Construction and Analysis of Corpus of College English Teaching and Testing

Dingshun Li[1,a*], Gailin Liu [2,b]

[1] School of Liberal Arts, Xi'an Technological University, Xi'an, China

[2] School of Freshmen, Xi'an Technological University, Xi'an, China

[a*]email: lidingshun@xatu.edu.cn, [b] luga888@sina.com

Abstract. With the increasing awareness of the importance of big data in almost all the fields comes widespread attention to the use of corpus in language learning again. In order for English learners, teachers and researchers to have a paradigm for corpus construction to refer to and understand the important role of Corpus of College English Teaching and Testing in English teaching and testing, this paper makes a detailed introduction to the corpus construction from the perspectives of the necessity for building the corpus, principles for corpus building, corpus composition and technical specifications in corpus construction with the further hope that they can have a good command of the whole process of building a corpus. Meanwhile, a comprehensive analysis of functionalities and significance of its sub-corpus in linguistic studies and practical applications is carried out to prove that this corpus is of high quality and great value to college English teaching and testing. Finally this corpus is also believed to play a significant part in college English teaching reform and research.

Keywords: Corpus of College English Teaching and Testing; Corpus construction; Corpus analysis; Application of corpus

1. Introduction

Since 2000, corpus construction has witnessed a great boom in the academic circle in China and overseas and various corpora for different purposes were built to facilitate the rapid development of English Teaching and research. In U.S. and U.K. alone, corpora for teaching purpose include British National Corpus (BNC), Corpus of Contemporary American English (COCA), Cambridge Learner Corpus (CLC), Michigan Corpus of Academic Spoken English (MICASE), Corpus of Learner English (CLE) and so forth. In China, several corpora for English teaching are also built such as China Learner English Corpus (CLEC), China Spoken English Corpus (CSEC), China Academic Spoken English Corpus (CASEC), and so on. In addition, some corpora about college textbooks and College English Test were constructed such as Corpus of New Horizon College English Textbooks (CNHCET) by Foreign Language Teaching and Research Press and Corpus of College English Test (CCET) by Duanhe Yang [1].

All of these corpora belong to monolingual corpora widely used for English teaching and research across the world and they are all designed for either specific or general use. For example, CLEC is intended mainly for English writing and CCET for college English test training. There is no a balanced corpus particularly targeted at college English teaching and testing. For this reason, Corpus of College English Teaching and Testing (CCETT) were built to satisfy the urgent needs of teaching in China.

2. Principles for Corpus Building

The corpus is created on the basis of the following fundamental principles.

2.1 Purpose and Typicalness. What the corpus we built is mainly used for is the first priority for the selection of texts. That is to say, the corpus materials shall be those most closely related to learners' learning and testing purposes. Also important is that the corpus shall be composed of different levels of texts in difficulty for different grades such as texts suitable for freshmen, sophomore, and so on. On the other hand, all these graded texts must be typical and authoritative enough to represent what

https://doi.org/10.61957/joerai-20230406

learners are involved in on everyday-academic-life basis. Consequently, all the bilingual texts are chosen from 6 well-known College English Textbooks and 10 English-Chinese dictionaries.

2.2 Practicality and functionalities. In order for this corpus to find wide use and satisfiy the different requirements of teaching, testing and research, corpus creation must factor in various applications. To begin with, for the training of translation and writing, this corpus includes bilingual sub-corpus with aligned English-Chinese sentences and for the testing purpose, it also entails a monolingual sub-corpus of all the proficiency test papers for College English Test Band-4 and 6 (CET-4 and CET-6), and for National Entrance Examination for Postgraduates (NEEP). Secondly, all the texts of monolingual sub-corpus are designed with multiple formats including .txt, .docx and .pdf. This is done for the sub-corpus to be concordance or searched by different corpus tools such as Word Smith Tools, Anteconc and so on or other searching soft wares such as File Locator, Any Text Searcher and so forth. Lastly, the monolingual sub-corpus is tagged by CLAWS, a widely-recognized annotation tool designed by Lancaster University. In doing so, we aim for successfully retrieving various linguistic features such as some basic lexical and syntactic constructions frequently tested in these test papers.

2.3 Openness and sharing. This corpus is designed as an open corpus due to the constant update on texts from test papers and textbooks. Texts for college English proficiency test papers increase annually and texts from college English textbooks may be updated every few years. Accordingly, it is necessary to add the new texts to the corpus. On the other hand, this corpus is prepared with web retrieving interface and local desktop concordance tools. These online and local applications are created or included to benefit more teachers and learners in college English teaching and learning [2].

3. Composition of CCETT

This Corpus preparation mainly aims at college English teaching and testing. For this reason, corpus texts we select are from six college English textbooks which are most widely used in colleges and universities, 10 well-known English-Chinese dictionaries and all the college English proficiency test papers ranging from 1980 to 2023. In addition, a large number of elaborately selected English-Chinese sentences are also included in the bilingual sub-corpus so as to provide sufficient bilingual examples at sentence level for teachers and students to further refer to. Corpus composition is seen in Fig. 1.

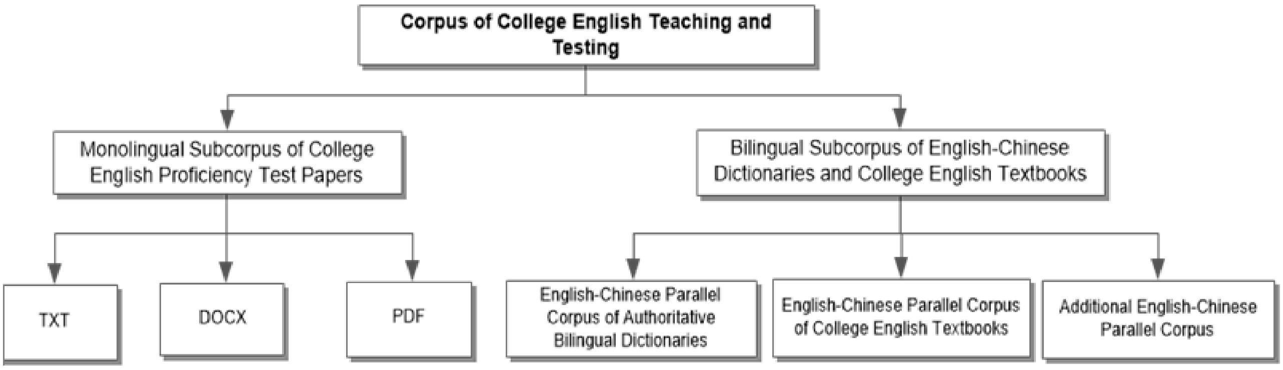
Corpus of College English Teaching and Testing
Monolingual Subcorpus of College English Proficiency Test Papers
Bilingual Subcorpus of English-Chinese Dictionaries and College English Textbooks
TXT
DOCX
PDF
English-Chinese Parallel Corpus of Authoritative Bilingual Dictionaries
English-Chinese Parallel Corpus of College English Textbooks
Additional English-Chinese Parallel Corpus

Figure 1. Corpus composition

3.1 Monolingual Sub-corpus of College English Proficiency Test Papers. Monolingual sub-corpus of CCETT consists of three types of texts from CET-4, CET-6 and NEEP, each with the number of 181, 112 and 82 respectively (See Table 1). The texts from CET-4 test papers range from 1989 to 2023, CET-6 from 1990 to 2023 and NEEP from 1980 to 2023.

Table 1. Composition of monolingual sub-corpus

Source	Range	Num. of Texts	Occurred Types	Expected Types	Tokens
CET-4	1989-2023	181	4665	4200	597,670
CET-6	1990-2023	112	5990	5500	839,342
NEEP	1980-2023	82	6844	5500	140,025

A brief look at this table finds the number of actual occurred types is 465 more than that of expected types, which represent vocabulary size a candidate should have for attending CET-4. Similarly, in CET-6 the number of occurred types is 490 more than that of expected types and in NEEP, the number is 1044 more. A simple analysis of the difference in the number types shows the vocabulary sizes tested in these proficiency tests are much larger than those a student is required to have, which is a further indication of an increase in test discrimination with increased test difficulty from CET-4 to NEEP [3].

3.2 Bilingual Sub-corpus of English-Chinese Dictionaries. As the most effective and authoritative learning materials, dictionaries rank the first in importance among all the corpus texts. And so in corpus construction, a selection of 10 dictionaries is made after a careful investigation into characteristics of current popular bilingual dictionaries available in the market. All the dictionaries find so wide use and enjoy so great popularity among English learners across China that their example sentences are selected as the primary bilingual corpus texts of CCETT. Detailed information on these dictionaries can be seen in Table 2.

Table 2. Composition of English-Chinese dictionaries

	Name	Num. of Sentences	Types
1	A Dictionary for English Expressions in College Writing (DEECW)	3,955	9,764
2	A New English-Chinese Dictionary (NECD)	9,172	9,576
3	Cambridge Advanced Learner's Dictionary (CALD)	34,326	26,011
4	Collins COBUILD Advanced Learner's English-Chinese Dictionary (CCALECD)	58,274	37,516
5	LANGMAN Dictionary of Contemporary English (LDCE)	52,405	25,960
6	Langman Language Activator (LLA)	44,313	25,573
7	Merriam-Webster's Advanced Learner's English-Chinese Dictionary (MWALECD)	50,716	26,474
8	Oxford Advanced Learner's English-Chinese Dictionary (OALECD)	36,079	20,775
9	The 21st Century English-Chinese Dictionary (21CUECD)	33,393	19,845
10	The Concise Oxford English-Chinese Dictionary (COECD)	77,317	22,090
Total		399,950	223,584

This portion of corpus consisting of example sentences from 10 bilingual dictionaries totals approximately 399,950 aligned sentences with vocabulary size up to 37,516. A comparison of these data with the corresponding data in Table 3 from college English textbooks enables us to find that the language data these dictionaries contain are much richer and larger than those included in 6 textbooks whether in vocabulary size or sentence varieties. This fact reveals the important role of dictionaries in language teaching and the necessity to take sentences from bilingual dictionaries as the fundamental bilingual corpus [4].

3.3 Bilingual sub-corpus of college English textbooks. The bilingual texts of this sub-corpus are from three sources: six college English textbooks, ten widely-used English-Chinese dictionaries and additional bilingual texts with aligned English-Chinese sentences (See Table 1).

Six college English textbooks are chosen because each of these textbooks is most widely used and highly appreciated among colleges and universities. To be specific, all the texts for intensive reading in these textbooks and their corresponding Chinese translations are selected. What is included about lexical information in these college English textbooks can be seen in Table 3.

Table 3. Composition of college English textbooks

Name	Num. of Texts	Num. of Sentences	Types of English Texts
21 Century College English	86	4,249	7,510
Experiencing English	36	1,350	6,452
College English New Concept	36	1,572	6,565
NEW Concept English	42	3,879	6,620
New Horizon College English	224	12,263	8,102
New standard College English	48	1,746	7,761
Total	472	25,059	43,010

Though the analysis of Table 2, we find vocabulary size differs in each series of textbooks from 6,452 to 8,102 but these number still ranges from 5,000 to 9,000, which indicates the necessary vocabulary size a fluent English speaker should have. The larger the number is the higher the level of an English learner in English is. The fact that New Horizon College English comes closer to the number of 9000 words shows this series of textbooks attempts to meet the maximum requirement of the Teaching Requirements for College English Course (2022) and its lexical richness is much higher than that in other series of textbooks (see Fig. 2).

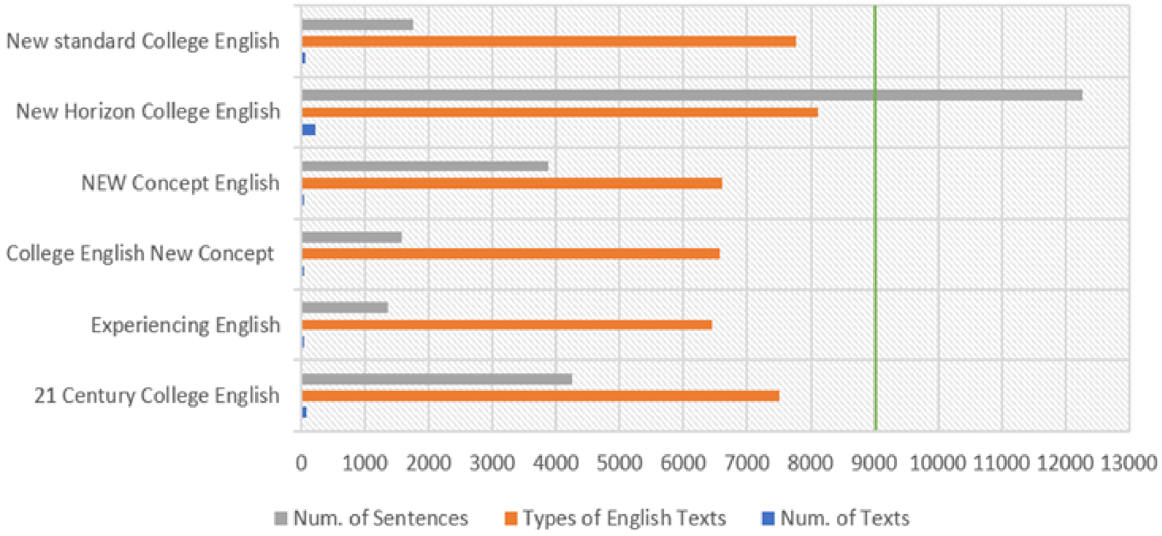

Figure 2. Analysis of lexical richness in textbooks

3.4 Bilingual Sub-corpus of Additional English-Chinese Aligned Sentences. Although corpus of dictionaries and textbooks provides a wealth of sufficient and authoritative language information, a huge amount of language data is sure to write more enormous and underlying language rules and laws. In view of this, an extra selection of English-Chinese aligned sentences is conducted to enrich CCETT. All these sentences stem from a wide range of textbooks for listening, speaking, reading, wring and translation, from several world-famous bilingual journals and Newspapers, from some subtitles of classical movies and TV series, etc. The sources are too varied and diversified to list here.

4. Major Technical Specifications in Building Corpus

In corpus construction, there are several important technical regulations employed to serve the various purposes including AI-based text cleansing, POS annotation and bilingual sentence alignment.

4.1 AI-based Text Cleansing. All the texts in the corpus are processed in three steps to remove unnecessary information and errors in format conversion. Part of corpus files are originally images unfavorable for corpus concordance tools to conduct a search and make a statistical analysis. As a result, these files in image format must be converted into text, docx, and searchable pdf files in the first step in which Abbyy FineReader 16 with AI functionalities is employed to undertake the accurate conversion from images to the three types of text formats. In the second step, Emeditor 22 and Black Horse software, along with a set of regular expressions are utilized to remove some format and word errors happening in the OCR conversion. Finally, a further error correction of all the files obtained through step 2 is performed with the help of GPT-4 to get the final raw corpus texts.

4.2 POS Annotation. For some linguistic features to be retrieved or searched, it is necessary to annotate raw corpus texts. As a consequence, CLAWS4 (the Constituent Likelihood Automatic Word-tagging System) tagger, the most powerful annotation tool so far developed by UCREL (University Centre for Computer Corpus Research on Language) at Lancaster is used to tag all the texts from monolingual sub-corpus of college English proficiency test papers so as to extract some special sentence patterns and grammatical constructions such as collocations and colligations [5].

4.3 Bilingual Sentence Alignment. Bilingual texts from college English textbooks and some additional corpus texts are aligned neither at sentence level nor at paragraph level. So we must perform sentence-level alignment first before making a concordance. Hence, Align Factory 4.0.3 and SCAT 2023 are combined to help carry out the high-efficiency and quality alignment.

5. General Analysis of CCETT

In this part, a general analysis of monolingual sub-corpus of college English proficiency test is made by using the corpus analysis tool Word Smart developed by the former School of International Studies, Xi’an Technological University [6].

5.1 Wordlist-based Analysis of Monolingual Sub-corpus. One of basic functions Word Smart has is wordlist-based statistical analysis which can perform an overall analysis of the monolingual sub-corpus from five major parameters: lexical density, mean word length, means sentence length, mean paragraph length and readability (See Table 4).

Table 4 Composition of College English Textbooks

N	Item	CET-4	CET-6	NEEP
1	Lexical Density	0.46	0.54	0.59
2	Mean word length	4.46	4.69	5.35
3	Mean sentence length	7.09	8.01	8.25
4	Mean paragraph length	99.11	112.94	139.79
5	Readability	10.51	11.11	11.76

Lexical density of CET-6 in Table 4 approximates the number for TEM-4 (0.487) [7] and mean word length is greater than average word length in spoken language. This indicates texts tested in CET-4 are very formal essays. And readability of CET-4 is as high as 10.51, which is equivalent to the level of high school graduates in English speaking countries. If we take a look at the various values in CET-4, CET-6 and NEEP, we can find all these values increase from CET-4 to NEEP. This changing trend can be seen in Fig. 3.

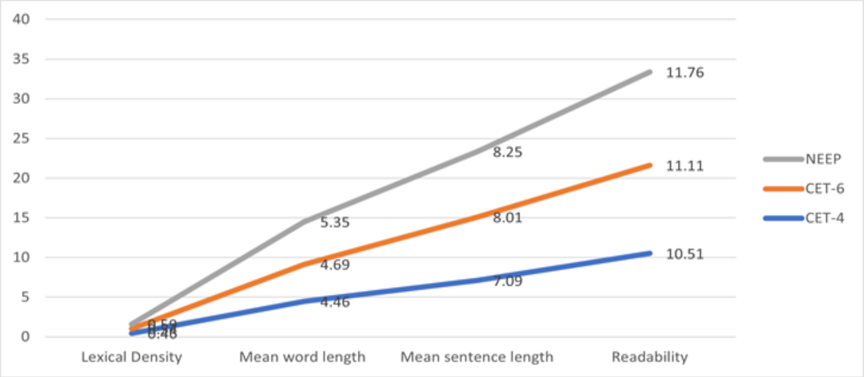

Figure 3. Shifting trend of wordlist-based parameters in texts

It is not difficult to understand these upward changes in the data because in reality, the intended difficulty for these tests from CET-4 through CET-6 to NEEP is sure to rise with increasing requirements of these tests in the complexity of words, sentences and essays. But here one of important findings is that the difficulty in English in CET-6 and NEEP stands the same level as that of texts used to test the third-year college students in English-speaking countries.

These analyses show us the great significance of monolingual sub-corpus for English learners in various language indicators. This is also further evidence that this sub-corpus has significant effects on college English teaching and testing.

5.2 Analysis of Bilingual Sub-corpus of English-Chinese Dictionaries and College English Textbooks. With the rapid development of machine translation (MT) in the last decade, high quality language assets have found increasing importance in raising the efficiency and quality of computer-aided translation (CAT). Large amount of language assets, bilingual translation memory in particular are used to train various translation models which translators can further finetune by using high-quality bilingual sentences, essentially the equivalence of translation memory. In view of this, bilingual sub-corpus of English-Chinese dictionaries and college textbooks are converted into translation memory to finetune a translation model to validate the quality of these bilingual language data. The translation model is provided by OPUS, an open source project concerning machine translation applications and research [8]. In this case, based on OPUS-CAT MT Engine, we choose the English-mandarin Chinese model with the name of opus+bt-2021-04-19 to be finetuned with the converted translation memory of about 410,000 aligned sentences. The finetuned results are shown in Fig. 4.

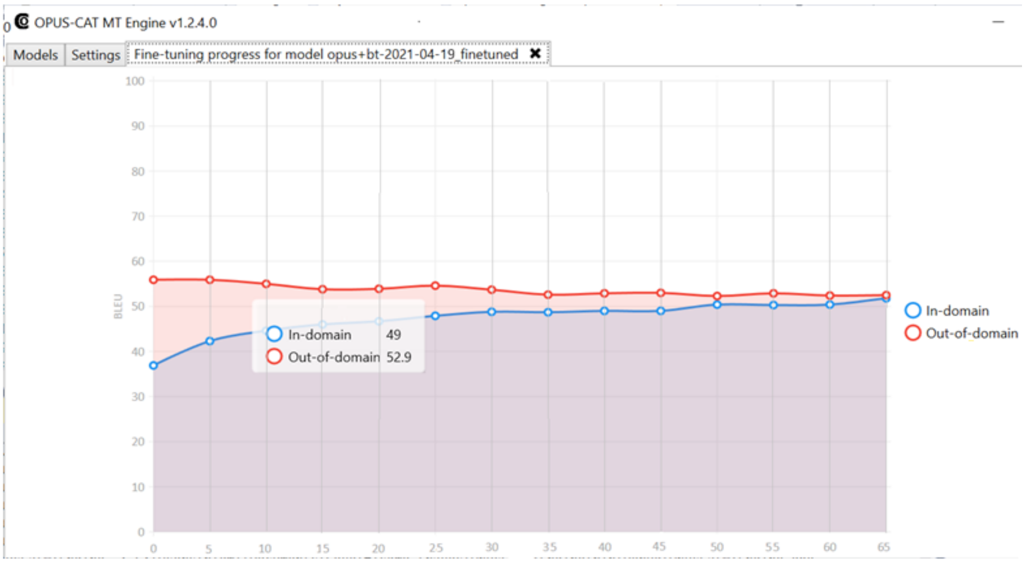

Figure 4. Finetuning progress for the model of opus+bt-2021-04-19

The results show out-of-domain line is gradually on the decline and in-domain line is on the rise so that both lines come closer and closer eventually. This reveals the positive role of translation memory in finetuning the translation model because the corpus data we collected is derived from bilingual dictionaries and college English textbooks and all of its translations for the corpus texts are contributed by professional translators or experts well-known in translation fields. Furthermore, the texts of this sub-corpus are very formal written language, so the rising trend of in-domain line approaching the out-of-domain line is a good indication that the finetuning has played positive part in improving the translation model. In other words, this also shows the bilingual sub-corpus is of high quality and great value in translation and teaching.

6. Conclusions

Through the above-mentioned description, discussion, statistical and practical analysis of Corpus of College English Teaching and Testing, a general workflow of purposes, principles, design, composition and technical methods and specifications in building corpus are presented in a detailed way in order to help establish a universal mode for corpus construction of this type. By following the process of corpus creation, learners, teachers or researchers can set up their own corpus projects to serve the purpose of teaching and research.

With the help of statistical analysis of wordlists from monolingual sub-corpus, some significant and valuable features in language are reflected and when the bilingual sub-corpus is used to finetune the language model, a considerable improvement to it is made. All of these practical applications show this corpus is of the great reliability, credibility and practicality in language teaching, testing and academic research as well.

References

[1] Z. Q. He, X. W. He, Overview of English corpus research: review, current status and Prospects, J. Foreign Language Education. Vol. 42, (2011) 5-15. (In Chinese)

[2] Z. L. Zhou, D. S. Ren, A corpus-based study on the translation and communication of ocean discourse with Chinese characteristics: framework, contents and principles, J. Journal of Ocean University of China. Vol. 4, (2022) 23-31. (In Chinese)

[3] D. H. Yang, A corpus built for college English band-4, band-6 tests, J. Technology Enhanced Foreign Language Education. Vol. 113, (2007) 23-31. (In Chinese)

[4] Nation: *Learning Vocabulary in Another Language*, (Cambridge University Press, UK 2013), p98.

[5] L. Wang, M. C. Liang, On POS Tagging reliability for EFL learners' transcribed spoken data, J. Foreign Language Education. Vol. 28, (2007) 48-51. (In Chinese)

[6] D. S. Li, G. L. Liu, A brief introduction to China English corpus of China Daily, J. English Square. Vol. Z3, (2011) 58-59. (In Chinese)

[7] T. He, A comparative study on readings in TEM-4 and TEM-8, based on Compleat Lexical Tutor J. English Square. Vol. 28, (2022) 58-59. (In Chinese)

[8] Information on https://helsinki-nlp.github.io/OPUS-CAT/