English Language & Literature Teaching, Vol. 20, No. 2 Summer 2014

Clustering English Textbooks for Vocational Education Based on Lexical Similarity

Yong-Hun Lee[*]
(Chungnam National University)
Kyung-Suk Chang[**]
(Korea Institute for Curriculum & Evaluation)

Lee, Yong-Hun & Chang, Kyung-Suk. (2014). Clustering English textbooks for vocational education based on lexical similarity. *English Language & Literature Teaching*, *20*(2), 193-210.

As occupation fields become more fine-grained and specialized in the globalized world, the importance of English for vocational education becomes greater than the past. Currently, under the government authorization, several English textbooks for vocational high schools have been developed for 5 main vocational areas, i.e. agriculture, industry, commerce, marine and fishery, and domestic industry. However, little is known about with what criteria the textbooks were classified and grouped for each specified field under the five areas. The goal of this paper is to provide a principled and scientific method to classify the English textbooks for the five vocational areas. For this purpose, we collected English textbooks from 13 vocational fields and made a textbook corpus with the textbooks. For quantitative and statistical analysis, we adopted previous studies on linguistic similarity and a hierarchical clustering method. Our study proceeded as follows. First, we made a word list for each English textbook. Then, we calculated the *percentage of lexical similarity* along the word lists of 13 textbooks. Finally, we applied the hierarchical clustering method to producing a dendrogram of English textbooks for vocational education. Along the resulting dendrogram, we could identify which textbook was closer to which one, and had to be grouped with which one. The finding suggests implications for vocational English material development, vocabulary learning and teaching, and research on textbook analysis.

[English for vocational education/ESP/corpus/NLPTools/textbook development]

[*] First author, [**] Corresponding author

교보문고 KYOBO Book Centre

I. INTRODUCTION

As occupation fields have become specifically-divided and specialized in the rapidly changing global society, the field of English for vocational education has sought for ways to better prepare students for such change in the global job market. Efforts at the government level have been made to meet the needs from the changing society, related to its competitiveness in the globalized world. The government proposed a new test system to measure key competencies of vocational high school students. The Test for Enhanced Employability and Upgraded Proficiency (TEENUP, henceforth) has been developed and some pilot tests have been administered to measure vocational school students' core occupational skills. The newly developed test TEENUP is composed of four areas: communication skill in Korean and English, mathematical skill, problem-solving skill, and job ability (MEST, KCCI & ORP, 2012). The new policy has influenced English learning and teaching at vocational high schools in the given context. Such changes in the society have drawn attention to English for specific purposes(ESP) in the field of English language teaching and learning(Dudley-Evans & St. John, 1998; Lee, 2014).

New textbooks were developed to support English learning and teaching at vocational high schools. Vocational high schools in Korea are conventionally classified into five areas: agriculture, engineering, commerce, marine and fishery, and domestic industry. Under the classification there are various schools, where students are trained for specific vocational fields. Table 1 shows what and how many textbooks were developed under each vocational area. 13 textbooks were authorized by the Ministry of Education or Local Offices of Education, and they were distributed to vocational high schools across the country. It is expected that vocational high schools under each vocational area adopt or adapt the authorized textbooks for English teaching for vocational training and preparing their students for the newly developed test, TEENUP.

TABLE 1

Classification of Vocational Textbooks

Vocational Areas	Vocational Textbooks(The number of textbooks)
Agriculture	Agriculture, Horticulture &Landscape (2)
Engineering	Architectural Engineering, Electrical Engineering, Electronic Engineering, General Engineering, Mechanical Engineering (5)
Commerce	Finance, Retail &Logistics (2)
Marine & Fishery	Maritime, Nautical (2)
Domestic industry	Culinary Art, Tourism (2)

교보문고 KYOBO Book Centre

Although vocabulary has an important influence on the textbook development for different vocational fields, little has been known about what criteria were applied to such categorization of English textbooks for vocational education, and by what criteria vocabulary was selected for the textbooks under each vocational area(Chang, Lee, Lee, & Jeon, 2013). To see how appropriately vocabulary selection was made to meet the needs from different vocational fields, it seems to be necessary to have a close examination into lexical properties and syntactic structures of English sentences in each textbook developed and authorized by the ministry or local offices of education.

The goal of this paper is to provide a principled and scientific method to classify the English textbooks for different vocational areas. For this purpose, we make a textbook corpus with the authorized 13 vocational English textbooks in Table 1. Then, we apply quantitative and statistical methods to each textbook in the corpus and group the textbooks based on the lexical similarities. Some implications and suggestions are made for material development for English for vocational education and research on the lexical approach to textbook analysis.

This paper is organized as follows. Section 2 introduces both theoretical and methodological backgrounds to this study. This paper is theoretically based on vocabulary learning and technical vocabulary. It is methodologically based on computational stylometry in corpus linguistics and machine learning in computational linguistics. Following Section 2 explaining each of these disciplines, Section 3 mentions the research method and procedure. Section 4 presents the analysis results, followed by the discussions based on the results. Section 5 summarizes this study and made some suggestions for further research.

II. LITERATURE REVIEW

1. Vocabulary Learning and Technical Words

Nation (2001) classified words into four groups: high-frequency words, academic words, technical words, and low-frequency words. It was found that each type of word occupies about 80%, 9%, 5%, and 5% of the running words in a text. It is also known that among these types of words, technical vocabulary determines the characteristics of texts(Dudley-Evans & St. John, 1998; Nation, 2001).

Several previous studies on vocabulary learning (Barber 1962; Cowan, 1974; Higgins, 1966) suggested that it was not the English teacher's job to teach technical words directly, and that the technical words should be handled and learned through study of each field. However, since many researchers (Finnocchiaro, 1964; Laufer, 1989, 1992; Nation 2001)

교보문고 KYOBO Book Centre

have mentioned the 95 percent coverage as the target level in order to gain reasonable comprehension of texts, the technical words cannot be ignored for the understanding of the vocational textbooks.

According to Nation (2001, p. 18), "technical words contain a variety of types which range from words that do not usually occur in other subject areas to those that are formally like high-frequency words but have special meanings." He classified the technical words, as shown in Table 2, with Category 1 the most technical and Category 4 the least(Nation, 2001, pp. 198-199)[1].

TABLE 2

Four Categories in Technical Words

Category	Characteristics
Category 1	' The word form appears rarely if at all outside of this particular filed.
Category 2	' The word form is used both inside and outside this particular field but not with the same meaning.
Category 3	' The word form is used both inside and outside this particular field, but the majority of its use with a particular meaning though not all, are in this field. ' The specialized meaning it has in this field is readily accessible through its meaning outside the field.
Category 4	' The word form is more common in this field than elsewhere. ' There is little or no specialization of meaning, though someone knowledgeable in the field would have a more precise idea of its meaning.

Among them, the words which differentiate the vocational textbooks are those in Category 1, since the words in the other categories may be common to all of the textbooks.

As Nation (2001, p. 204) pointed out that, considering the situation that large numbers of technical words occur in specialized texts, it is necessary for language teachers to prepare learners to deal with the technical words in that learning technical words is closely related to learning the relevant vocational field. Accordingly, although English teachers are not equipped to teach vocational texts to the students, they can help the students be accustomed to the technical words in the relevant vocational field. Nation (2001, p. 205) also suggested that "learners should approach specialized vocabulary strategically, considering whether particular words are worth learning, and how they can be most efficiently learned."

[1] The algorithm by which we can extract the technical vocabulary in the vocational textbooks was developed in Lee, Chang, Lee and Kim (2012).

교보문고 KYOBO Book Centre

2. Computational Stylometry

Computational Stylometry is an area used in both corpus linguistics and computational linguistics. It is an attempt to capture the essence of the style of the particular author by reference to a variety of quantitative criteria, usually lexical criteria, called *discriminators* (Oakes, 1998, 2009). It is also used for genre identification and text grouping, based on the lexical and/or syntactic properties (discriminators) of each text.

There have been a few studies on approaches which employed statistical methods to measure linguistic relationships. Among the studies, Ellegård (1959) supposed that there existed a language *L* consisting of just 20 roots and used the *product-moment correlation coefficient r* to measure how much its descendants had been developed from the common language *L*. Dyen et al. (1967) adopted the *portion of cognate pairs* to obtain estimates of language divergence and word retention from determinations of the number of cognate words found in pairs of Austronesian languages. Dyen et al. (1992) took a quantitative approach to judging historical relatedness. They adopted the *percent cognate words*, and took the number of shared cognates in a list of 200 cross-linguistically common words as a measure of overall similarity among 84 Indo-European languages and then used a cluster analysis to produce a dendrogram of Indo-European(Swadesh 1952).

3. Machine Learning

Machine learning is defined as the study of computational systems that improve performance on some task with experience (Langley, 1996; Mitchell, 1997). The methods represent learned knowledge in a declarative, symbolic form as opposed to more numerically oriented statistical or neural-network training methods (Mooney, 2003).

There are several types of machine learning models, and they can be classified, as shown in Table 3, on the basis of types of variables and methods of learning (Abney, 2008). According to Abney (2008), there are a few more machine learning methods which cannot be contained into the classification in Table 3, including reinforce learning and semi-supervised learning.

TABLE 3

Types of Machine Learning Models

Learning / Variable	Unsupervised	Supervised
Nominal	Clustering	Classification
Real	Density Estimation	Regression

교보문고 KYOBO Book Centre

Among these types of machine learning, this paper adopted the *clustering* method, which is both nominal and unsupervised. The clustering method which is used in this paper is unsupervised in that the corpus data have no annotations and that the similarity matrix is calculated purely based on the similarity on the lexical items, without any previous knowledge on the corpus data. The method is also nominal in that the variables which used for the calculation are nominal, not real.

There are several different types of clustering methods, and the clustering method used in this paper is hierarchical clustering. *Hierarchical clustering* is a method of cluster analysis which seeks to build a hierarchy of clusters and uses an agglomerative method. The analysis results are conventionally represented by a dendrogram. This paper adopted a hierarchical clustering method, since a dendrogram is a useful representation where we can easily capture the overall tendencies of data.

III. RESEARCH METHODS

1. Corpus Data and Analysis Tools

The corpus linguistics has made a contribution to the systematic textbook analysis in English learning and teaching(Kim, 2009; Kim & Suh, 2006). The corpus-based approach has been used for analyzing textbooks the English textbooks used at the vocational high schools. We compiled a textbook corpus for the analysis of English textbooks for vocational high schools as follows. First, we had access to the PDF version of 13 textbooks under investigation. Then, we converted the PDF files into plain text files. However, the text file contains Korean characters (*Hangul*) in the content. In order to properly analyze the textbook, it is necessary to remove all the characters except English alphabet, numbers, and punctuation markers. We made a software tool which automatically removes all the Korean characters from the text file. In this software tool, we used the American Standard Code for Information Interchange (ASCII) code to exclude all the characters except alphanumerics and punctuation marks. Accordingly, all the Korean characters were removed from the text file. The tool was included in *NLPTools for English Education Experts*, which will be explained below. Using this tool, we converted the text file to contain only English alphabet, numbers, and punctuation markers. We applied the same process to the other 12 vocational textbooks and a compiled textbook corpus was constructed with the resulting text files. The compiled textbook corpus contained 13 plain text files and 231,768 word tokens.

This paper used three tools for analyzing the data: *NLPTools for English Education Experts*, *NLPTools for Corpus Experts*, and R. *NLPTools for English Education Experts*

교보문고 KYOBO Book Centre

was used for basic pre-processing of vocational textbooks, for making a word list for each textbook, and for comparing the word lists of 13 vocational textbooks. *NLPTools for Corpus Experts* was used for producing a similarity lexical matrix, which was produced by comparing the word lists of 13 vocational textbooks. The algorithm for calculating the lexical similarity was illustrated in Section 3.3. The program R was used for generating a dendrogram based on the similarity matrix which is produced by *NLPTools for Corpus Experts*.

There are four different versions of NLPTools: *NLPTools for General Users, NLPTools for Corpus Experts*, *NLPTools for English Education Experts*, and *NLPTools for Speech Experts*. Among these versions, *NLPTools for General Users* is distributed by Cambridge University Press with Lee (2007). This program contains basic corpus-analyzing tools such as a frequency counter, a concordancer, and an English tagger; and this program is consisted of 44 menus. *NLPTools for Corpus Experts*, which is used in this paper, is an advanced version of *NLPTools for General Users*. This advanced version contains 35 more menus in addition to 44 basic menus. *NLPTools for English Education Experts* contains 48 more menus in addition to 44 basic menus and includes more tools that have been developed to analyze lexical and syntactic properties of English textbooks and test materials and to evaluate those materials. *NLPTools for Speech Experts* includes more tools that have been developed to handle various types of speech corpora. The algorithm for lexical similarity introduced in this paper is also implemented in *NLPTools for Corpus Experts*.

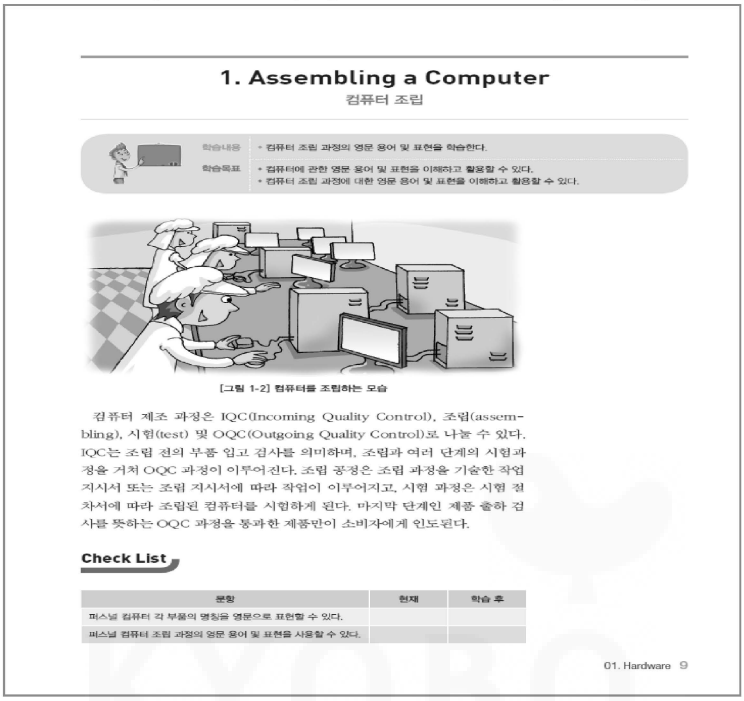
2. Research Procedure

To elaborate on the research method, our study proceeds as follows. First, we selected 13 authorized vocational textbooks, which are currently used at high schools in Korea. Figure 1 shows an example of an English textbook for vocational education.

교보문고 KYOBO Book Centre

FIGURE 1

A Sample Textbook (English for Electronic Engineering)



1. Assembling a Computer
컴퓨터 조립

학습내용 • 컴퓨터 조립 과정의 영문 용어 및 표현을 학습한다.

학습목표 • 컴퓨터에 관한 영문 용어 및 표현을 이해하고 활용할 수 있다.
• 컴퓨터 조립 과정에 대한 영문 용어 및 표현을 이해하고 활용할 수 있다.

[그림 1-2] 컴퓨터를 조립하는 모습

컴퓨터 제조 과정은 IQC(Incoming Quality Control), 조립(assembling), 시험(test) 및 OQC(Outgoing Quality Control)로 나눌 수 있다. IQC는 조립 전의 부품 입고 검사를 의미하며, 조립과 여러 단계의 시험과정을 거쳐 OQC 과정이 이루어진다. 조립 공정은 조립 과정을 기술한 작업지시서 또는 조립 지시서에 따라 작업이 이루어지고, 시험 과정은 시험 절차서에 따라 조립된 컴퓨터를 시험하게 된다. 마지막 단계인 제품 출하 검사를 뜻하는 OQC 과정을 통과한 제품만이 소비자에게 인도된다.

Check List

분항	현재	학습 후
퍼스널 컴퓨터 각 부품의 명칭을 영문으로 표현할 수 있다.		
퍼스널 컴퓨터 조립 과정의 영문 용어 및 표현을 사용할 수 있다.		

01. Hardware 9

Since all the textbook files had the PDF formats, we converted the PDF files into plain text files. However, since the text files contained Korean characters (*Hangul*) in the contents, we used a tool in *NLPTools for English Education Experts* to remove all the characters except English alphabet, numbers, and punctuation markers (pre-processing). Finally, we made a textbook corpus with these 13 textbook files.

Second, we made a word list for each textbook. These processes were done by *Wordlist Maker* as shown in Figure 2. During the process, we eliminated all the numbers from the lists, and we applied a token/type-based analysis (not a lexeme-based analysis or a family-based analysis) in order to maximize the lexical differences among the textbooks. One may apply a lexeme-based analysis or a family-based analysis instead. However, it does not change the overall grouping and tendencies.

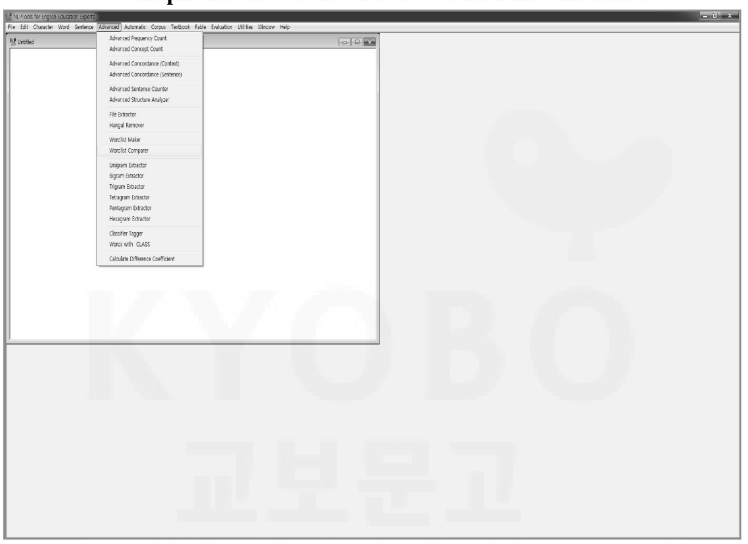
Third, we used a word-list compare tool in *NLPTools for English Education Experts* to identify which word appears in which textbook. Figure 2 is the screenshot for running the tool in the program, and Figure 3 illustrates the analysis results. Currently, up to 50 textbooks or word lists can be compared simultaneously using *Wordlist Comparer* in *NLPTools for English Education Experts*. However, 20~30 word lists have usually been

교보문고 KYOBO Book Centre

compared in the actual analyses. There are two versions of comparing methods, whose options can be set in the subsequent menu. In the first option, as you can see in Figure 3, the program only shows the occurrence/absence of the given word in the given word list. In that case, the number of the second column refers to how many textbooks contain the given word. In the second option, which is not shown in Figure 3, the program enumerates the actual frequencies of the given word in the given word list. In that case, the number of the second column refers to the total number of frequencies of the given word in all of the word lists of all the 13 vocational textbooks.

FIGURE 2

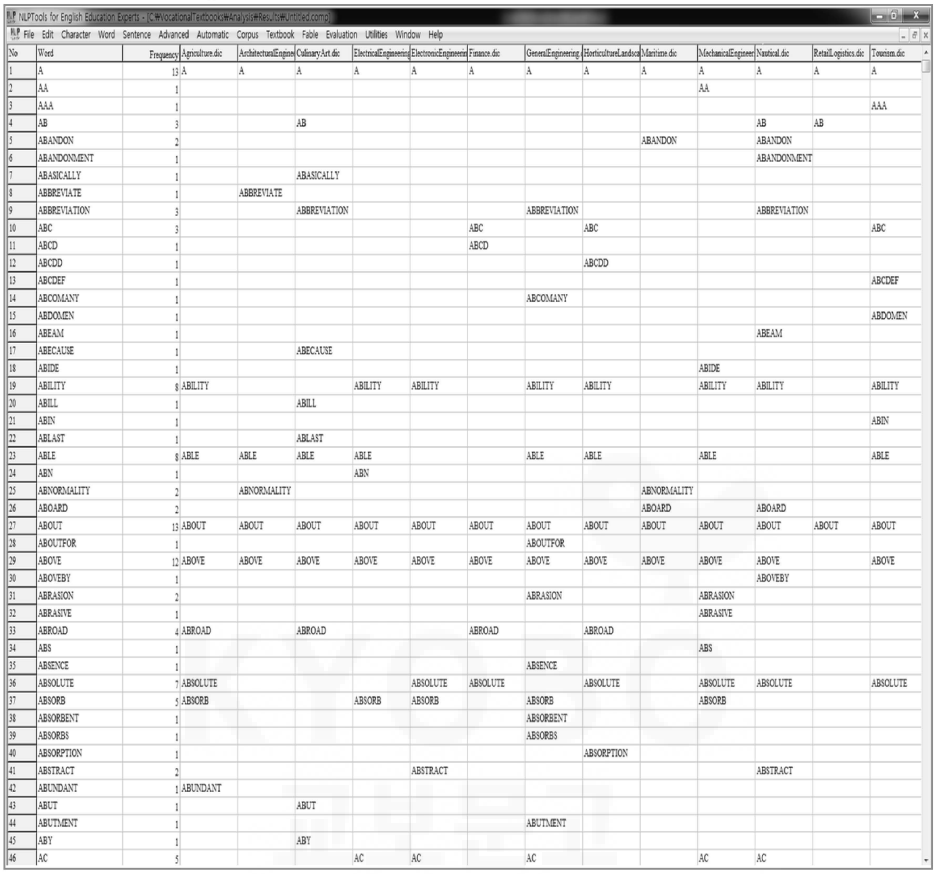
Comparison of Word Lists for Vocational Textbooks



교보문고 KYOBO Book Centre

FIGURE 3

Comparison Results of 13 Word Lists for Vocational Textbooks



No	Word	Frequency	Agriculture.dic	ArchitecturalEngineering.dic	CulinaryArt.dic	ElectricalEngineering.dic	ElectronicEngineering.dic	Finance.dic	GeneralEngineering.dic	HorticultureLandscape.dic	Maritime.dic	MechanicalEngineering.dic	Nautical.dic	RetailLogistics.dic	Tourism.dic
1	A	13	A	A	A	A	A	A	A	A	A	A	A	A	A
2	AA	1										AA			
3	AAA	1													AAA
4	AB	3			AB								AB	AB	
5	ABANDON	2									ABANDON		ABANDON		
6	ABANDONMENT	1											ABANDONMENT		
7	ABASICALLY	1			ABASICALLY										
8	ABBREVIATE	1		ABBREVIATE											
9	ABBREVIATION	3			ABBREVIATION				ABBREVIATION				ABBREVIATION		
10	ABC	3						ABC		ABC					ABC
11	ABCD	1						ABCD							
12	ABCDD	1								ABCDD					
13	ABCDEF	1													ABCDEF
14	ABCOMANY	1							ABCOMANY						
15	ABDOMEN	1													ABDOMEN
16	ABEAM	1											ABEAM		
17	ABECAUSE	1			ABECAUSE										
18	ABIDE	1										ABIDE			
19	ABILITY	8	ABILITY			ABILITY	ABILITY		ABILITY	ABILITY		ABILITY	ABILITY		ABILITY
20	ABILL	1			ABILL										
21	ABIN	1													ABIN
22	ABLAST	1			ABLAST										
23	ABLE	8	ABLE	ABLE	ABLE	ABLE			ABLE	ABLE		ABLE			ABLE
24	ABN	1				ABN									
25	ABNORMALITY	2		ABNORMALITY							ABNORMALITY				
26	ABOARD	2									ABOARD		ABOARD		
27	ABOUT	13	ABOUT	ABOUT	ABOUT	ABOUT	ABOUT	ABOUT	ABOUT	ABOUT	ABOUT	ABOUT	ABOUT	ABOUT	ABOUT
28	ABOUTFOR	1							ABOUTFOR						
29	ABOVE	12	ABOVE	ABOVE	ABOVE	ABOVE	ABOVE	ABOVE	ABOVE	ABOVE	ABOVE	ABOVE	ABOVE		ABOVE
30	ABOVEBY	1											ABOVEBY		
31	ABRASION	2							ABRASION			ABRASION			
32	ABRASIVE	1										ABRASIVE			
33	ABROAD	4	ABROAD		ABROAD			ABROAD		ABROAD					
34	ABS	1										ABS			
35	ABSENCE	1							ABSENCE						
36	ABSOLUTE	7	ABSOLUTE				ABSOLUTE	ABSOLUTE		ABSOLUTE		ABSOLUTE	ABSOLUTE		ABSOLUTE
37	ABSORB	5	ABSORB			ABSORB	ABSORB		ABSORB			ABSORB			
38	ABSORBENT	1							ABSORBENT						
39	ABSORBS	1							ABSORBS						
40	ABSORPTION	1								ABSORPTION					
41	ABSTRACT	2					ABSTRACT						ABSTRACT		
42	ABUNDANT	1	ABUNDANT												
43	ABUT	1			ABUT										
44	ABUTMENT	1							ABUTMENT						
45	ABY	1			ABY										
46	AC	5				AC	AC		AC			AC	AC		

As observed, one can identify which words appear in which textbook. The number of the second column refers to how many textbooks contain the given word. For example, the word *a* is contained all of the 13 textbooks, while the word *abandon* is contained only two textbooks 'Martime.dic' and 'Neutical.dic'

Fourth, we calculated the *percentage of lexical similarity*, which was similar to the *percent cognate words* in Dyen et al. (1992).[2] The basic algorithm was mentioned in the next section. This process was automatically done by a tool in *NLPTools for Corpus Experts*. Figure 4 is the screenshot for running the tool in the program, and Figure 5 illustrates the analysis results.

[2] Note that it was impossible to use the *percent cognate words* since the scope of this paper did not include the historical changes.

교보문고 KYOBO Book Centre

FIGURE 4

Calculating a Similarity Matrix

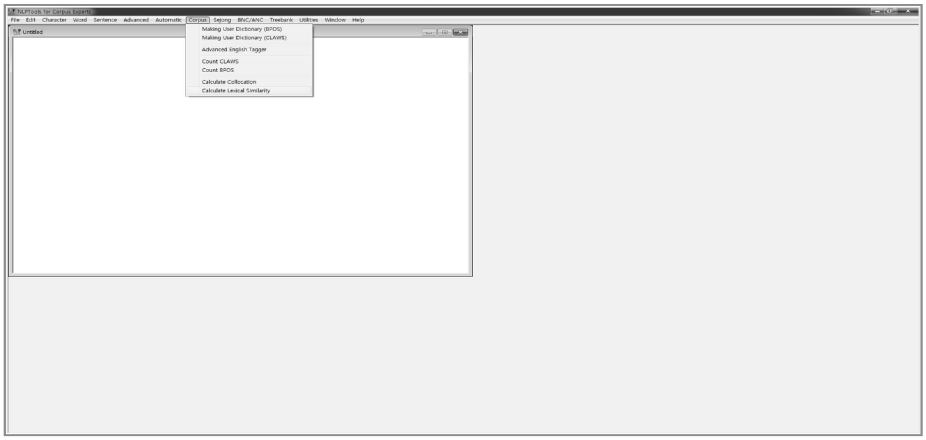
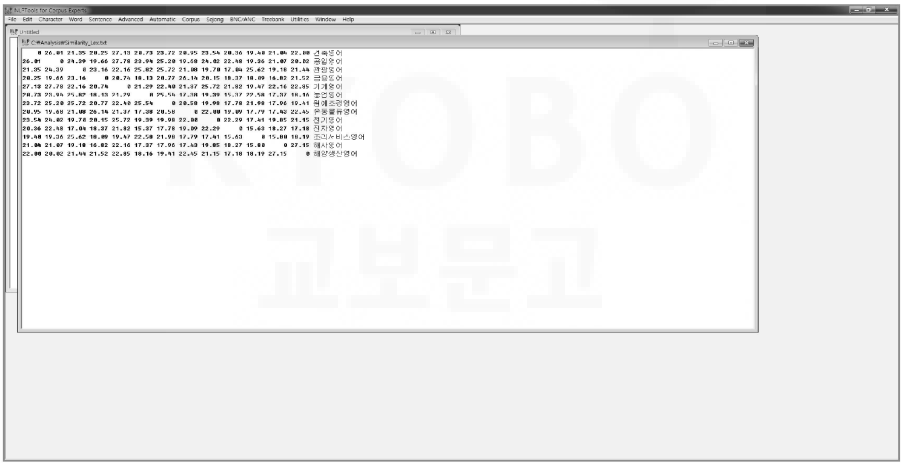


FIGURE 5

A Similarity Matrix for Vocational Textbooks



Here, the names textbooks are enumerated at the last column. The same order is applied to both the row and the column. That is, *Architectural Engineering* is applied both to the first row and to the first column. Likewise, *General Engineering* is applied both to the second row and to the second column. Accordingly, the cell with the first row and second column (26.01 in Figure 5) refers to the similarity between *Architectural Engineering* and *General Engineering*. Likewise, the cell with the second row and first column (26.01 in Figure 5) refers to the similarity between *Architectural Engineering* and *General Engineering*.

교보문고 KYOBO Book Centre

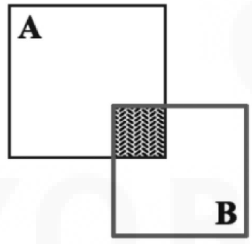
Finally, based on a matrix of the *percentage of lexical similarity*, we adopted a clustering method (one of the unsupervised machine learning algorithms) and produced a dendrogram using R (Baayen, 2008; Johnson, 2008).

3. Calculation of Lexical Similarity

The *percentage of lexical similarity* was automatically calculated in this paper using a tool in *NLPTools for Corpus Experts*. The basic idea of the calculation came from Jaccard similarity coefficient (Jaccard, 1901, 1912). In order to understand the basic ideas of Jaccard similarity coefficient, let' see the following figure.

FIGURE 6

Calculation of the Percentage of Lexical Similarity



Here, we have two sets A and B, and the shaded part is the part common to both the set A and the set B. Jaccard similarity coefficient is calculated as $c(A \cap B)/c(A \cup B)$, where $c(X)$ refers to the number of members in the set X. Since the *percentage of lexical similarity* was calculated based on Jaccard similarity coefficient in this paper, it can be said to be calculated based on set-theoretical operations.

In a matrix for the *percentage of lexical similarity*, we have the same values in the diagonal axis ($M[i,j]=M[j,i]$). We have $n(n\text{-}1)/2$ pairs for n word lists. Since we have 13 textbooks, we have 13(12-1)/2=78 textbook pairs. For each pair, we have five set-theoretical operations. The five operations are $c(A)$, $c(B)$, $c(A \cap B)$, $c(A \cup B)$, and $c(A \cap B)/c(A \cup B)$. Therefore, we have the total 78×5=390 operations to get a matrix for the *percentage of lexical similarity*. One may reduce the number of operations by counting $c(A)$ and $c(B)$ only once and re-using the data. However, even in that case, we have 78×3+13=247 operations, since you cannot calculate $c(A \cap B)$, $c(A \cup B)$, and $c(A \cap B)/c(A \cup B)$ in advance. 247 operations are still too many to be handled manually or just with a calculator. This is why we made a tool for these calculations in *NLPTools for Corpus Experts* and automatically calculated the values. Since this number is too many

교보문고 KYOBO Book Centre

calculations to be handled manually, we made a tool for these calculations in *NLPTools for Corpus Experts* and automatically calculated the values.

It may be asked why we did not use the WordNet:Similarity functions, which was frequently used to calculate the lexical similarity in WordNet (Felbaum, 1998). Currently, 10 components have been developed in Wordnet:Similarity (Pedersen et al., 2004; Pedersen, 2008). Those 10 components are *hso*, *jcn*, *lch*, *lesk*, *lin*, *path*, *random*, *res*, *vector*, and *wup*. However, there is one important problem to use the WordNet:Similarity functions in our study. As it is generally known, the concepts (words) are hierarchically organized in the WordNet, whereas the words in our study are not hierarchically organized in the word lists. Accordingly, it is questionable to use the WordNet: Similarity functions to calculate the lexical similarity among the word lists. Jaccard coefficient is one of the archaic but fundamental methods and it is still used nowadays in a few areas. This is why the efficient in the analysis was chosen for the present study. One may ask why the present study did not use document clustering algorithms which are frequently used in computational linguistics as in Manning and Schütze(1999). It was assumed that English textbooks are different from the usual texts in that the pedagogical aspects have to be considered in the analysis of the textbooks. This is why the algorithms was not used in our study.

IV. FINDINGS AND DISCUSSIONS

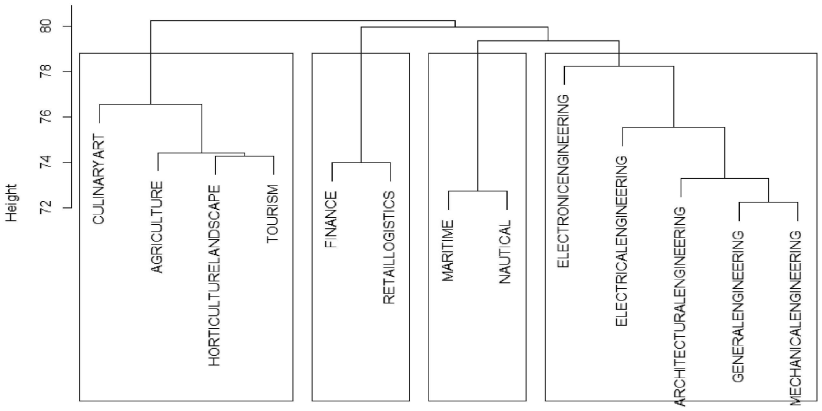
After we made a matrix of *percentage of lexical similarity* for each pair of textbooks in Figure 5, we used a hierarchical clustering method and produced a dendrogram using R. As the result is shown in Figure 7, the grouping is automatically performed using a function in R. As it is shown in this dendrogram, the 13 vocational English textbooks are grouped into four groups, purely based on the lexical similarity.

Comparing the result with the classification in Table 1, we may find the following facts. The second, third and fourth groups in Figure 7 correspond to Commerce, Marine & Fishery, and Engineering Areas respectively in Table 1. These three groups are clearly identified in the statistical analysis of vocational textbooks as well. However, the vocational textbooks in Agriculture and Domestic Industry Areas are interwoven in the dendrogram in Figure 7. As we can observe, two textbooks in Agriculture Area i.e. Agriculture, and Horticulture & Landscape are properly included in the textbooks for Domestic Industry Area. There is a gap between the current classification of the textbooks shown in Table 1 and the classification based on the lexical similarity of textbooks shown in Figure 7.

교보문고 KYOBO Book Centre

FIGURE 7

Cluster Dendrogram of Vocational Textbooks



It is also found that the distribution of lexical similarity is not identical even within a vocational area. For example, the vocabulary in Mechanical Engineering is the most closest to that of General Engineering; Architectural Engineering is the next; Electrical Engineering is the next; and Electronic Engineering is the most different from General Engineering.

The dendrogram we got from the systematic analysis may provide some implications for the teaching of vocabulary in vocational English textbooks. First, it is suggested that the identified gap in grouping the vocational English textbooks should be narrowed by regrouping them using the lexical similarity. The data on the lexical similarity can be used for the use of the authorized textbooks for English teaching and learning for vocational education. The lexical data of the current vocational English textbooks can be also used for the English curriculum development and its revision. It is expected that the findings of the present research have implications for further material development for English teaching and learning for vocational education. They can provide lexical information to those who develop the Test for Enhanced Employability and Upgraded Proficiency (TEENUP) to measure students' key skills for vocation. The findings provide implications for research on the vocabulary for English for vocational education.

Second, the dendrogram can be used for constructing a word list for a specific vocational area or a specific vocational field, which is contained within the vocational area. As Nation (2001, p. 205) mentioned, students should approach specialized vocabulary strategically. Even though teachers cannot teach the technical vocabulary directly, they can provide a technical word list for each vocational area or each vocational field. Then, as the Academic Word List (Coxhead, 2000) was helpful to all the academic texts, the technical word list is

교보문고 KYOBO Book Centre

also expected to be helpful for students to understand each vocational area or each vocational field.

As discussed in Section 2, technical vocabulary occupies about 5 percent of the running words and it is impossible to fully understand the vocational textbook without mastering the technical vocabulary in the given textbook. It needs to be borne in mind that 95 percent coverage is the target level in order to gain reasonable comprehension. Technical vocabulary is also important in that it determines the characteristics of the textbooks in the given vocational field. Accordingly, providing a word list for a specific vocational area or a specific vocational field will surely be helpful for students to approach specialized vocabulary strategically. However, since the distribution of lexical similarity is not identical even within a vocational area, it is suggested that we consider these characteristics during the construction of the word lists.

In this paper, we investigated how the vocational English textbooks could be classified scientifically based on the lexical similarity. A clustering method was applied to the lexical similarity analysis and we classified 13 authorized vocational English textbooks which were authorized by the government or local offices of education, and used at the vocational high schools across the nation.

Our study proceeded as follows. After compiling a corpus using the 13 vocational textbooks, a word list was made for each textbook, and the wordlists of 13 textbooks were compared by using a tool in *NLPTools for English Education Experts*. Then we produced a matrix for the *percentage of lexical similarity* using a tool in *NLPTools for Corpus Experts*. Finally, based on the matrix of *percentages of lexical similarity* for each pair of textbooks, a dendrogram was produced using R.

The dendrogram from the analysis showed that the textbooks under Commerce, Marine and Fishery, and Engineering Areas were appropriately grouped under the current classification. It was also found that there was a gap between the current grouping of the textbooks for Agriculture and Domestic Industry Areas, and the grouping on the basis of the lexical similarity of the textbooks.

The dendrogram in this study has the following implications for the teaching and learning vocabulary in English for vocational education. First, the identified gap in grouping the vocational English textbooks should be narrowed by regrouping them using the lexical similarity for effective technical vocabulary learning to take place. Second, the dendrogram can be used for constructing a technical word list for specific vocational areas or fields, which enables students and teachers to approach specialized vocabulary strategically in English learning for vocational purposes. The corpus-based approach to textbook analysis the current study employed provides useful information to developing teaching and learning materials for ESP at the vocational high schools.

교보문고 KYOBO Book Centre

REFERENCES

Abney, S. (2008). *Semisupervised learning for computational linguistics*. New York: Chapman & Hall/CRC.

Baayen, R. (2008). *Analyzing linguistic data*: *A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Barber, C. (1962). Some measurable characteristics of modern scientific prose. In *Contribution to English syntax and philology* (pp. 21-43). Goteburg: Acta Universitatis Gothoburgensis. Reprinted in J. Swales (Ed.), (1985). *Episodes in ESP* (pp. 3-16). Oxford: Pergamon Press Ltd.

Chang, K., Lee, B., Lee, Y., & Jeon, Y. (2013). *A study of developing English remedial teaching and learning materials at the vocational school*. Research Report CRI 2013-17. Seoul: KICE.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213-238.

Cowan, J. (1974). Lexical and syntactic research for the design of EFL reading materials. *TESOL Quarterly*, *8*, 389-400.

Dudley-Evans, T., & St. John, M. J. (1998). *Developments in English for specific purposes: Multi-disciplinary approach*. Cambridge: Cambridge University Press.

Dyen, I., James, A., & Cole, J. (1967). Language divergence and estimated word retention rate. *Language*, *38*, 38-46.

Dyen, I., Kruskal, J., & Black, P. (1992). An Indo-European classification: a lexicostatical experiment. *Transactions of the American Philosophical Society*, *82*(5), 1-132.

Ellegård, A. (1959). Statistical measurement of linguistic relationship. *Language*, *35*(2), 131-156.

Fellbaum, C. (1998). *WordNet*: *An electronic lexical database*. Cambridge, MA: MIT Press.

Finnocchiaro, M. (1964). *English as a second language*. Tokyo: Kinseido.

Higgins, J. (1966). Hard facts. *ELT Journal*, *21*, 55-60.

Jaccard, P. (1901). Étude Comparative de la Distribution Florale Dans Une Portion des Alpes et des Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, *37*, 547–579.

Jaccard, P. (1912). The distribution of the flora in the Alpine zone. *New Phytologist*, *11*, 37–50.

Johnson, K. (2008). *Quantitative method in linguistics*. Oxford: Blackwell.

Kim, N. B. (2009). A corpus-based lexical analysis of the speech texts: A collocational approach. *English Language & Literature Teaching, 15*(3), 151-170.

교보문고 KYOBO Book Centre

Kim, Y. M., & Suh, J. (2006). Usage analysis of vocabulary in Korean high school English textbooks using multiple corpora. *English Language & Literature Teaching, 12*(4), 139-157.

Langley, P. (1996). *Elements of machine learning*. San Francisco, CA: Morgan Kaufmann.

Laufer, B. (1989). What percentage of text lexis is essential for comprehension? In C. Lauren & M. Nordman (Eds.), *Special language*: *From humans thinking to thinking machines* (pp. 316-323). Clevedon: Multilingual Matters.

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In P. Arnaud & H. Bejoint (Eds.), *Vocabulary and applied linguistics* (pp.126-132). London: Macmillan.

Lee, J. Y. (2014). General English or ESP/EAP? Rethinking college students' needs for GE and ESP/EAP. *English Language & Literature Teaching, 20*(1), 133-156.

Lee, Y. (2007). *Corpus analysis using NLPTools and their applications*: *Applications to linguistic research, English education, and textbook evaluation*. Seoul: Cambridge University Press.

Lee, Y., Chang, K., Lee, B., & Kim, M. (2012). A corpus-based study of extracting ESP vocabulary for vocational education. *Proceedings of the 2012 International Conference for the Korean Association for Corpus Linguistics* (pp. 50-51).

Manning, C., & Schütze, H. (2002). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

MEST, KCCI & ORP. (2012). *Framework of the test for enhanced employability and upgraded proficiency*. Seoul: ORP Publication.

Mitchell, T. (1997). *Machine learning*. New York: McGraw-Hill.

Mooney, R. (2003). Machine learning. In R. Mitkov (Ed.), *The Oxford handbook of computational linguistics* (pp. 376-394). Oxford: Oxford University Press.

Nation, P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

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

Oakes, M. (2009). Corpus linguistics and stylometry. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics*: *An international handbook*, Vol. 2 (pp. 1070-1090). Berlin: Gruyter.

Pedersen, T. (2008). Empiricism is not a matter of faith. *Computational Linguistics*, *34*, 465-470.

Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004). WordNet: Similarity - measuring the relatedness of concepts. In *Proceedings of the 19th National Conference on Artificial Intelligence* (*AAAI*), 1024-1025.

교보문고 KYOBO Book Centre

Swadesh, M. (1952). Lexico-stating dating of prehistoric ethnic contacts: With special reference to North American Indians and Eskimos. *Proceedings of the American Philosophical Society*, *96*, 452-463.

Examples in: English
Application Languages: English
Applicable Levels: Secondary

Yong-Hun Lee
Dept. of English Language & Literature
Chungnam National University
99 Daehak-ro, Yuseong-gu
Daejon 305-764, South Korea
Tel: 042-821-5114/ 010-7797-7844
E-mail: yleeuiuc@hanmail.net

Kyung-Suk Chang
Korea Institute for Curriculum & Evaluation
Jeongdong 15-5, Jung-gu,
Seoul 100-784, South Korea
Tel: 02-3704-3504/ 010-9491-3557
E-mail: kschang@kice.re.kr

Received in April 15th, 2014
Reviewed in May 26th, 2014
Revised version received in June 5th, 2014

교보문고 KYOBO Book Centre