E-ISSN: 2706-8927
P-ISSN: 2706-8919
www.allstudyjournal.com
IJAAS 2023; 5(12): 06-09
Received: 06-09-2023
Accepted: 15-10-2023

Chang Wei
School of Foreign Languages and Culture, Panzhihua University, Sichuan, China

Corresponding Author:
Chang Wei
School of Foreign Languages and Culture, Panzhihua University, Sichuan, China

The construction and application of vanadium-titanium academic English corpus

Chang Wei

DOI: https://doi.org/10.33545/27068919.2023.v5.i12a.1083

Abstract
Corpus construction is the basis for in-depth language research, and the standardized operation of the corpus construction process is a necessary condition to ensure the quality of the corpus. Taking the building process of Vanadium Titanium Academic English Corpus as an example, this paper explores the construction process and some ideas for standardized construction of the corpus, and puts forward corresponding solutions to common problems in the process of building the corpus, aiming at forming a mature database building scheme. The results show that the process of corpus construction provided in this paper is feasible.
Background: English for specific purposes (ESP) has unique structure and characteristics in terms of expression, idioms and textual rhetoric, and contains professional information in specific fields and disciplines. ESP corpus is not only an important object of corpus linguistics research, but also an important carrier of professional knowledge in specific industries. The development of science and technology increasingly requires language learning and research to be more integrated into the professional field. For a professional document, the largest information carrier is professional vocabulary, not ordinary vocabulary [1]. Therefore, the establishment of ESP English corpus is a necessary condition for the in-depth study of English for specific industries.
Subjects and Methods: In order to build a qualified Vanadium Titanium Academic English Corpus, a construction process was designed to build this corpus. The whole construction workflow was divided into four steps. The first step is to collect and download articles. This part determines the source, content quality and size of the corpus. The second step is to clear up the corpus. The purpose of this part is to discuss how to form a pure language text. The third step is the naming of the corpus text file. This part mainly provides convenience for future information retrieval through scientific naming. The fourth step is text annotation. Through the annotation of the key information of the text, the future retrieval becomes possible, and the text has added value. The whole four steps provide a common means for the construction of corpus.
Results: Corpus construction planning with clear planning, reasonable division of labor and scientific content arrangement can effectively realize the purpose of corpus construction.
Conclusions: The successful completion of corpus construction shows that this construction process is feasible. At the same time, it also shows that the specific construction operation under this process can ensure the successful realization of the objectives of each stage. This general corpus building method can be used to build various other types of corpora. The tagging of the corpus improves the research value of the corpus and makes the corpus widely used in scientific research, dictionary editing, teaching and other fields.

Keywords: Vanadium-titanium academic English, corpus, vanadium-titanium research

1. Introductions
With the continuous development and progress of modern economy and society, rare metal resources begin to play an important role in more and more fields. Vanadium is "the seasoning of contemporary industry" and titanium is known as "strategic metal", which is widely used in aerospace, chemical industry, steel, medical and other fields. To establish a high-quality vanadium titanium academic English Corpus (ESP) and carry out English language research on vanadium-titanium industry has become an important way to promote the academic development of vanadium-titanium industry and accelerate the spread of domestic vanadium-titanium research results to the world.

2. Process and Methods

2.1 Reasonable design for the feasibility of the project

ESP corpus includes English language knowledge, language knowledge and skills, as well as professional English and professional knowledge in a specific field [3]. Therefore, for language research in specific industry, the construction of ESP corpus is the first choice. At the same time, corpus builders must do self-evaluation to the objective conditions required in the construction and ensure the feasibility of the project implementation. For example: Whether the training work is well organized, whether the organization planning is reasonable, whether the division of labor is clear, and whether the source of materials is reliable. Once the project is implemented, any small adjustment will lead to a huge waste of workload.

The vanadium-titanium academic English corpus is established for the purpose of vanadium-titanium English research. The project has clear content orientation. We take English academic articles on vanadium and titanium materials as the content. Therefore, the construction of this corpus takes the ESP corpus as the basic principle and goal, and all the articles selected for the corpus must be research on vanadium and titanium materials.

2.2 Construction scale

With the development and application of computer technology, the construction of corpus becomes much easier than before, and the scale of corpus is getting larger and larger. Corpora with tens of millions or even hundreds of millions of word scale have been established, such as brown corpus, COBUILD corpus, Longman corpus, etc. Therefore, the construction scale of corpus should not be too small. However, corpus size is not the only standard to measure the quality of the corpus. The size planning of ESP corpus should depend on the richness of text resources available in the research field.

Based on the timeliness of vanadium and titanium research and the richness of texts, the construction of vanadium-titanium academic English corpus takes in English academic papers published internationally in recent three years, which determines the construction scale of 2 million words.

2.3 The requirements on content balance

The quality of corpus is an important prerequisite for the accuracy of corpus-based research results. A corpus that can comprehensively reflect the true face of language use must ensure the inner content balance of the corpus while pursuing the amount of content. Although it is difficult to truly achieve the internal balance of the corpus, according to the core "vanadium" and "titanium" of this vanadium-titanium academic English corpus, it is found that a certain degree of balance can be achieved in the research involving vanadium and titanium. The specific ideas are: For the vanadium-titanium academic English corpus, vanadium and titanium belong to two research directions. Therefore, firstly 200 articles focusing on the study of vanadium and titanium respectively are scheduled, keeping a balance between vanadium and titanium. Secondly, on the selection of the research content of the article, it is found that the research topics around vanadium, titanium, vanadium oxide, titanium oxide, vanadium steel and titanium steel are relatively common. Therefore, on the selection of the article content, totally 400 articles are scheduled, including 100 articles with vanadium and vanadium oxide as the key words, 100 articles with titanium and titanium oxide as the key words, 100 articles with vanadium steel as the key words and 100 articles with titanium steel as the key words. This fully ensures that the content of vanadium and titanium research has a certain balance in the corpus.

2.4 The source and quantity of corpus

At the beginning of the construction of the corpus, the overall planning of the source and quantity of the corpus is very important, which is not only the guarantee of the quality of the corpus content, but also the key to ensure the research quality based on corpus. Corpus requires sufficient amount of corpus text, but it cannot be a random accumulation of language text materials, let alone text materials made up by people. Research results based on small number of source text materials may not be convincing.

The source of vanadium-titanium academic English Corpus is mainly the collection of electronic PDF of academic papers published on the world-famous journal websites Web of Science and Science Direct, both of which can provide ideal articles in terms of quantity and quality. The author confirms articles provided by the webs for free with the search keywords Vanadium, titanium, Vanadium oxide, titanium oxide, Vanadium steel, and titanium respectively, and then selects and downloads the papers mainly published in the latest three years. A total of 400 articles are processed to form vanadium-titanium academic English Corpus.

Table 1: Corpus content and text name coding list

Serial number	Keyword	Code	Text name encoding format	Quantity
1	Vanadium/Vanadium oxide	VVT	VVT + year + period / month + natural number	100
2	Titanium/Titanium oxide	TVT	TVT + year + period / month + natural number	100
3	Vanadium steel	VS	VS+ year + period / month + natural number	100
4	Titanium steel	TS	TS + year + period / month + natural number	100

2.5 Corpus Collection and cleaning

2.5.1 Collection and download

Modern computer technology and abundant network resources make it easy to obtain text materials for corpus. At present, the most common method is to use crawler tools, such as Python, Houyi collector, etc. With the help of crawler tools, it becomes more convenient to obtain text data in specific fields on the Internet.

The collection work is done through keyword retrieval on ScienceDirect website, then use crawler tool to grab all the required article names and their DOI numbers in batches, and finally manually download PDF documents one by one on the site Web of Science.

2.5.2 Document cleaning

It is an important principle for the content of non-multimodal corpus to retain text information as much as possible and remove non-verbal text elements that will interfere with later text retrieval. Generally speaking, the more nonverbal text elements contained in the source text,

the more complex the cleaning process is. Non-text elements and redundant format information, such as: formulas, equations, charts, watermarks, footnote (or endnote) serial numbers, headers and footers, spelling errors, etc. are basically unable to be identified and processed automatically by computers in batches, and can only be identified and processed manually. The common tools used in this step are: PDF editor, Editplus, Powergrep, etc.

The articles downloaded for the vanadium-titanium academic English corpus are all in PDF format. In order to facilitate text cleaning, we use PDF editor software to convert PDF files into word format in batches, and conduct manual text cleaning on the basis of word format. In this study, the symbol <table> is inserted in the corresponding position to replace the table information after a table is deleted in the corpus. Photos, tables, formulas and professional terms are replaced by <picture>, <chart>, <equation> and <professional term> respectively at the same place. After all the word files are cleaned up, they are converted into TXT format in batch. The main purpose of the cleaning work is to make the text become clean plain text, which is convenient for later text annotation.

2.6 Naming and organization of corpus

The regularity and scientificity of text naming is one of the standards to measure the quality of corpus. In view of the fact that the corpus text is different in source, category, time and other information, in order to facilitate future research and make researchers clearly grasp the basic information of each text in the corpus. The text should be scientifically named on the basis of following principles: First, the name should be as short as possible, try to control it within eight characters due to some analysis software limits the length of the file name. Second, it is named after the combination of letters and numbers. Letters are used to bear meta information such as corpus category and subject, and numbers can reflect the number of text. Finally, avoid the use of special symbols. Special symbols are easy to be displayed incorrectly in some analysis software, and it is also easy to cause interpretation ambiguity. The text organization of the whole corpus should be clearly classified according to the hierarchical structure.

Based on the above principles, the naming of vanadium-titanium academic English corpus has formed the naming rule of "category + year + period/month + natural number". For example: The article published in May 2019 under the category of vanadium and vanadium oxide will be named as VVT-201905. Due to there will be more than one articles in this category within this time, a natural number is added after the file name. For example: the name VVT-201905-01 refers to the first of many articles published in May 2019 under the category of vanadium and vanadium oxide.

2.7 Text annotation

Text information annotation refers to the annotation of the basic information of the text at the beginning of the text, such as background information, source, topic or type, time, etc. Corpus tagging has made a key contribution to the application of corpus, enriched the corpus information as a language resource for future research and development [4]. All the annotation information in this corpus consists of two parts, all of which are placed in angle brackets (see Table 2). The annotation position of header information is placed at the beginning of the text document. The annotation adopts XML language, and the information annotation appears in pairs, using XML format (example: <Author></Author>).

Table 2: Annotation information and meaning

Annotation Information	Meaning
<CORPUS_NAME> </CORPUS_NAME>	Corpus name
<ARTICLE_TITLE> </ARTICLE_TITLE>	Article title
<JOURNAL_TITLE> </JOURNAL_TITLE>	Journal name
<VOLUME> </VOLUME>	Volume number
<ISSUE> </ISSUE>	Issue number
<DOI> </DOI>	DOI number
<PUBLICATION_YEAR> </PUBLICATION_YEAR>	Year of publication
<AUTHOR> </AUTHOR>	Author
<KEY_WORDS> </ KEY_WORDS >	Key word
<ABSTRACT> </ABSTRACT>	Abstract

The richer the marked information of the corpus, the greater the application value of the corpus [5]. Therefore, the annotation information should be as comprehensive and complete as possible. For the absence of some annotation information, the measure taken in this study is to label it as missing, such as "<DOI>Missing </DOI> ". After all the annotation work is completed, the word text is converted into TXT text in batches, and stored according to the text organization to form a corpus.

3. Problems in the construction of vanadium-titanium academic English special corpus (ESP)

The construction of corpus is a systematic project, which involves not only the reasonable selection and arrangement of existing corpus resources, but also the consideration of future research needs. On each step of all the processing work, different construction planners often make different decisions based on objective conditions and their own knowledge and judgment. This decision will bring a series of problems to corpus construction.

There are three main problems in the construction of vanadium-titanium academic English Corpus: First, as far as vanadium-titanium academic English is concerned, there is no public corpus available at present, so the construction work is lack of reference in the construction process. The selection on article and text cleaning are in exploratory stage. In the process of text cleaning, there is no absolute certainty about what information can be deleted and what information cannot be deleted. Text can only be cleaned under conventional understanding, which may result in the loss of some important information. Second, the research fields and contents with vanadium and titanium as the key words are extremely extensive, and the industry span is also very large. Strictly speaking, it is necessary to further refine and classify the corpus, and it will be more scientific to form an industry-specific vanadium- titanium research English academic corpus. However, this may cause the problem of insufficient sources of research materials on some industries. Third, due to the different difficulties in obtaining and processing corpus and the imbalance of language research in each industry, there is also a great imbalance in the construction of special corpus [6]. The follow-up problems caused by this imbalance often seriously interfere with the results of research. As far as the vanadium-titanium academic English corpus is concerned, the annotation content cannot cover all meta information.

It's possible that a more optimized and detailed annotation method will be adopted in the future in order to seek a deeper research.

On the whole, the construction of English corpus is still in extensive construction stage. There is a lack of unified norms and standards in the process of ESP corpus construction, especially the guiding standards from industry professionals. This will easily lead to the loss of information that may be valuable in the future in the process of building the corpus. Cross industry in-depth exchanges and cooperation among various disciplines can not only achieve the high-quality corpus construction, but also achieve the ultimate goal of resource sharing.

4. Application value of vanadium-titanium academic English Corpus

4.1 Providing reference information for the academic paper writing on vanadium-titanium research

The real language materials and contextual information provided by the corpus can provide learners with a variety of real materials related to writing topics, which not only makes up for the lack of language input in classroom teaching, but also provides a basis for learners' writing[7]. The research on vanadium and titanium industry is still relatively weak in China, and there is a certain gap with developed countries in terms of the quality and quantity of papers. For domestic scholars whose mother tongue is Chinese, it is extremely difficult to write high-quality English academic articles. In order to reach the international industry writing standard, we must learn and master the writing rules and general expressions of academic papers in the same industry. The corpus can enable vanadium-titanium industry scholars to retrieve the required information from the corpus, which can not only improve the cultivation of language accuracy for researchers, but also provide language reference for researchers from a professional perspective, which greatly saves the task of consulting materials in the writing process. The amount of targeted information increases in the retrieval process, laying the foundation for in-depth research.

4.2 Serving the compilation of bilingual core-word dictionary of vanadium- titanium academic English

Corpus based lexicography technology has become the mainstream method of modern lexicography [8].An ideal bilingual dictionary should contain the words and phrases that readers want to find. An accurate translation should be provided for each word and effective and practical examples should be provided to reveal the aggregation and combination of vocabulary. Only a large-scale corpus can provide enough real cases and examples for compilation, which greatly reduces the workload of lexicographers. The practical application value of ESP corpus is mainly reflected in the research of professional vocabulary [9]. The content of vanadium-titanium academic English corpus can provide real sentences and actual definition and usage of core words in bilingual dictionary of vanadium titanium academic English, and objectively reflect the research trends, research frontiers and research results related to vanadium and titanium. Therefore, the vanadium-titanium academic English Corpus serves the compilation of vanadium-titanium academic English reference books, which is one of the most important application values of the corpus.

4.3 Providing assistance for vanadium-titanium English Teaching in Colleges

Corpus research has changed the traditional mode of language teaching and learning. Corpus plays an important role in language teaching and research, and has a great impact on what and how teachers teach. Vanadium-titanium English is an important part of professional English in Colleges. How to give full play to the role of corpus in professional English classroom teaching has been a topic of concern for professional English teachers in recent years. Compared with traditional classroom teaching, the teaching method of introducing corpus is more intuitive and visual, and it is also more convenient to find language rules that are not easy to find by traditional methods. At the same time, it provides a reference for improving teachers' traditional teaching methods and students' English writing habits.

5. References

1. Chen Mingyao. ESP and corpus construction. Foreign Language Research. 2000;2:60.
2. Li Fensheng, Ran Xianglin. Research on the development strategy of Sichuan vanadium and titanium industry under the new economic norm. Science and Technology Entrepreneurship. 2019;32(12):32-35.
3. Jiang Yunlei, Kang Guangming. Construction and application of multimodal corpus based on ESP. Journal of Wuhan Vocational and Technical College. 2017;06:53.
4. Leech G. Introducing corpus annotation. In: Garside R, *et al*., editors; c1997. p. 1-18.
5. Li Wenzhong. Corpus marking and annotation: Taking the Chinese English corpus as an example. Foreign Language Teaching and Research. May 2012;340.
6. Dong Aihua. Construction, application, problems, and development trend of special purpose corpus. Journal of Beijing Institute of Printing. 2013;21(05):59-62 + 74.
7. Conrad S. Quantitative Corpus-based Research: Much More than Bean Counting. TESOL Quarterly. 2001;2.
8. Xiao Huayun, Chang Baobao. Serve as a retrieval platform for bilingual dictionary compilation. Computer Engineering and Application. 2005;15:117-119.
9. Hutchinson T, Waters A. English for Specific Purposes. Cambridge: Cambridge University Press; c1987.