A Research on the Linguistic Features of Scientific English Corpus Based on Java Web

Nan Li[1,2], Zhange Meng[3]

[1] Xingtai University, Xingtai 054001, China

[2]Communication University of China, Beijing 100024, China

[3] Xingtai University, Xingtai054001, China

Abstract

Nowadays, countries are becoming ever closer in exchanges and cooperation. English as an international language is widely used in various fields, and the level of English learning has become an important measure of a man. It is of great practical significance to study the linguistic features of English in depth, which can provide important information technology support for Chinese English learners. However, influenced by a variety of subjective or objective factors, China's text corpus software is scarce and its effect is poor, which leads to the low level of linguistic research in China - even the most renowned English language scholars in China also need to use foreign corpus software for research. Unfortunately, these software applications are often incompatible, which has greatly restricted the development of English linguistics in China. Therefore, in this paper, based on Java Web, the author studies the linguistic features of scientific English corpus and puts forward ideas for the construction of scientific English corpus system, which aims to be used as a reference for the development of scientific English corpus in China.

Keywords: Java Web, Scientific English, Corpus.

1. SUMMARY OF RESEARCH

1.1 Background

Information technology is not only an important means of sustainable development of scientific research institutions, but also an inevitable trend of the development of human civilization. Compared with traditional research methods, information technology is more comprehensive and accurate, which also leads to the development of traditional research means in the direction of precision. Besides, information technology is faster in data processing, and also can be used for independent analysis and collection of the data obtained, the effect of which is beyond the reach of traditional research methods. Nowadays, the English corpus has been developed for decades, its theory and technology have been basically mature and it has made great progress in research of various types of English common features, which is an important foundation of English teaching and research focus. But for now, the construction of scientific English corpus is still to be improved, and the trend of English interdisciplinary development is becoming more and more obvious, which makes a big space for the demand for and development of scientific English corpus. At the same time, according to the records of related research, many research achievements in scientific English phrases have not officially recognized by the academic community; therefore, it is necessary to pay attention to the construction of scientific English corpus, so that it can play a proper role in scientific English teaching and research.

1.2 Literature review

In the research of linguistic features of scientific English corpus, the key aspects are as follows: firstly, the vocabulary of the same or similar meaning in scientific English and the different collocation features; secondly, the heterogeneity of the base form of notional word and its corresponding inflectional deformation in scientific English; thirdly, the reference of semantic frames of cardinal numerals in understanding of the heterogeneity of words and phrases collocation in scientific English; and fourthly, the characteristics of extemalization and simplification of the evaluative meaning of semantic prosodies in scientific English (Xing and Yang, 2014). The

construction of scientific English corpus is of great significance to promote the development of science English. It is mainly embodied in the following aspects:Firstly, the construction of scientific English corpus can contain more technical English vocabulary and provide phrases, interpretation, synonyms, antonyms, and practical examples of applications for scientific English vocabulary. In addition, because of the continuous development of technology, new words often emerges. The construction of scientific English corpus can also include more emerging words, e.g. botnet, which is a network of computers infected by a program that communicates with its creator in order to send unsolicited emails, attack websites, etc. Secondly, due to the inclusion of some vivid words or phrases with distinctive cultural characteristics, it is able to understand the cultural characteristics of the West timely in the process of scientific language learning, so as to enhance the level of scientific English learning. Thirdly, it can include terminologies of a number of areas, e.g. geoglewhack, which is using Google to find fewer results, or exactly one result. Fourthly, the establishment of a scientific English learning column for a certain type of vocabulary can help learners to better learn and organize relevant knowledge of scientific English. Fifthly, the construction of a column of analysis of meaning of words and grammar plays an important supporting role in the learning of scientific English. Sixthly, the construction of a column of collocation can clearly point out several common mistakes of collocation. Lastly, adding a column of linguistic varieties can help learners choose the right scientific English presentation in each context (Wei, 2014). At present, there are some shortcomings in the research of linguistic features of scientific English corpus:Firstly, it lacks necessary technical support for the technical English corpus and the theoretical basis. Secondly, the vocabulary retrieval and calculation is mainly based on the frequency and is relatively simple. Thirdly, the method of language research based on scientific English corpus is relatively ineffective and it is difficult to carry out comprehensive test. The results are not universally applicable. Fourthly, the scientific English corpus used is mainly based on foreign ones, and its devices are relatively backward and often incompatible (Wu and Lv, 2014).

2. OVERVIEW OF KEY THEORIES AND TECHNOLOGIES

2.1 Overview of Java Web

Java Web, the sum of a set of technologies for solving related problems in the Web via Java language, is mainly composed of two parts, JSP and Servlet. Among them, JSP technology is mainly responsible for front-end development to achieve a variety of front-end functions, and Servlet is mainly responsible for dealing with a series of background services (Liang and Li, 2011). In the development of a Web project, regardless of what language or frame is chosen, it is necessary to combine these two technologies to realize the important role of the Web. The Web structure is shown in Figure 1:

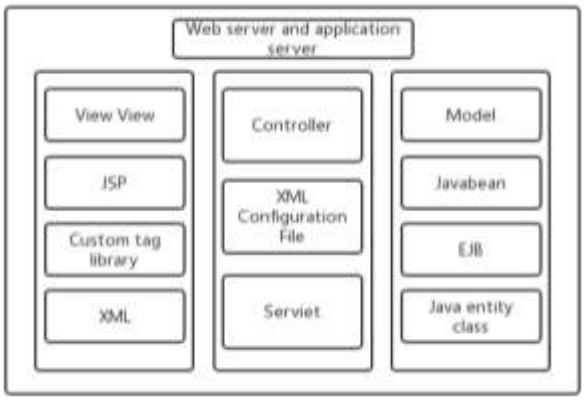

Figure 1. Web Server Structure Diagram

Moreover, the Web also has a 3-tier architecture, that is, the view - interface layer, the controller - business logic layer, and the model - data access layer (Liu, 2011). Under the influence of high cohesion and low coupling, the Web consists of different layers according to the hierarchical structure, which is an important structure of the Web.

2.1.1 Servlet

Servlet technology is mainly responsible for some front-end processing work. In the Web, the Servlet is mainly used to communicate with front-end users and respond to the needs of the client. Whenever the client sends an instruction to the Servlet, the service side of the Servlet will work according to the instruction and return the resulting processing results to the client to meet the client's initial request (Shi and Lu, 2015). This way is of high stability and security. Typically, there are only two situations where a service can not be responded to - one is the client request timeout caused by poor network conditions and the other is network disconnection. The specific work flow is shown in Figure 2:

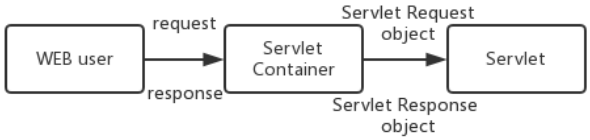

Figure 2. The Request Process of Graph 2:Servlet

The basic process is that when the Web user sends a request to the Servlet container, there will be different effects according to the different requests, e.g. the Servlet Request and Servlet Response, under different conditions, can produce different containers, and then the corresponding calculations and adjustments can be made according to the container. Next, the Servlet will store the processing results in a certain form in the relevant properties of Servlet Response. This stored content can be directly parsed by computer and used by the user (Wu, 2015).

The next is mainly the introduction of different effects produced by different technical standards, as shown in Table 1:

Table 1 Different Effects under Different Technical Standards.

Name	Description
HttpServletRequest	The Servlet container contains the HTTP request information In the HttpServletRequest object, the Servlet component comes from the Read user request data in Request object. In addition, HttpServletRequest can store Request range Shared data in.
HttpServletResponse	User generated HTTP response results.
HttpSession	The Servlet container for each HTTP session is created A HttpSession instance, HttpSession can exist Sharing data in session range.
ServletContext	The Servlet container creates one for each Web application ServletContext examples, ServletContext can Storing shared data of application.

2.1.2 JSP

JSP technology and php, NET and other principles have the same purpose, is a tool for developing Web based on HTML technology. Judging from the server side, the most important difference from other items is that JSP file is a single file, which does not only contain a variety of content of JSP itself but also HTML and other server-side code content (Yin, 2016). And in addition to the form of a single file, JSP technology can also be in the form of several files, that is to separate the contents of JSP itself from the contents of HTML and other server-side code contents. In this mode, JSP recognizes java codes and adds special tags before coding, so that the different parts of the work are clear and easy to distinguish (Chen, 2013).

Whenever the system gets information from a browser or user, each unit in the JSP container is activated and transformed into an open source file of Servlet. And what needs to pay special attention to is that if only sent a request, it will not result in code compiling, which will only start when the requests are sent continuously. And each time a request is sent, the system will self-examine the progress of code compiling and obtain new results

of compiling. If the results of compiling are usable, the computation can finish directly; otherwise, it continues until the results of compiling are usable. Figure 3 shows the JSP parsing process in container

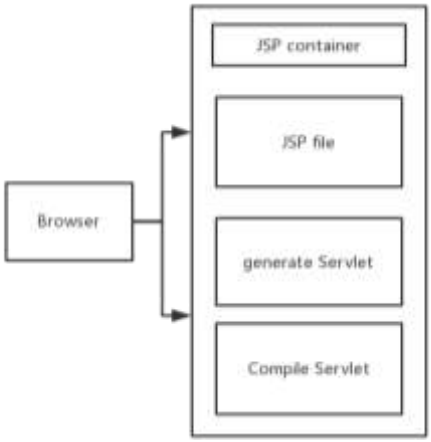

Figure 3. Analytical Process of Container to JSP

3. OPTIMAL DESIGN OF SCIENTIFIC ENGLISH CORPUS SYSTEM

3.1 Analysis of functional module requirements

The scientific English corpus system is an important functional system of linguistic analysis. It is of great significance to carry out in-depth research of this system, which can be applied in scientific English and promote the improvement of research on scientific English in China. The functional modules of the system are as follows:

The linguistic feature analysis system consists of 3 parts. Among them, the common function module mainly includes two functions. On the one hand, it is the word search including exact search and regular search. On the one hand, it is data analysis, which is to evaluate the MI score, MI3 score, Z score, density analysis, likelihood function score, T score, etc., which can be reflected in the black-and-white location map. The second part is the characteristic function module. Comparatively speaking, scientific English is of certain difficulty and complexity. As for the function module, the corpus provides substantive retrieval. At the same time, attention should be paid to the chart production and the body language, thus creating a good environment - that is special function module. The special function module mainly consists of three parts, one of which is segmentation - percentage segmentation and content segmentation. Besides, tagging and article upload are required for the special function module.

3.2 Analysis of authentication module requirements

3.2.1 Visitor role

Users need to log in to use a linguistic feature system. It can only provide visitors with a small amount of services, e.g. login interface or instructions, etc. Those who want to get more services must contact the administrator to get an account and password.The schematic diagram is shown in Figure 4:

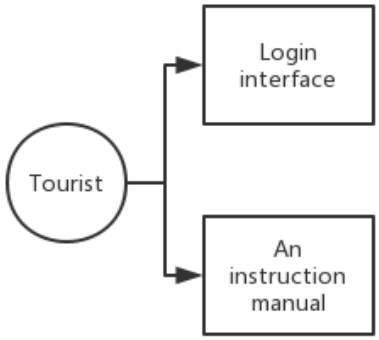

Figure 4. Schematic Diagram of Visitor Login Function

3.2.2 User role

After obtaining an account and password, users can use services and all features provided by the linguistic feature system. In this mode, any user who wants to change the account and password can contact the administrator. After login, the services available to users mainly contain three categories - the first category is common function. The common function is divided into two parts - the first is data analysis, including analysis of density, MI score, Z score, T score, NT3 score, likelihood function score, etc. Among them, the density analysis contains a black-and-white location map. The second is word search, including exact search and regular search. The second category is characteristic function, mainly consisting of consistency search, marks, keywords and so on. The third category is special function, including article upload, segmentation, tagging, etc., of which, segmentation includes content segmentation and percentage segmentation (Fang, 2014).

3.2.3 Administrator role

In the services of linguistic feature system, the role of administrator is to maintain the normal operation of the system rather than to use the system services. The administrator has the top authority, and his main job is to manage users' personal account and information. The schematic diagram is shown in Figure 5:

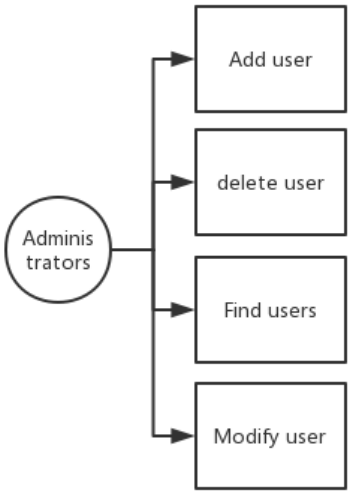

Figure 5. Administrator Authority Diagram

3.3 Analysis of system functional requirements

3.3.1 Exact search

Exact search is the fundamental function of a linguistic feature system, and its core function is to search the corresponding article or research factors for a certain keyword or several keywords. Through the analysis of the

needs of different users, it can be seen that the function of exact search is not limited to the search of a certain word, but also to search the word combinations based on the keyword. The amount of match varies greatly and the exact search can also play a fundamental role in other follow-up functions in the linguistic feature system (Huand Pang, 2015).

3.3.2 Regular search

Compared with exact search, the biggest difference of regular search is that no accurate word is required, and even the search of a prefix or suffix, etc. can also have rich results. This method can satisfy the study of various usage characteristics of English vocabulary. For example, if you want to search all the words that end with "-able", you can enter "\w+able" in the corresponding search box, then you will get all the words in the article ending with "-able" (Xiao, 2016). This method, in essence, is to search the characteristics of words, which is a more advanced search mode. According to the different research content, you can choose the corresponding method to get a better result (Liu, 2015).

3.3.3 Calculation of MI score

MI score, namely mutual information score, is used to describe the matching strength between different words. MI score is between 0 and 1, the higher the score, the higher the matching strength between the two words. The method of calculation is as follows:

$$I(a,b) = \log_2 \cdot \frac{P(a,b)}{P(a)\cdot P(b)\cdot 2s} = \log_2 \cdot \frac{F(a,b)\cdot W}{P(a)\cdot P(b)\cdot 2s} \quad (1)$$

Where W represents the total storage capacity of corpus, and W, S represents the span between words. In addition, MI score can also be obtained by the following formula:

$$I(a,b) = \log_2 \cdot \frac{W\cdot F(a,b)}{F(a)\cdot F(b)} \quad (2)$$

3.3.4 Calculation of Z score

Assuming that the total storage capacity of corpus is W, and for a certain word, its collocates appear in the corpus for C1 times, it can be seen that the frequency of the collocates appearing in each lexeme is $\frac{C_1}{\mathrm{W}}$(Chen, 2015). If the span is set as S, then the frequency of presence of collocates relative to the target word is $C_1 \cdot (2S+1)/W$. For a target word (N), the probability (P) of the presence of collocates is as follows:

$$P = \frac{C_1\cdot(2\mathrm{S}+1)}{W} \cdot \frac{N}{\mathrm{W}} \quad (3)$$

In addition, P can be represented in another way as follows:

$$SD = \sqrt{(2\mathrm{S}+1)\cdot N\cdot(1-\frac{C_1}{W})\cdot\frac{C_1}{W}} \quad (4)$$

After solving, the formula can be:

$$E = \frac{C_1\cdot(2\mathrm{S}+1)\cdot N}{W} \quad (5)$$

3.3.5 Special functions

In addition to the requirements above, a linguistic feature system should also meet a number of special requirements, such as content segmentation, that is, segmenting different content, abstract, introduction, material methods, conclusions, etc. in an article to make them research factors independent of each other, by which the learners can grasp the overall content of the article more accurately, so as to enhance the level of application. In addition, it is the function of article upload. A linguistic feature system must be supported by a strong corpus. In the context of the continuous development of society, many old resources have lost their value and become

encumbrance to the corpus. It requires to include more new materials to enhance the function of linguistic feature system, which can be done by the function of article upload (Chen, 2014).

ACKNOWLEDGMENTS

This work was supported by the Humanity and Social Science Youth Research Grants 2015 from Chinese Ministry of Education (Project: A Study on the Effective Building of Multimodal English Classroom Discourse in Primary and Secondary Schools; Grant No. 15YJC740140).

REFERENCES

Chen K.Q. (2014). On some linguistic features of English scientific and technological style, Shanghai science and technology translation, (03), 47-49.

Chen M. (2015). On grammar and stylistic features of scientific English, Science and technology information of China,(12), 251+253.

Chen Y.J. (2013). Research on the Skopos Theory of medical English translation based on the stylistic features of science and technology, Journal of Huaibei Normal University (philosophy and social sciences edition),34 (01), 153-156.

Fang F. (2014) Petroleum Technology Xu Wen Qiao Xiu, English stylistic features and translation strategies, Chinese translation for science and technology,27 (04), 7-10.

Hu J.Y., Pang K. (2015) Research on the features and translation of new English words in science and technology from the perspective of frame semantics, Journal of Northeast Agricultural University (social sciences edition), 13 (06), 78-82.

Liang W.H., Li X. (2011). Tourism Strategy Research of Macao University of science and technology - Based on the behavioral characteristics of English learning in Tourism Management Specialty, Professional English teaching optimization of tourism management,26 (02), 89-94.

Liu A.Q. (2015). Lexical features of EST and their translation strategies, Journal of Henan University of Urban Construction,(01), 79-81.

Liu J. (2011). Science and technology English syntactic features and translation, Science Journal (a),(03), 222-223.

Shi C., Lu H. (2015). Vocabulary features and translation principles of popular English for science and technology, Foreign language teaching in Shandong,(03), 80-83.

Wei M.F. (2014). Lexical features and translation of English technical terms, Chinese science and technology translation,27 (01), 5-7+23.

Wu J.L. (2015). Characteristics and translation of vocabulary and syntax in scientific English, Journal of Anhui University of Technology (social sciences edition),(06), 77-79+89.

Wu J.Y., Lv L. (2014). Syntactic features and translation techniques of long sentences in scientific English, Literature of Anhui (second half),(08), 53-54+64.

Xiao C.Y. (2016). Materials science and technology English language features and translation, Technology and its market analysis,23 (08), 238-240.

Xing F.F., Yang Y.C. (2014). Discussion on technical English features and translation skills of Chinese, Foreign language teaching, (01), 50-51.

Yin J.P. (2016). Conceptual metaphor: an effective way to achieve stylistic features of EST texts, Chinese science and technology translation, (02), 40-43.