners of English is excluded as their English is regarded as an interlanguage instead of a legitimate variety of English.

To make the corpus as representative as possible, the data is collected from a wide range of sources, covering the major genres as other corpora of native English except the spoken data. For practical reasons, it is difficult to collect spoken data of China English. However, written data represents the actual use of English in China by proficient Chinese speakers of English, and it can meet the needs of linguistic investigation of China English.

Size matters a lot to a corpus. Nowadays it is not uncommon to see a corpus with more than 100 million tokens. The most influential corpora of native English are British National Corpus (BNC) and Corpus of Contemporary American English (COCA), each of which contains 100 million or 560 million words of text respectively. However, the corpora of other English varieties are generally much smaller in size. ICE contains over 20 sub-corpora, each of which has about 1 million tokens. Therefore, we plan to collect about 13 million words of text, 13 times as large as the sub-corporus of ICE. It

---

**Table 1: Data sources of CEC**

| Category       | Number of texts | Number of tokens | Number of types |
|----------------|-----------------|------------------|-----------------|
| Natural science| 1324            | 3174622          | 110980          |
| Applied science| 546             | 953951           | 45425           |
| Social science | 4494            | 16079207         | 248045          |
| International affairs | 2358 | 2535967         | 57448           |
| Economy       | 1553            | 2003378          | 61934           |
| Art           | 1944            | 1346122          | 55524           |
| religion      | 998             | 562475           | 27772           |
| Entertainment | 2599            | 1444268          | 62133           |
| Literature    | 283             | 8653464          | 117561          |

The composition of the corpus is shown in Table 1.

### 2.2 The Comparable Corpus of China English (CCCE)

It is a project led by Alex Fang in Hong Kong, which "aims at a grammatical description of China English as an emergent variety"[5]. CCCE is comprised of two components: a Corpus of Chinese Media English (CCME) and a Corpus of British Media English (CBME). CCME collects 2,226 texts from newspapers, magazines and the Internet covering the following five domains: news report, business, social life, culture and arts, and editorials. CBME collects 1,770 texts mainly from the British National Corpus and partially from other electronic sources as the BNC "does not cover the genre of the Internet and magazine editorial"[6]. Each sub-corpus has 1,125,267 and 1,133,643 tokens respectively [6].

At present, two modules were competed: Chinese Medial English and British Media English. Another module is being built: a corpus of abstracts of academic papers. The corpus enables researchers to make comparative and contrastive studies of China English and British English. However, the corpus only covers two genres: newspapers and magazines, which is actually a corpus of Chinese Media English.

### 2.3 Corpus of English Text about China (CETC)

This is a collection of written text which is about or in connection with China. One of the primary aims of the corpus is to help researchers and lexicographers find and extract words and expressions that denote things or concepts in China or Chinese culture. Therefore, texts written by both Chinese speakers of English and other native or non-native speakers of English are collected in the corpus. For the first type of text, the text mainly comes from English media in China, English literary works by Chinese or overseas Chinese, English magazines in China and academic publication in China. For the second type of text, the text is mainly from English media in America and England, books and stories on the subject of China or Chinese culture by native or non-native speakers of English outside China.

With about 200 million tokens, the corpus is by far the largest one of the China English corpora. A pilot version is available for public use [7].
is large enough to serve as a reliable source for linguistic studies of China English, and at the same time it is not too large to slow the speed of searches. In other words, the corpus is manageable in size and architecture and fast to queries.

3.2 The Design
The design of CCE is made up of three components: sources, texts and annotations as is shown in Figure 1.

The design of CCE is made up of three components: sources, texts and annotations as is shown in Figure 1. The data of the corpus comes from a variety of sources. In terms of media, both printed and electronic texts are to be collected for inclusion in the corpus. A large portion of the data is from the Internet as the data is freely available online, such as news reports and journal articles. However, some literary works and academic writings are scanned as they are not available online.

As for the texts and domains, only written texts are collected. They cover a wide range of domains, such as news, financial, literary which will be introduced below in details.

The corpus is to be annotated with POS tags. With raw and POS-tagged data, primary studies can be done to look at linguistic features of China English at semantic and grammatical levels.

3.3 The Genres and Sources of CCE
CCE is composed of the texts from the following four genres: fiction, newspapers, magazines and academic publications. The texts of newspapers come from the online versions of the three national newspapers (China Daily, People’s Daily and Global Times) and one regional newspaper (Shanghai Daily). The texts of magazines come from the online versions of one national monthly magazine (China Today), and one regional weekly magazine (Beijing Review). Both national and regional newspapers and magazines were chosen in order to make the source of the corpus more varied. Fiction texts include novels, short stories and dramas written by Chinese. Academic texts include academic books and magazine articles. Fiction includes novels, short stories and plays written by overseas Chinese from different countries. Academic publications include academic books and journals. The genres and sources of CCE are shown in Figure 2.

3.4 The Text Composition of CCE
Texts were sampled for CCE from the genres and sources as shown in Figure 2. In this section, the text composition of each component parts will be expounded.

3.4.1 The newspaper part of CCE. It was reported that there were about 20 English newspapers in China [8]. Out of them, four newspapers were selected as the sources of the text for the corpus. They are China Daily, People’s Daily, Global Times and Shanghai Daily. The first three are national newspapers in China with a very large circulation. For example, People’s Daily claims to have a circulation of three to four million. The last one is a regional newspaper which will provide the latest information in Shanghai to expats living in Shanghai, visiting foreigners, and local white-collar workers proficient in English. So both national and regional newspapers were included in the corpus. Text was sampled from the four newspapers published during the period of 2006 to 2013. The number of texts and tokens collected from China Daily, People’s Daily, Global Times and Shanghai Daily was 3,634/1,335,750, 3,040/902,633, 3,370/1,273,404 and 1,444/1,224,595 respectively. In all, a total of 11,488 texts and 4,736,382 tokens were sampled for CCE. The texts were at varying length with an average of 412 words. Efforts were made to keep a good mix between specific domains: business (730,569), politics (705,487), environment (632,109), social (730,569), culture (689,205), technology (663,812), and sports (584,631). The percentages of each domain are shown in Figure 3 As is shown in Figure 3, the texts are basically evenly divided between the seven domains with business (16%) at the top and sports (12%) at the bottom of the list in terms of percentage.

3.4.2 The magazine part of CCE. China Today is a monthly magazine founded by National Honor Chairman Soong Ching Ling. Beijing Review is a weekly magazine published by the China International Publishing Group. Both magazines focus on topics on China’s economy, environment, politics, culture and lifestyle [9]. The texts were sampled from the two magazines published during the period of 2006 to 2013. The number of texts and tokens collected from China Today and Beijing Review was 2,136/1,388,972 and 1,341/949,040 respectively. In all, a total of 3,477 texts and 2,338,012 tokens were sampled for CCE. The average length of each text was 672 words. The texts of the magazines were selected from the same domains as those of newspapers: business (371,743), politics (295,522), environment (327,693), social (383,341), culture (377,732), technology (315,409), and sports (266,572). The percentages of each domain are shown in Figure 4.

As is shown in Figure 4, the texts are basically evenly divided between the seven domains with business (17%) at the top and sports (12%) at the bottom of the list in terms of percentage.

3.4.3 The fiction part of CCE. The texts of fiction are made up of published novels (e.g. Hajin’s The Bridegroom), plays (e.g. David Henry Hwang’s M. Butterfly) and short stories (Xinran’s The Good Women Of China) written by Chinese speakers of English. Most of the authors are overseas Chinese from different regions of the world. Some of the authors are: Ha Jin, Anchee Min and Nien Cheng from America, Harry Huang and Lien Chao from Canada, Pearl Buck and Li Cunxin from Australia and Jung Chang and Xinran from England. In all, a total of 34 texts and 3,406,968 tokens were collected for CCE.

3.4.4 The academic part of CCE. The academic sub-corpus of CCE is composed of texts from books and academic journals. The books were selected from social science, covering topics on education, history, philosophy, politics and economy. The journals were selected from natural science, covering topics on mathematics, physics, computer science and communications. The titles of some
journals are Journal of Control Theory and Application, Frontiers of Mathematics in China, Frontiers of Computer Science, China Communications, Chinese Physics Letters, etc. In all, a total of 200 texts and 3,476,610 tokens were collected for the academic sub-corpus.

3.5 Tokens, Types and Type-Token Ratio
CCE is composed of more than 13,962,102 tokens in 15,333 texts. The number of tokens, types and type-token ratios (TTR) is shown in Table 2. TTR is calculated as the total number of types divided by the total number of tokens multiplied by 100.

In Table 2, it is found that fiction and academic writings have a greater TTR than newspapers and magazines do. This may suggest that the former tend to have a more varied vocabulary.

3.6 The Annotation of the CCE
CCE was tagged with the freeware TagAnt [10], which tags texts with POS using the TreeTagger engine. By POS tagging, the corpus can be used for linguistic analyses of China English.

4 CONCLUSION
In the paper, the design and text composition of CCE were introduced in detail. One of the primary objectives of the corpus is to support linguistic studies on China English. It is argued that with more than 13,962,102 tokens the corpus is a good and reliable resource for researchers to investigate the lexical, semantic, syntactic and pragmatic features of China English. By contrastive and comparative studies of the linguistic features between China English and other varieties of English it is possible to identify the distinctive features of China English as a emergent variety of the English languages.

Moreover, the corpus is a helpful resource for English language teaching in the context of English as a Lingua Franca (ELF) in which the theories and practice of English language teaching shift from teaching the “Standard” English to English as a lingua franca. With the corpus evidence, English teaching textbooks and materials can be developed with due consideration given to China English.

The corpus is also beneficial for dictionary making, especially English-Chinese dictionaries. Many of them have a certain bias in language description. They tend to describe the so-called “standard English” and are “Anglo-American centered. That is to say, China English is rarely included in these dictionaries. One of the reasons might be that corpora of China English are not available for lexicographers. With corpus evidence and software tools, “lexicographers are able to describe the linguistic and cultural features of words
and phrases from China English adequately in their dictionaries [7].

Finally, the corpus is useful in the field of computational linguistics, natural language processing and corpus linguistics as a convenient and ready-to-use database.

However, the corpus has some limitations. First, it does not include any spoken data. Although written data can represent the general situation of the English language used in context of China, it will reveal more distinctive features of spoken China English with spoken data. Next, the fiction and academic parts of the corpus collect too much data from a single source. More varied data should be collected from different sources and different authors. Future efforts can be devoted to these areas. However, the representativeness and balance of a corpus are matters of degree. With careful design and data selection, CCE could represent the real situation of the English language used in China as a variety

ACKNOWLEDGMENTS

This work was supported in part by the Laboratory of Language Engineering and Computing of Guangdong University of Foreign Studies under Grant No. LEC2017ZBKT002: A Study on Grammatical Nativization of China English. It is also part of the research achievements funded by the Humanities and Social Science Planning and Funding Office of Ministry of Education of the P. R. C. under Grant No. 15YJA740048: China English or Chinglish: A Study Based on China English Corpus, the Department of Education of Guangdong Province under Grant No. 2017WTSCXZ2: A Study on the Definitions of Chinese-English Dictionaries for CFL Learners and Guangdong Planning Office of Philosophy and Social Science under grant No. GD14XWW20: A Corpus-based Study on China English. We would like to express our gratitude to them for their support.

REFERENCES

[1] Wei, R. and Su, J. 2012. The statistics of English in China. English Today, 28(3): 10-14.
[2] Li, W. 2012. Corpus markup and annotation: China English Corpus as an example. Waiyu Jiaoxue Yu Yanjiu (Foreign Languages Teaching and Research), (3): 336-345.
[3] Liang, J. 2015. Collocational Features of China English: A Corpus-based Contrastive Study. Ph. D. Thesis. The Hong Kong Institute of Education.
[4] Sun, H. and Lv, J. 2010. The construction and application of China English Corpus. Henan Shifan Daxue Xuebao (Journal of Henan Normal University),37(1): 219-223.
[5] Fang, A. C. and Le, F. 2012. Building a comparable corpus for comparative and contrastive studies of China English. http://xueshu.baidu.com/usercenter/paper/show?paperid=04bf37036d9f4163bf74a7a2492428&site=xueshu_se. retrieved on March 16, 2018.
[6] Fang, A. C., Le, F. and Cao, J. 2012. The Design, Establish and Primary Study of a Comparable Corpus of China English. Yuyan Yanjiu (Studies in Language and Linguistic),2:113-127.
[7] Xia, L., Xia, Y., Zhang, Y and Nesi, H. 2016. The corpora of China English: Implications for an EFL dictionary for Chinese learners of English. Lexikos, 26:416-435.
[8] Zhang, X. 2012. Zhongguo yingyu baokan fazhan xianzhu ji celue fenxi (On the current situation and development strategies of China’s English newspapers). Xinwen Fenxi (News Analysis), 9: 103-104.
[9] Peking Review. 1958. Introducing Peking Review. Peking Review,1(1): 3.
[10] Anthony, L. 2015. TagAnt (Version 1.1.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/software.