CSCanada
Studies in Literature and Language
Vol. 17, No. 1, 2018, pp. 60-64
DOI:10.3968/10526

ISSN 1923-1555[Print]
ISSN 1923-1563[Online]
www.cscanada.net
www.cscanada.org

The Preliminary Study on the Construction of the Energy Power Corpus

ZHANG Qian[a],*; WANG Chunyan[b]

[a]Lecturer. School of Foreign Languages, North China Electric Power University, Beijing, China.
[b]School of Foreign Languages, North China Electric Power University, Beijing, China.
*Corresponding author.

Supported by the Fundamental Research Funds for the Central Universities of Ministry of Education of China (2017 MS086)

Received 16 May 2018; accepted 5 August 2018
Published online 26 August 2018

Abstract

Under the severe development background of the energy power corpus, this passage is meant to make efforts to do some preliminary study for it from the aspects of the criteria for corpus design, principles for the text selection and releasing platform for the corpus and its maintenance. Based on some successful experience of the copra such as BNC, this corpus would make some specific adjustments for its unique use. For practically using goal, this corpus aims to benefit the teachers and students, researchers, translators and people who have certain needs.

Key words: Energy power corpus; Corpus tagging; Corpus storage; Sharing platform

Zhang, Q., & Wang, C. Y. (2018). The Preliminary Study on the Construction of the Energy Power Corpus. *Studies in Literature and Language, 17*(1), 60-64. Available from: http://www.cscanada.net/index.php/sll/article/view/10526
DOI: http://dx.doi.org/10.3968/10526

1. INTRODUCTION

1.1 The Research and Development Background of the Corpus

Nowadays, the construction of the corpus brings people's research on language from the intuition—based to the corpus—based, which triggered a revolution in the language research field. (Haiyan, 2007) Established in the 1960s to 1970s, the Brown Corpus has collected one million words in American written English while the Lancaster-Oslo/Bergen Corpus (LOB) focus on the written English in Britain on the same scale; and the London—Lund Corpus of Spoken English (LLC) established in the 1980s is the first Spoken English Corpus which has collected half of one million words in its infancy. (Jianxin, 1996)When it comes to 1990s, the British National Corpus[BNC] has appeared whose capacity reached 100 million, and the International Corpus of English[ICE] even has collected the English corpus from all over the world more than 20 countries and districts which can be used English contrastive study beyond the borders, culture and language field. (Jianxin, The Talk about the Design and Content of the English Countries' Corpus, 1999) In China, researches on the English corpus linguistics started from the late1970s which was in step with the world. In addition, one of the most influential corpora is JDEST whose capacity has extended 4 million words in the charge of Huang Renjie and Yang Huizhong of Shanghai Jiao Tong University in the 20th century. (Huizhong, 2002)

What is more, the development of corpus grows rapidly since 1990s and its situations of research trend are as follow: (1) the large- scale construction of corpus with massive variety; (2) the deep research for corpus; (3) the corpus is used widely all kinds of fields related with languages. (Haiyan, The Future of the Legal Corpus Construction, 2007) It is not difficult any more to build a large-scale corpus but how to build a large-scale corpus characterized by the geography, style and even the text and a further exploration and study based on the current corpus. (He, 2001)

There are three different English corpus in domestic: Chinese Learner English Corpus, Parallel Corpus and Special English Corpus. (Jianjiang, 2014) This energy power corpus would contain two sub-corpora, one is the monolingual corpus which is the basis of the other corpus--

the parallel corpus. Parallel corpus is a bilingual corpus that consists of the original text and its corresponding translation. Both of the sub-corpora are targeted for practical usage like the teaching purpose and translation and academical usage like the contrastive linguistic study, some linguistic researches and so on. In China, Chinese-English bilingual corpus focuses on the different fields respectively in Institute of Computational Linguistics of Peking University, Harbin Institute of Technology, Foreign Language Teaching and Researching Press, Institute of the Language Application of the State Language Commission, Institute of Software Chinese Academy of Sciences, Institute of Automation, Chinese Academy of Sciences. (Feng, 2002)

1.2 The Current Development Situations of the Energy Power Corpus and the Objectives of the Corpus-Construction

Although we promote to develop corpus centering around some certain fields, and different subjects have different development conditions that fields in computer, law and something else are going well. However, with the rapid development of the corpus linguistics, there is not even a special corpus for the electricity which is the "blood of the industry". And specific purpose for the electricity even in the very famous corpus is rarer than hen's teeth. The British National Corpus is classified with seven categories. They are corpora about the spoken English, fiction, magazine, newspaper, academic source, non-academic source and miscellaneous source. And the first five categories are also included in the Corpus of Contemporary English. What is more, the categories even in the Lancaster-Oslo/Bergen Corpus are classified as the fiction, general texts, academic sources and news. And in respect of the papers published in Chinese journals or in the foreign documents, the number of the papers mainly about the electricity also makes up a small portion. The numbers in the chart 1 is enquired from the main database from home and abroad. There are several requirements to be listed for the paper numbers. Firstly, the researching with the keywords "Corpus Electricity" are achieved with the function search-within-results of the "Corpus" which is not as well as the "energy power corpus". Secondly we have to be aware of that the same passage may be embodied in the different journal databases that the actual number of the paper is smaller than the number below. Thirdly, no more explanations after the numbers mean that the keywords appear in the full text while not limited in the titles. In addition, the keywords like "Electricity Corpus" may not appear as a phrase.

Table 1
Data as of 2017/12/02

Key words of searching websites		Corpus	Electricity Corpus	Energy Power Corpus
Google Scholar	In Chinese（paper numbers）	About 145,000	About 2070	232
		About 10,600 (appearing in title)	6 (appearing in title)	0 (appearing in title)
	In English（paper numbers）	About 11,100	44	76
		312 (appearing in title)	0 (appearing in title)	0 (appearing in title)
Chinese National Knowledge Infrastructure（CNKI）	In Chinese（paper numbers）	52,259	5	0
		6144 (appearing in title)	0 (appearing in title)	0 (appearing in title)
	In English（paper numbers）	103,466	8	24
		4949 (appearing in title)	0 (appearing in title)	0 (appearing in title)
Medalink	In Chinese（paper numbers）	17,543	145	1
		8,883 (appearing in title)	6 (appearing in title)	1 (appearing in title) Ps: the keyword is part of a name of a college.
	In English（paper numbers）	248,743	3,153	184
		57,4696 (appearing in title)	946 (appearing in title)	1 (appearing in title)
WAN-FANG	In Chinese（paper numbers）	13,032	86	0
		6,058 (appearing in title)	5 (appearing in title)	0 (appearing in title)
	In English（paper numbers）	25,614	13	38
		4,415 (appearing in title)	0 (appearing in title)	0 (appearing in title)
CQVIP		16,536	43	0
		9,618 (appearing in title)	2 (appearing in title)	0 (appearing in title)
JSTOR		237,387	2,824	13,991
		4,167 (appearing in title)	0 (appearing in title)	0 (appearing in title)

In conclusion, there is not even one paper for the energy power corpus and several corpus with rich capacity concerns little about the energy power.

However, the demand for English of energy power would become bigger and bigger, especially for the implementation of the "One Belt, One Road" Initiative. And the corpus construction of the energy power would definitely have great value for the scientific research personnel who work in the universities, electric industry, research institutes and so on. And the final end is practically useful.

2. ENERGY POWER CORPUS DESIGN

2.1 Criteria for Corpus Design

The quality of the corpus results from its design which also leads to the quality of the research results. (Jianjiang, 2014) Therefore, it is important to decide the criteria for the corpus and its scope, text content, representative and balance must be included without questions.

2.1.1 The Representative of the Corpus

As Yang Huizhong mentioned that the representatives of the corpus influence the study which based on the corpus and the reliability and universality of conclusions about the study directly. (Yang & Wei, 2005) And its representatives can be reflected as the following aspects: (1)its capacity that the bigger the capacity is, the more representative the corpus is; (2)its length of the time that whether it has contained content varied from the very early time to the updating world or not; (3)its width of the text that whether it has contained the content to satisfy the needs for the different community.

2.1.2 The Openness of the Corpus

The system of the corpus itself is an open system rather than a close one. (Jia, Li, & Han, 2010) English for special purpose is applied for a certain industry that the words do not have various meanings like the universal English. Therefore, the real point for the energy power corpus is the openness rather than its scope. There are two reflections for this: (1)updating in time in order to have a good knowledge of the newest information and increase the capacity of the corpus. To be honest, it is to be required with the dramatic growth of the electricity demand; (2) combing the corpus. It needs much time to construct a corpus that the cooperation would save much time and improve the efficiency.

2.1.3 The Flexibility of the Corpus

Different from the corpus of the law, the energy power corpus does not need to pursue the scale of the corpus while the various legal activities leads to the dynamic principles in the law corpus. Making sure the reasonable portion of different text is enough to operate a study or draw a conclusion.

In addition, during the corpus construction of the energy power, pushing it forward in the path of the intelligence is necessary. It is one of the functions of the "Monitor corpus" to find out the new phenomenon and new change of the language application, and there is a dynamic element in COCA as a form of the monitoring corpus which make the new words can be included constantly and analyzed by the corresponding software which could find out and spot the new words. (Ren & Yang, 2011) Only keep pace with the time shall we not be thrown away by the time and it is true of the corpus, especially for the information explosion age.

2.2 Principles for the Text Selection

2.2.1 Corpus Collection

Corpus collection would directly affect the width of the resources. And as mentioned above, energy power corpus aims to serve both the academically and practically usage which means these resources have to be distributed in many aspects. Another element have to be defined is the time length. The earliest resources for the energy power corpus are collected from 1990s. Besides, there are two methods of the words' collection that one is from the reliable websites and the other one is from the offline sources that typed with hands or by means of the software of OCR from books or pictures. The specific resource distribution is illustrated in the following table.

Table 2

The source of corpus	Detailed information	Comments
Books	Some teaching materials, reference books by some experts and so on	Because books are not easy to take and refer to. It only counts a small portion that about 30%.
Websites	Google Scholar, Chinese National Knowledge Infrastructure（CNKI）,Medalink, WAN-FANG, CQVIP, JSTOR and so on.	Websites is the main resources for the corpus because of its convenience. And the portion would be 70%.

Nevertheless, the resources are included but not limited to the above ways, and the newest information acquired in an all-round way is also necessary. There is no doubt that the quality of the words should be at a high-level that has the value to refer to.

2.2.2 Corpus Processing

There is no doubt that there are many mistakes in the raw materials from the websites which exist in some pictures, tables, italic types, footnotes, some special notations and so on. And sometimes there are the gibberish and messy format. And these mistakes would not only influence the lexical analysis but the correctness of the collocation. Then the bilingual words need to be aligned. Alignment is a method that put the original language words and target language words in two different texts and the words have to be corresponded by sentence to sentence or paragraph to paragraph. (Lang, Li, & Fan, 2014)

Aligner embedded in the CAT and MS Office, especially the function of search-and–replace in the word are the primary tools. Above all, this process is in order to get a clean text.

2.2.3 Corpus Tagging

Corpus tagging is a way to reprocess the language resources that decides the degree of refinement. In other words, the material is analyzed more specific, the results would be more distinct. According to Leech(1993), there are seven principles in the corpus tagging: (1) tagging can be deleted to recover the original words; (2) tagging can be separated to be stored additionally; (3) the final user of the corpus is clear about the principles and meanings of the tagging; (4) it has to be illustrated in the operating documents that the tagger and the way he has applied; (5) it has to be stated that tagging is not perfect and it is a probably useful method; (6) a neutral model which accepted by most people has to be the choice as much as possible; (7) any tagging model cannot be considered as the head standard. (Leech, 1993) XML is the most popular technology choice for the corpus tagging nowadays which can allow users design his own tags. In accordance with the Leech's principles and the energy power's features, this corpus would make its own tags which is presented below:

Sentence Model=(title, author, source, keywords)

Document Model=(title author, source, sentence, keywords)

Above all, this tag is divided into the words, the sentences and the documents. Every word in the brackets is necessary.

2.2.4 Corpus Storage

Because of the target users' various demands, the energy power corpus would be designed to contain two sub-corpora with both of them covering the relevant content. And in order to store and update the corpus conveniently, every sub-corpus can collected materials from different channels based on each characteristics.

3. RELEASING PLATFORM FOR THE CORPUS AND MAINTENANCE.

3.1 Releasing Platform for the Corpus

It is recommended to use the current software instead of developing a certain corpus for the energy power corpus at the beginning. And there are some sharing software existed, such as Wordsmith Tool v4, Concordance v3, Monoconcord and so on. However, only the independent software or the websites can support and push the large--scale corpus going further. the ASP can be the ideal releasing platform which using many webpages and tools of code development. To be specific, every sub-corpus would have a certain number files which are distinguished by the different research demands and named by the principle of convenience. In order to have a more convenient research in the future, every file in the sub-corpus would be linked by the hyperlink.

3.2 A Sharing Platform

A resource-sharing platform not only saves the unnecessary same labour and the time but also convenients the researchers who would collect and analyse all the information they want. To help with the development of the corpus, a Web-based sharing platform is presented. Similar to many corpora, this corpus also intends to be achieved to combine the resources and corpus in some universities and the research institutes on account of the target users.

In the British National Corpus, listing some of the most notable corpora: Brigham Young University, BNCWeb at Lancaster University, BNCWeb at Oxford University, Intellitext(University of Leeds) and Phrases in English. However, the Lancaster University requires the users to registrate and the Oxford University is only for the Oxford University users. Besides, Corpus Linguistics at Beijing Foreign Studies University needs users to own his number and passwords. In order to disseminate the content and benefit more users, the Energy Power Corpus' main features are:

·a classifying system and its category;
·membership management;
·the security and patent;
·the researching and filtering technology.

The technology needs updating consistently in accordance with the market demands meanwhile this corpus is open to individuals or organizations.

And because of the issue of the copyright, the resources in the corpus can be divided into two categories, one is free and the other would charge according to the negotiation between the authors and the manager of the corpus in consideration of the allocation of the profits and rights. And the charging system would adopt the method of the membership model which has already taken by many libraries, institutes and so on.

After the construction of the energy power corpus, the maintenance personnel still play an quite important play. In view of the uniqueness of the energy power corpus, the best location for the corpus would be the electric power university. It is also necessary to choose a chief manager to take command of the corpus as well as lead a group which consists of the technicians, the administrative persons and personnel for corpus collection. And most of the employees would be chosen from the student candidates on account of the various institutes in universities and the low-cost labour. Furthermore, students would get more practical experience from this corpus work.

CONCLUSION

All in all, based upon the current trends in the world, energy power corpus would play a tremendous role. And its representative, openness and flexibility can make sure the scale, quality and the further development of this corpus while the principles for the text selection are all the most basic ones for any corpus since content-driven is the root way to promote and develop. In addition, sharing platform and the design of the link definitely is bound to improve users' experience.

In pace with the advocation of China that going out and One Belt and One Road Initiative, establishing an energy power corpus is necessary. Besides, different groups, like the teachers and students, researchers, translators and people have needs all can find what they expect in the corpus.

REFERENCES

Feng, Z. W. (2001). *The history and current situations of the Chinese corpus study*. *Journal of Chinese Language and Computing* (Singapore), *1*, 43-62.

He, A. P. (2001). *The Introduction*. *Study the language with the corpus* (pp.23-24). Thomas, J. & Short. Foreign Language Teaching and Research Press.

Jia, J. Z., Li, C. G., & Han, X. (2010). The construction of the internet corpus for the legal framework. *Information Studies: Theory & Application*, *33*(5), 88-91.

Lang, Q. Y., Li, H. S., & Fan, Q. S. (2014). The application of the electric English corpus in the electric majors. *Theory Research*, *11*, 205-206.

Leech, G. (1993). Corpus annotation schemes. *Literary and Linguistic Computing*, *8*.

Ren, W., & Yang, P. (2011). Step to the internationalization: The current development situation and trend of the Chinese interpretation study. *Chinese Translators Journal, 32*(01), 29-32.

Wang, J. X. (1996). *The introduction of the three current English corpus* (Vol.3, p.37). Foreign Language Teaching and Research Press.

Wang, J. X. (1999). *The talk about the design and content of the English countries' corpus*. *Special Issue of Foreign Language and Culture in Journal of PLA University of Foreign Languages, 6,* 44-46.

Wu, J. J. (2014). *The principles, procedures and methods of the energy power corpus* (Vol.5, pp.72-77). Hua Zhang Press.

Yang, H. Y. (2007). *The future of the legal corpus construction* (Vol.1). Beijing: China National Institute of Standardization.

Yang, H. Y. (2007). The future of the legal corpus construction. *Terminology Standardization & Information Technology,* (1), 40-43.

Yang, H. Z., & Wei, N. X. (2005). *The construction and study of the chinese learners' spoken English corpus*. Shanghai: Shanghai Foreign Language Education Press.

Yang, H. Z. (2002) *The introduction of the corpus linguistics*. Shanghai: Shanghai Foreign Language Education Press.