CSCanada
Studies in Literature and Language
Vol. 13, No. 5, 2016, pp. 41-49
DOI:10.3968/9031

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

A Corpus-Based Study on Collocation of Technical Words in EST

WEI Jiejing[a],*

[a]School of Foreign Languages, Shanxi Normal University, Linfen, China.
Corresponding author.

Received 21 August 2016; accepted 17 October 2016
Published online 26 November 2016

Abstract
This paper adopts the corpus linguistic approach in the morphologically-based study of technical words of collocations in the EST. In this research, the writer extracts all the collocates of two pairs of technical words with reference to their base forms and inflectional plural forms; calculate their MI values; analyze their collocations behaviors and then conduct cluster analyses of all the collocates. The results show that each pair of the node words does not invite quite a number of the same collocates both in form and in meaning and the shared collocates are still identified as different cluster members based on their proximity, which shows heterogeneity of language and gets implications for EST teaching and lexicography.

Key words: Corpus; Technical words; EST; Collocation; Heterogeneity

Wei, J. J. (2016). A Corpus-Based Study on Collocation of Technical Words in EST. *Studies in Literature and Language, 13*(5), 41-49. Available from: http://www.cscanada.net/index.php/sll/article/view/9031
DOI: http://dx.doi.org/10.3968/9031

INTRODUCTION

Nowadays collocations are gradually beginning to occupy a more prominent place in linguistics and applied linguistics. In the flourishing area of corpus linguistics, the study of collocations has always been the central focus of researchers. However, it is found that the previous studies on collocation are mostly based on the general English corpus and the learner corpus, making comparison and contrast between authentic English and EFL (English as a foreign language) for the guidance of students in general English learning, while not too much work has been done on ESP (English for specific purposes).

In this paper, the writer focuses on the collocation of technical words in EST (English for Science and Technology), with the following considerations: Development in science and technology not only brings more convenience in daily life, but also influences language deeply. EST vocabulary can be classified into three categories according to its meaning and purpose. The three categories are technical words, semi-technical words and non-technical words respectively. The most obvious characteristics of technical words in EST are objectivity, accuracy and conciseness, so it is quite natural to say that EST with technical words has the linguistic features on its own compared with general English and other registers of English. Furthermore, the mushrooming growth of English vocabulary results from the rapid development of modern science and technology, and a large number of newly-coined English words are technical words related to science and technology. In EST teaching, although the teaching of vocabulary has been given weight, there seems to be lack of awareness of the importance of collocation. Collocation of technical words in EST has come to be seen as being of great significance in EST teaching and lexicography, particularly in production-orientated dictionaries.

In the present study, a corpus-based study is adopted, the writer uses the evidence from corpora to inform the analysis, refine the categories, resolve the ambiguities and descriptive problems, and provide a fund of authentic examples from which the writer can choose those that make the points of view with greatest force. The corpus used is JDEST, which is the first academic English corpus and has been commended by the famous British linguist J. Sinclair as “the pioneer” in specialized

corpus construction. In order to provide a reliable description of technical words collocation in EST texts, both qualitative analyses and quantitative measurements are adopted. The study will steadfastly follow the Item-Environment Method throughout, and use statistical software to get the collocational patterning and MI value of node words, and cluster analysis would be carried out, for the purpose of generalizing categories and functions according to the statistical results together with linguistic knowledge.

A distinguishing point about the current research from other ones in the field is worth noticing: it takes the base form and inflectional plural form of one lemma as a pair of node words for comparison and contrast study, which has not been studied deeply before.

The purpose of the present study is to explore the nature of vocabulary and its collocation in EST texts, to find out whether there exist different collocational behaviors between base form and plural form of one lemma and therefore the research itself will shed new light on further lexicography and EST teaching.

1. DEFINING COLLOCATION AND PREVIOUS STUDY

There are many differing views as to exactly what collocation entails and it is notoriously difficult to give a precise definition of the term "collocation" despite its frequent occurrence in the language.

A historical review shows that the word "collocation" was first used as a technical term by the British linguist J. R. Firth, along with the famous explanatory slogan: "you shall know a word by the company it keeps" (Firth, 1957, p.11), by which he means that a word receives its meaning from the word or words it co-occurs with.

Firth's definition and explanation of collocation offer linguists and researchers some important insights for understanding the subject. And based on his explanation of collocation, a large number of related definitions have been put forward with different focuses. Extensive investigation shows that collocations can be broadly defined as well as narrowly defined.

Linguists in broad sense approach, notably M. H. K. Halliday and J. Sinclair, propose to study collocation at the autonomous linguistic level of lexis. Halliday defines collocation in the following way (Halliday, 1976, p.25):

> Lexis seems to require the recognition merely of linear co-occurrence together with some measure of significant proximity, either a scale or at least a cutoff point. It is this syntagmatic relation which is referred to as "collocation"…

Similarly, Sinclair defines collocation as "the occurrence of two or more words within a short space of each other in text" (Sinclair, 1991, p.21). In their studies, co-occurrence is seen as the most significant criterion and collocation seems to be independent of other factors, such as whether there is grammatical relationship between the co-occurring words. He distinguishes two types of collocations, namely "significant" collocation and "casual" collocation, and sometimes reserves the term "collocation" for the former type. According to him, significant collocations are co-occurrences of words "such that they co-occur more often than their respective frequencies and the length of text in which they appear would predict" (Ibid, p.22).

Besides focusing on co-occurrence, another characteristic of this approach to collocation is its emphasis on naturally occurring data. They have developed a set of ideas and notions, including node, span and collocates, in connections with corpus-based collocational studies. According to J. Sinclair, the node word in a collocation is the one whose lexical behavior is under examination. It is normally presented with other words to the left and right and these are called collocates. And span is the measurement, in words, of the co-text of a word selected for the study. A span of –4, +4 means that four words on either side of the node word will be taken to be its relevant verbal environment. This is now well known as the Key Word in Center (KWIC) pattern.

The approach in narrow sense holds that collocation study without grammar restrictions is insufficient, and it is heavily influenced by grammatical structures.

Mitchell integrates lexis, syntax and meaning into the study of collocation, arguing for "the essential one-ness of grammar, lexis and meaning". He proposes that collocation is a recognizably association of "roots", which is defined by him the abstracted lexeme for the common elements of a scatter of words, and also concludes that "making grammatical generalizations concerning collocations is the most valuable approach" (Mitchell, 1975, p.108).

However, Mitchell's idea that collocation is of roots may suffer some weakness. Different word forms of a lexeme or a lemma, in many cases, may have different collocational behaviors. And perhaps his idea of "roots" is true to a certain extent.

Some analysts treat word forms which are inflectionally or derivationally related to each other as if they are instances of a single word family, known as a lexeme or a lemma (Kennedy, 2000, p.207). They study collocational behavior through lemmatization. Lemmatization is a process of classifying together all the identical or related forms of a word under a common headword, for instance, *go, gone, goes, went* can be classified under the headword *go* (Ibid.).

However, some linguists oppose to study collocation on the basis of lemmatization. Renouf (1986), Sinclair (1991), Stubbs (1996), Tognini-Bonelli (2001) and Zhang (2008, 2012) have all argued against conflating items sharing a common lemma on the grounds that each word has its own special collocational behavior. They believe that conflation often disguises collocational patterns.

Williams (1998) notes that in the context of molecular biology research papers the collocates of the word *gene* and those of the word *genes* are quite distinct, both prior and subsequent to the node word. Doyle (2003) likewise shows that there are few shared collocates between grammatically related forms of lemmas in scientific textbooks; he looks, for example, at *amplifier*, *amplifiers* (only three shared collocates), *circuit, circuits* (only two shared collocates), *frequency, frequencies* (only one shared collocates) and *shift, shifts* where he finds no shared collocates at all. Hoey (2005) proposes the definition of collocation that "it is a psychological association between words (rather than lemmas) up to four words apart and is evidenced by their occurrence together in corpora more often than is explicable in terms of random distribution" (Hoey, 2005, p.5). In his definition, it is obvious that he prefers to study collocation on words.

In this study, the writer shall follow the definition of Sinclair (1970, 1991) where the concept of "significant collocation" between two or more words occurs when those words occur more frequently than would normally be predicated from the type and length of text as the research is based on the huge quantities of corpus data. The idea that collocational analysis should be done on words rather than lemmas would be followed. Collocation of the base form and the inflectional plural form of a lemma would be made the comparison and contrast in the following study.

2. RESEARCH METHOD

The study is corpus-based and both quantitative and qualitative method is to be adopted to sort out the data. For the frequency of node words, the number of collocates, the MI value between node word and its collocates, the different clusters of collocates, quantitative approaches are adopted to ensure objectivity. As to dealing with collocations that are not significant, qualitative method is used, since this study is focused on the collocates that fall into the category of content words. Data retrieval and analysis are mainly performed by the software Wordsmith Tools 4.0 and DPS (Data Processing System).

2.1 Wordlist

The writer uses Wordsmith 4.0 to make a wordlist in JDEST corpus, according to which the higher-frequent technical words are picked out for detailed observation and comparison.

2.2 Concordances and Dispersion Plot

A concordance is a collection of the occurrences of a word-form, each in its own textual environment (Sinclair, 1991, p.32). Concordance is one of the most primary analytical methods in corpus research. With Wordsmith 4.0, the word form under examination appears in the center of each line, with extra space on either side of it (Ibid., p.33). In this study, the writer will use Dispersion Plot tool in concordance to see where the search word occurs most for selection of sub-corpus for further study.

2.3 Collocates and MI Value

After sub-corpora are selected for each word, concordance lines would be conducted again and collocates of the search word would be presented. Collocates are the words which occur in the neighborhood of the search word. The span in this study is 5, 5 (5 to left and 5 to right) and as for collocates here, the writer means the content words (nouns, verbs and adjectives etc.) in the span. The present research adopts the frequency of co-occurrence, MI value as the statistical measures. The frequency of co-occurrence refers to the number of occurrences of a certain node in a corpus; it is equal to the number of concordance lines that are extracted from the corpus. And MI value can ensure the representativeness and collocative significance of the collocates.

MI (Mutual Information) value can show the collocational strength between words. It calculates the probability of the two co-occurring words by comparing the product of their relative frequencies in the corpus with the observed frequencies of their co-occurrences. The higher the MI value is, the stronger the collocational strength between the two words is. Suppose two word forms *a* and *b* are searched, if *a* and *b* tend to occur in conjunction, their mutual information will be high. If they are not related and occur together only by chance, their mutual information will be very low. Finally, if the two events tend to "avoid" each other, their MI value will be negative (Oakes, 1998, pp.63-64).

2.4 Cluster Analysis

Cluster analysis is an exploratory data analysis for solving classification problems. Its object is to sort cases (people, things, events, etc.) into groups, or clusters, so that the degree of association is strong between members of the same cluster and weak between members of different clusters. Clustering is a method of unsupervised learning, and a common technique for statistical data analysis used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. In this study, collocates of each node would be classified into different clusters through DPS.

3. COLLOCATIONAL ANALYSIS OF TECHNICAL WORDS

Perform the research procedures mentioned above, and we collect five pairs of technical words occurring with high frequency in the JDEST corpus, with both base form and inflectional plural form of one lemma in each pair. Table 1 below gives the list of the five pairs. The frequencies of these words are also listed.

Table 1
Frequencies of Researched Words in JDEST

Base forms	Frequency	Plural forms	Frequency
Antibody	180	Antibodies	113
Gene	255	Genes	150
Laser	380	Lasers	157
Neutron	474	Neutrons	291
Virus	315	Viruses	92

In the following study, two of the five pairs would be treated as a node word to conduct collocational researches.

3.1 Extractions and Analysis of *Antibody* and *Antibodies*

Through dispersion plot in Wordsmith 4.0, the frequency of *antibody* and *antibodies* are listed respectively to make a contrast for selecting sub-corpus in JDSET to carry out collocational analysis. In the two figures listed below, *file* presents serial number of different sub-corpus in JDEST and *hits* shows frequency of the search word in corresponding sub-corpus. According to the two plots listed below, we choose t191.lib, t271.lib, t272.lib and t273.lib to carry on concordance research and cluster analysis. The texts in the four sub corpora are on biology and medicine.

N	File	Words	Hits	per 1,000	Dispersion	Plot
1	t191.lib	106,599	38	0.36	0.454	
2	t271.lib	114,883	81	0.71	0.414	
3	t272.lib	112,772	36	0.32	0.540	
4	t273.lib	120,310	25	0.21	0.502	

Figure 1
Dispersion Plot of *Antibody* in JDEST

N	File	Words	Hits	per 1,000	Dispersion	Plot
1	t191.lib	106,599	31	0.29	0.378	
2	t271.lib	114,883	39	0.34	0.430	
3	t272.lib	112,772	24	0.21	0.453	
4	t273.lib	120,310	19	0.16	0.469	

Figure 2
Dispersion Plot of *Antibodies* in JDEST

After selection of sub-corpora is made, concordance is conducted and collocates of *antibody* and *antibodies* are presented. As has mentioned in previous chapter, collocates in this paper are the words which occur in the neighborhood of the search word and the span is 5, 5 (5 to left and 5 to right). Following are the MI values of each collocate with *antibody* or *antibodies*. In order to ensure the representativeness and collocative significance, only the collocates with an MI value > 3 are listed.

Table 2 and Table 3 show the collocates of *antibody* and *antibodies* respectively. As has mentioned above, the collocates in this study are content words which have close relationship with the node word, so collocates belonging to functional words are not presented in the tables. From the two tables we can see that there is a tendency for nominal forms to dominate, which proves the domination of nouns in EST texts.

Table 2
Collocates of *Antibody* With MI Value>3

N	Word	MI	Total	Left	Right	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
1	Antibody										180					
2	Antigen	8.743	30	24	6	1	1	4	5	13	0	0	0	0	5	1
3	Specific	6.625	14	13	1	0	2	2	4	5	0	0	0	0	1	0
4	Levels	6.716	13	1	12	0	1	0	0	0	0	9	0	1	0	2
5	Binding	7.576	13	6	7	1	3	1	1	0	0	3	0	0	1	3
6	Complexes	8.806	9	0	9	0	0	0	0	0	0	8	0	0	1	0

To be continued

Continued

N	Word	MI	Total	Left	Right	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
7	Ferritin	9.690	8	5	3	1	1	0	0	3	0	1	2	0	0	0
8	Production	6.836	8	3	5	1	0	0	2	0	0	5	0	0	0	0
9	Anti	7.391	7	6	1	1	0	2	3	0	0	0	0	0	0	1
10	Serum	6.470	7	6	1	2	0	1	1	2	0	0	0	0	0	1
11	Used	5.041	7	2	5	1	0	0	1	0	0	2	1	1	1	0
12	Titers	10.391	6	0	6	0	0	0	0	0	0	6	0	0	0	0
13	Response	5.609	6	2	4	0	1	1	0	0	0	4	0	0	0	0
14	Reaction	6.899	6	2	4	0	1	1	0	0	0	2	0	0	1	1
15	Bound	7.618	6	5	1	0	0	1	3	1	0	1	0	0	0	0
16	Labelled	9.806	6	5	1	1	2	0	0	2	0	0	0	1	0	0
17	Conjugate	9.583	6	1	5	0	1	0	0	0	0	4	0	0	0	1
18	Ligand	8.098	5	4	1	1	0	2	1	0	0	0	1	0	0	0
19	Mediated	7.465	5	0	5	0	0	0	0	0	0	2	0	2	0	1
20	Neutralizing	10.391	5	5	0	0	0	0	1	4	0	0	0	0	0	0

Table 3
Collocates of *Antibodies* With MI Value > 3

N	Word	MI	Total	Left	Right	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
1	Antibodies										113					
2	Antigen	6.830	5	0	5	0	0	0	0	0	0	0	0	3	1	1
3	Antigens	8.290	6	2	4	1	0	0	1	0	0	0	0	1	2	1
4	Conjugated	10.840	6	6	0	1	1	0	1	3	0	0	0	0	0	0
5	Ferritin	10.169	7	6	1	1	0	1	4	0	0	0	0	0	0	1
6	Idiotypic	11.799	5	5	0	0	0	0	0	5	0	0	0	0	0	0
7	Levels	6.009	5	5	0	2	2	1	0	0	0	0	0	0	0	0
8	Monoclonal	10.741	6	6	0	0	0	0	0	6	0	0	0	0	0	0
9	Purified	9.062	6	4	2	0	0	1	2	1	0	0	0	1	0	1
10	Serum	6.920	6	6	0	2	1	0	1	2	0	0	0	0	0	0
11	Specific	6.659	9	6	3	0	1	2	0	3	0	1	0	1	1	0
12	Using	6.357	6	6	0	1	1	1	0	3	0	0	0	0	0	0
13	Anti	9.162	15	14	1	1	2	2	9	0	0	0	0	1	0	0

In EST texts, *antibody* belongs to the group of technical words, which has a unique meaning that is exclusive and explicit in a certain field. According to *Longman Dictionary Contemporary English* (2004, p.65), it means "a substance produced by your body to fight disease." After extraction and statistical analysis of collocates of *antibody* and *antibodies*, it is found that the collocates of each form within the span 5 are 19 and 12 respectively, while they share 6 of them. It can be inferred from the MI value that the collocational strength of the shared 6 words of *antibody* and *antibodies* are roughly the same, which indicates the collocational similarities

of the base form *antibody* and its inflectional plural form *antibodies*.

However, there exist 19 collocates that they do not share, showing existence of different collocational behaviors. Additionally, the shared 6 collocates, together with each node word's collocates fall into different clusters after cluster analysis is conducted.

As has been mentioned in Table 4, cluster analysis is the assignment of a set of observations into subsets (called clusters) so that observations in the same cluster are similar in some sense. In this study, collocates of each node word would be classified into different clusters through DPS according to their association with each other. Collocates of *antibody* and *antibodies* in different clusters are listed below in Table 4 and Table 5 respectively.

Table 4
Collocates of *Antibody* in Different Clusters

Cluster	Collocates of *antibody* in different clusters
1	Specific, Anti, Bound, Neutralizing
2	Levels, Binding, Complexes, Ferritin, Production, Serum, Used, Titers, Response, Reaction, Labelled, Conjugate, Ligand, Mediated
3	Antigen

Table 5
Collocates of *Antibodies* in Different Clusters

Cluster	Collocates of *antibodies* in different clusters
1	Idiotypic, Monoclonal
2	Antigen, Antigens, Conjugated, Ferritin, Levels, Purified, Serum, Specific, Using
3	Anti

From Table 4, we can see that collocates in the first cluster of a*ntibody* are semi-technical words and formal non-technical words. Collocates in the second cluster are mainly technical words and semi-technical words, and the third cluster is technical word. In contrast, in Table 5, technical words fall into the first cluster, and the second cluster are mainly technical words and non-technical words, while the third cluster is a semi-technical word. Further more, as for the 6 shared collocates, 3 of them fall into different clusters of the two node words respectively. The collocate *antigen* is in the third cluster of *antibody*, while it is in the second cluster when collocating with *antibodies*. When collocating with *antibody*, *specific* is in the first cluster, while it falls into the second cluster when collocating with *antibodies*. And the collocate *anti* is also different when it collocates with the two node words respectively. It falls into the first cluster with *antibody* but the third cluster with *antibodies*. Different associations of the collocates of each node word prove that there exist different collocational behaviors between *antibody* and *antibodies* though they are morphological forms of the same lemma.

3.2 Extractions and Analysis of *Virus* and *Viruses*

The extractions and analysis of *virus* and *viruses* are similar with that of *antibody* and *antibodies* demonstrated above. Through dispersion plot in Wordsmith 4.0, the frequency of *virus* and *viruses* are listed to make a contrast for selecting sub corpus in JDSET. According to Figure 3 and Figure 4 listed below, we choose t191.lib, t271.lib and t272.lib to carry on concordance research and cluster analysis. The texts in the three sub corpora are on biology and medicine.

N	File	Words	Hits	per 1,000	Dispersion	Plot
1	t171.lib	121,743	2	0.02	0.300	
2	t191.lib	106,599	77	0.72	0.345	
3	t201.lib	135,089	1		-0.069	
4	t251.lib	100,326	12	0.12	-0.069	
5	t271.lib	114,883	73	0.64	0.425	
6	t272.lib	112,772	132	1.17	0.443	
7	t273.lib	120,310	17	0.14	0.126	
8	t71.lib	110,253	1		-0.069	

Figure 3
Dispersion Plot of *Virus* in JDEST

N	File	Words	Hits	per 1,000	Dispersion	Plot
1	t191.lib	106,599	38	0.36	0.385	
2	t201.lib	135,089	1		-0.069	
3	t251.lib	100,326	13	0.13	0.022	
4	t271.lib	114,883	16	0.14	0.353	
5	t272.lib	112,772	20	0.18	0.508	
6	t273.lib	120,310	4	0.03	0.596	

Figure 4
Dispersion Plot of *Viruses* in JDEST

After selection of sub corpora is made, concordance is conducted and collocates of *virus* and *viruses* are presented. Following are the MI value > 3 of each collocates with *virus* or *viruses*. It is presented in Table 6 and Table 7 respectively.

Table 6
Collocates of *Virus* With MI Value > 3

N	Word	MI	Total	Left	Right	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
1	Virus										282					
2	Infection	6.839	22	7	15	2	3	1	1	0	0	12	0	1	2	0
3	Measles	8.845	20	20	0	1	0	0	0	19	0	0	0	0	0	0
4	Cells	4.260	14	6	8	0	2	2	2	0	0	0	2	1	4	1
5	Infections	7.151	10	5	5	2	1	2	0	0	0	3	1	0	1	0
6	Isolated	6.680	9	0	9	0	0	0	0	0	0	0	3	2	3	1
7	Cell	4.150	9	4	5	2	0	0	2	0	0	1	1	1	2	0
8	Disease	5.099	9	3	6	1	1	0	1	0	0	0	0	0	4	2
9	Mosaic	9.135	8	7	1	0	1	0	0	6	0	0	0	0	1	0
10	Infected	6.293	8	4	4	0	1	3	0	0	0	1	0	2	1	0
11	Plant	6.181	7	3	4	2	0	0	0	1	0	0	0	2	2	0
12	Strain	6.790	7	1	6	0	1	0	0	0	0	2	2	1	1	0
13	Antigen	5.720	7	1	6	1	0	0	0	0	0	4	2	0	0	0
14	Replication	7.430	6	3	3	0	0	3	0	0	0	2	0	0	0	1
15	Rubella	8.430	6	6	0	0	0	0	0	6	0	0	0	0	0	0
16	Vaccine	8.135	6	2	4	2	0	0	0	0	0	2	0	0	2	0
17	Tobacco	7.007	6	4	2	1	0	0	3	0	0	0	0	2	0	0
18	Norwalk	9.082	6	6	0	0	0	0	0	6	0	0	0	0	0	0
19	Human	5.038	6	5	1	2	1	0	2	0	0	0	0	0	0	1
20	Immunity	7.680	6	4	2	2	1	0	1	0	0	0	0	1	1	0
21	Type	4.940	5	3	2	0	1	1	0	1	0	1	0	0	1	0
22	Acute	6.151	5	4	1	1	0	0	3	0	0	0	0	1	0	0
23	HIV	7.582	5	3	2	2	0	0	0	1	0	2	0	0	0	0
24	Hepatitis	8.167	5	5	0	0	1	0	4	0	0	0	0	0	0	0
25	Spread	6.604	5	2	3	0	0	2	0	0	0	2	0	0	1	0
26	Related	5.323	5	4	1	0	0	3	0	1	0	0	0	0	1	0
27	Particles	7.456	5	0	5	0	0	0	0	0	0	5	0	0	0	0
28	Induced	5.819	5	1	4	0	1	0	0	0	0	4	0	0	0	0
29	Infectious	7.341	5	5	0	1	1	0	0	3	0	0	0	0	0	0

Table 7
Collocates of *Viruses* With MI Value > 3

N	Word	MI	Total	Left	Right	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
1	Viruses	74														
2	Plant	7.889	6	6	0	0	0	0	0	6	0	0	0	0	0	0
3	Agents	8.344	6	2	4	0	0	1	1	0	0	0	1	1	2	0
4	Use	6.377	5	4	1	0	2	0	2	0	0	0	1	0	0	0
5	Group	6.596	5	2	3	0	0	1	1	0	0	0	1	0	2	0

According to *Longman Dictionary Contemporary English* (2004, p.2218), *virus* is "a very small living thing, smaller than BACTERIA, that causes infectious illnesses." It is found that the collocates in the span 5 are 28 and 4 of the two node words, with only one collocate shared.

Table 8
Collocates of *Virus* in Different Clusters

Cluster	Collocates of virus in different clusters
1	Measles, Mosaic, Replication, Rubella, Norwalk, Type, Hepatitis, Spread, Related, Particles, Induced, Infectious
2	Cells, Infections, Isolated, Cell, Disease, Infected, Plantstrain, Antigen, Vaccine, Tobacco, Human, HIV, Immunity, Acute
3	Infection

As *viruses* only have four collocates here, cluster analysis is only conducted for collocates of *virus*. From Table 8, we can see that most collocates fall into the first cluster and the second cluster, and the proportion of technical words, semi-technical words and not-technical words are roughly the same, while the word in the third cluster is a technical word. In contrast, the four collocates of *viruses* are semi-technical words and non-technical words, without technical words. It is obvious that there exist different collocational behaviors between *virus* and *viruses*.

As space is limited, collocational analysis of the other three pairs are not listed. From the analysis of the five pairs, it is found that each pair of the node words does not invite quite a number of the same collocates and the shared collocates are still identified as different cluster members based on their proximity, which does not support their lemmatization to some extent.

CONCLUSION

Through large amount of authentic data from JDEST, we have found that there exist different collocational behaviors between different word forms of one lemma in EST, which does not support their lemmatization to some extent. After investigation of collocation of base form and inflectional plural form of one lemma and cluster analysis of these collocates, it is found that collocates of each form are not identical and they gather in different clusters, which show heterogeneity of language.

Although corpus linguistics has processed authentic language effectively and it is easy to find, sort and count items, either as a basis for linguistic description or for addressing language-related issues and problems with the increasing scale of corpus, it is found that the heterogeneity of human language is not concerned enough in this field. A linguistic unit consisting of multiple items may have different structural variations, which should not be neglected. And it is necessary to consider the heterogeneity as well as similarity in the study of language. For students of English for academic purposes, a balance of similarity and heterogeneity is of major importance in building their mental lexicons in a balanced way.

Modern lexicography is now firmly based on corpora as a source of words. Word frequency is an important factor in choosing which words into a dictionary and in isolating the senses of those words. However, frequency should not be the only criterion. Great importance should be attached to the study of collocation related to the register, and this needs to be reflected in dictionaries as well. It is hoped that through the study of collocation of register specific corpora a new light may be shed on the development of language, thereby assisting in the development of new dictionaries that will make full use of the organizing and storage capacities of the computer. Being sub language and genre specific, such dictionaries should fulfill a valuable pedagogic as well as a descriptive role.

In EST teaching, teachers should draw learners' attention to the language phenomenon, making students more aware of collocations, helping them understand that collocations are neither freely combined nor completely fixed, but to some degree arbitrary. Therefore, collocations have to be learned with specific attention. A good knowledge of collocation will be helpful for the learners with academic purposes to understand words and sentences in EST texts.

This study has been exploratory in nature and some difficult issues have not yet been tackled adequately. Due to the focus and the space of the study, it has to restrict the investigation to collocational behaviors of a manageable range of technical words. The characteristics of collocation in EST are far from being exhaustively

described. More work is needed in the area and it calls for a larger corpus and more extensive investigations, for it plays an important role in the study of English lexicon. The research on the very topic of this study is expected to be complemented by follow-up studies in the field.

REFERENCES

Biber, D., Conrad, S., & Reppen, R. (2000). *Corpus Linguistics*. Cambridge: Cambridge University Press.

Daskalovska, N. (2015). Corpus-based versus traditional learning of collocations. *Computer Assisted Language Learning,* (28), 130.

Doyle, P. G. (2003). *Replicating corpus linguistics: A corpus-driven investigation of lexical network in text* (Unpublished doctoral dissertation). Lancaster: University of Lancaster.

Fang, M. Z. (2011). *EST: Its paradigms and translation*. Beijing: National Defense Industry Press.

Firth, J. R. (1957). *Papers in Linguistics 1934-1951*. Oxford: Oxford University Press.

Halliday, M. A. K. (1976). *System and function in language: selected papers*. In G. Kress(Ed.). Oxford: Oxford University Press.

Hoey, M. (1991). *Patterns of lexis in text*. Oxford: Oxford University Press.

Hoey, M. (2005). *Lexical priming: A new theory of words and language*. London and New York: Routledge.

Kennedy, G. (2000). *An introduction to corpus Linguistics*. Beijing: Foreign Language Teaching and Research Press.

Longman Dictionary Contemporary English. (2004). Beijing: Foreign Language Teaching and Research Press.

Mitchell, T. F. (1975). *Principles of firthian Linguistics*. London: Longman.

Oakes, M. P. (1998). *Statistics for corpus Linguistics.* Edinburgh: Edinburgh University Press.

Sinclair, J. (1966). Beginning the study of lexis. In C. E. Bazell, J. C. Catford, M. A. K. Halliday, & R. H. Robins (Eds.), *In memory of J. R. Firth* (pp.410-430). London: Longman.

Sinclair, J. (1991). *Corpus, concordance, collocation.* Oxford: Oxford University Press.

Wei, N. X. (1999). *Towards defining collocations: A practical scheme for study of collocations in EAP texts* (Unpublished doctoral dissertation). Shanghai: Shanghai Jiaotong University.

Williams, G. C. (1998). Collocational networks: Interlocking patterns of lexis in a corpus of plant biology research articles. *International Journal of Corpus Linguistics, 3*(1), 151-171.

Yang, H. Z. (2002). *An introduction to corpus*. Shanghai: Shanghai Foreign Language Education Press.

Zhang, J. D., & Liu, P. (2008). Collocation study on EST texts based on word form. *Foreign Language Research, 6*(112), 28-35.