International Journal of Social Policy and Education
Vol. 1, No. 2; December, 2019. pp 135-142
ISSN 2689-4998 (print), 2689-5013 (online)

Published by International Center for Promoting Knowledge
DOI: 10.61494/ijspe.v1n2a13
URL: https://doi.org/10.61494/ijspe.v1n2a13

Construction of CivDEAP Corpus and Its Application in Academic Research and Teaching Practice in Civil Engineering

Baicheng Zhang
Professor
School of Foreign Studies
Chongqing Jiaotong University
China
&
Ling Yang
Lecturer
School of Civil Engineering
Chongqing Jiaotong University
China

Abstract

Civil engineering academic English corpus CivDEAP is a sub-corpus of the DEAP (Database of English for Academic Purposes) Corpora. The paper first introduces the construction of CivDEAP in terms of construction objective and scheme, text collecting, naming, cleaning, and annotating, and then explores the application of CivDEAP in both academic research and teaching/learning practice. Lastly, the paper offers some tentative suggestions on the follow-upsize-expanding construction of CivDEAP.

Keywords: academic English; civil engineering; specialized corpus; corpus construction

Introduction

A corpus is a collection of language use instances which are usually collected according to the sampling criterion serving the purpose of corpus construction. In the 1950s, American structuralists conducted linguistic study through collecting authentic samples of people's language use, and the notion of "collecting real data" therefore came into being (Leech 1992). Structural linguists, who were regarded as the forerunners of corpora, attached great importance to language data in real communication, and devoted themselves to the "commitment to putting real language data at the core of what linguists study" (Michael & O'Keeffe 2010: 4). In the late 1950s, as the improvement of computer technology, authentic language data were processed and indexed by utilizing computers (Parrish 1962). 1970s saw considerable improvements of computer processing capability and information technology. The Key Word in Context (KWIC) concordances replaced the prior catalogue indexing cards, and automatic subject analysis was realized (Hines et al. 1970). In the 1980s and 1990s, corpora became tools for linguists and applied linguists (McCarthy & O'Keeffe 2010).

According to the construction objective, features of language samples, corpora can be categorized into different types, such as general corpora, specialized corpora, synchronic corpora, diachronic corpora, spoken corpora, written corpora, native speaker corpora, learner corpora, monolingual corpora, parallel/bilingual corpora and multilingual corpora (Liang, Li & Xu 2010). Compared with large general corpora, specialized corpora serve certain research purposes, and are usually composed of language samples collected in specific domains. Academic English corpora, as one of the sub-types of specialized corpora, have been playing great roles in theoretical study on the features of academic discourse and practical application in the teaching and learning of academic writing.
Internationally, among the already-built specialized academic English corpora, some have been exerting enormous academic influence, including Michigan Corpus of Academic Spoken English (MICASE), British Academic Spoken English Corpus(BASE), Michigan Corpus of Upper-level Student Papers (MICUSP), Louvain Corpus of Native English Essays(LOCNESS), International Corpus of Learner English(ICLE), etc. Domestically in China, some representative academic English corpora are the English Corpus of Science and Technology (JDEST) build in Shanghai Jiaotong University in 1980s, and several others build in early this century: Chinese Learner English Corpus (CLEC), College Learners' Spoken English Corpus (COLSEC), Spoken and Written English Corpus of Chinese Learners (SWECCL).

The DEAP (Database of English for Academic Purposes) corpus, as a sub-type of specialized academic English corpus construction project supported by Foreign Language Teaching and Research Press and Huge Mind (Beijing) Education & Technology Co., Ltd. was launched in 2016, aiming to build an academic English corpus of over 100 million words covering about 30 disciplines in humanities, social sciences, and natural sciences. So far, 21 sub-databases have been approved, among which, MedDEAP (Medical Database of English for Academic Purposes) corpus, BioDEAP (Biology Database of English for Academic Purposes), and another 4 sub-databases have been built, and 11 sub-corpora are about to be completed, including InfoDEAP (Information Database of English for Academic Purosese) corpus, EduDEAP (Pedagogy Database of English for Academic Purposes) corpus, and some others.

The construction of academic English corpus in engineering in China also shows initial results. CNKI (China National Knowledge Infrastructure) search shows that the construction and research of engineering corpus mainly involve English corpus of mechanical engineering (Zhang 2019), English corpus of agricultural engineering (Qi 2019); Fan 2019), English corpus of architectural engineering (Liu 2016; Zhang 2016), English corpus of abstracts of master's thesis in transportation engineering (Qi 2017), English corpus of aircraft manufacturing engineering (Wang 2015), etc. In terms of the construction and application of civil engineering academic English corpus, Liu Guocong and Gao Jun (2016) extracted 120 abstracts of civil engineering academic papers respectively from databases of CNKI(Chinese National Knowledge Infrastructure) and ASCE (American Society of Civil Engineers), and investigated the linguistic features of the genre. Tong Xing and Qiu Pengcheng (2016) briefly introduced the construction of a small-scale civil engineering English corpus and the effect evaluation of its application in teaching practice, but did not specify any information of the corpus, and no subsequent research based on the corpus was reported or published. Against this background, we are able to see the urgent need for the construction of civil engineering academic English corpus, which is the rationale for the construction and application of CivDEAP.

Corpus-Construction Objective

According to the design scheme of the DEAP Corpus, the text sampling taxonomy for DEAP follows the Chinese National System of Level One Disciplines for Degree Education (The Academic Degrees Committee of the State Council 2013), which also stipulates the Level Two sub-disciplines within each Level One disciplines. Following these requirements, CivDEAP aims to cover six Level-Two sub-disciplines within the Level-One discipline of civil engineering. The corpus contains three types of texts (research article, review article and editorial materials) published in 24 high-quality English academic journals. Consequently, CivDEAP will bea balanced, representative and timely academic English corpus with a size of over 5 million words, and serves academic discourse research and curriculum teaching practice.

Corpus-Construction Scheme

Disciplines

According to the Catalogue of Academic Degree Granting and Talent Cultivation Disciplines (Version 2011) issued by the Academic Degrees Committee of the State Council and the Chinese Ministry of Education, the Level-One discipline Civil Engineering (Code 0814), within the discipline category of Engineering (Code 08), has set up six Level-Two sub-disciplines, namely: Geotechnical Engineering (081401), Structural Engineering (081402), Municipal Engineering (081403), Heating, Gas Supply, Ventilation and Air Conditioning Engineering(081404), Disaster Prevention and Mitigation Engineering and Protection Engineering (081405), and Bridge and Tunnel Engineering (081406). The CivDEAP corpus covers all the six sub-disciplines, and is therefore discipline-balanced as well as broadly representative.

Source Journals

Through literature searching in WOS (Web of Science), we listed10 high-quality international research journals for each Level-Two sub-discipline (60 in total) within the Level-One discipline of Civil Engineering. Then, we consulted some experts and scholars working in civil engineering research, and initially selected 5 journals in each sub-discipline (30 journals in total). Finally, after referring to the Impact Factor and QC (Quartile in Category) in JCR (Journal Citation Reports) of the source journals, 4 journals in each sub-discipline (24 journals in total) are determined as the source journals for sampling and collecting CivDEAP corpus texts (see Table 1).

Table 1. Source journals of CivDEAP

No.	Level-Two sub-discipline	Journal title
1	Geotechnical Engineering	*Computers and Geotechnics*
2		*Engineering Geology*
3		*International Journal of Rock Mechanics and Mining Sciences*
4		*Soil Dynamics and Earthquake Engineering*
5	Structural Engineering	*Cement and Concrete Research*
6		*Engineering Structures*
7		*Journal of Composites for Construction*
8		*Journal of Engineering Mechanics*
9	Municipal Engineering	*Construction and Building Materials*
10		*Environmental Modelling & Software*
11		*International Journal of Greenhouse Gas Control*
12		*Water Research*
13	Heating, Gas Supply, Ventilating and Air Conditioning Engineering	*Building and Environment*
14		*Energy and Buildings*
15		*Renewable Energy*
16		*Solar Energy Materials and Solar Cells*
17	Disaster Prevention and Reduction Engineering and Protective Engineering	*Earthquake Engineering & Structural Dynamics*
18		*International Journal of Solids and Structures*
19		*Journal of Sound and Vibration*
20		*Natural Hazards Review*
21	Bridge and Tunnel Engineering	*Journal of Bridge Engineering*
22		*Journal of Structural Engineering*
23		*Tunnelling and Underground Space Technology*
24		*Structural Control & Health Monitoring*

Literature Type

According to characteristics of the 24 source journals, CivDEAP selected three types of literature/texts, namely research article, review article and editorial material. Considering the publications of the source journals, CivDEAP roughly balances the number of texts in each of the three types of literature in 6 Level-Two sub-

disciplines within the Level-One discipline of Civil Engineering, with research articles accounting for about 90%, review articles accounting for about 6%, and editorials accounting for about 4%.

Publication Year

In order to ensure the timeliness and recency of the corpus data, CivDEAP principally selects texts published in 2018. However, due to the inadequacy of review articles and editorial materials published in 2018 in some source journals, we also collected s number of review articles and editorials published in 2019(6), 2017(1) and 2015(2) in order to achieve a roughly balanced ratio of three types of literature in the 6 sub-disciplines.

Text Collecting

Firstly, each of the selected source journal titles is searched in the database of Web of Science to obtain the literature lists of the journal included in WOS. Then, relevant types of documents were screened out by setting 2018 as the "Publication Year", and selecting "Article", "Review" and "Editorial" respectively. Finally, according to the order of "Times Cited", a certain number of texts are selected for each journal: 35 research articles, 3 review articles and 2 editorials. If certain types of texts are inadequate, then "Publication Year" is respectively extended to 2019, 2017, 2016 or 2015 until the number of documents meets the requirements.

When collecting texts, given the huge workload brought by a large number of charts and formulas in most civil engineering research papers, we did not adopt the method of downloading PDF documents and converting them. Instead, we prepared WORD templates first, then obtained full texts in webpage format by selecting "Full Text from Publisher", and copied the Title, Author, Abstract, Keywords, Text, Acknowledgements of the documents and pasted them into the prepared templates according to the internal structure order of the documents (references and appendices were deleted according to the design scheme of CivDEAP).

In the first stage, we collected 960 texts (160 for each sub-discipline, and 40 for each journal), with a size of more than 6. 2 million words. In the second stage, according to the design scheme of DEAP Corpus, about 5 million words in each sub-corpus, we tried deleting some texts to satisfy the size requirement, and finally 780 texts (130 in each sub-discipline, and 30-35 in each journal) were retained, of which 712 were research articles, 42 were review articles and 26 were editorials. The size of unannotated CivDEAP is 5040349 words [1], and the annotated one (with header and structural information annotation) is 5088754 words (See Table 2 for more specific information)

Table 2. An overview of CivDEAP

Literature Type	Number of Texts	Size (Raw)	Size (Annotated)
Research Article	712	4638002	4682516
Review Article	42	339698	342181
Editorials	26	62649	64057
Total	780	5040349	5076876

Text Naming

All the texts of the CivDEAP Civil Engineering Academic English Corpus were named in the way of "Level-One Discipline Code+ Level-Two Disciplines Code+Literature Type Code+Text Number". The code of Level-One Discipline of Civil Engineering is denoted by the first letter C of the discipline title; the six Level-Two sub-disciplines are encoded respectively as the initials of the first two content words of the discipline title (see Table 3 for details). Among the three types of literature, research articles are encoded as RA, review articles as RV, and editorial materials as ED. Texts are numbered with three digits (from 001 to 780) in the order that research articles are numbered first (from 001 to 712), followed by the review articles (from 713 to 754), and finally the editorial materials (from 755 to780). Therefore, for example, CBTRA001 indicates that the text is a research article, belonging to the Level-Two Discipline of Bridge and Tunnel Engineering under the Level-One Discipline of Civil Engineering, and it is the first text of CivDEAP corpus, while CSEED780, the 780th text of CivDEAP, is an editorial of the Level-Two Discipline of Structural Engineering.

It needs to be noted that, following such a naming scheme, Journal Title is not included in the text name of CivDEAP, but is reflected in the header information of each text. For example, < journal_title>Journal of Bridge Engineering</journal_title> indicates the literature published in Journal of Bridge Engineering.

Table 3. Titles and codes of the 6 level-two disciplines and text numbers

	Level-Two Discipline Title	Code	Text Number
1	Bridge and Tunnel Engineering	BT	001-118
			713-319
			755-759
2	Disaster Prevention and Reduction Engineering and Protective Engineering	DP	119-239
			720-724
			760-763
3	Geotechnical Engineering	GE	240-355
			725-730
			764-771
4	Heating, Gas Supply, Ventilating and Air Conditioning Engineering	HG	356-474
			731-738
			772-774
5	Municipal Engineering	ME	475-592
			739-747
			775-777
6	Structural Engineering	SE	593-712
			748-754
			778-780

Text Cleaning

Text cleaning has been conducted in two stages: cleaning WORD texts and TXT texts respectively. In WORD-text cleaning stage, we first conducted Searching and Replacing to delete line breaks, extra spaces and a large number of graphs (with graph titles kept) in batch operations. Then compiled and executed the macro command of Microsoft WORD to eliminate all tables (with table titles kept) [2]; Finally, WORD texts are converted into TXT format and saved as UTF-8 texts. When cleaning up TXT texts, we used Text Editor to set up a batch cleaning scheme to deal with the problems of initial and end spaces of paragraphs, blank lines between paragraphs, full-width punctuation and letters, and unreadable codes. Then we performed "Boundary Searching" and "Replacing" to delete in batches the download links of web pages beneath the graphs.

After the above two stages of work, we had 780 clean texts of CivDEAP, well prepared for text annotation.

Text Annotating

XML(Extensible Markup Language)has been used for CivDEAP text annotation. Following XML marking-up scheme, certain content is placed between the start-tag (marking the beginning of a specific area) and the end-tag (marking the end of a specific area) pairs. The names of the start-tag and end-tag are placed between a less-than symbol (<) and greater-than symbol (>). The end-tag is preceded by a slash (/). <Author>, for instance, indicates the start-tag of the author information of the article, while</Author> the end-tag of that. Text annotation in the CivDEAP Corpus includes two parts: text header information and test structural information.

Header Information

Adding text Header information started at the very beginning when the texts are collected: we adopted XML format and preset specific header information at the beginning of each WORD template, starting with <header> and ending with </Header> (see Table 4 for details). When collecting each text, we copy the meta-information and paste them in original order into the prepared WORD template, such as: <Publication_Year>2018</Publication_Year>, <Volume>23</Volume>, etc.

Table 4. Header information labeling of texts in CivDEAP

Label	Meaning
<Header>	Start of Header Information
<Publication_Year></Publication_Year>	Publication Year
<Domain></Domain>	Domain
<Discipline></Discipline>	Discipline
<Contributor></Contributor>	Data Collector
<Journal_Title></Journal_Title>	Journal Title
<Volume></Volume>	Volume Number
<Issue></Issue>	Issue Number
<Pages></Pages>	Pages
<DOI></DOI>	Digital Object Identification
</Header>	End of Header Information

Literature published in some of the 24 source journals are conventionally categorized into "Volume "but not "Issue", and some of the literature are only numbered but do not have pages. For the missing Issue and Page number information, we marked them all as "unknown". In CivDEAP Corpus, 525 texts only do not have Issue number; 189 texts have no Page information; all texts have DOI(Digital Object Identification) code and other meta-information.

Structural Information

Structure information of the literature was also marked in XML format, which was carried out on TXT texts. Power GREP was used to add structural information of relevant literature[3] in batches in the "Regular Expression" retrieval mode, and then manual inspection was carried out to correct labeling errors. As the texts of CivDEAP corpus involve three types of literature, structure of which may differ from each other in one way or another. Even for the same kind of literature published in the same journal structures may be more or less different. Therefore, we only labeled the structural information shared by most literature (see Table 5). According to the natural state of the literature published in the source journals, the structural information we marked mainly includes Title, Author, Abstract, Keywords, Introduction, Methods, Results and Discussions, Conclusions, and Acknowledgements. Due to the limitation of time and energy, we did not conduct detailed distinction and annotate the parts of some literature that do not belong to the above-mentioned structural elements, or the contents of a part contains several structural types.

Table 5. Structural information labeling of texts in CivDEAP

Label	Meaning
<Article_Title></Article_Title>	Article Title
<Author></Author>	Author
<Abstract></Abstract>	Abstract
<Keywords></Keywords>	Keywords
<Introduction></Introduction>	Introduction
<Methods></Methods>	Research Methodology
<Results_and_Discussions></Results_and_Discussions>	Research Findings and Discussions
<Conclusions></Conclusions>	Conclusions
<Acknowledgements></Acknowledgements>	Acknowledgements

Application of CivDEAP

The CivDEAP Corpus can be applied both in academic research and teaching/learning practice. In academic research, the corpus can be used to investigate the characteristics of civil engineering academic discourse at the levels of vocabulary, phrase, syntax, rhetoric, discourse, etc.

It is also feasible to carry out comparative studies on characteristics of academic discourse between two sub-disciplines, or among different types of literature in the same journal(s). Comparative studies of academic discourse between relevant disciplines can also be conducted among the sub-corpora of CivDEAP.

In terms of teaching and learning practice, CivDEAP corpus can be used in the teaching and learning of courses like English for Special Purposes, Academic Writing, Discourse Analysis, etc., providing adequate authentic samples of academic English in research articles in civil engineering, and enabling teachers and students to be familiarized with features of civil engineering academic discourse to facilitate their academic writing and international publication. Further, by comparing CivDEAP with a large-scale general English corpus, lexical items frequently used in civil engineering academic discourse can be extracted, which would contribute enormously to the compilation of an Academic English Dictionary of Civil Engineering that will benefit teachers, students as well as other practioners working in civil engineering research.

Concluding Remarks

This paper introduces the construction of civil engineering academic English corpus CivDEAP in terms of construction objective, construction scheme, text collecting, naming, cleaning and annotating. As a sub-corpus of DEAP Academic English Corpus, CivDEAP, together with other sub-corpora, will play an important role in further promoting the construction of specialized academic corpus as well as theoretical and applied research. However, the corpus construction and application research are not accomplished overnight. In order to make better use of CivDEAP, the follow-up size-expanding construction needs to be considered. Specifically, the follow-up construction of CivDEAP can be designed and carried out from three perspectives:

To begin with, from a diachronic perspective. According to the existing data collection scheme, on one hand, literature published in the 24 source journals before 2018 can be traced back, and the three types of literature (research articles, review articles and editorials) can be collected at intervals, every five years, for instance, i. e. tracing back from 2018 to 2013, 2008, 2003, 1998, 1993, etc., until to the years when the journals started publication. On the other hand, literature published in those source journals after 2018 will also be collected every five years, i. e., 2023, 2028, etc. In this way, a diachronic corpus CivDDEAP (Civil Engineering Academic English Diachronic Corpus) can be built to facilitate diachronic research to capture possible diachronic changes of relevant academic discourse features.

Next, the follow-up expansion construction can be realized from the perspective of language types. Based on the text collection scheme of CivDEAP, 24 high-quality Chinese journals in civil engineering can be selected, and three types of literature will be collected, to build a CivBDAP (Civil Engineering English-Chinese Bilingual Academic Corpus), which can be used for comparative research, investigating similarities and differences of certain academic discourse features in academic journals at home and abroad.

The third perspective is the construction of a parallel corpus. Both the English translation and the Chinese form of the Titles and Abstracts of the literature published in Chinese civil engineering research journals mentioned above will be collected, and a CivPDAP (Civil Engineering Parallel Database for Academic Purposes), therefore, can be built, which is convenient to investigate characteristics of the academic discourse in Title and Abstract parts of Chinese and English academic papers as well as the translation strategies.

Notes

[1] The specific data of CIV DEAP corpus capacity (number of words) are obtained in Power GREP using regular expression [A-ZA-Z0-9-].

[2] I would like to thank teacher Li Liang of the foreign language school of Jilin normal university for providing technical support, writing macro commands and deleting forms in batches.

[3] I would like to thank Professor Peng Gong at the Chinese Academy of Sciences, Professor Xu Jiajin in Beijing Foreign Studies University and Dr. Li Liang in Jilin Normal University for their support and assistance in using Power GREP and compiling regular expressions.

Funding

This work was supported by Chongqing Municipal Education Commission under Grant [19SKGH058]; and School of Foreign Studies, Chongqing Jiaotong University under Grant [2017WZ01].

References

Hines, T.,J. Harris & C. Levy. 1970. An experimental concordance program. Computers and the Humanities 4(3): 161-171.

Leech, G. 1992. Corpora and theories of linguistic performance [A]. In J. Svartvik (ed.). Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82 [C]. Berlin and New York: Mouton de Gruyter. 125-148.

Michael, M. & A. O'Keeffe. 2010. The Routledge Handbook of Corpus Linguistics [M]. London and New York: Routledge.

Parrish, S. M. 1962. Problems in the making of computer concordances [J]. Studies in Bibliography 15: 1-14.

Fan, Xuyan. 2019. A Corpus-based Study on Co-selection of Keywords in Agricultural Engineering English Research Papers [D]. Unpublished MA thesis, Northeast Agricultural University.

Liang, Maocheng, Li, Wenzhong, &Xu, Jiajin. 2010. Using Corpora: A Practical Course book [M]. Beijing: Foreign Language Teaching and Research Press.

Liu, Guocong & Gao, Jun. 2016. A corpus-based analysis of abstract translation in civil engineering [J]. Crazy English(Theoretical Edition)(2): 161-164+208.

Liu, Jia. 2016. Construction standard and application of a small English corpus of construction engineering [J]. China Standardization (11): 113-114.

Qi, Linlin. 2017. A corpus-based Error Analysis of English Abstracts of Transportation Engineering Masters' Theses [D]. Unpublished MA thesis, Dalian Maritime University.

Qi, Yusi. 2019. A Corpus-based Study on Nominalization in Agricultural Engineering English Texts [D]. Unpublished MA thesis, Northeast Agricultural University.

Tong, Xing & Qiu, Pengcheng. 2016. A study on ESP-based private college students English corpus construction - A case of civil engineering English [J]. Crazy English (Theoretical Edition)(2): 141-142.

Wang, Xiaoying. 2015. A study of the construction and application of small-aircraft manufacturing engineering English corpus [J]. Journal of Huaibei Vocational and Technical College (5): 94-95+112.

Zhang, Chi. 2016. A discussion on the construction and application of a small-scale construction engineering English corpus [J]. Journal of Chifeng University (Philosophy and Social Science Edition) (10): 253-255.

Zhang, Yafeng. 2019. Construction and application of English corpus of mechanical engineering (2009-2017) [J]. Journal of Huanghe University of Science and Technology (4): 94-99.