International Journal of English Linguistics; Vol. 8, No. 6; 2018
ISSN 1923-869X E-ISSN 1923-8703
Published by Canadian Center of Science and Education

First Exploration into the Feasibility of the Construction for Energy Power Corpus

Shuaicheng Tao[1], Qian Zhang[1] & Liangqiu Lyu[1]

[1] School of Foreign Languages, North China Electric Power University, Beijing, China

Correspondence: Shuaicheng Tao, School of Foreign Languages, North China Electric Power University, No.2 Beinong Road, Changping District, Beijing, China. Tel: 86-156-5284-9698. E-mail: sunshine1012@163.com

Received: August 1, 2018 Accepted: August 20, 2018 Online Published: August 22, 2018

doi:10.5539/ijel.v8n6p139 URL: https://doi.org/10.5539/ijel.v8n6p139

Abstract

With the rapid development of science and technology, people are more likely to resort to technological products to tackle new problems in life. Associated with some famous theories, corpus, as Professor Mona Baker's theory suggests, can be utilized in many areas and simplify the process of multilingual transformation. Since it has been rising for a certain period of time, corpus related to energy power has still not been built yet. Despite some potential problems to be solved in the actual exploration, this thesis aims to study on the feasibility of the construction and development for energy power corpus on the basis of the ways, tools, overall design and planning, etc. of the construction so as to make up the lack of data and provide more possibilities in the research field of energy power, and help to broaden the scope of corpus database for ease of more researches and findings in the future study.

Keywords: corpus, corpus linguistics, energy power, translation studies

1. Preface

Nowadays, with the rapid development of big database and artificial intelligence, more and more things which used to be impossible can serve to people's needs in modern society. In other words, we are now living in an unprecedented world surrounded by various technological means. Corpus, we can say, is a technological product which may date back to the middle and later periods of the 20th century. Professor Mona Baker, for the first time, suggests that there exist the universal features of translation based on corpus-based investigations of translated texts (Ma & Miao, 2009). As Maria Tymoczko put it as follows,

Corpus translation studies enable us, for example, to encode in compact and efficient forms, to access and interrogate vast quantities of data—more data than any single human being could ever manage to gather or examine in a productive lifetime without electronic assistance (1998).

For this reason, a large number of studies have applied corpus to various aspects due to the practical use of it since the late 90s, when Professor Mona Baker and Yang Huizhong both elaborated their findings as harbingers in or abroad.

2. Literature Review

The electric power industry is one of the fundamental and principal industries of the national economy. It is developing gradually into a huge industry serving the needs for the people and its country. The electric power experts and workers are busy making all kinds of breakthroughs in different aspects day and night. For example, people working in this industry mainly focus on power generation, power supply, power development and power reform.(Lyu, 2018; China Electricity Council, 2017) Meanwhile, they also care about non-fossil energy, energy conservation, electric energy replacement and issues on environmental protection (Lisin, Shuvalova, Volkova, & Strielkowski, 2018; Riva, Ahlborg, Hartvigsson, Pachauri, & Colombo, 2018; Li, Kang, & Gao, 2017; Li & Fan, 2018), which really make a difference to the long-term development of the industry. But it is the case that few studies have been made in the related aspect like corpus in this field as an output tool easing the burden of scholars' composing academic achievements in a different language.

There are some well-known corpora having been built around the world, among which are English-Norwegian parallel corpus, German-English parallel corpus of literature texts, multidisciplinary corpus of academic journal paper built by Ken Hyland, Babel Chinese-English parallel corpus by Professor Xiao Zhonghua, General

Chinese-English parallel corpus by Beijing Foreign Studies University, etc. At present, corpus, concerned with many specific fields of study, has sprung up like mushrooms after rain. Some scholars have expounded the idea and feasibility of building corpus in different areas, such as business English textbook corpus (You, 2016), aviation English corpus (Fu, 2011), TCM (traditional Chinese medicine) English corpus (Xue, 2004), etc.

3. Introduction to the Construction of the Energy Power Corpus

Energy power corpus, as it is literally, is the corpus built with the actual use of language and linguistic data related to energy power. Once in most cases categorized in the field of EST (English for science and technology), the energy and power industry is gaining a firm foothold and on its booming way. Thus, it is about the time to build an energy power corpus for the convenience of the scholars and workers in the relevant study areas.

3.1 Significance of the Study

According to Peter Newmark (2001), there are three main functions of language, the expressive, the informative and the vocative function. The informative text, which is also called content-focused text, as he put in *A Textbook of Translation*, "The format of an informative text is often standard: a textbook, a technical report, an article in a newspaper or a periodical, a scientific paper, a thesis, minutes or agenda of a meeting. (p. 40)" From the skopos theory, we know that "it is the purpose of the translation which determines the translation methods and strategies that a translator may adopt in order to produce a functionally adequate translation. (Ma & Miao, p. 81)"

The energy power corpus we are building includes targeting those new researchers, or the students pursuing further degree. Obviously, the purpose of their doing translation in the research with the aid of corpus is to spread the new ideas and findings and make themselves understood in a different language. The texts related to energy power are scientific, and basically in the informative type, as is in accordance with Newmark's categorization. Therefore, the building of the energy power corpus is to collect the informative texts, which are based more on content rather than the modifier, opposite to the so-called form-focused texts. Professor Hu Kaibao (2016) from Shanghai Jiaotong University once made some research on machine translation (MT). He put forward that MT can and should be taken advantage of to reduce the time and cost in translation. Both MT and human translation are complementary. Also, MT is rather suitable for the stylized and informative text, whose terms are relatively fixed, the meanings are clear and the repetitive rate of the sentence structure is relatively high. Generally speaking, corpus is one segment of MT. Supported by bilingual parallel corpus, MT is able to function that well. On account of this, the paper and texts concerned with energy power are quite appropriate to be put in corpus storage owing to the benefit brought by it.

In this area, few research and study can be found globally, while Lang, Li and Fan from Shenyang Institute of Engineering (Liaoning Province, China) have once made some exploration into the assumption of specialty English corpus for electric power in 2014, but no further progress has been made so far. Thus, we have good reason to build a corpus specialized in paper and texts related to energy power.

3.2 Methodology of the Research

From Baker's notion of corpus and her classification of corpus types, we are inspired to make a combination of corpus and translation studies in the field of energy and electric power. By accumulating a certain amount of related texts and performing a series of text processing steps, we make the energy power corpus, which is filled with actual use of both English and Chinese in the context right of this area, offering great convenience in retrieving the specified contents. Basically, it is another way of interpreting big data. It can help those who are willing to make some corpus-informed, corpus-based and corpus-driven researches when constructed.

3.3 Research Materials and Tools

Since the energy power corpus being constructed is aiming at doing a favor to those scholars and students in the field and industry, we tend to build it with the first hand materials and well-known articles published in the Science Citation Index (SCI). Our working group is located in North China Electric Power University in Beijing, China, which makes it convenient for us to acquire the corresponding materials.

With the development of science and technology, we inevitably need to resort to some electronic software living in the information age. It is a must in the corpus study besides the corpus material and texts. Generally speaking, we ought to clean the texts with text processing tools, mark the texts and metadata with annotation tools and search for data with concordancers and query tools. From this point, we select the following representative research tools, which are suitable for the needs and easy to access to.

3.3.1 Text Processor

Once the text is collected, it should be cleared up. As there exist some input mistakes and extra characters when

it is shifted from other material in the process of text collecting, the text should be cleared up in case of the later malfunction in the operation. Text processor is a nice application for free, with which the function of shifting from SBC case to DBC case, extra space elimination at any place, nonstandard characters rectification, etc. can be realized. Single text can be tackled, and batch processing is available. On top of this, EditPad Pro can also function as a text processor, which can serve the needs in dealing with the text.

For further operation, another tool is suggested. PowerGREP is the one that can fully check the visible characters in the text with the aid of regular expression. Compared with EditPad Pro and PowerGREP, text processor is much easier to get access to. (See Figure 1)

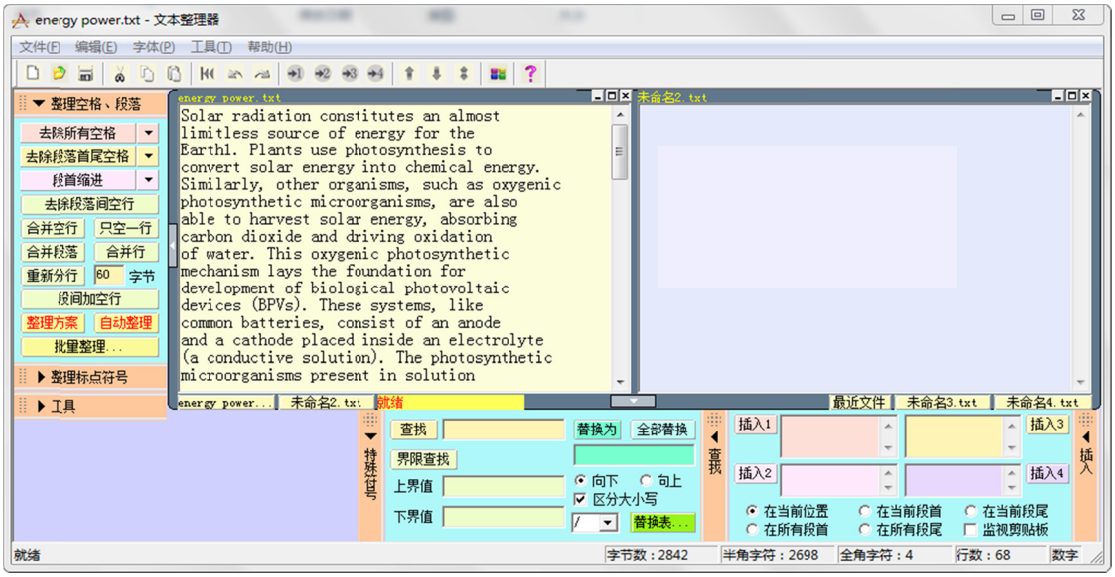

Figure 1. Main interface of the text processor

3.3.2 Annotation Tools

In order to set up condition and provide some ground for searching and analyzing, metadata will be annotated and part-of-speech tagging will be made. POS tagging is a way to show the grammatical features and structures in corpus study. TreeTagger is a multilingual version of automatic POS tagger program. It takes advantage of the markup language and tags every one of the word in the text for ease of retrieve and language processing. (See Figure 2)

In addition, the PowerGREP tool mentioned previously is also available for manual segmentation of words in batches in multiple texts and folders. It is a different form of realization for the operation.

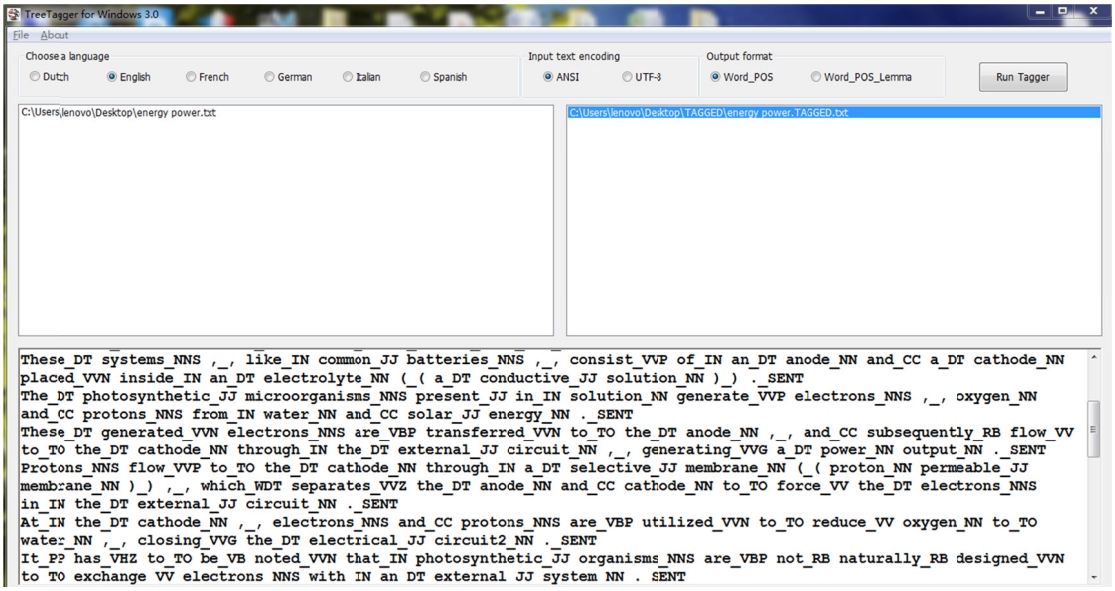

Figure 2. Main interface of TreeTagger

3.3.3 Concordancers and Query Tools

When the clearing and annotation work is done, the texts can be used for search and research with some tools. AntConc and PowerGREP are both concordancers and query tools. As some other functions of PowerGREP have been introduced, actually, it is a multifunctional tool in corpus study. AntConc is a freeware, multiplatform tool for carrying out corpus linguistics research and data-driven learning. It contains about seven useful tools as follows: a. Concordance tool; b. Concordance plot tool; c. File view tool; d. Clusters (N - Grams); e. Collocates; f. Word list; g. Keyword list. (See Figure 3)

From the retrieval results, data can be found out and analyzed in a visualized way. By observing a large batch of language materials and linguistic phenomena in actual use, the regular patterns can be drawn among them. And accordingly, that is the key of the corpus study.

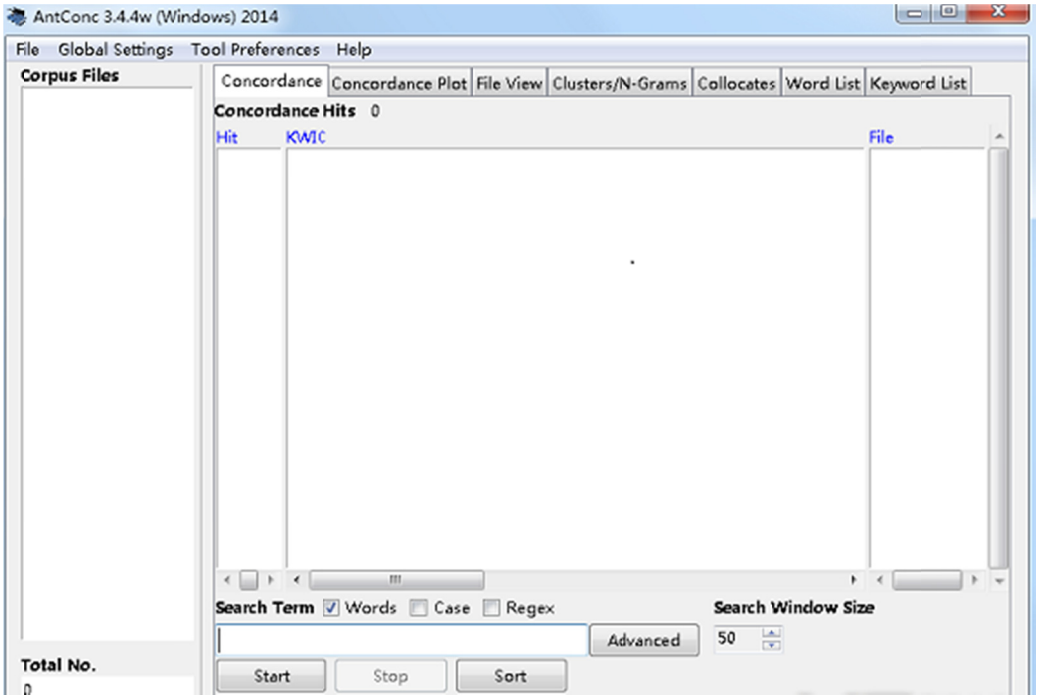

Figure 3. Main interface of AntConc

4. The Overall Designing of the Corpus

The construction of an energy power corpus really makes sense. The study actually begins in 2018, and may last for about two years as initially estimated. Located in a national university with energy power as its distinguishing feature, the working group consists of professors and teachers specialized in English linguistics and energy power. Some Master degree candidates are also involved in the group. The study will be carried out by the whole working group.

We know that professional literature on a specific subject related to science and technology is mostly composed of the specialized vocabulary as a major carrier. It is a way of information conveying rather than playing with words. In this regard, the core feature of this type of paper is that the sentence pattern is relatively monotonous, with few grammatical structures and a large number of buzz words and terminology being used. For the students who major in energy power, the corpus can provide them with guidance and direction in the process of their professional learning. By means of the tools in corpus study, the word frequency can easily be worked out, that is to say, the highest use of words list appearing in the professional writings. In addition, the matched terminology data bank and translation memories will be extracted out of the corpus data. Therefore, it makes it easier for students to grasp the core words and sentence patterns right in this area, and helps them save a large amount of time which can be utilized in the professional in-depth study. It also applies to the study for the scholars and workers of the same or related areas when doing research. They can work on exploring new findings and get them written on paper in English from the data of corpus for the convenience, especially for the work urgently needed on a large scale. In this way, scholars in and abroad can easily find out the academic achievements and what's going on in the academia. Energy power corpus, to some extent, plays a role of relieving the burden and promoting communication.

4.1 The Design Ideas

Now that a bilingual parallel corpus is what we are aiming to achieve, there is much need for us to make clear why bilingual and why make it a parallel one. Since the first bilingual parallel corpus founded in Canada, the Canadian Hansard Corpus has played some role in language study from all over the world. Some widely known corpora, take BNC (British National Corpus), a monolingual one for example. As "a 100 million-word collection of samples of written and spoken language from a wide range of sources", "BNC has, despite its large size, serious limitations as a translation aid if you are translating contemporary specialized texts." (Wilkinson, 2006) According to professor Wang Kefei and Liu Dingjia, compared with monolingual corpus, the English-Chinese parallel corpus not only includes the linguistic data of both languages, but also the interrelationship between English and Chinese in translation. Thus, when extracting information from the bilingual corpus data, we need to distinguish and extract the congruent relationship between lines of the two languages as well. (2017:4) Wang also pointed out that the relationship between the two languages in corpus and the comparison study on them is kind of natural. The corresponding materials make it the most reliable data for dictionary editors, especially in machine translation and natural language processing. A parallel corpus, if aligned, can provide empirical model for the system of machine translation based on the illustrative sentences and statistics. It also serves to supply validation to the rule-based machine translation, with a large amount of translation memory provided. (2012:23)

In view of these, professor Wang has been devoting himself to the design and construction of the China English-Chinese Parallel Corpus, a super-large-scale parallel one founded at home, so as to make deeper research and address some problems which small corpus cannot deal with. Also, the energy power corpus, to some extent, can act as a good complement to the former in a specific field, as a section of corpus of Language for Specific Purpose (LSP). Mr. Huang Libo (2017) generalized a research overview of the LSP corpus-based translation studies in and abroad in recent years and provided a summary of the characteristics in this area, pointing out that the scale of the corpus should be controlled in a certain amount so that the researcher can better rein the data in an easier way if the corpus is representative and balanced enough. Beyond that, we need to take two more factors into special consideration in building a specialized corpus, one being the specific purpose of construction and contextualization of the texts, the other being genre, text type, theme and variation of English. An open and dynamic medium sized corpus is the aim, through which retrieval and sharing is no longer a technological problem in the future accessing and developing process.

The bilingual corpus, when built, is available to be used in the future classroom. For example, for students of the related majors at university, an exploratory lesson can be set up to teach them how to use the corpus in hand to finish the research paper writing in a foreign language, like English. By learning the operations of software retrieval, every one of them can handle the language transformation if the corpus is built large enough. It helps the teachers in specialty English teaching for energy power, and teachers of the relevant courses to get the

students' papers of research findings published on international journals. It acts as a way to expedite the achievements. Thus, it is bound to help the whole industry to step forward and the energy power corpus can be part of the training of the personnel in this area in the future, helping them understand the words and terms in English.

4.2 Main General Principles

The corpus building will be at a steady and stable pace. For the language setting, it is intended to be a bilingual parallel corpus. In order to fulfill the target functions, two languages are needed and should be corresponding to each other. At the very beginning, a monolingual corpus of each of the two languages (Chinese and English) may get established beforehand to offer some kind of reference and guidance for the later bilingual one.

The capacity of the corpus is neither to be large nor to be small. In order to serve the needs of scientific research, the corpus can't be too small. As the first exploration, it is impractical to make it large enough. According to the present situation, we had better make it to a certain amount, then sum up the experience before scaling it up to a larger one.

4.3 Advantages in the Objective Environment

After years of research and development, corpus research has made some progress lately. Since many fields have set foot in corpus building and development, they have accumulated much experience for us to follow. In different areas, we have some similar ways, methods and procedures in collecting data, cleaning and marking texts, retrieving information, etc., which we can learn from and avoid the possibly alike mistakes.

Located in a national university that is co-built by State Grid Corporation of China and other six central enterprises in electric power, and is well-known for its professional achievements in electric power, our working group has more access to the faculty and staff, and even the scholars right in the relevant field. After years of teaching and scientific research practice, most of them are rather familiar with the corresponding English words, terms and sentence patterns used in international academic and research paper. With their help, the energy power corpus we are working on is supposed to be authentic, authoritative and useful.

With the technological means getting improved, a large volume of collected papers online are on the rise and available for our construction of corpus. It also serves the need for us to access to some top research findings in the industry under the guidance of the experts.

5. The Difficulties and Limitations in the Corpus Construction

As is mentioned above, at this time when everything seems to be in connection with corpus data, it is high time that energy power corpus should be involved in the big data analysis. In this regard, besides some advantages in the corpus building process that we can make use of, there really exist some difficulties for us to tackle and handle well in the trial exploration.

5.1 Copyright

First of all, the copyright of the data we are to collect should be given great attention due to the fact that most of the data have already been published, as universally exist in corpus construction. Since we mainly take it for academic purpose at the present time, we need to take account of the future use of it in the corpus building process and get the legal copyright when necessary.

Additionally, some of the paper data and documents selected to resort to are confidential or strictly confidential for a period of time, which makes it inaccessible to be put in storage. The copyright problem tops all the difficulties and also marks the limitation of the study.

5.2 Selection of the Linguistic Data

In the process of sorting out data, a whole standard is needed to make sure the selected data are accurate, authentic and normative expression of English. Then here comes the question. Should we sort out paper only from Science Citation Index? Then which country should we choose for the paper? Are there any boundaries in between? If we only allow SCI papers, what about those from other periodicals? Can we get the concrete and persuasive rule from the set data? Actually, from the preciseness point of view, it's a matter of delimitation, i.e. in which way can we get the most accurate and authentic usage and rules. At the very beginning, there are some necessities for us to demonstrate the sampling and verify the standard with a view to the design capacity, corpus sources and the balanced sampling.

As is elaborated in the studies of the corpus-based translation progress made in the recent 15 years (Wang & Huang, 2008), the energy power corpus also faces the problem of imbalance in terms of translation universal.

From the available data resources, the E-C data make an apparently larger proportion to the C-E ones, which is often the case. Basically, it is the result of a shortage of cooperative translators of both languages and native English translators. To tackle the problem, great efforts should be made to balance the data material and narrow the influence by widening the way data are collected. Qualitative and quantitative selection is also needed to be taken into consideration to make sure that the corpus covers contents of all varieties.

5.3 Categorization

Once the data is collected, the corpus needs to be categorized into different branches. Though the corpus is mainly about energy and power, it includes many small directions, which is hard to tell them apart from each other. Therefore, how to make scientific classification between different branches is a challenge to be faced with. As different people hold different views on the categorization, it needs to be negotiated further in case that any small research direction may be fallen into the wrong branch or even omitted.

5.4 Data Sharing

The magic of corpus lies in its big data and data sharing. When corpus is being built, its reliability and validity is to be made quite sure, making it of great value in the field. The realization of data sharing is another problem to solve when the corpus building task is finished. As is mentioned in 4.1, most data collected are under copyright protection, making it harder to share online or on other platform. In other words, if the collected data cannot be shared by the academia, its beauty and utility would be greatly discounted.

5.5 Updated

Things have been changing. The energy power corpus, when established, should be kept fresh and up to date. The linguistic data has its timeliness, and the scientific data and research methods also pass by along with the time. When all change with each passing day, the data in corpus should keep pace with the time.

In addition, as time goes by, some previous data cannot reflect the language trend in the current use. It may lead to inaccurate result if people go on using the outdated data. New materials should be involved in the corpus to get it updated and consistent with the reality. But the problem is how can it be kept renewed all the time？ If it is a must, how often should it be renewed? Failing to answer this question may result in the limitation of the study.

All of these difficulties and limitations in the corpus construction have to be taken into full account, seeking for satisfying answers in the actual operation.

6. Conclusion

According to Monzó (2003), being in translation makes the students “feel much more confident in translating. To see what others have done before provides them with patterns and solutions accepted by clients and the market.” By virtue of the corpus, whether parallel or comparable, monolingual or multilingual, students not only know the rules of the language system, but also recognize the features in translation itself. Hence, associated with energy power, the corpus as a tool can give full scope to the development of scientific research in this area. From the word frequency led out from the tools, we can easily conclude the features and tendencies of a language in a specific field. When the rule is formed, most of the difficulties in reading and writing can be overcome.

As Sinclair once said in 2003 in the International Conference of Corpus Linguistics that the progress for the construction of large corpora has been getting slowly, instead, a large number of small corpora are on the rise. When some certain supersized corpus is under construction, building more specialized and relatively small corpus will be a great trend in the future development of linguistics, which is what we are striving for.

In brief, the tools and techniques have laid a good foundation for corpus construction. The popularity of the network makes it available to share first hand international academic resources online. The experience from former scholars will get the energy power corpus going and give us the boost that we need. The members in this project group have years of experience in English translation related to energy power and higher research level of English for specific purposes and corpus linguistics. All of these give adequate feasibility to achieve a new form of data, energy power corpus. It is an extension of LSP corpus, and is bound to enrich the type of the corpus so as to play a more crucial role in the future study.

Acknowledgments

Supported by the Fundamental Research Funds for the Central Universities of Ministry of Education of China (2017 MS086). Thanks for the supervisor of this project Ms. Zhang who makes explorations along with me and gives me a lot of guidance in the paper writing. And I really appreciate my academic advisor Mr. Lyu for his professional suggestions and guidance for me. Finally, I’d like to thank those who have offered me personal assistance from the very beginning. Many thanks!

References

China Electricity Council. (2017). Annual Development Report on China's Electricity Industry Released. *State Grid*, (9), 16-17.

Fu, X. Y. (2011). Construction and Application of Aviation English Corpus. *Journal of Civil Aviation Flight University of China*, *22*(5), 52-55, 58. https://doi.org/10.3969/j.issn.1009-4288.2011.05.015

Hu, K. B., & Li, Y. (2016). Study on Features of Machine Translation and Its Relationship with Human Translation. *Chinese Translators Journal*, *37*(5), 10-14.

Huang, L. B. (2017). An Overview of Translation Studies Based on Corpus of LSP. *Journal of Beijing International Studies University*, *39*(2), 70-82, 136. https://doi.org/10.12002/j.bisu.042

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. https://doi.org/10.3969/j.issn.1002-2589.2014.11.097

Li, J., & Fan, H. X. (2018). Review of China's Development in Electric Power in 2017 and Vista for 2018. *Energy of China*, *40*(1), 14, 19-22. https://doi.org/10.3969/j.issn.1003-2355.2018.01.004

Li, J., Kang, X. W., & Gao, H. (2017). Review of China's Development in Electric Power in 2016 and Vista for 2017. *Energy of China*, *39*(3), 27-32. https://doi.org/10.3969/j.issn.1003-2355.2017.03.005

Liang, M. C., Li, W. Z., & Xu, J. J. (2010). *Using Corpora: A Practical Coursebook*. Beijing: Foreign Language Teaching and Research Press.

Lisin, E., Shuvalova, D., Volkova, I., & Strielkowski, W. (2018). Sustainable Development of Regional Power Systems and the Consumption of Electric Energy. *Sustainability*, (10), 1111. https://doi.org/10.3390/su10041111

Lyu, J. P. (2018). Exploration on the New Trend of Electric Power Development. *Science & Technology Industry Parks*, (10), 152.

Ma, H. J., & Miao, J. (2009). *Selected Readings of Contemporary Western Translation Theories*. Beijing: Foreign Language Teaching and Research Press (pp. 81, 356-358).

Monzó, E. (2003). Corpus-based teaching: The use of original and translated texts in the training of legal translators. *Translation Journal*, *4*(7), 1-5. Retrieved from http://accurapid.com/journal/26edu.htm

Newmark, P. (2001). *A Textbook of Translation*. Shanghai: Shanghai Foreign Language Education Press. (pp. 39-42).

Riva, F., Ahlborg, H., Hartvigsson, E., Pachauri, S., & Colombo, E. (2018). Electricity access and rural development: Review of complex socio-economic dynamics and casual diagrams for more appropriate energy modeling. *Energy for Sustainable Development*, (43), 203-223. https://doi.org/10.1016/j.esd.2018.02.003

Tymoczko, M. (1998). Computerized Corpora and the Future of Translation Studies. *Meta*, *43*(4), 652-660. https://doi.org/10.7202/004515ar

Wang, K. F. (2012). On the Design and Construction of the Super-large-scale China English-Chinese Parallel Corpus (CECPC). *Foreign Languages in China*, *9*(6), 23-27. https://doi.org/10.3969/j.issn.1672-9382.2012.06.005

Wang, K. F., & Huang, L. B. (2008). Corpus-based Translation Studies: Progress in Recent 15 Years. *Foreign Languages in China*, (6), 9-14.

Wang, K. F., & Huang, L. B. (2012). Construction and Application of Parallel Corpora: Issues and Comments. *Technology Enhanced Foreign Language Education*, (6), 3-10. https://doi.org/10.3969/j.issn.1001-5795.2012.06.001

Wang, K. F., & Liu, D. J. (2017). Concordance and Application of the Super-Sized English-Chinese Parallel Corpus: A Big Data Perspective. *Technology Enhanced Foreign Language Education*, (6), 3-11.

Wilkinson, M. (2006). Compiling Corpora for Use as Translation Resources. *Translation Journal*, *10*(1). Retrieved from http://translationjournal.net/journal/35corpus.htm

Xue, X. Y. (2004). Conception of Establishing English Corpus of Traditional Chinese Medicine. *Journal of Guangzhou University of Traditional Chinese Medicine*, (6), 482-485. https://doi.org/10.3969/j.issn.1007-3213.2004.06.022

You, J. G., & He, J. N. (2016). Constructing Business English Textbook Corpus for Dictionary-making. *Foreign Language Learning Theory and Practice*, (4), 59-64, 98.