Mining Lexical Co-occurrence Features Based Research On Construction Engineering English

Jinghui Ren*

1Department of Public Foreign Languages, Jilin University of Finance and Economics, Changchun City, Jilin Province, 130000, China

*Corresponding author’s e-mail: 327218871@163.com

Abstract. The application of Artificial Intelligence techniques on data mining as well as semantic studies has given rise to new discoveries over the research of lexicons in English for Specific Purposes (ESP). The present paper, assisted by the corpus software, carries out a data mining research on the lexical co-occurrence features and explores an empirical study of semantic preferences in Construction Engineering English. The semantic co-occurrence research has been conducted from three aspects involving lexical categories, lexical collocations and lexical colligations. With data retrieval and context analysis, the present research has found out that the lexical categories of Construction Engineering English are limited to seven fields, which are chemical substance, construction materials, structures, positions & measurement, natural environment, living environment and force. Besides, lexicons in different categories demonstrate distinctive semantic preferences and collocation features. And the semantic colligations of Construction Engineering lexicons are majorly neutral in semantic preference, which further proves the neutral and objective characteristics of technology articles.

1. Introduction

Traditional linguistic studies are overall considered to be conducted under the qualitative framework which is featured with individual experience, subjective impressions and limited textual analysis. However, with the development of interdisciplinary studies, computers and corpus have been introduced to the applied linguistics, which have provided quantitative approaches for data mining and semantic analysis, making it possible to explore linguistic features in a large scale objectively.

Kennedy [1] once defined corpus as “a body of written texts or transcribed speeches, serving as basis for textual analysis and descriptions.” In current research, the author has built monolingual corpora of Constructional Engineering English including four sub-corpus which are constructional materials and structures, plumbing engineering, environmental engineering and civil engineering. The tools in the corpus software, such as concordancer, frequency calculation, key word index and collocation, have assisted a massive-scale data mining in a liner sequence. The corpus will be adopted to elicit data in analyzing lexical co-occurrence relations of lexicons among the self-built corpora.

Lexicons are the basic component of a text and lexical collocations have been one of the ever-discussed hot spots in contemporary corpus-based linguistic studies. Just as Firth [2] once noted “you shall know a word by the company it keeps”, which has demonstrated the significant roles lexicons are playing in the coherence and cohesion of a text. Besides, John Sinclair, a world renowned expert of Corpus Linguistics, has discovered that “lexicons usually demonstrate certain co-occurrence patterns of co-occurrence relations with the words, related grammatical structures, semantic meanings as well
as pragmatic contexts, which generally form the ‘extended unit of meaning’ from the perspective of Corpus Linguistics.[3]” In other words, Sinclair’s “extended unit of meaning” is comprised by five parts, which are node words, collocations, colligation, semantic preference and semantic prosody.[4] To be specific, from the perspective of Corpus-based linguistic study, his theory is deemed as a model of meaning constructions, mining lexical features from four aspects, which is facilitated by the data mining and frequencies. The co-occurrence theory among different linguistic elements can be summarized as follows in Table 1. [5]

| Five Elements          | Interpretations                                      | Co-occurrence                       |
|------------------------|------------------------------------------------------|-------------------------------------|
| Node word              | Independent meaning unit                             | Key words                           |
| Collocation             | Company of words                                     | Between lexicons and lexicons       |
| Colligation             | Company of grammatical structures                    | Between lexicons and grammar        |
| Semantic Preference     | Company of collocations within the syntactic span    | Among lexicons, grammar and meanings |
| Semantic Prosody        | Pragmatic levels in context                           | Among lexicons, grammar and functions|

From the unit model in Table 1, it can be displayed that Sinclair’s “extended units of meaning” has integrated computational linguistics, phraseology, corpus linguistics and statistics, enriching the research paradigm and providing a data-based approach for lexical and semantic studies. In the current study, the research methodology of corpus and data statistics have been adopted to make an empirical study of lexical co-occurrence features in Construction Engineering English from three lexical levels so as to get a better understanding of semantic preferences.

The present paper constitutes five sections. The first part is the introduction to the topic and related theoretical backgrounds and the second part demonstrates the research design, exhibiting how computer technologies have assisted the current study in data mining and corpus construction. What’s more, the fourth part is the data-based analysis of semantic preferences in lexical co-occurrence from three aspects which are lexical categories, lexical collocations and lexical colligations, taking the corpora of Constructional Engineering English as the case study. The last section summarizes primary findings of the research.

2. Material and Methods

As is universally acknowledged, the construction of a corpus is supposed to meet the research demand. Based on the above principle, the author has constructed a self-built corpora involving four sub-corpus which are constructional materials and structures, plumbing engineering, environmental engineering and civil engineering. Just as the author in another published EI index paper Mining Textual-Feature Based Research On Translator’s styles noted it “the one with most information which can represent different cultural backgrounds is intended to be selected”[6]. In order to make the present corpora more comprehensive and representative which might cover primary semantic features of ESP lexicons, we have carefully selected wide ranges of materials from a series of college textbooks for Architecture published by the China Architecture and Building Press. After careful selection and collections, the corpora then have undergone several proceedings, such as electronic-version transformation, classification and tagging for parts of speech.

The corpus software utilized are WordSmith Tools 3.0 and AntConcordance 2.0 to help with data retrieval and data mining, among which functions like Concord, Wordlist and Keywords are mainly adopted. The overall scale of the present self-built corpora can achieve as many as 400,000 characters. Besides, the four sub-corpus are almost at the same level in lexical density, which are 37.1, 36.8, 37.3 and 34.5 respectively by mining their type-token ratios. The author has has drawn the research method adopted by He Anping [7] in his semantic studies on medical English, which is to extract core words or key words from the text and to analyze the co-occurrence features they have kept with the
surrounding collocations. The current research is supposed to explore the semantic preference from three aspects, which are lexical categories, lexical collocations and lexical colligations.

3. Results and Discussions

3.1. Lexical Categories
The present research is going to discover lexical categories based on the self-built Construction Engineering English Corpus. Based on high frequency words calculated by the computer software among the four sub-corpus involving constructional materials and structures, plumbing engineering, environmental engineering and civil engineering, the research has collected high frequency words, especially core lexicons, of each sub-corpus by excluding function words like at, with, the, over, in, to, on, etc. Among the high frequency lexicons, the research has sampled first 20 key words to infer semantic preferences so as to summarize main lexical categories of Construction Engineering English. After data retrieval and calculation by corpus, major semantic categories as well as their corresponding proportion of the present Construction Engineering Corpus could be elaborated in Table 2. As is demonstrated, seven basic categories have been concluded, including chemical substance, construction materials, structures, positions & measurement, natural environment, living environment and force. Obviously, lexicons belonging to the living environment takes up the highest proportion with 32%. Besides, the percentage of structures, position & measurement as well as natural environment are quite close to each other around 15%. What’s more, chemical substance and construction materials account for 10% and 9% correspondingly. Last but not least, force referring to depictions about various stress states occupies only 5%. The sample key words of each semantic category has been selected and listed in Table 3.

| Semantic Category      | %     | Definition                                                                                           |
|------------------------|-------|------------------------------------------------------------------------------------------------------|
| Chemical Substance     | 10%   | refers to any molecule including organics and inorganic substances                                   |
| Construction Material  | 9%    | refers to materials that can be used for civil engineering and architectural constructions            |
| Structure              | 16%   | Refers to all the components that can withstand forces for supporting buildings                      |
| Position & Measurement | 15%   | refers to directions, positions and dimensions                                                       |
| Natural Environment    | 13%   | refers to atmospheric physics phenomena related with seasons and temperatures that include wind, cloud, rain, snow, etc. |
| Living Environment     | 32%   | refers to the environment closely attached with humanities                                           |
| Force                  | 5%    | refers to depictions about various stress states                                                    |

Table 2. Basic semantic categories of Construction Engineering lexicons.

| Semantic Category      | Key words for illustration                                                                 |
|------------------------|---------------------------------------------------------------------------------------------|
| Chemical Substance     | flocculation, filtration, photocatalytic, cellulose, aldehyde, biological, coagulation, anionic, bioreactor, aluminium, organic, chloride |
| Construction Material  | fiber, rubber, plastics, steel, wood, cement,                                               |
3.2. Lexical Collocations

As is known, you shall know a word better by the company it keeps. The lexical collocation refers to the co-occurrence relationship between lexicons and lexical clusters centered upon key words found above. After locating key words, we have utilized the “collocates” function with the collocation span 5L to 5R and the collocation frequency is supposed to be no less than 10 times. The present research has taken “planning, administration, land, built, design, energy and heat” as sample key words. According to our study, the basic semantic collocations of the high frequent key words selected from the Construction Engineering English corpus could be generalized into seven semantic categories, which are attributes, meteorology, civil planning, direction&measurement, environment, structures and applications. To be specific, keys words belonging to the meteorology are usually collocated by lexicons related with weather and application. Key words belonging to civil planning are majorly accompanied by near all the seven categories. Besides, environmental key words like land are mainly semantically accompanied by modifications, planning and application. What’s more, it could be easily noticed that lexicons belonging to the same semantic categories might experience different semantic collocations and semantic preferences. We shall take “planning, administration and design” for illustrations. Though the above three words share certain connotations in design, the word “planning” is the core one compared with the other two lexicons which has the widest collocations in semantics like adjectives related with ecological, competent, heat, outer and frame. What comes next in semantic ranges is “design” and the word “administration” is the narrowest one in semantic collocations.

3.3. Lexical Colligations

In terms of the lexical colligation, it refers to the co-occurrence relationship between lexicons and grammar, which aims to explore semantic preference on the grammatical level. In order to observe the colligation features, the present research have selected several sample words with high frequencies whose T Scores are between 4 to 11. The sample lexicons involve planning, land, design and energy, and their corresponding colligation features are shown in Table 4. As can be seen, generally most of the lexicons in the context of Construction Engineering have show neutral semantic preferences, which demonstrates that the discourse styles and semantics of technology articles are to be neutral. Apart from those, the colligation structures of “planning + be supposed to” and “design + be done” reflect that the utilization of passive voices are quite frequent and typical in Construction Engineering
English, which further prove the neutral and objective characteristics of technology articles. What’s more, another semantic feature of lexical colligation in current study is the utilization of dashes “-” has been adopted to create compound words, thus the semantic meanings get connected. For illustration, the word “land-use” means the utilization of land resources and the word “energy-saving” denotes energy conservation.

Table 4. Colligation Features of Construction Engineering lexicons.

| Sample Words | T Score | Colligation Features |
|--------------|--------|---------------------|
| planning     | 10.7   | city/urban... + planning |
|              | 4.7    | planning + be supposed to |
| land         | 6.4    | land-use |
|              | 6.9    | the state/country...owned land |
| design       | 6.7    | the city/urban...+design |
|              | 4.2    | design+ be done |
| energy       | 4.3    | build/promote...+energy |
|              | 4.6    | energy-saving |

4. Conclusions
The present research has taken a quantitative study of semantic co-occurrence features of Construction Engineering English based on the self-built corpus which includes constructional materials and structures, plumbing engineering, environmental engineering and civil engineering from three aspects covering lexical categories, lexical collocations and lexical colligations. With data retrieval and context analysis, the study has figured out that the lexical categories of Construction Engineering English are limited to seven fields, which are chemical substance, construction materials, structures, positions & measurement, natural environment, living environment and force. Besides, lexicons in different categories demonstrate distinctive semantic collocations which include attributes, meteorology, civil planning, direction & measurement, environment, structures and applications. Apart from those, the research also has discovered that the semantic colligations of Construction Engineering lexicons are majorly neutral in semantic preference, which further proves the neutral and objective characteristics of technology articles.

Acknowledgments
The present research is sponsored by Social Science Foundation of Jilin Provincial Department of Education (grant No. [2017] JJKH20170145SK) and the Key Project of Jilin Higher Education Association (grant No. JGJX2017C101) . Sincere acknowledgement goes to the author of WordSmith Tools, Prof. Mike Scott from Aston University.

References
[1] Kennedy, G., 2000. An Introduction to Corpus Linguistics. Foreign Language Teaching and Research Press, Beijing.
[2] Firth, J. 1957. A synopsis of linguistic theory, 1930-1955. In J. Firth et al. (eds). Studies in Linguistic Analysis (Special Volume of the Philological Society). Blackwell, Oxford.
[3] Sinclair, J. and R. Carter. 2004. Trust the Test: Language, Corpus and Discourse. Routledge, London.
[4] Sinclair, J. Corpus, Concordance, Collocations. 1991. Oxford University Press, Oxford.
[5] Stubbs, M. 2001. Words and Phrases: Corpus Studies of Lexical Semantics. Blackwell, Oxford.
[6] Jinghui Ren and Lixin Wang. 2012. Mining Textual-Feature Based Research On Translator’s Styles. International Conference on Machine Learning and Cybernetics. Xi’an.
[7] He Anping. 2013. YULIAOKU DE DUANYU LINIAN JIQI JIAOXUE JIAGONG. Guangdong Higher Education Press, Guangdong.